A Tutorial on Using Generative Models to Advance Psychological Science:

Lessons from the Reliability Paradox

Nathaniel Haines*
The Ohio State University, Department of Psychology,
1835 Neil Ave., Columbus, OH., 43210
Email: nathaniel.b.haines@gmail.com; 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
University of Notre Dame, Department of Psychology,
E312 Corbett Family Hall, Notre Dame, IN 46556
Email: tbeaucha@nd.edu; Website: https://psychology.nd.edu/people/theodore-beauchaine/

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 & AI Institute
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/

*Corresponding authors

Abstract

Theories of individual differences are foundational to psychological and brain sciences, yet they are traditionally developed and tested using superficial summaries of data (e.g., mean response times) that are disconnected from our otherwise rich conceptual theories of behavior. To resolve this theory-description gap, we review the *generative modeling* approach, which involves formally specifying how behavior is generated within individuals, and in turn how generative mechanisms vary across individuals. Generative modeling shifts our focus away from estimating descriptive statistical "effects" toward estimating psychologically interpretable parameters, while simultaneously enhancing the reliability and validity of our measures. We demonstrate the utility of generative modeling in the context of the "reliability paradox", a phenomenon wherein replicable group effects (e.g., Stroop effect) fail to capture individual differences (e.g., low test-retest reliability). Simulations and empirical data from the Implicit Association Test, Stroop, Flanker, Posner, and Delay Discounting tasks show that generative models yield (1) more theoretically informative parameters, and (2) higher test-retest reliability estimates relative to traditional approaches, illustrating their potential for enhancing theory development.

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

Introduction

Across the social, behavioral, and brain sciences, the researcher's primary agenda is to provide an explanation for *why* observable measures (i.e., data) exhibit systematic patterns (Hempel & Oppenheim, 1948). Typically, this process begins with the development of a theory or hypothesis about what should happen in specific situations (e.g., experiments) if our assumptions about the phenomena are correct. To test our theory, we design an environment that places the agent under consideration (e.g., human, rat) into a specific situation, and we collect data that describe the agent's experience (e.g., decisions, neural activity). To evaluate the fidelity of our hypothesis, we use statistical models such as *t*-tests, ANOVAs, regression, and factor analytic models to connect our explanatory theories to the data observed in our experiment (Guest & Martin, 2020; Kellen, 2019; Suppes, 1966). We use these statistical models to make inferences, such as inferring population-level effects or estimating out-of-sample predictive accuracy using in-sample data, and relate these inferences to claims about theory. However, underlying our statistical models are causal or distributional assumptions that are too often mis-aligned with the substantive theories of interest, a situation we refer to as a *theory-description gap*. When assumptions are misaligned, theories become "divorced" from the statistical tests meant to validate or invalidate them, and the disconnect impedes progress in science (Michell, 2009; Szollosi & Donkin, 2019; Yarkoni, 2019).

In this article, we review and demonstrate how *generative modeling*—a theory-driven data analysis approach where formal model specification is guided by hypotheses about latent processes—can alleviate the pressures associated with theory-description gaps. The goal of this demonstration is twofold. First, we show that the inferences made with the traditional statistical analyses applied in behavioral and brain sciences generate misleading conclusions about the utility of behavioral tasks. Second, we show that this problem can be addressed using modern computational modeling and hierarchical Bayesian model fitting methods to analyze behavioral data. It complements recent efforts examining methods for reliably measuring cognition (Zorowitz & Niv, 2022) by examining how the choice of data analysis approach fundamentally affects psychometric properties, but also theoretical power, of behavioral measures. Put together, this article is an appeal for wider use of these more sophisticated statistical methods, showing that the theoretical advantages of generative modeling approaches translate directly to practical benefits in terms of improved measurement properties like reliability and validity.

Specifically, we focus on how generative modeling resolves a vexing theory-description gap that threatens the integrity of research on individual differences: the *reliability paradox*. The reliability paradox refers to the counterintuitive result that person-level measures of behavior across a broad range of tasks (e.g., the Implicit Association Test, Stroop, Flanker, and Posner Cueing tasks) and modalities (e.g., accuracy, response times, task-based and resting-state fMRI) show poor test-retest reliability despite task manipulations showing consistent and robust effects at the group level (Chen et al, 2015; Elliott et al, 2020; Enkavi et al, 2019; Gawronski et al, 2017; Hedge et al, 2017; Noble et al, 2019). The reliability paradox is a major challenge to behavioral research because low test-retest reliability implies that the psychological constructs on which theories are based are unstable across time, making the task or modality untrustworthy for validating theories based on temporally stable individual differences (Elliot et al, 2020; Dange et al, 2020; Schimmack, 2019). As we demonstrate, however, the reliability paradox arises from the implicit, overly restrictive assumptions that researchers make when using standardized statistical models that fail to connect their theories to data in meaningful ways. To illustrate how generative models can overcome the reliability paradox, we use response time measures as our observable data from several tasks. However, we additionally apply this approach to response / preference data from the Delay Discounting task, showing that the principles and benefits of generative modeling are much more general and apply to any measure extracted from data (e.g., accuracy).

A Brief Overview of 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. 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). In terms of understanding the aggregate effects of manipulations, traditional statistical approaches appears to provide good results. However, reliability of dependent variables has essentially no effect on the outcomes of hypothesis tests looking for differences between groups, and in fact low reliability may be desirable when performing significance tests (Overall & Woodward, 1975; 1976; Nicewander & Price, 1978).

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 Implicit

Association Test (IAT), also show similarly low test-retest correlations (r = .01-.72; average of r ≈ .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 reliability is low, 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 blood oxygenation level dependent (BOLD) responses in functional magnetic resonance imaging (fMRI) studies is quite sobering. In an optimistic setting where brain and behavioral measures have test-retest reliabilities of .5, assuming that 1 hour of fMRI scanning is $750 (USD), a study powered to detect a true effect of r = .3 costs $750 × 346 ≈ $259,500 for data collection alone (see Haines, Sullivan-Toole, & Olino, 2023). 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 measures 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 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.

The Gap Between our Substantive Theories and Descriptive Statistical Models

The reliability paradox paints a bleak picture for any research across the psychological and brain sciences that relies on behavioral data. However, we believe that the paradox arises not due to some fundamental flaw with behavioral tasks, but instead due to how we typically derive inference from behavioral data. Below, we break inference down into three components: (1) the behavioral paradigm itself, (2) the behavioral model, and (3) the group model (Figure 1). We then discuss how the assumptions underlying our most common statistical approaches often misalign with our theories regarding how behavior arises, resulting in a “theory-description gap” that impedes scientific progress.

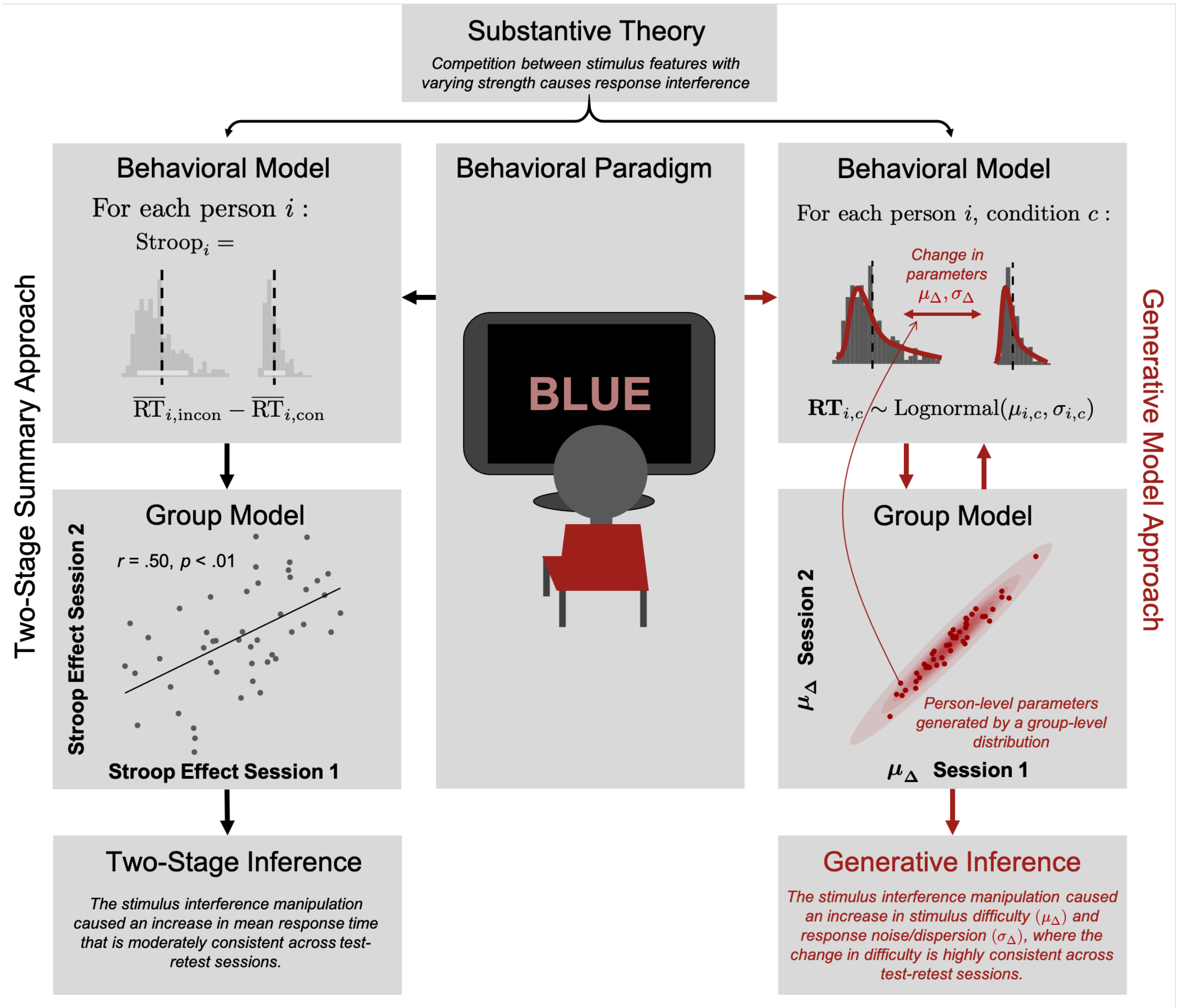
Substantive Theory
Competition between stimulus features with varying strength causes response interference
Two-Stage Summary Approach
Behavioral Model
For each person i :
$\text{Stroop}_i =$
$\overline{\text{RT}}_{i,\text{incon}} - \overline{\text{RT}}_{i,\text{con}}$
Behavioral Paradigm
BLUE
Behavioral Model
For each person i, condition c :
Change in parameters
$\mu_\Delta, \sigma_\Delta$
$\text{RT}_{i,c} \sim \text{Lognormal}(\mu_{i,c}, \sigma_{i,c})$
Generative Model Approach
Group Model
r = .50, p < .01
Stroop Effect Session 2
Stroop Effect Session 1
Group Model
μ_Δ Session 2
μ_Δ Session 1
Person-level parameters generated by a group-level distribution
Two-Stage Inference
The stimulus interference manipulation caused an increase in mean response time that is moderately consistent across test-retest sessions.
Generative Inference
The stimulus interference manipulation caused an increase in stimulus difficulty (μ_Δ) and response noise/dispersion (σ_Δ), where the change in difficulty is highly consistent across test-retest sessions.

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 group model is used to summarize and generalize 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 group-level 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.).

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. For example, 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, which makes it challenging to generalize results, including estimates of reliability, across studies (Rouder & Haaf, 2019).

Theoretical issues also arise when comparing different behavioral tasks that are intended to measure the same phenomenon. 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*.

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. Referring back to the Stroop task, our observable measures are distributions of responses and response times for each person, each condition, and perhaps each session in the case of test-retest measurements. Formally, the behavioral model for the Stroop task 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 person participant, where a positive Stroop effect indicates that average response time is longer for incongruent than congruent trials. This difference in mean response time serves as what we call the *behavioral model*, which is the mathematical model used to make inferences about individual differences in latent cognitive processes. 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 the Data Analysis section). 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, 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 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 as a difference in constructs rather than incomplete behavioral models (e.g., Cyders & Coskunpinar, 2011; Duckworth & Kern, 2011; Bosson et al., 2000).

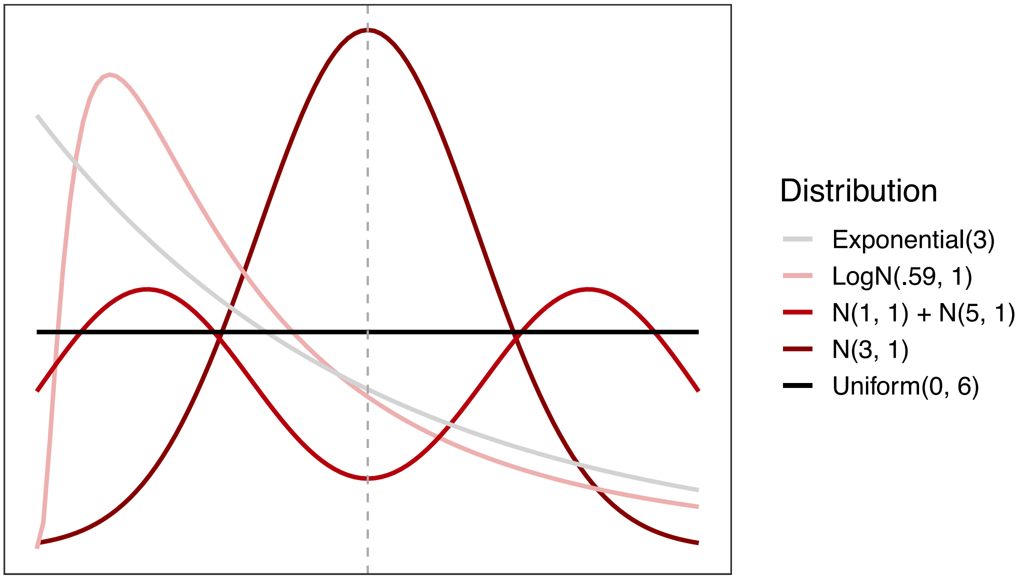

Figure 2. Five distributions all sharing the same mean. An equivalent mean can be derived from five different data generating distributions: a typical normal distribution (N, blue), a lognormal distribution (LogN, red), a sum of two normal distributions (yellow), an exponential distribution (Exp, purple), and a uniform distribution (Unif, green). Because all of these distributions have exactly the same mean, they would produce the same conclusions if analyzed with the behavioral model from Equation 1, regardless of how different their actual data-generating process may be.

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 person. 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; Kvam & Turner, 2021), or skew (Kvam & Busemeyer, 2020; Leth-Steensen et al., 2000). Ignoring these aspects of individual-level distributions has practical implications as well—in the supplement, we show that analyses ignoring distribution information in favor of mean differences are vastly inferior to ones that incorporate estimates of distribution-level differences. Without employing a behavioral model that captures distributions, 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).

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). Rotello et al. argued that the error in inference was directly attributable to the fact that the behavioral model did not properly account for differences in participants' unwillingness to choose a suspect between the two conditions. Therefore, the model failed to capture the intended effect because the difference in detection accuracy was 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 Haines, Kvam, & Turner, 2023; 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.

The Group Model

Given a behavioral model, inference in psychological and brain sciences typically then proceeds using some form of null hypothesis significance testing (Tong, 2019). For example, researchers might apply a *t*-test or multiple regression to the output of Equation 1, yielding a *p*-value that is subsequently interpreted with respect to the substantive theory of lexical interference. We term this second-stage model the “group” model to better reflect its purpose—as shown in Figure 1, the group model is used to make generalizable inferences in the face of uncertainty, using the group sample of parameters estimated from the behavioral model as input. 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 and then used in a secondary statistical model—makes a strong assumption that contributes to poor test-retest reliability and low validity more generally when it is violated (Ly et al., 2017; Rouder & Haaf, 2019; Turner et al., 2017). Specifically, it ignores uncertainty (i.e., measurement error) associated with each person’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 person-level Stroop effects are estimated with infinite precision (i.e., no error), or that we have an infinite number of trials for each person. 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). By contrast, using group models that account for individual-level uncertainty leads to more powerful group- and person-level inferences (e.g., Brown et al, 2021; Haines et al., 2020; Romeu et al. 2019), as we will describe in the next section.

Interim Summary

Our central premise is that atheoretical behavioral models that rely on the two-stage approach described above produce an impoverished and incomplete view of rich individual differences underlying behavioral data. In the following section, we describe how generative modeling is better suited to detect and understand individual differences in behavioral data compared to traditional approaches. Throughout, we focus our attention on how generative modeling affects inference on 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.

Bridging the Theory-Description Gap with Generative Modeling

The Generative Modeling Perspective

A theory-description gap arises when we continue to refine our verbal or conceptual theory to the point at which it is no longer amenable to simple experimental designs, summary statistics, and standard inferential modeling. 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 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.

Fortunately, theoretical frameworks that allow us to better characterize behavioral and neural data 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., 2020), 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 can simulate data of varying consistency with behavioral observations at the level of individual participants[1]. Generative models differ from statistical models (e.g., ANOVA, generalized linear modeling) discussed above because they specify a data generating process that is instantiated from a theory about behavior. Thus, mean contrasts do not qualify as generative models because they reduce person-level data to a single estimate that cannot capture a full distribution of behavior (Equation 1). We will refer to this approach toward modeling behavior-generating processes as the *generative perspective*. We present simulations (see section titled Simulated Demonstration) to provide a concrete example of why this approach is useful.

A Generative Model of Behavior

To build a generative model, we return first to the *behavioral model*. In our case, we need a model that can capture a person's distribution of response times across trials (and condition type, etc.). Our choice of behavioral model should be informed by theoretical assumptions and other background knowledge. For response time data, we know that response time distributions (1) must take on only positive real values, (2) have some central tendency with corresponding variability around this central tendency (e.g., a mean and variance), (3) are often right skewed,

[1] 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). Formally, a generative model can be viewed as a model that specifies the joint probability distribution $p(y, \theta)$ between observable data y and unobservable model parameters θ.

(4) are shifted away from 0 due to “non-decision” factors (e.g., visual encoding time, motor response time), and (5) typically show a linear relation between the sample mean and standard deviation such that increases in mean response time are accompanied by increased variability of response times from trial-to-trial (the *law of response time*; Wagenmakers & Brown, 2007).

There are many different distributions that can incorporate our knowledge about response times, and thus the process(es) that generated them. Here, we limit our discussion to the normal, lognormal, and shifted-lognormal distributions (Figure 3). There are, of course, many generative models of response times that can meet the criteria outlined above (Ratcliff et al, 2016; Busemeyer et al, 2019; Heathcote & Matzke, 2022). Our goal is not to compare these models, but rather to build successively more generative assumptions into models of response times to demonstrate why they are necessary. It is our hope that this simple tutorial-like approach will inspire readers to adopt more sophisticated and mechanistic process models that are more actively studied areas such as mathematical psychology.

The Normal Model. 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 of response times, but it can still be useful for exemplifying the shift away from the behavioral model in Equation 1 and toward 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 person’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. The Stroop effect can be characterized by within-person changes in the shape of each person’s response time distribution across trials within conditions. When using a normal 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 person. 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 described by 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 $\sigma_{i,1,t} = \exp(\sigma_{i,base,t})$.[2] 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).

[2] 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.

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 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 follow a 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 also has a very helpful property in how the mean and standard deviation parameters interact (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 3A and 3B 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 a person'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).

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 person 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 3C 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 because the response modality remains constant across conditions within each session. 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.

Naturally, the shifted lognormal will not be the best generative model of behavior for all tasks. By its nature, a one-size-fits-all solution to behavioral modeling will not capture the minutiae performance on all the tasks we examine here. The goal instead is to examine the degree to which we can improve upon the two-stage approach by incorporating (ostensibly) more valid assumptions about the behavior we are trying to model. The shifted log-normal provides a statistical distribution that reflects several realistic assumptions we can make: that people will have different average processing speeds, that response times will be right-skewed and vary across trials, and that some component of response times can be attributed to processes that are not directly related to decision making. For specific tasks, we could build upon these assumptions by adding an account of accuracy (as with diffusion / accumulator models), across-trial variability in stimulus effects, collapsing response boundaries, or across trial shifts in attention, for instance. However, doing so treads out of clear-cut assumptions that should obviously be made a priori and into complicated statistical model comparisons among many theories with plausible assumptions (Navarro, 2019). To avoid this complication, we stop at the shifted lognormal as a simple model embodying clear-cut, desirable properties of a generative account of response times. With these caveats in place, it is also worth mentioning that, even despite its simplicity, there are several studies in mathematical psychology that collectively establish a solid theoretical basis for the mechanisms that could create lognormally distributed response time distributions (Ulrich and Miller, 1993; Heathcote and Love, 2012). As one example, Ulrich and Miller (1993) show that a simplified version of the cascade process used in connectionist network models (McClelland, 1979) produces lognormally distributed finishing times. Regardless, we encourage the reader to consider and embrace more realistic and modern extensions of these and other generally accepted frameworks for evidence accumulation models.

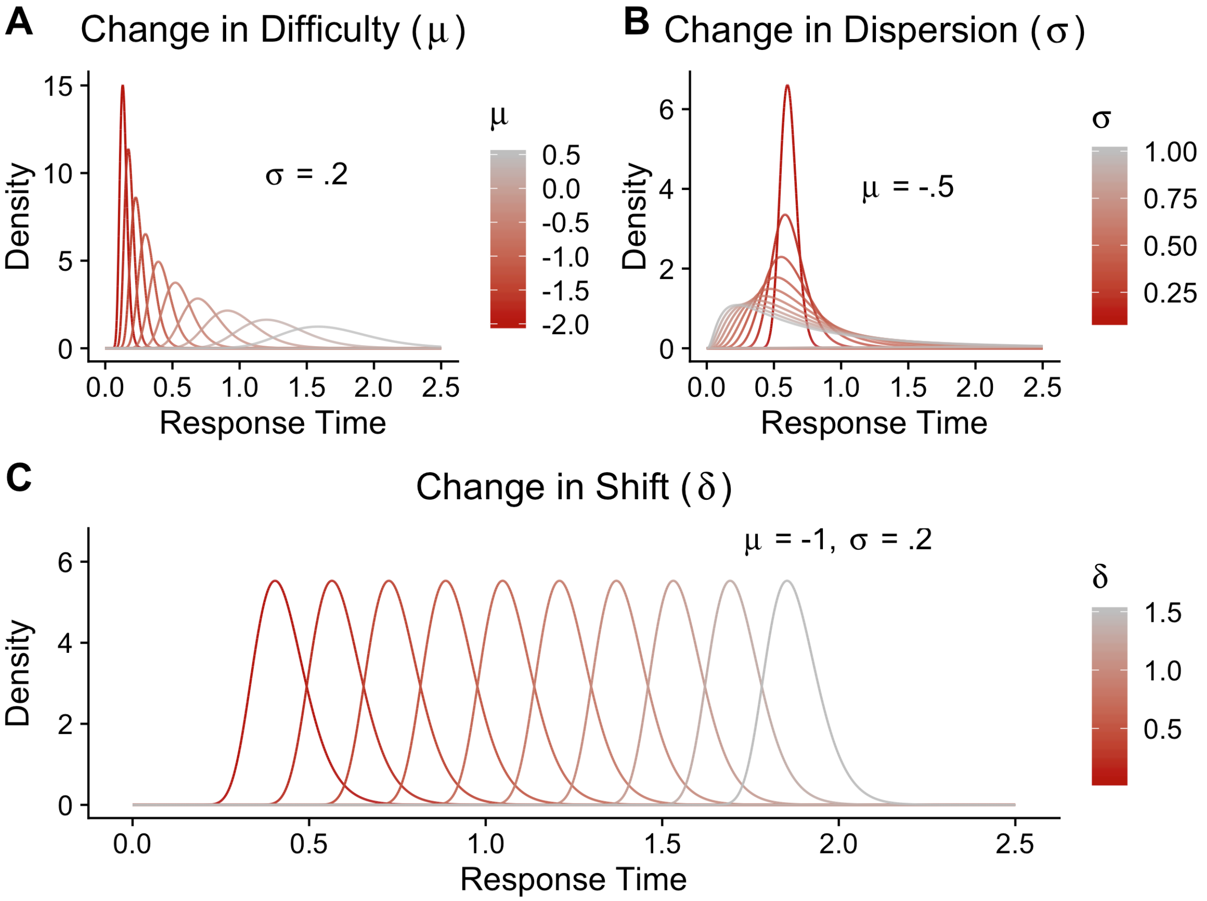

Figure 3. Lognormal and shifted lognormal generative distributions. (A) For the lognormal and shifted lognormal distributions, changes in the μ parameter (interpreted as “stimulus difficulty”) produce changes in both means and variances of response time distributions. (B) The σ parameter controls dispersion (interpreted as “decision noise”); changes in σ affect means and ranges of likely response times, but medians remain constant. (C) For the shifted lognormal distribution, the shift parameter δ translates the lognormal distribution forward in time without

changing the shape of the response time distribution. The lognormal distribution is a special case of the shifted lognormal distribution wherein the shift parameter (δ) is set to 0. The shift parameter is interpreted as “non-decision time”, as it can capture individual differences in components of the data-generating process that are not relevant to decision-making but nevertheless produce variation in mean response times (e.g., visual encoding time, motor responses).

More generally, many of the advantages of generative modeling in the context of behavioral data are derived from their ability to capture distributional information about responses or response times beyond the first and second moments (mean and variance) as well as how these distributions change across people. There are several clear examples of cases where the wrong analysis—based on mean response times, or ignoring person-level information in favor of modeling aggregate data—has resulted in the wrong conclusions about psychological processes (Evans et al, 2018; Heathcote et al, 1991; Andrews & Heathcote, 2001; Kvam et al, 2022). In the supplementary material, we examine the importance of response and response time distribution by comparing two methods of analysis: one that analyzes reliability using mean response time differences and one that analyzes reliability using distribution-level information (K-L divergence) on a simulated data set where the true correlation between parameters at different timepoints is perfect (r = 1). These simulations and analyses illustrate that using distribution information commonly allows us to account for 2-16 times the variance in behavior across time points—meaning there is a great deal of room for improvement upon the use of atheoretical summary statistics as behavioral models in principle.

A Generative Group Model

With the behavioral model defined, the next step is to estimate each person's behavioral model parameters, which we use in combination with a group model to make inferences about the target population that our sample represents. Because the group model makes assumptions about the distribution of person-level behavioral model parameters, our choice of group model should reflect our substantial knowledge to the same extent that our choice of behavioral model does. Notably, the group model implied by the two-stage approach it is not a "safe" default—the implicit generative model it assumes is highly constrained and most likely inappropriate, assuming that either the within-person sample size (number of trials) is infinite/arbitrarily large, or that the within-person standard deviation of response times is approximately zero (the latter is depicted in Figure 4; see the Methods section and Supplementary Note 2 for details). In the case of the Stroop task, these assumptions respectively correspond to us collecting an infinite number of responses in both conditions before taking the mean difference in response times, or alternatively, that every response time is exactly the same for a given person within conditions. From the perspective of the group model, this latter assumption is equivalent to assuming that the distribution of variability in response times across people is a point mass at 0—or that every person shows no variability in response times whatsoever. Of course, we know *a priori* that neither of these assumptions are met—a generative group model should allow for us to relax these overly restrictive assumptions, freeing up parameters to vary in ways that are consistent with our theories or background knowledge (see Figure 4).

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 person 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 people who are themselves nested within groups, as well as designs where amounts of person-level data are limited. Key for our purposes, *hierarchical Bayesian analysis* offers us the flexibility to specify arbitrarily complex group models, including those that allow us to estimate the shifted lognormal model parameters across people in a hierarchical fashion. A history of work in both mathematical statistics and psychometrics shows that such models confer more precise estimates of the “true” person-level parameters by pooling information across people (Efron & Morris, 1977; Gelman, 2006; Williams et al, 2020). In our case, uncertainty arises from both the inherent probabilistic nature of behavior (i.e., variation in response times within persons, or σ), and the limited number of trials that we observe within persons/conditions.

Here, we assume that person-level behavioral model parameters follow correlated multivariate normal distributions across test-retest sessions, and we use this same specification for all three behavioral models (see the Methods section for mathematical details and visual depictions). The test-retest correlation obtained from this generative model spanning from within- to between-person variation is akin to the “true” or “dis-attenuated” correlation that one obtains when using classical corrections for attenuation due to unreliability (Haines, Sullivan-Toole, & Olino, 2023; Williams et al, 2020; Rouder & Haaf, 2019). With the group model specified, we now have three fully specified generative models. Next, we use these models to 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 resolving the theoretical gap and providing practical benefits in terms of the (apparent) reliability of behavior.

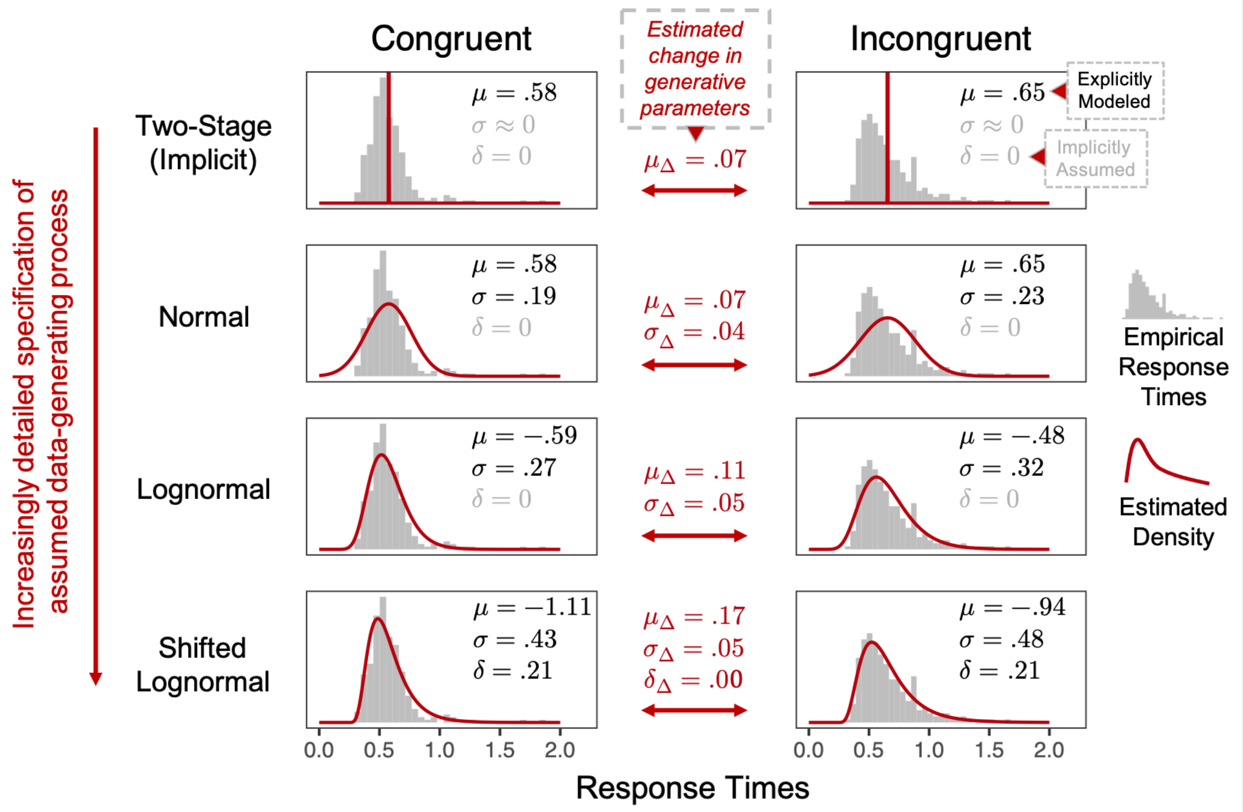

Figure 4. Building generative models consistent with theory. The two-stage approach is often used by default and chosen without reference to an underlying theory. By contrast, the generative approach begins with a model of behavior at the person level (e.g., a lognormal distribution), and inferences are made by interpreting changes in model parameters across conditions, people, 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 in Figure 3). 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?

Method

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 to show how generative modeling principles generalize beyond response time data. We include the delay discounting results in the main text for completeness, but relegate details on model specification to the supplementary materials for brevity. 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 March 2023), the collective citation count of the original research pertaining to these tasks is over 66,000 (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 supplementary materials. 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 supplementary material]), 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.

Data Analysis

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[3]. 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 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; Parsons, 2020). However, in the supplementary materials, we provide a more detailed examination of how to incorporate contaminant response times with a simple mixture model, and our results show that this simple adjustment improves both model fit and reliability (see Figure S10).

Two-Stage Approach

[3] 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.

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.

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[4]. 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.

Delay Discounting Task. 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 supplementary material for details)[5]. We compare these estimates to a hierarchical Bayesian estimation approach described below.

Generative Modeling Approach

[4] 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.

[5] 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 supplementary material for details).

Response Time Models. First, we would like to describe in more detail how the behavioral model underlying the two-stage approach can be viewed as a special case of the normal generative model. The sample mean is the analytical maximum likelihood estimator for the μ parameter in a normal distribution. However, the two-stage approach assumes the mean is estimated without measurement error. There are only two cases in which this assumption is valid. Specifically, recall that the standard error on the sample mean ($\sigma_{\hat{\mu}}$) (and correspondingly, the μ parameter of a normal distribution) is $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$. In our case, σ corresponds to the standard deviation of a person's response time distribution ($\sigma_{i,c,t}$ in Equation 2), and n to the number of trials they underwent in the given condition. Assuming no measurement error is equivalent to assuming that $\sigma_{\hat{\mu}} = 0$, which only occurs as either $\sigma \rightarrow 0$ or $n \rightarrow \infty$. Of course, we know *a priori* that each person's distribution of response time will have variability from trial-to-trial (i.e. $\sigma > 0$), and that we have a limited number of trials (i.e. $n \ll +\infty$).

Given the generative behavioral models defined 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 task condition at each of the two sessions t are drawn from a normal group-level distributions with unknown means and standard deviations:

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

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 Figure 8). Although we do not demonstrate it here, the group-level model (i.e., Equation 7) can be extended to 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:

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

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 supplementary material 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 supplementary material.

Note that we tested multiple different group-level models to determine how sensitive our results were to changes in group-level generative assumptions. We report details in the supplementary materials, but offer a brief overview of results here. First, we used parameter recovery simulations to show that we can accurately recover the "true" underlying test-retest correlation using the full generative models as described in the main text (see Supplementary Note 4 and Figure S2). Second, we tested an alternative group-level model wherein both the person-level base and change parameters were drawn from separate multivariate normal distributions, as opposed to only the change parameters (the "Joint Separate" model in Supplementary Note 7 and Figure S8). Finally, we tested another group-level model wherein we directly estimated the $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ parameters as opposed to estimating baseline and change parameters. For this model, we assumed that person-level parameters were drawn from a single multivariate normal distribution (but separate for μ versus σ) across conditions and sessions (the

"Joint Single" model in Supplementary Note 7 and Figure S9). Both models produced very similar results to the model specification presented throughout the main text.

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 choice sensitivity (c) parameters (for details, see supplementary materials).

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[6]. 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 supplementary material.

Results

[6] 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 supplementary material.

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.

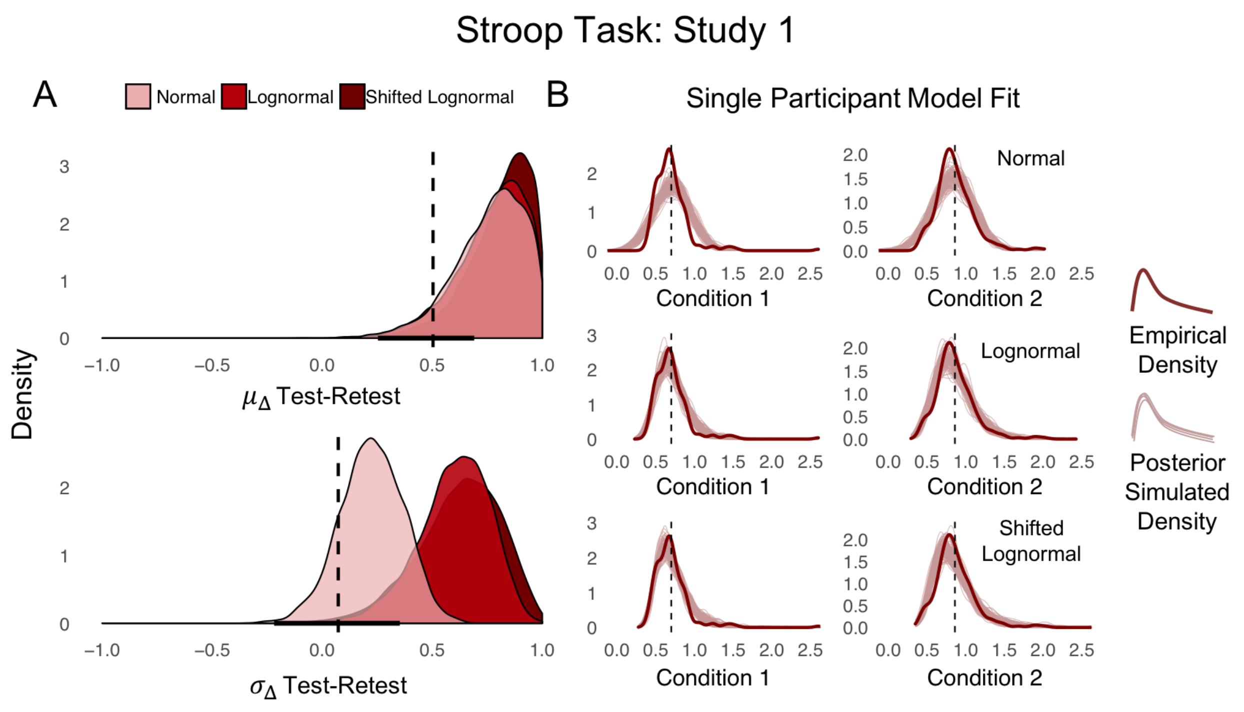

Figure 6. Estimates of reliability for each modeling approach. (A) Posterior distributions for the test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 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 participant.

We now direct attention to 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 (Kruschke, 2014). 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 statistical 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. By 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 supplementary materials 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 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 supplementary materials (see Figures S2-S6).

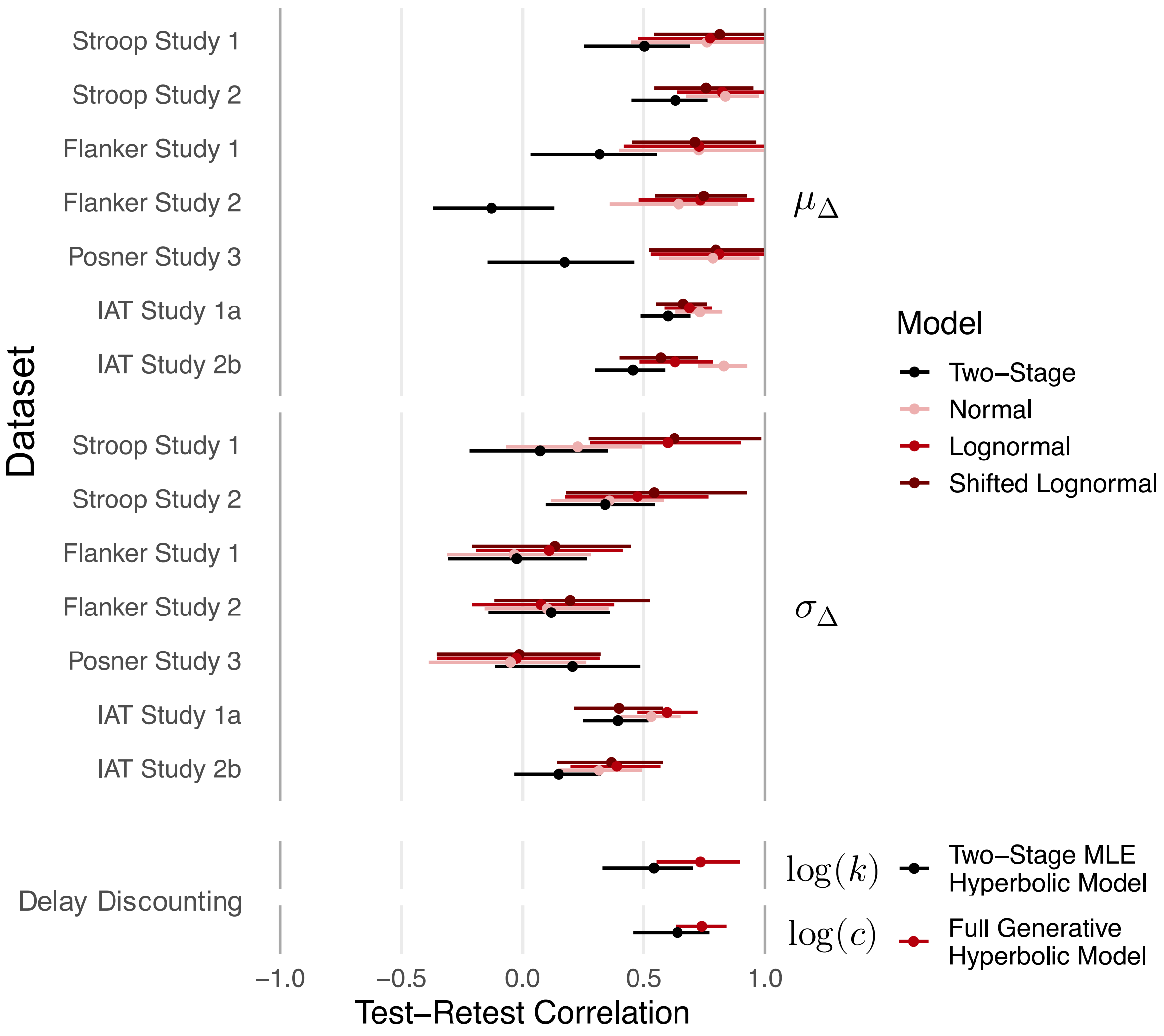

Figure 7. Test-retest correlations for all tasks re-analyzed in the current study. The points and intervals are posterior means and 95% highest density intervals on the test-retest correlation estimated per Equation 8. See Table 1 and the supplementary materials for more detailed figures and description of each task.

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]
		σ_Δ	.07	[-.22, .37]
	Shifted-Lognormal	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 supplementary materials.

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 an assumed data-generating model and not the task per se.

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 7 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 7. 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 8). 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 supplementary materials).

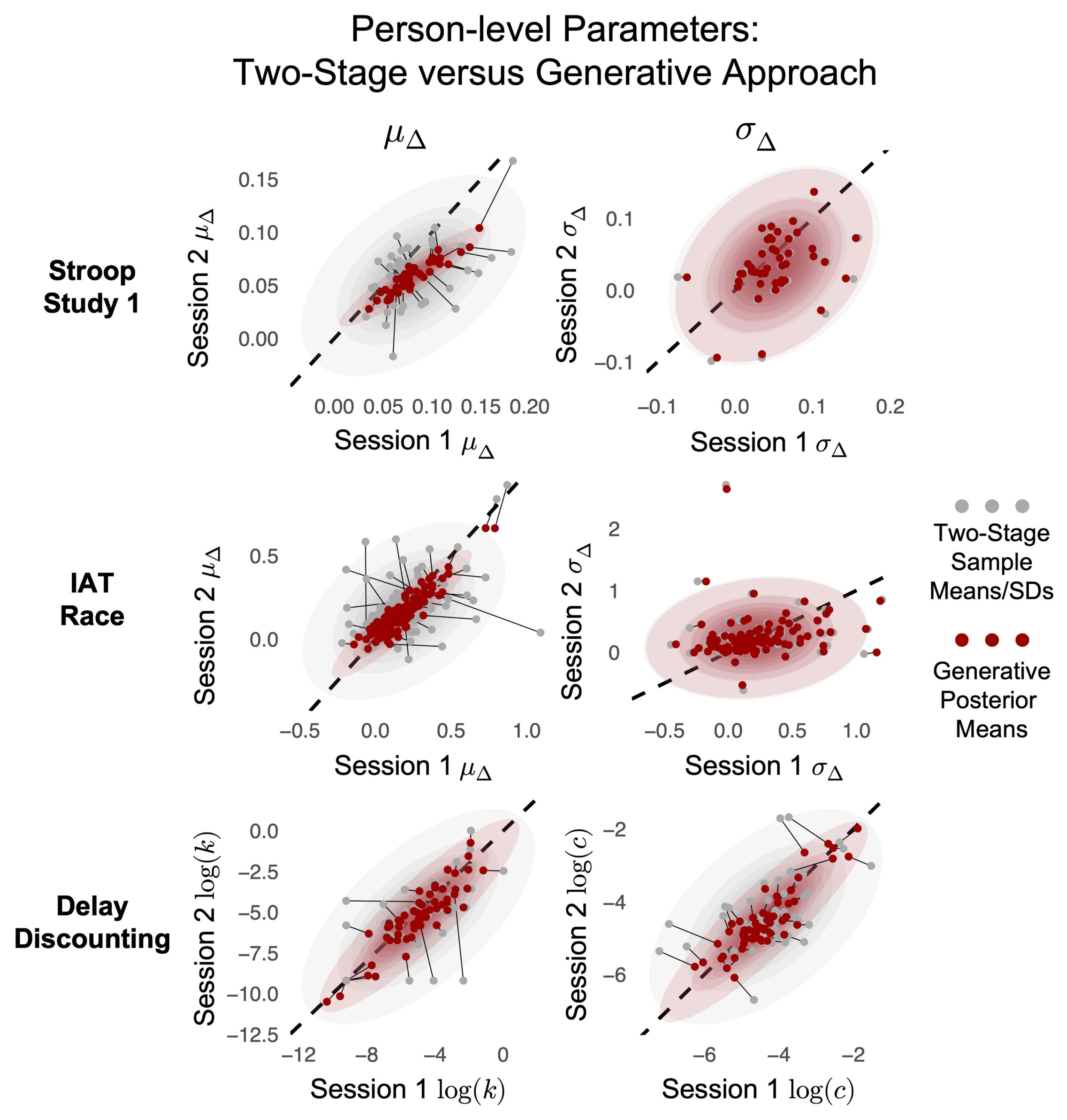

Figure 8. Relationship between two-stage estimates and generative model parameters. For the response time models (Stroop & IAT tasks), two-stage estimates are the sample mean and standard deviation contrasts for each participant and retest session (i.e. estimates from the summary statistic approach); 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.

Discussion

Our results illustrate how better generative assumptions can improve both the reliability of our parameter estimates and the validity of conclusions we can draw from behavioral data. Further, they run counter to mounting claims that behavioral tasks are poorly suited for developing theories of individual differences, which has been (erroneously) attributed to the behavioral measures themselves as opposed to the methods used to derive inferences from them (e.g., Dang et al., 2020; Enkavi et al., 2019; Gawronski et al., 2017; Hedge et al., 2017; Wennerhold & Friese, 2020). Comparisons across five popular tasks demonstrate how incorporating assumptions about data-generating processes allows for a more precise characterization of individual differences in behavior by better isolating sources of variability across experimental conditions. 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.

By contrast, traditional methods of analyzing behavioral data are largely atheoretical—they make implicit data-generating assumptions of which researchers seem not to be aware, 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. Importantly, this two-stage approach inappropriately assumes that the resulting individual-level summary measures contain no measurement error.

Further Improvements

Beyond the models we used in the current study, 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 the mechanisms underlying 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; Haines, Kvam, & Turner, 2023; 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. One extension of the shifted lognormal model is the lognormal race model (Rouder et al., 2014), that jointly describes choice response time distributions via a competitive shifted lognormal accumulation. 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 (DDM) 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 simplification of the DDM, such as the EZ diffusion model (Wagenmakers et al, 2007), and then compare parameters between conditions (congruent, incongruent). Another, more recent method for applying even complex models is to use machine learning methods for training neural networks to perform parameter estimation (Radev et al, 2020; Sokratous et al, 2023). Trained networks can then be embedded in simple point-and-click online tools that make generative modeling more accessible than ever (Kvam et al, 2023).

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 effects 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).

Early indications are that these models, which incorporate even more generative assumptions, are able to produce even better reliability and predictive validity than the lognormal models used in this paper (see Kvam et al, 2023). The message is clear: deeper theories and more appropriate models – ones that incorporate the structure of the task, stimuli, and responses – yield better conclusions about individual differences in performance. The trade-off that a theoretician or modeler must make is then between the richness and complexity of a model, and thus its interpretability and proclivity to overfit the data, and its ability to provide a simple, coherent, and parsimonious explanation of behavior (Myung, 2000). Beyond allowing for model comparison and quantitative evaluations, a coherent parametric structure arising from principled generative assumptions makes models and theories more straightforward to communicate and discuss.

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; Haines, Kvam, & Turner, 2023; Muthukrishna, & Henrich, 2019; Regenwetter & Robinson, 2017; Ross et al., 2020; Rotello et al., 2014; Szollosi & Donkin, 2019).

As we illustrated here, generative modeling is an iterative process, and throughout model development, model parameters can always be assessed to determine their psychometric properties. Although we focused on test-retest reliability, there are many other properties worth exploring including parameter identifiability (Spektor & Kellen, 2018), parameter recovery (Ahn et al, 2011; Haines et al, 2018; Miletic et al, 2017), tests of selective influence (a form of construct validity where experimental manipulations are related to expected changes in specific parameter values; Criss, 2010), as well as parameter convergence among/between behavioral models and models derived at other levels of analysis (Turner et al, 2017; Haines et al, 2020; Haines, Kvam, & Turner, 2023; Kvam et al, 2021). Bayesian analysis facilitates joint estimation of all model parameters and their hypothesized relations, thus allowing for proper calibration of uncertainty in key parameters, such as test-retest correlations. Although we focused on what we believe are uncontroversial generative assumptions in the generative models we tested here, we also discuss extensions to these generative models, including using more sophisticated evidence accumulation models of choice and response time behavior, in the supplementary materials. These models can be further improved by methods that take advantage of adaptive experimental designs to further maximize informativeness of behavioral data in Supplementary Note 9 (Cavagnaro et al, 2011; Myung et al, 2013; Yang et al, 2020).

Learning More about Generative Modeling

One potential barrier to the use of generative models is that they necessitate a single joint model across all data modalities when making inference. In the context of the current study, this means that one could not simply fit the generative models proposed in our work to new data, extract the person-level parameters, and expect that such parameters should show better correspondence with other individual difference measures. Instead, one would need to construct a joint model of the behavioral and external measures of interest, which can be challenging without extensive experience with generative models. However, advances in computational statistics have recently made generative modeling much more accessible.

We anticipate that generative modeling approaches 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, so they can apply these techniques to their own work. Resources include introductions to the philosophy and utility of generative or computational modeling for theory development (var Rooij & Baggio, 2020; Guest & Martin, 2020), tutorials on building generative models from first principles (Wilson & Collins, 2019; van Rooij & Blokpoel, 2020), practical textbooks that combine introductions to both behavioral model development and hierarchical Bayesian modeling (Farrell & Lewandowsky, 2018), tutorials and case examples on developing joint generative models of behavior and brain activity (Turner et al, 2017; Palestro et al, 2018), open source R and Python software packages that allow both beginners and advanced users to apply popular generative models of behavioral to their own data using hierarchical Bayesian modeling (Ahn et al, 2017; Mathys et al, 2014; Wiecki et al, 2013), and now machine learning tools for estimating and comparing generative models (Radev et al, 2020; 2021; Soktratous et al, 2023). These and related sources make it clear that there is much more to building sound generative models than just capturing the shape of empirical data distributions, although it is clear from our results that even minimal assumptions can drastically improve the measurement of individual differences. Computational models that incorporate domain-specific assumptions about generative processes are a good example (Guest & Martin, 2020; Wilson & Collins, 2019; Jarecki et al, 2020), which facilitate a level of mechanistic inference not provided by the simple, broad-based behavioral models used in the current study. We hope our message has come through clearly—that the landscape for generative model development is vast, whereas traditional two-stage summary approaches are inherently limited in scope and application.

Concluding Remarks

We end with a cautionary yet hopeful note: As history has revealed, heuristic use of summary statistics absent generative models impedes scientific progress, leading us down paths we would have never explored had we been made aware of the implicit assumptions that directed us there. Although our generative models may be wrong or mis-specified, their explicitness forces us to specify theoretical assumptions regarding how behavior arises, thereby requiring us to spend time thinking about the mechanisms that underly the brain, behavior, and their inter-relations. By embracing their incompleteness, we can strive to build generative models that are precise and thus meaningfully incorrect, rather than relying on vague, heuristic theories whose verisimilitude can be deceptive because of the theory-description gap. Identifying and knowing where our assumptions are wrong provides a natural path toward deepened understanding of the mechanisms underlying behavior. In this vein, the generative models we propose here are not intended as a final step. Quite the opposite—we encourage researchers to start from such models, and iteratively extend them to better meet the particulars of their theory, task, or application.

Funding Details

T.P.B. was supported in part by Grants UH2DE025980 and UL1TR002733 from the National Institutes of Health, P.D.K. was supported in part by a grant from the National Institutes of Health (R01DA021421) and a SEED grant from the University of Florida Informatics Institute, W.-Y.A. was supported in part by Artificial Intelligence Graduate School Program Grant (No. 2021-0-01343) and the Seoul National University Creative Pioneering Researchers Program, and B.M.T. was supported by a CAREER award from the National Science Foundation. This work was accepted for presentation at the 32nd annual convention for the Association for Psychological Science, presented at the 62nd annual meeting of the Psychonomic Society, and was previously hosted on a preprint server PsyArXiv at https://osf.io/preprints/psyarxiv/xr7y3.

Disclosure Statement

The authors declare no conflicts of interest.

Data Availability Statement

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

Arslan, R. C., Brümmer, M., Dohmen, T., Drewelies, J., Hertwig, R., & Wagner, G. G. How people know their risk preference. *Scientific Reports, 10*, 1-14 (2020).

Bahg, G., Evans, D., Galdo, M., and Turner, B. M. (2020). Gaussian process linking functions for mind, brain, and behavior. *Proceedings of the National Academy of Sciences*, *117*, 29398-29406.

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

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, *5*(6), 601–609. https://doi.org/10.1016/j.bpsc.2019.12.019

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

Efron, B. & Morris, C. Stein's paradox in statistics. *Scientific American 236,* 119–127 (1977).

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.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222, 309–368.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. Science Advances, 3, e1701381.

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. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics* **48,** 432–435.

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., (2020). Reliability_2020. GitHub repository, https://github.com/Nathaniel-Haines/Reliability_2020.

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., Myung, J. I., Turner, B. M., & Ahn, W.-Y. (2020). Anxiety modulates preference for immediate rewards among trait-impulsive individuals: A hierarchical Bayesian analysis. *Clinical Psychological Science, 8*, 1017-1036. doi:10.1177/2167702620929636

Haines, N., Kvam, P. D., & Turner, B. M. (2023). Explaining the description-experience gap in risky decision-making: Learning and memory retention during experience as causal mechanisms. *Cognitive, Affective, and Behavioral Neuroscience. 23*, 557 – 577. doi: 10.3758/s13415-023-01099-z.

Haines, N., Sullivan-Toole, H., & Olino, T. (2023). From Classical Methods to Generative Models: Tackling the Unreliability of Neuroscientific Measures in Mental Health Research. *Biological Psychiatry: Cognitive Neuroscience & Neuroimaging*. doi:10.1016/j.bpsc.2023.01.001

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

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2020). Self-reported impulsivity does not predict response caution. *Personality and Individual Differences 167,* 110257.

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

Kvam, P. D., Romeu, R. J., Turner, B. M., Vassileva, J., & Busemeyer, J. R. (2021). Testing the factor structure underlying behavior using joint cognitive models: Impulsivity in delay discounting and Cambridge gambling tasks. Psychological Methods, 26(1), 18–37. doi:10.1037/met0000264

Kvam, P. D., Irving, L. H., Sokratous, K., & Smith, C. T. (2023). Improving the reliability and validity of the IAT with a dynamic model driven by similarity. *PsyArXiv.*

Kvam, P. D., and Turner, B. M. (2021). Reconciling similarity across models of continuous selections. *Psychological Review*, 128(4), 766–786.

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.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287–330.

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. *Perspectives on Psychological Science (2021).* doi:10.1177/1745691620974769

Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of significance tests. Psychological Bulletin, 85(2), 405–409. doi:10.1037/0033-2909.85.2.405

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

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 82(1), 85.

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. Psychological Bulletin, 83(5), 776–777.

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

Radev, S. T., Mertens, U. K., Voss, A., & Köthe, U. (2020). Towards end-to-end likelihood-free inference with convolutional neural networks. *British Journal of Mathematical and Statistical Psychology, 73*(1), 23-43.

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P. C. (2021). Amortized bayesian model comparison with evidential deep learning. *IEEE Transactions on Neural Networks and Learning Systems.*

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

Ross, C. T., Winterhalder, B. & McElreath, R. (2020). Racial Disparities in Police Use of

Deadly Force Against Unarmed Individuals Persist After Appropriately Benchmarking Shooting Data on Violent Crime Rates. *Social Psychological and Personality Science 16,* 194855062091607.

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

Sokratous, K., Fitch, A. K., & Kvam, P. D. (2023). How to ask twenty questions and win: An automated model of risk preferences from small samples of willingness-to-pay prices. *PsyArXiv.*

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

Strickland, J. C., & Johnson, M. W. (2021). Rejecting impulsivity as a psychological construct: A theoretical, empirical, and sociocultural argument. *Psychological Review, 128*, 336–361.

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

Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Pedroni, A., ... & Mata, R. (2020). Brain–behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience, 14*, 194.

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

Ulrich, R., & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*, 513–525.

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., Martin, S. R., DeBolt, M., Oakes, L. & Rast, P. (2020). A fine-tooth comb for measurement reliability: Predicting true score and error variance in hierarchical models. *PsyArXiv Preprint*. doi:10.31234/osf.io/2ux7t

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

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences,* 45, e1.Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). ADOpy: A Python package for adaptive design optimization. *Behavior Research Methods*. doi:10.31234/osf.io/mdu23

Zorowitz, S., & Niv, Y. (2022). Improving the reliability of cognitive task measures: A narrative review. *PsyArXiv.* doi:10.1016/j.bpsc.2023.02.004.