Uncertainty In Cognitive Science

Jesper Fischer Ehmsen
School of Communication and Cognition, University of Aarhus,
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

03/06/2024
Cognitive Science MA.
Riccardo Fusaroli

Summary

Understanding human cognition and behavior is the primary aim of Cognitive Science. To achieve this, quantitative methods are frequently used. When these quantitative methods are employed assumptions and simplifications must be made, which are embedded in the models used. This thesis explores some of these assumptions herein, propagation of uncertainty and validation of the models themselves, in a simulated settings. Uncertainty propagation stems from the fact that when quantitative data is collected uncertainty is embedded in these measurements. A failure to account for these uncertainties can have a substantial impact on the inferences made based on these measures. With a focus in how uncertainties in statistical and cognitive models propagate, the thesis will further investigate how the validation process of cognitive models can be improved, by embracing and quantifying the inevitable uncertainties associated with our cognitive models. The framework proposed revolves around simulating agents with known properties which are then fitted to the cognitive model to access the model's ability to detect these simulated properties, while accounting for uncertainties embedded in these estimations. The thesis will explore these considerations through the three-parameter psychometric function, a widely used cognitive model.

It will be shown that a correlational approach to determine internal model validity is at best quite insensible compared to a more sophisticated approach based on the intra class coefficient. The thesis will then demonstrate how uncertainties in parameter estimates and in these two metrics can be minimized through more sophisticated methods. This will be done without a need for increasing the number of trials or subjects, which is the standard approach. In this regard the thesis highlights two important methods for minimizing uncertainty. Firstly, optimizing the design of the experiment such that each trial will contain the most information possible. Secondly, incorporating already collected data, such as reaction times, into the cognitive model as a means of decreasing the uncertainties in the measures of interest. The thesis goes on to explore and re-analyses published data using the psychometric function. Here it is demonstrated that incorporating structural assumptions of how the data was collected, as well as incorporating reaction times, does not only decrease uncertainty in the reliability, but also describes the data well. Lastly the thesis highlights and demonstrates novel opportunities for conducting power analyses using. Here it is demonstrated, based on the re-analysis of the published data, that by using simulations it is possible to build predictive-models that accurately estimate the number of trials, subjects and effect-size needed for the psychometric function to find group differences in a particular parameter estimate.

This highlights an avenue for researchers building cognitive models to inform others about their models' strengths and weaknesses in estimating parameters of interest. Lastly with this novel way of generalizing power analyses it is shown that the number of trials in a cognitive science experiment is highly relevant in estimating the psychometric model's ability to pick up on group differences in parameters, which is completely neglected by commonly used power analysis soft wares.

Acknowledgement

part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

All scripts, code, models, and simulated data can be found on the Github page. Here all analyses as well as the entire manuscript are documented.

Github: https://github.com/JesperFischer/Master-thesis

Introduction

Most scientific inquiry revolves around measurements of the physical world, whether that it is the time it takes for a cup to fall to the ground or for a person to react to a visual stimulus on a computer screen. These measurements will be associated with uncertainty, as repeatedly measuring the same thing will result in different measurement values. This makes uncertainty a fundamental aspect of scientific inquiry and theories. It is therefore also a role of science to quantify such uncertainties.

In this thesis, I will investigate uncertainty handling in Cognitive Science and provide ways to properly account for these uncertainties. The thesis will rely on Monte Carlo simulations, which provide a robust method for accounting for uncertainties, in analyses and models. Specifically, the thesis will introduce a partially novel approach to testing and validating the parameters of cognitive models. This approach is going to be used with a focus on the psychometric function, a commonly used cognitive model. It will be shown that the parameters of the model and their uncertainties can be reduced by several different interventions. These include optimization of the experimental design, and by incorporating additional information already available in most experiments. The thesis will demonstrate how modeling and incorporation of such concepts can decrease uncertainty in the estimation of parameters of already published data. Lastly, using this re-analysis of published data, the thesis will highlight opportunities to conduct power analyses, utilizing a novel modeling framework. This framework can help make power analyses for a particular model more rigorous by incorporating and propagating uncertainty. Comparison of this novel power analysis framework will be compared to popular tools such as G*power.

Uncertainties in science

Science can be thought of as a systematic way to organize knowledge in hierarchies, leading to testable hypotheses. Knowledge can be hard to define, but most often it is something that is achieved though experience. Imagine a cup being dropped, most people will have the knowledge that it will fall towards the ground and reach our foot at a particular speed because of our previous experiences with dropping a cup. This is to say that knowledge is the relationships that we believe to be true with differing amounts of certainty. Even though we might say we are completely certain that the cup will fall to the ground, and reach it at a particular speed, this assumption is only true most of the time. Given that the natural world is bound on probabilities, complete certainty is unwarranted, both in the assumption of the cup hitting our foot, but especially the speed at which

it hits our foot with. Here, the interest is not in the unforeseen events, but instead in the predictability and (un)certainty of the expected. Science would normally be interested in the acceleration of the cup and the uncertainty in this estimate. Scientists have shown that objects dropped on Earth will accelerate towards the ground with an acceleration of $9.81 \frac{m}{s^2}$ (Johannes & Smilde, 2009). However, this number does not mean anything without an estimate of the uncertainty, and an understanding of the assumptions entailed with these numbers The first proposition is well studied and the 95% confidence interval of the value is estimated to be $[9.78; 9.84] \frac{m}{s^2}$ (Johannes & Smilde, 2009). The second proposition is also well studied, as we believe that the density, the shape, and weight of the cup if dropped outside of a vacuum are important. To estimate this acceleration, measurements must be made. These measurements include the distance the cup travels and the time it takes. With these measures of distance and time, uncertainty is introduced and propagated to get an estimate for the acceleration, but also the uncertainty associated with it. This example highlights two main points that this thesis will explore. Firstly, uncertainties are organized in hierarchies and are just as important as beliefs. Secondly, taking these uncertainties seriously and herein estimating and propagating them should not be a choice, or something that can be avoided. After examining these potential issues, the thesis will propose methods for incorporating and minimizing uncertainties through simulations. The goal is to shed light on the often-overlooked uncertainties in the data collected on human behavior and cognition, while also offering strategies for addressing them. It will be argued that accounting for uncertainties is more important than ever, especially in research of complex systems such as humans. This urgency arises due to the availability of computational resources, having made it possible to easily develop more sophisticated analyses and models that have dependencies on lower-level analyses. This hierarchical dependency underscores the necessity for proper uncertainty handling. Without propagating uncertainties, resulting estimates or beliefs become overly confident. Furthermore, these computational resources enable uncertainty propagation without a deep understanding of the underlying mathematics, thus making it accessible to most researchers, with some coding experience. To effectively convey both the statistical models, as well as the underlying uncertainties associated with doing computations on data, the thesis will start by exploring different types of uncertainty.

Levels of uncertainty and uncertainty propagation

I will here broadly define three different types of uncertainty, namely measurement, estimation, and test-retest uncertainty. These definitions are going to be used throughout the thesis. These definitions are not exhaustive and will be centered around how experimental studies in cognitive science are conducted, from data collection to data analysis. Before examining these three types of uncertainty it is imperative to acknowledge that uncertainties can be defined in hierarchies, and that uncertainty propagates through these hierarchies. This uncertainty propagation means that as calculations are completed based on measures with uncertainty, the uncertainty propagates to the results of the calculations. Simulations will be used to show how uncertainty propagation can be understood, and handled, without a need for rigorous mathematical proofs. For a more mathematical treatment see (Saccenti et al., 2020).

The lowest level of uncertainty is in the measurements themselves i.e. measurement uncertainty. Measurement uncertainty reflects the uncertainty in how well one can, for instance measure the reaction time on a computer or the time it took a falling cup to reach the ground. This level of uncertainty is sometimes neglected in Cognitive Science when applying statistical models, because they are sometimes thought to be minuscule. An example could be reaction time tasks, commonly used in cognitive science (Sternberg, 1969; Stone, 2014). These reaction time experiments may or may not have minuscule measurement uncertainty, depending on the experimental setup (Crocetta & Andrade, 2015; Holden et al., 2019; Ohyanagi & Sengoku, 2010). This is not to say that cognitive scientists do not care about measurement uncertainty, as moving towards more sophisticated measurement methods is an ongoing endeavor. For instance, using better and more sophisticated computers to measure reaction times (Crocetta & Andrade, 2015). Minimizing this kind of uncertainty most often revolves around getting better tools to measure the variable(s) of interest.

The next level of uncertainty arises when calculations are done on data. This is true for calculations done on measures with and without measurement uncertainty. This uncertainty will be referred to as estimation uncertainty. Estimation uncertainty is often quantified by the statistical model, such as in the form of a standard error of a correlation coefficient (CC) or the width of a posterior distribution of a parameter. Estimation uncertainty is always present and influenced by measurement uncertainty. Minimizing estimation uncertainty is a primary concern for scientists, particularly as many cognitive science experiments center around null hypothesis testing. This null

hypothesis testing will typically involve testing whether parameter estimates include a particular value, typically, 0. Therefore, minimizing estimation uncertainty serves to highlight underlying effects. The standard approach to minimize estimation uncertainty is to get more data, provided it originates from the same population. In cognitive science this often includes increasing the number of trials and/or subjects to get more precise estimates, thereby reducing estimation uncertainty. However, this method of minimizing estimation uncertainty is not without its drawbacks or inherent uncertainties. Increasing the number of trials in a cognitive task may inadvertently exacerbate estimation uncertainty. This can occur due to various factors, including boredom, habituation, fatigue, and lack of engagement, particularly when experimental tasks become very long (Jeong et al., 2023; Meier et al., 2024). Increasing the number of trials could also make participants more prone to switching between cognitive strategies. If these switches are not properly accounted for in the analysis, they might be interpreted as additional noise by the model and its parameters. Increasing the number of subjects will decrease estimation uncertainty on the population level estimates, if the sample population is homogeneous. The tradeoff between subjects and trials in an experiment is therefore quite important to minimize estimation uncertainty, but also minimize the overuse of resources. In addition to these traditional approaches, there are alternative methods for minimizing estimation uncertainty, such as modifying the task design or including more information in the model (Baldi Antognini et al., 2023; Stone, 2014). This modification or optimization of the task design involves individualizing the task design, to maximize the informativeness of each presented stimulus. This task design optimization is frequently used in psycho-physical experiments, where adaptive algorithms like PSI, QUEST and ADOPY are used to select stimuli that minimizes the uncertainty in the estimated parameters (Prins, 2013; Watson, 2017; Yang et al., 2021).

The final layer of uncertainty to be addressed is the test-retest uncertainty, which arises from the variability of parameter estimates over time. In cognitive science the additional uncertainty on retesting the estimates of a parameter stems from the fact that humans vary over time. This variation is influenced by various behavioral factors like learning, but also psychological factors such as mood and arousal (Schurr et al., 2024).

Measurement uncertainty in a correlational design

In order to demonstrate and account for these types of uncertainties, the thesis will here show how it is possible to add measurement uncertainty to a correlation analysis. This will be accomplished

by using simulations. A correlation analysis is chosen as the example for three main reasons. Firstly, a significant portion of published literature in Cognitive Science revolves around conducting correlational analyses on measures that have quantifiable uncertainties. These measures typically involve estimated parameters, structural properties of the brain or even reactions times (Berker et al., 2016; Luijcks et al., 2015; J. Wu et al., 2021). Secondly, this example can be generalized, such that instead of estimating a CC on data with measurement uncertainty, it could equivalently be done for more complex models. Supplementary analysis 1 shows how these uncertainties interact in a linear model, in a test-retest paradigm. Lastly, this type of analysis is going to be used extensively throughout the thesis serving as a primer for the upcoming sections.

Initially, a the special case where little to no uncertainty exists in the data is examined. This case allows for comprehension of the estimation uncertainty of the CC in isolation, without the influence of measurement uncertainty. Analytical solutions exist to calculate this estimation uncertainty and is incorporated in most statistical software (Makowski et al., 2023). However, this can also be shown by simulations, or more accurately by re-sampling. To estimate the uncertainty in the CC, the data is re-sampled with replacement, a technique known as bootstrapping (Efron, 1983). Iterating this process of resampling gives a distribution of CCs, which with enough iterations will converge towards the analytic solution. It is generally recommended to have least 30 data points, to ensure convergence to the analytical solution (Efron, 1983; Efron & Tibshirani, 1994; C. F. J. Wu, 1986). For the simplest case of recalculating the CC (without measurement uncertainty) and its uncertainty, the process might seem tedious compared to using the direct analytic solution. However, once implemented and understood, this approach allows for adding and propagating all types of uncertainty, coming from various kinds of distributions. One of the advantages of having an analytic solution to the case of estimating the uncertainty of the CC is to ensure proper setup of code and scripts. This first step therefore serves as a validation step, before venturing into areas where analytic solutions are scarce or nonexistent.

The initial step is therefore to demonstrate that the two approaches of simulating and analytically estimating the uncertainty of the CC is identical across ranges of correlations and sample sizes. To achieve this, simulated data from a multivariate normal distribution with the following parameters are produced.

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x \cdot \sigma_y \cdot \rho_{xx} \\ \sigma_x \cdot \sigma_y \cdot \rho_{xx} & \sigma_y^2 \end{bmatrix}\right)$$

Where:

$$\mu_x = 50, \quad \mu_y = 100, \quad \Sigma = \begin{bmatrix} 10^2 & 10 \cdot 10 \cdot \rho_{xy} \\ 10 \cdot 10 \cdot \rho_{xy} & 10^2 \end{bmatrix}$$

The multinormal distribution produces random variables with a means μ_x a standard deviation σ_x and crucially, with a CC between the random variables ρ_{xy}. This distribution is ideal for understanding how the CC changes as it is a parameter of the distribution. Demonstrating the equivalence of bootstrapping and the analytic solution to the estimation uncertainty, involves simulating CCs from the set $\rho_{xy} \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$ with the total number of samples per random variable being $N \in \{50, 100, \ldots, 500\}$ (Makowski et al., 2022; R Core Team, 2024). See supplementary Figure 1 for demonstration of the similarity of these two approaches.

After having established the equivalence between the two approaches, one can now proceed to add measurement uncertainty to each observation. To add measurement uncertainty, one can instead of randomly re-sampling pairs of data points from the original data-set, as done in the case without measurement uncertainty, one can re-sample these pairs using an error distribution. The original data is then inserted as the mean of this error distribution and the uncertainty (standard deviation) of this distribution, is the measurement uncertainty. A straightforward choice of error distribution would be the normal distribution, which would reflect that the uncertainty is assumed to be bidirectional, with no preferred direction. This can be formulated as the following:

$$\begin{pmatrix} \hat{x}_i \\ \hat{y}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix}, \begin{bmatrix} m_x^2 & m_x^2 \cdot m_y^2 \cdot \rho_m \\ m_x^2 \cdot m_y^2 \cdot \rho_m & m_y^2 \end{bmatrix} \right)$$

where $\hat{x}_i$ and $\hat{y}_i$ represent the observed estimates of x and y on a particular simulation given their measurement uncertainty m_x and m_y and the correlation between them ρ_m. Of note, is that one might re-sample the original data from other error distributions. For instance, if values are strictly positive, then simulating from a truncated normal or strictly positive distributions like a lognormal, would be preferred to avoid sampling values that cannot be obtained i.e. negative reaction time values.

Figure 1 A shows a scatter plot of x_i and y_i, these might represent different measurements from a Cognitive Science experiment, say reaction times and time spent awake. For visualization purposes only measurement uncertainty was added to the y-coordinates meaning that from the above equation $m_x = 0$ and $m_y \in \{1, 2, \ldots, 10\}$, here this uncertainty is depicted as error bars on

individual data points. Figure 1 B displays how the estimated CC distribution obtained by bootstrapping changes based on these measurement uncertainties. The figure demonstrates that the CC estimated via bootstrapping is attenuated, while its distribution widens with increasing measurement uncertainty, mirroring what one would find using analytical solutions (Saccenti et al., 2020).

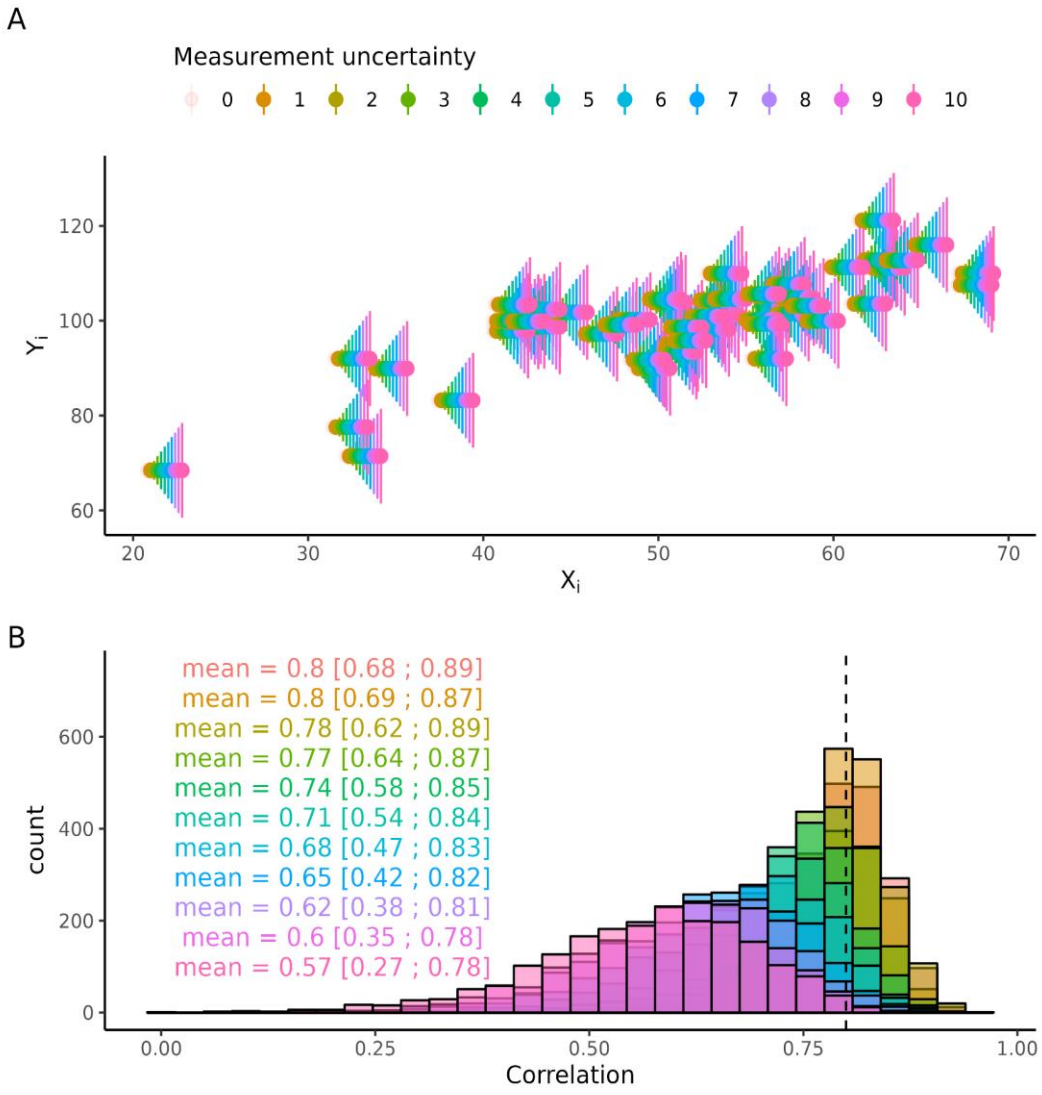


Figure 1 Measurement uncertainty on correlation coefficient. (A) Displays a scatterplot with varying amounts of measurement uncertainty. (B) displays how the correlation coefficient distribution, obtained through bootstrapping changes with increasing measurement uncertainty. Vertical line is the simulated correlation coefficient without uncertainty.

Here it was shown with normally distributed noise that decreased the size and width of the CC, later nontrivial types of noise are added where the simulation approach used here is necessary to properly propagate the uncertainty. This example does not demonstrate how test re-test uncertainty interacts with these lower levels i.e. measurement and estimation uncertainty. One can imagine measuring the CC from figure 1 B twice, and getting different results, even if the measurements were infinity precise, and there was no uncertainty in the estimation process due to infinity many data-points. For a more concrete and elaborate example, see supplementary analysis 1. The main message here is that to get reliable estimates and, in the end, to make reliable inference one needs to account for all sorts of uncertainties and the lower in the hierarchy you move the more fundamental and important they become. Having a parameter estimate that is stable over time will not matter if you cannot estimate or measure it reliably in the first place.

Modeling definitions and validation.

Modeling definitions

The rest of the thesis will revolve around refining, testing, and designing models of cognition, Cognitive modelling will be deployed for this purpose. Cognitive modeling serves as an intermediate level in a hierarchy of computational models on top, and statistical models in the bottom. These types of models can be thought to differ in their flexibility, assumptions, and scope of investigation. It should be noted that these models have many commonalities, such as being mathematical representations of a data generating process. This makes these definitions operational and should be thought of as having fuzzy boundaries (Durstewitz et al., 2016).

Statistical models are the models primarily used in medical and social sciences. These models mostly consist of linear and generalized linear (mixed) models (Bahadori et al., 2023; Maravelakis, 2019). These models are linear combinations of independent variables which are sometimes transformed (making them generalized). The mathematical representation of such models follows:

$$F(y) = \beta \cdot X + \epsilon$$

Where y is a vector of dependent variables of N elements, F is a link function that maps the conditional mean unto a particular space, common link function are the logit and log transformations which maps unto domains of [0 ; 1] and [0 ; ∞] respectively. These domains could be probabilities and strictly positive values like reaction times. β is a vector of regression coefficients with P predictors which gets estimated, X Is a matrix of independent variables of size [N, P]. Lastly ϵ is a vector of N elements containing the errors of the model. The benefit of these statistical or regression models is that maximum likelihood estimators are available, making parameters estimation fast and efficient. However, one limitation is that they put quite big constraints on the types of models that can be fit, i.e. there must be a linear mapping between all independent variable and the dependent variable, in a domain that can be mapped with a link function. This constraint will in many instances make theories hard or impossible to test as human behavior and cognition can be nonlinear (Ivanova et al., 2022). It should be noted that the CC examined in the previous section, can be thought of as a special case of this linear model, where β is a single value (the CC) and y and x are z-transformed vectors, see supplementary figure 2.

Cognitive models are models that are meant to resemble the generative processes of human behavior more closely. These models are generally more theoretically driven as the constraint of linear combinations is avoided, by employing different optimization schemes. In many cases cognitive models are estimated in a Bayesian framework due to the flexibility with which models can be specified. The main advantage of these models, in this context, is the added freedom in model specification.

Computational models are the upper most level of the hierarchy, which here will be used to refer to the generalization of cognitive models to other scientific domains, such as physics, biology chemistry etc. These models are outside the scope of this thesis.

These three categories are arbitrary, and many methods and models will fall between them. However, these arbitrary definitions do add value in communicating, the general framework being worked in and thereby what methods are used. The next section will describe a particular cognitive model, which will be the focal point for the rest of the thesis.

The Psychometric function

Here the psychometric function (PF) will be explored, as this type of function has been a stable corner stone in the cognitive science literature across many different sub fields (Bahrami et al., 2012; Coates & Chung, 2014; Courtin et al., 2023; Gold & Ding, 2013; Ma et al., 2024). The PF is a continuous function that maps real inputs (here called intensity values) onto probabilities, i.e. the domain $[-\infty; \infty]$ with the range of $[0; 1]$. In most cases the PF is identical to a logistic regression in statistical modeling and is commonly used in perceptual research where probabilities are then converted into binary forced choices through a Bernoulli or binomial distribution. The PF is usually a cumulative density function such as the cumulative logistic or normal distribution, amounting to a logistic or probit regression in the statistical framework. The main difference between the statistical and cognitive frameworks usage of the PF is the number of parameters.

The least number of parameters used to describe the PF is two, the threshold (α) and the slope (β). These two parameters describe the center of the curve, with α being the intensity of the stimulus at probability 0.5, with β being the steepness of the function around the threshold. In the cognitive modeling framework one or two more parameters are introduced, the lapse (λ) and guess rates (γ). These two parameters together handle the tails (i.e. the far ends) of the psychometric function and essentially making the probability in these tails non-deterministic i.e. the upper and lower bounds become (λ) and (γ) instead of 0 and 1 (figure 2). These additional parameters help fit the PF to data where attentional slips or wrong button presses happen. It can even be shown that including these parameters will greatly improve the estimation of the slope of the PF, if lapses and or guesses are present in the data (Wichmann & Hill, 2001). Figure 2 depicts how all these parameters change the shape of the PF. For the sake of this thesis, the cumulative normal distribution is used to map stimulus values to probabilities with a single lapse rate. This lapse rate will govern the distance between the upper and lower bound, essentially making it equally likely to have an erroneous response for high and low stimulus values. The mathematical formulation of the function is as follows:

$$p(x|\alpha, \beta, \lambda) = \lambda + (1 - 2 * \lambda) * \left(0.5 + 0.5 * erf\left(\frac{x - \alpha}{\beta * \sqrt{2}}\right)\right)$$

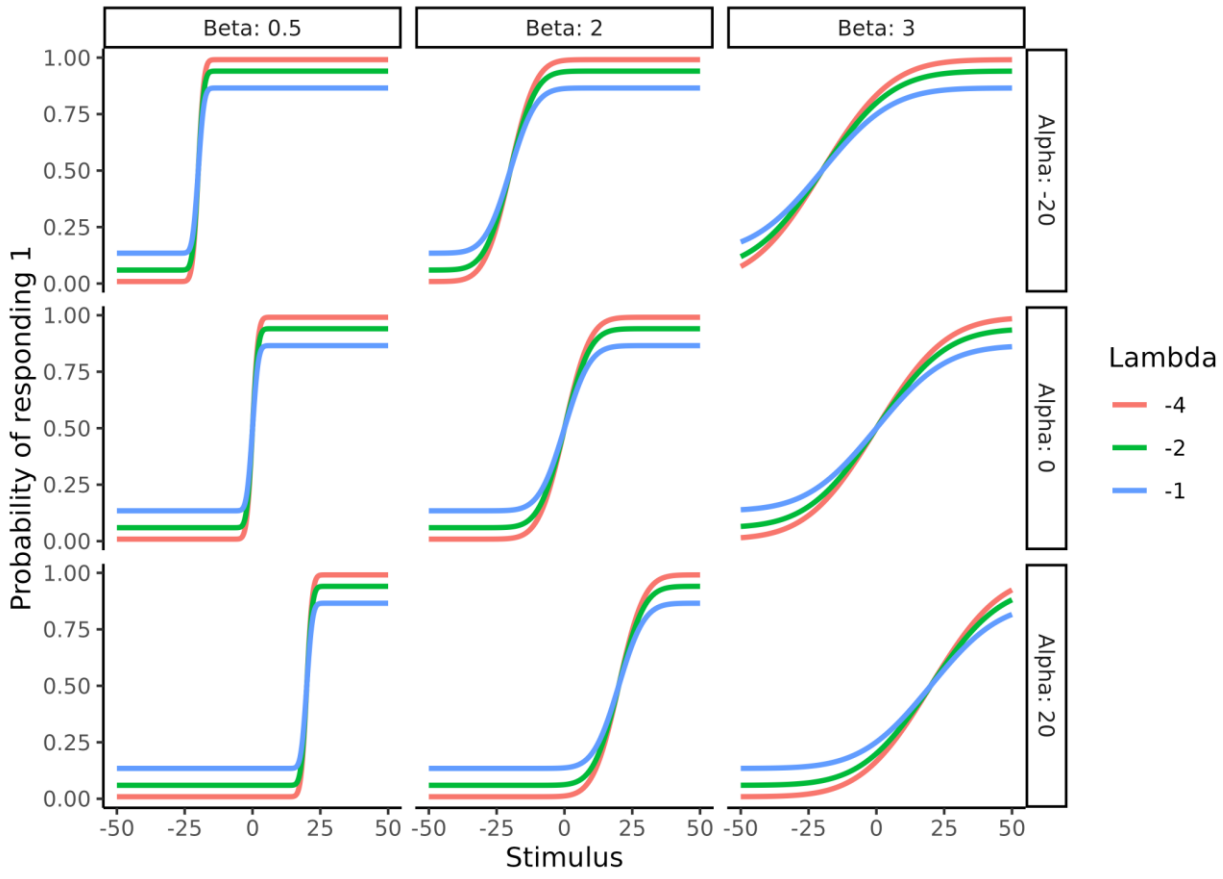


Figure 2 Psychometric parameters. Displays how the parameters of the psychometric function i.e., the threshold (alpha), the slope (beta) and the Lapse rate (lambda) of the psychometric function change its shape. Columns display how the beta parameters changes the slope of the function. Rows show how alpha changes the location of the center of the function changes. Lastly, colors in the plot depict how lambda changes the asymptotes in extreme stimulus (x) values.

Model validation.

In the same vein of validating the bootstrapping approach with the analytical solution, in the previous section about measurement uncertainty, cognitive models themselves are many times validated. This is to ensure that at least in principle the parameters of the model can be estimated with increasing accuracy with increasing number of trials. This section will highlight some of the emerging ways in which computational models in the cognitive science literature are being tested and validated and takes inspiration in the seminal paper from Wilson and Collins (Wilson & Collins, 2019) describing 10 simple rules of computational modeling, which is commonly cited when validation of computational models are described (Hess et al., 2024). There are at least three

main challenges when building and validation cognitive models, which are particularly relevant when writing novel models. How do we know that our models do what we think they do (identifiability)? How do we know that they accurately estimate the parameters of interest (Internal validity)? And lastly, how do we know that we can distinguish between competing models (external validity)? The last challenge is beyond the scope of the current thesis and is well covered elsewhere (Wilson & Collins, 2019). The answer to the first two challenges must be found in simulations, especially when our models become more and more complex and analytical solutions are sparse or even non-existent.

The simulation practice revolves around selecting an appropriate range of parameter values and using these to simulate data from our models. These simulated data are then used to refit the model to examine how well the model approximates the simulated parameter values. Ensuring that in these approximations are close simulations is desirable for the model to perform well, such that one can have faith in them when the real underlying process is unknown, i.e. analyzing real experimental data. An appropriate range of parameter values for a particular model can be difficult to select, as this is exactly the problem of identifiability. Several lines of information can help gauge this. Firstly, looking at mathematical constraints of the model formulations can reduce the possible ranges of parameter values. For the case of the PF this amounts to ensuring that the slope is strictly positive, as this ensures that increasing levels of stimuli will produce greater probabilities of responding 1. This also ensures that the standard deviation of the underlying probability density function is strictly positive. The lapse rate of the PF should be constrained between 0 and 0.5, again to ensure a particular the shape of the PF. Lapse rates below 0 and above 1 will produce probability values outside the [0; 1] range and values above 0.5 will flip the shape of the PF, as negative slope values would Not constraining the PF in this way could lead to two distinct solutions, as negative slope values and lapse rates above 0.5 would be able to produce the same mathematical transformation, making the solution non unique. From a more theoretical level an appropriate range of parameter values can be narrowed down by investigating whether the observed behavior (given the simulated parameter values) is physically or biologically plausible. For the PF we might expect a few of our participants to not be particularly interested in the task, this would amount to having a high lapse rate i.e. above 0.2 (amounting to 20% lapse responses) or shallow slopes. This behavior, however, is quite unlikely and expecting only few lapses in the experiment, given that it is conducted in a quiet environment is likely. Lastly using empirical

knowledge from the literature can help narrow the parameter space further. For the sake of argument, one might investigate the detection threshold for cold stimulation on the skin. Just given this information alone it is possible to narrow down the threshold for cold detection to being below the skin temperature of around 30-34 degrees˚ and above the absolute zero of -273 degrees (Courtin et al., 2023). However, knowledge from the scientific literature would suggest that thresholds between 28 and 33˚ would capture most of the population (Lithfous et al., 2020). These same arguments would also apply for the slope. This practice of investigating the assumptions of the simulated parameter values is closely related to prior predictive checks in Bayesian workflows (Kruschke, 2021).

The next challenge is about internal validity, i.e. can our model estimate the parameter values used to simulate the data on which the model estimates the parameter values. To test and validate our models in this regard, we simulate data from pre-specified parameter values, which have been deemed to be appropriate, using the first step described above. We then fit our models on this simulated data and investigate how well the model can estimate the latent simulated parameters (i.e. those that produced the data). This exercise of simulating behavior and then re-estimating the parameter values from the simulated behavior, is commonly known as parameter recovery. Generally, if this procedure succeeds, then the parameters are said to be recovered. The satisfactory criterion and metric used to access this procedure, often refers to some CC between the estimated and simulated parameter values (Wilson & Collins, 2019). Parameter recovery can thus be thought of as an internal validation of a model, which if done properly should increase the faith in the parameter estimates, when the model is fit to experimental data. The argumentation is thus, had the parameter estimates been known beforehand (i.e. simulated them), then we know that they are close to the estimated parameter values obtained. The assumption of this argument is that if our model recovers the parameter values well in a simulated setting, then it must also do so when fitted to experimental data, where the underlying parameters are unknown. This assumption rests on auxiliary assumptions. These auxiliary assumptions include, that the model that generated the data is the same or close to, as the one used to model the data. The process of parameter recovery thus assumes that we know the underlying generative model, which is not the case when fitting experimental data.

This point, of ensuring that we are selecting the right generative model is the challenge of external validity. The challenge is that infinitely many generative models exist, that are compatible

with the observed behavior. This challenge cannot easily be solved, as ensuring that we are using the right generative model would entail testing all generative models. That would mean being able to compare them, while ensuring that all these models are distinguishable In the Cognitive Science literature the common practice is to use theoretical framework(s) to build competing models with different assumptions of the underlying generative process and then compare this subset of models (Berker et al., 2016; Hess et al., 2024). This comparison of models is usually done on how well the models can describe the data, using statistical metrics such as information criteria or leave one out cross validation, with some penalization for complexity (Vehtari et al., 2017). This highlights two important aspects; first, our models reflect our theories and are therefore at best as good as our theories, and second, we are likely missing the true generative model in. This point of missing the true generative process can be partly mitigated by ensuring that the tested models are distinguishable, at least in principle. This challenge has been addressed by simulating data from all tested models and then refitting all models to the data simulated by each individual model. Returning to the example of the PF, we might have two competing theories of how stimulus values are translated into binary choices, one involving the lapse rate and one without. To conduct model recovery, data would be simulated from these two distinct models, each model would then be fit to all simulated datasets, and the best-fitting model would be determined for each case. The result of such model recovery is an N-by-N matrix, with N being the number of models. The rows in this matrix would indicate which model was used for the data simulation, and columns indicate which model was used to fit the data. The entries of the matrix would then be the frequency by which a particular model wins in model comparison, given the data simulating model. An identity matrix would represent that the models are completely distinguishable. Any deviation from an identity matrix would entail that for some of the simulations, the best fitting model was not the model that simulated the data (Wilson & Collins, 2019).

Limitations of current internal model validation steps

The model validation steps described above serve to increase faith in our models, their parameters, and the comparison between them. However, I will here argue that some of the metrics used to access these validations have notable flaws, with a particular focus on the problem of internal validity. First, the metrics used can be misleading, to show good model validation, by masking the actual poor validation. This problem can thus lead to false confidence in the model and overconfidence in the inference made based on it. Additionally, the metrics often lack sensitivity

or specificity to provide the person building the model with information about how and where, in parameter space, the model performs well, thereby leaving valuable insights hidden. In the following section, the thesis will highlight some of the metrics commonly used in the literature for model validity, which are described in Wilson & Collins (2019). As mentioned above internal validity is often accessed by simulating data from a model given a set of parameters. This simulated data is then fitted to the model, which then optimizes for the parameters. Subsequently, the CC between the estimated and simulated parameters is often computed as an estimate of internal validity (Hess et al., 2024; Schurr et al., 2024; White et al., 2018). In their seminal paper Wilson & Collins (2019) describes that ideally, the estimated and simulated parameters should be tightly correlated, without any bias. They also highlight that a weak correlation could mean bugs in the code, or an underpowered study i.e. too few trials. They also emphasize the importance of plotting a scatter plot of simulated vs estimated parameter values, to access if ranges of parameter values are problematic, but also to access whether there might be biases.

I will here argue that the CC is an inappropriate metric and that a version of an intra class correlation (ICC) is better suited for the task of interest. Acknowledging two important things; neither metric is perfect, and visually inspecting the simulated vs estimated parameter scatter plot is crucial. Failing to ensure sensitive and specific metrics for internal validity of the models, may result in significant resources being invested in a model that ultimately fails to perform adequately, hindering scientific progress. It could take years before researchers realize that a model is flawed, even in simulated settings, posing a significant roadblock to scientific advancement.

Current problems with internal model validity (parameter recovery)

The first and perhaps biggest problem of internal validation of computational models, is that it is not universally done. This makes it hard or even impossible to know if the generative model in question, can be trusted. The second, almost ubiquitous problem in the literature using parameter recovery is the oversight of interactions between parameters. This is less of a concern for individuals using an established cognitive model but should be a big concern in the methods papers describing and formalizing them. A prime example of this is the Hierarchical Gaussian filter (C. Mathys et al., 2011 ; C. D. Mathys et al., 2014). Where after having laid out the equations of the model, two of the most crucial parameters (κ and θ) of the model, that sets this model apart from the Kalman filter, are held constant when performing parameter recovery. Even in much simpler models, as in the PF described above, there are trade offs and interchangeability between

parameters. The last problem with parameter recovery is the reliance of CCs to access it. As has been suggested elsewhere, that the CC is at best insufficient and at worst misleading (Schurr et al., 2024). Three primary problems exist with using CC to examine internal validity, namely invariance to linear transformations, the domain and the interpretation of the estimate.

CCs are invariant to linear transformations. The means that two sets of variables i.e. [1,2,3] and [1,2,3] have the same correlation after transforming one of the sets with linear transformation. Consider the transformation $f(x) = 2 \cdot x + 3$, resulting in the sets [1,2,3] and [5,7,9]. The CC between these two sets will have the same CC as before the transformation. In terms of model validation these two instances would be very different. In the first, one would have perfect internal validity, whereas in the second it would be severely lacking. This lack of sensitivity to linear transformations does not make sense for parameter recovery, as we want a metric that penalizes this behavior.

The domain of correlations is [-1; 1]. However, this directionality is nonsensical for internal validity. A CC of -1 would mean perfect recovery of the parameters of the model, with a negative sign, meaning that you do recover the value (or the linear transformed value) just not the sign. Ideally, we seek a metric that ranges from no recovery to perfect recovery, rather than perfect recovery without the sign to perfect recovery with the sign.

Lastly, the interpretation of the CC in terms of parameter recovery poses challenges. Determining what is a sufficiently large CC for parameters and identifying what types of uncertainty is causing the correlation to be less than ideal, is not obvious. Attempts to make such distinctions have been made without much traction (White et al., 2018). All these issues resemble what researchers have encountered when trying to estimate the stability and/or test -retest reliability. Here the widely used solution was to use the ICC as the metric instead of the CC (Schurr et al., 2024).

ICC Parameter recovery

Because the idea of using the ICC as a metric for parameter recovery is relatively new and has only been suggested, and not implemented anywhere in the literature, to the authors knowledge (Schurr et al., 2024). I will here outline what the ICC is and how it can combat some of the shortcomings of the CC in accessing model validity.

The ICC, in its simplest form, is a ratio of irreducible variances (uncertainties) to the total variance in the data. In practical terms, the irreducible uncertainty is the uncertainty between

subjects, whereas the total uncertainty can have several components. To calculate the ICC, these variances need to be estimated such that their ratio can be computed. The estimation of the variances can be achieved using a model that properly accounts for these different types of variances. These models are typically hierarchical models, where known structure of the data is embedded.

Taking the of a researcher trying to determine the test-retest reliability on the detection threshold of cold stimulus. The researcher will have his subjects come in for x sessions and do the same task each visit. We will now assume that all subjects come from the same underlying distribution (i.e., the population), which is governed by a population mean and a population variance, i.e. the between subject variance. From this population level an individual subject level distribution is drawn, here each subject has their own mean and variance (within subject variances). Now for each session that each subject participates in a parameter value is drawn. This parameter value is drawn from subject level distribution, which then governs the participants' behavior on that session. This nested hierarchical structure is demonstrated in figure 3, where each of the levels are governed by the levels above and each level has an associated variance. The between subject variance is the variance of the population level distribution, and the within subject variance is the variance of each of the participant level distributions. The ICC as mentioned above is the ratio between this within and between subject variances. This can mathematically be expressed as.

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

Where $\sigma^2_{between}$ is the between the subject variance and σ^2_{within} is the within subject variance. Given the of interest the model's performance, we can simulate agents that have no within subject variance i.e. the same true parameter values for each session. Then its possible to examine how the number of trials and or subjects of the cognitive task will influence the model's ability to capture that there is no within subject variation. Note the number of sessions could also be examined.

This approach has one clear problem, it does not tell explicitly investigate how well the model estimates the true parameter values, for each participant at each session. The ICC described above only estimates how close each parameter is to itself between sessions. To capture the difference between the true simulated value and the estimated parameter value of the model, one might use the mean squared errors (MSE) between the simulated and estimated parameter values. This MSE would serve as a residual error of the model on the parameters. Including this into the

ICC formulation above is straightforward, as this is just another source of variance which can be added into the denominator. This also highlights one of the advantages of the ICC, i.e. it is a partitioning of variance (uncertainty) in the model. This partitioning of variance is valuable when building and validating models, as this gives clues to where the model fails and where it might excel. In figure 3 the MSE would amount to taking the difference between the estimated parameter value (distribution) of a particular subject at a particular session and the simulated value. Formally we add the MSE into the ICC equation.

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within} + \sigma^2_{\epsilon}}$$

Where σ^2_{ϵ} is the MSE. This conceptualization allows for putting parameter recovery for a model, into a single value for each parameter, that ranges from 0 to 1. This metric is also going to be trial and subject dependent. Here it should be noted that this formulation of the ICC would imply a nested hierarchical structure, as described above, and depicted in Figure 3. This is not necessarily the case for the CC. Using the CC, one could simulate numerous subjects and then calculate the CC on each of these subjects fit individually. Alternatively, the ICC used for parameter recovery could also be calculated in a non-nested hierarchical manner, where only a single session for each subject is simulated, i.e., mathematically $\sigma^2_{within} = 0$.

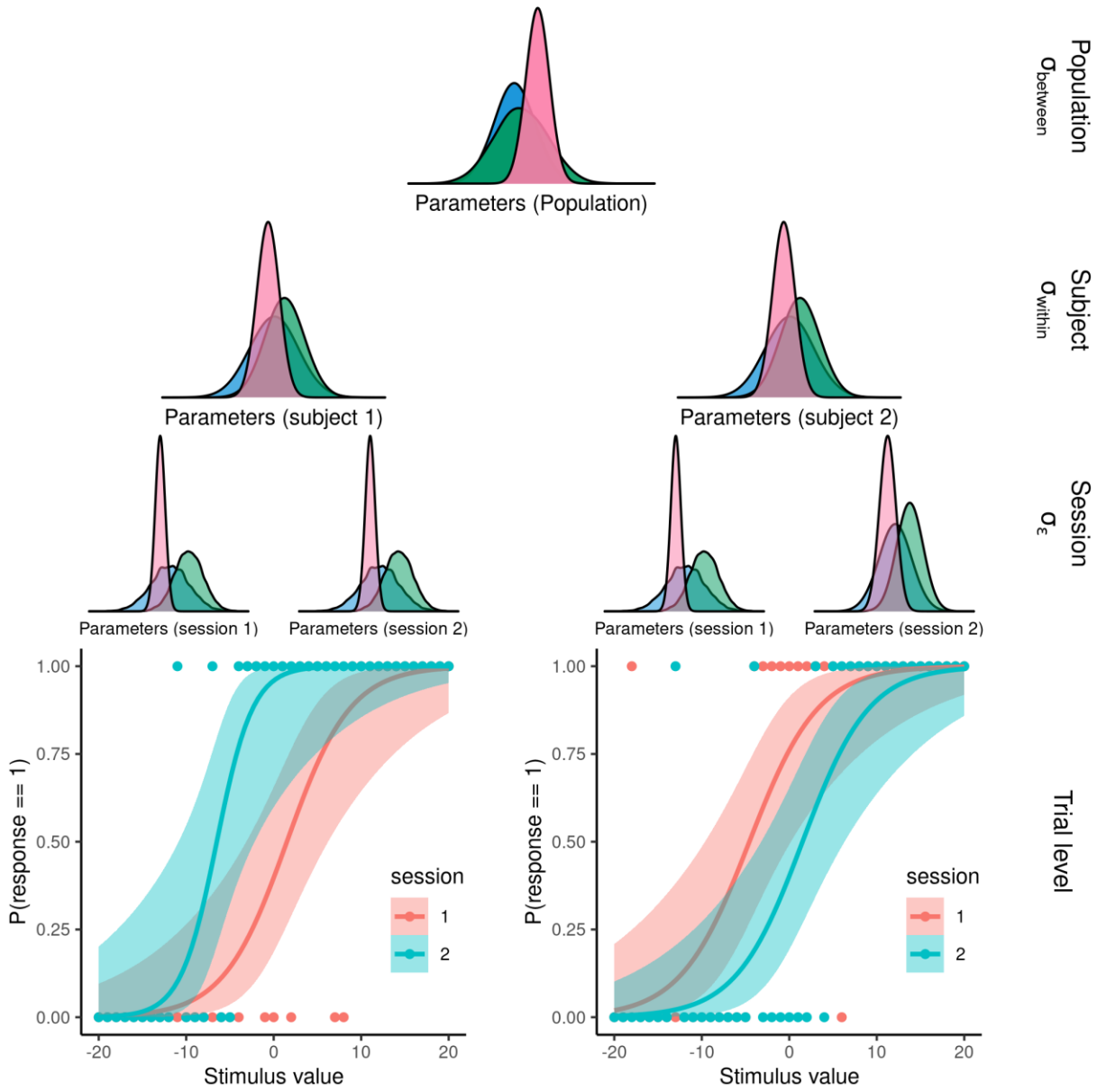


Figure 3. Visualization of a nested hierarchical model with sessions nested within subjects within a population of parameter estimates from a psychometric function. Displaying from the top down how the nested hierarchical model assumes structure of the underlying data. Data (points in the bottom plot) is entered at the lowest level where the cognitive or statistical model is fitted (here a psychometric function). The parameters of this model are drawn from a session distribution which is nested within a subject distribution which again is nested within a population distribution. This nesting of the parameters of the model allows for seamless estimation of the partitioning of variance within the model. For the sake of parameter recovery within subject variance is simulated to be 0, and the mean squared difference between the session level distribution and the simulated parameter value is calculated and put into the denominator of the ICC metric.

Figure 3 displays the conceptualization of the ICC with the additional MSE. Parameters propagate from the population level distribution to the subject level distribution and into the session level distributions, which then forms the cognitive model at the trial level. Figure 3 displays this as a psychometric function for two sessions.

The concept of parameter recovery using this framework aims to assess the degree to which the whole model can distinguish between the types of uncertainty. In this framework the within subject variance can be simulated to be zero, and the MSE can be calculated as the difference between the session level parameter estimates and the simulated parameter values. This entails that in the simulated setting the ICC-value from the above equation is 1, i.e. the only source of variation is between subjects. However, when running simulations, one can investigate how the model itself ascribes this variation, as uncertainty will be inferred within subjects or session, but also in the parameter estimates themselves.

Standard parameter recovery.

Turning the attention back to the three parameter PF. This cognitive model will be used to demonstrate this novel way of conducting parameter recovery. After having specified the model, one can simulate data from different ranges of parameter values, to select an appropriate range of parameter values. Parameter ranges are selected and simulated in accordance with table 1 and figure 4. Using the probabilistic programming language Stan and its interface with R, Rstan (Gabry et al., 2024), one can invert the model from the data to derive the estimates of the latent parameters, which were used to simulate the data in the first place. Note, that for all models displayed and estimated their convergence was accessed by ensuring Rhat values were below 1.03, and that no divergent transitions were present. Ideally all chains would have been inspected but given the vast amount of simulation presented throughout the thesis, visual inspection of each model was infeasible and summary diagnostics were used instead. Furthermore, all priors for all models presented were weakly informed. This would typically entail that most of the prior distributions were set as normal distributions with means of 0 and standard deviations of 3 between 5, in the unconstrained space. For a comprehensive list of all priors used, readers are referred to the supplementary material or the GitHub repository.

Table 1: parameters of the normal distribution used to simulate agents. Columns depicts the parameter type for the psychometric function, the mean and standard deviation of the normal distribution used for simulating the parameters and lastly the transformations for each of the parameters, from left to right.

Parameter	Mean	Sd	Transformation
Alpha	0	10.0	x
Beta	2	0.6	Log(x)
Lambda	-4	2.0	$Logit^{-1}(x)$ / 2

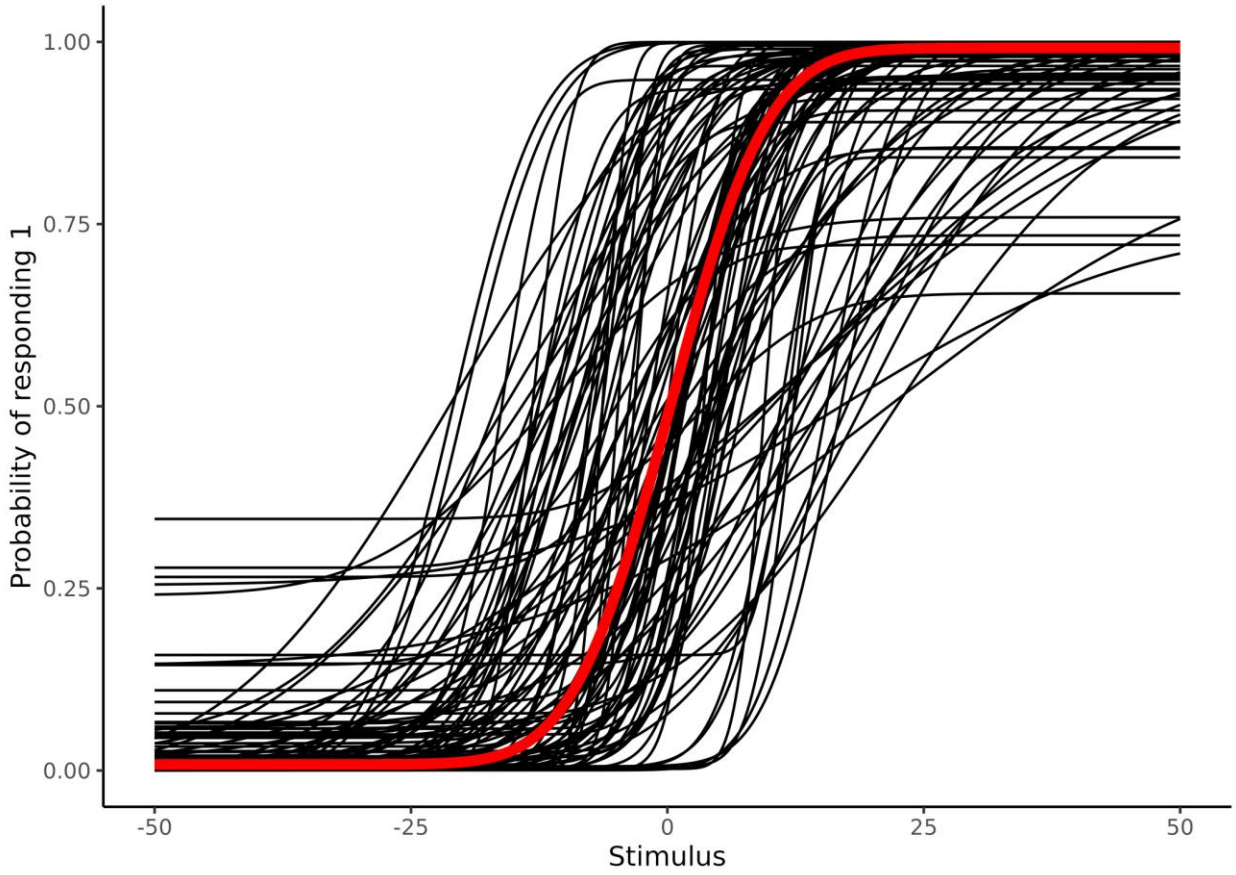


Figure 4. Displays 100 samples of the parameters of the psychometric function from table 1. Visualization of the implications of the simulated parameters of table 1. Black lines depict individual subjects, while the red line depicts the group mean.

100 pairs of parameters are simulated based on the values in table 1, amounting to simulating 100 subjects. The data obtained from these 100 subjects are then refitted using the same model. The pairwise scatter plot of estimated vs simulated parameter values are depicted in figure 5. Here the estimation uncertainty of each parameter is added as vertical lines. This simulation was done for 100 subjects over 100 trials, where each of the stimulus values were selected as the set $x_i \in \{-50, -49, \dots, 49, 50\}$. Figure 5 also displays how adding estimation uncertainty to the CC changes the resulting size and uncertainty estimate of the CC (i.e. its own estimation uncertainty). This influence of uncertainty on the CC resembles what was also shown in section on measurement uncertainty. It should be noted that the addition of this uncertainty does not necessarily have to decrease the size and or uncertainty of the CC. One could imagine a couple of points falling way off the identity line, with high uncertainties. These points would have less weight, when accessed with uncertainty, meaning that adding estimation uncertainty could in principle also increase the CC and decrease its own estimation uncertainty. This highlights the non-trivial and nonlinear link when uncertainties from fitted models are added to further analyses.

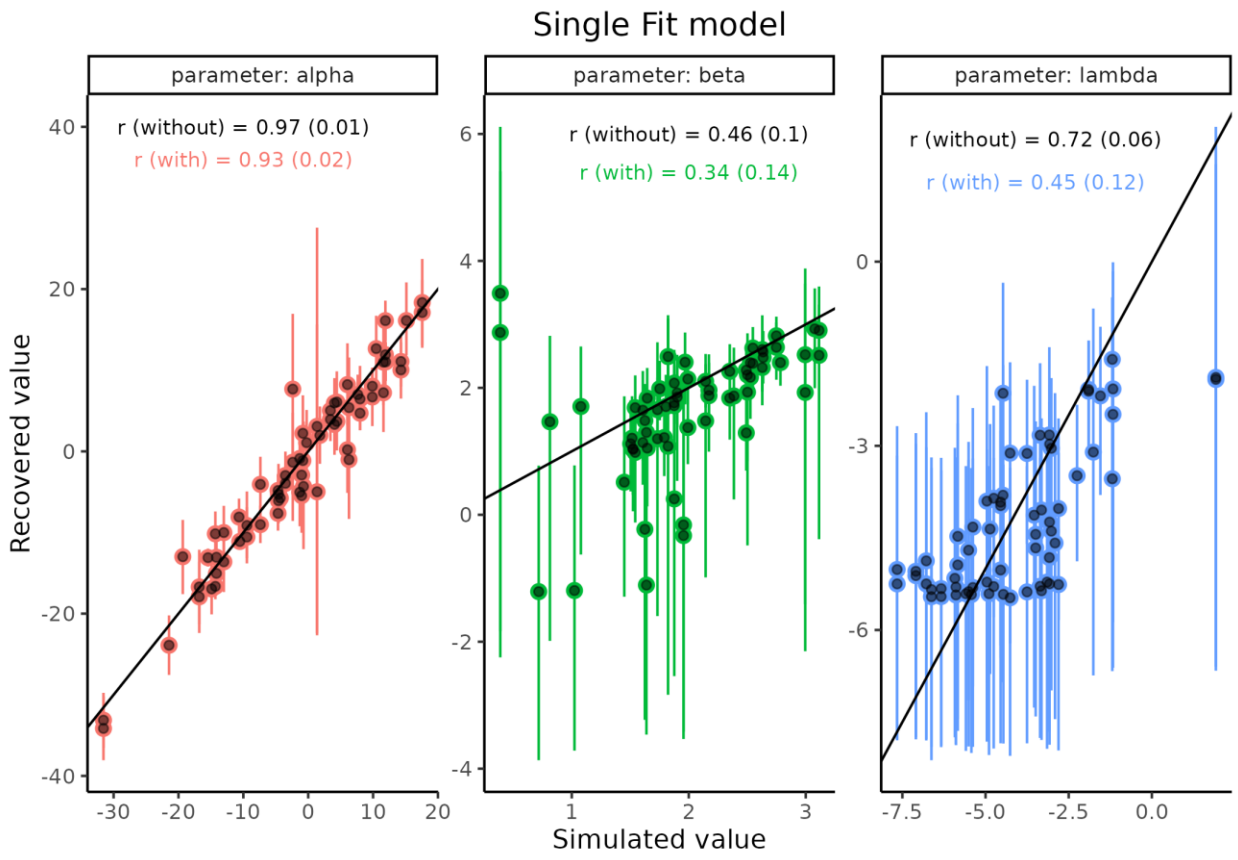


Figure 5. Parameter recovery for the three parameters of the psychometric function in the unconstrained space. Scatter plot of simulated vs recovered parameter values, with error bars displaying the 95 % highest density interval for that parameter on that simulation. Text on each facet shows the estimated correlation coefficient with its standard error (estimation uncertainty) (with) and (without) accounting for estimation uncertainty in the individual parameter estimates (data points), i.e., propagating uncertainty.

To evaluate the proposed ICC metric alongside the more standard parameter recovery approach, the same data-set as above was utilized. Crucially the data set above was simulated using only 50 simulations that were duplicated, making it eligible to compare the above standard parameter recovery with the ICC proposed. This simulation therefore implies that there is no within subject variation, as the first 50 data-sets were duplicated. One difference between the above single fit models and the proposed model depicted in figure 3, is the hierarchical structure embedded in the model. The hierarchical structure serves to shrink parameter estimates in relation to their distance and uncertainty from the mean of the higher-level distribution, which they are drawn from. This shrinkage, sometimes called pooling, has been shown to improve predictive capacity

and these models are commonly used in the Cognitive Science literature (Bates et al., 2015; Boekel, 2021; Gomes, 2022). To ensure a fair comparison the two internal validity metrics, the CC and the ICC, were calculated from this hierarchical model. Two ICC values were calculated now referred to as ICC_1 and ICC_2, referring to excluding and including the MSE respectively.

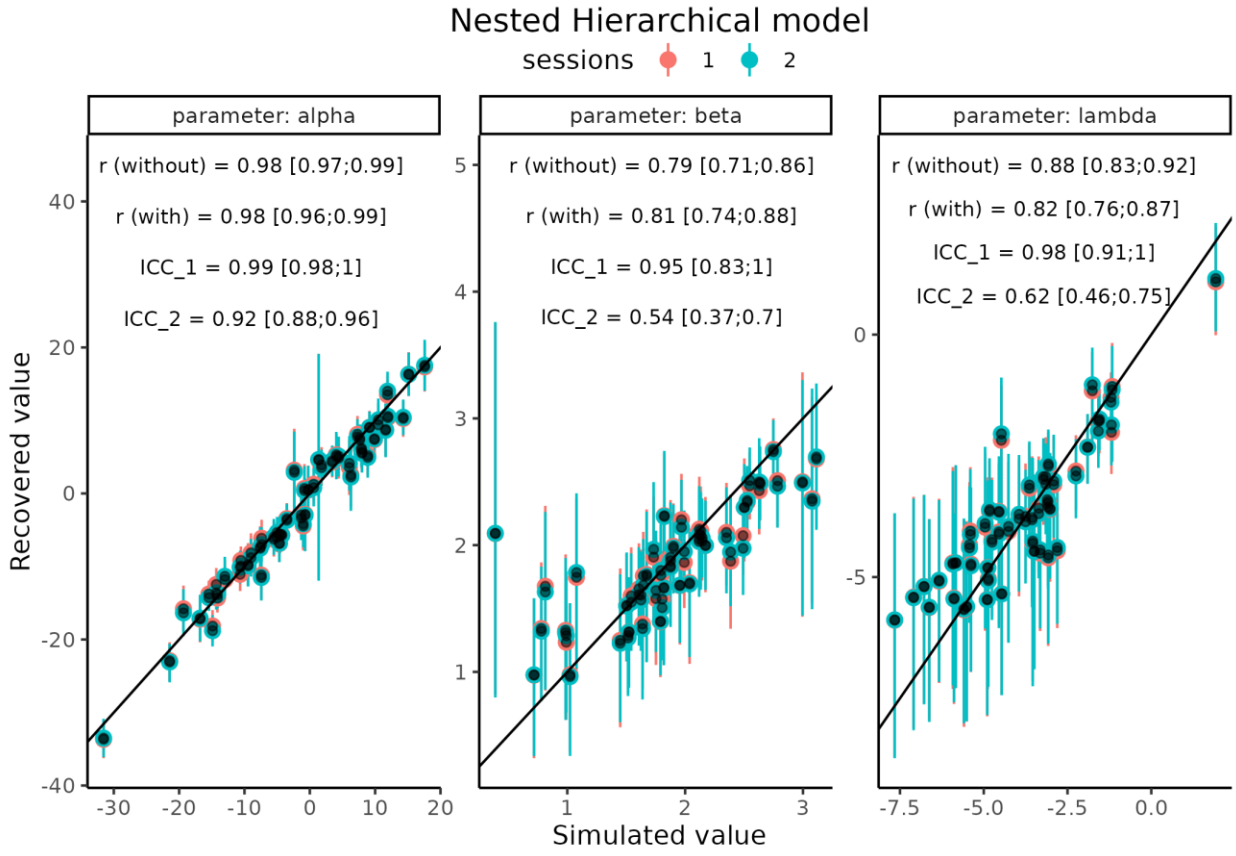


Figure 6. Parameter recovery for the three parameters of the psychometric function in the unconstrained space, using the nested hierarchical model. Scatter plot of simulated vs recovered parameter values, with error bars displaying the 95 % highest density interval for that parameter on that simulation. Text on each facet shows four metrics of internal model validity, correlation coefficient (without) and (with) accounting for estimation uncertainty, and the purposed ICC metric without and with including the mean squared error, ICC_1 and ICC_2 respectively.

Figure 6 illustrates that the hierarchical model fit improves the parameter recovery metrics. This is evident from both from a visual inspection of the points falling closer to the identity line, with less estimation uncertainty, and quantitatively by comparing the correlation estimates between figure 5 and 6. This finding helps to underscore why hierarchical models in general are preferred to single fit models, as the partial pooling improves estimation of the parameters (Bates

et al., 2015; Boekel, 2021; Gomes, 2022). This finding is further demonstrated in figure 7, where estimation uncertainty, here the 95% credibility interval, of each parameter at each session is plotted as histograms. It is quite important to note that a single simulation like this would be insufficient to ensure that the parameters are recovered. A good example of this is the lambda parameter. Investigating the pairwise scatter plot, of the nested hierarchical model, one might suggest that this parameter is quite well recovered. However, back transforming a lambda value of -5, on the unconstrained scale, would entail to a lapse rate of around 1.3%. This should be difficult, if not impossible, for the model to accurately estimate, especially given the 100 trials for each subject. Supplementary Note 2 describes this in more depth.

Turning the attention to the ICC values of figure 5. It is observed that ICC_1 on each of the three parameters has an upper bound at the maximum value of one. This can also be visually inspected by looking at the variation between pairs of data-points. Here the session one estimates are hidden behind the session two estimates, with only a few estimates deviating slightly. The ICC_2 estimate is crucially the lowest for all three parameters. Visual inspection of the pairwise scatter plot makes this clear as well. This metric is penalized for both the degree to which the points fall away from the identity line, but also by the estimation uncertainty associated with these points and the variation between pairs. This also explains why the alpha parameter is close to being asymptotic at 1, but with a little to be desired for simulated values between 0 and 10.

In the next section it will be shown how these metrics, especially the ICC_2, is influenced by different factors. It will be shown that by reducing estimation uncertainty, it is possible to increase this metric. Four different different strategies will be introduced to minimize estimation uncertainty, these include adaptive optimization design, increasing the number of trials, assuming different group mean slope values and lastly, jointly modeling of the binary responses with response times.

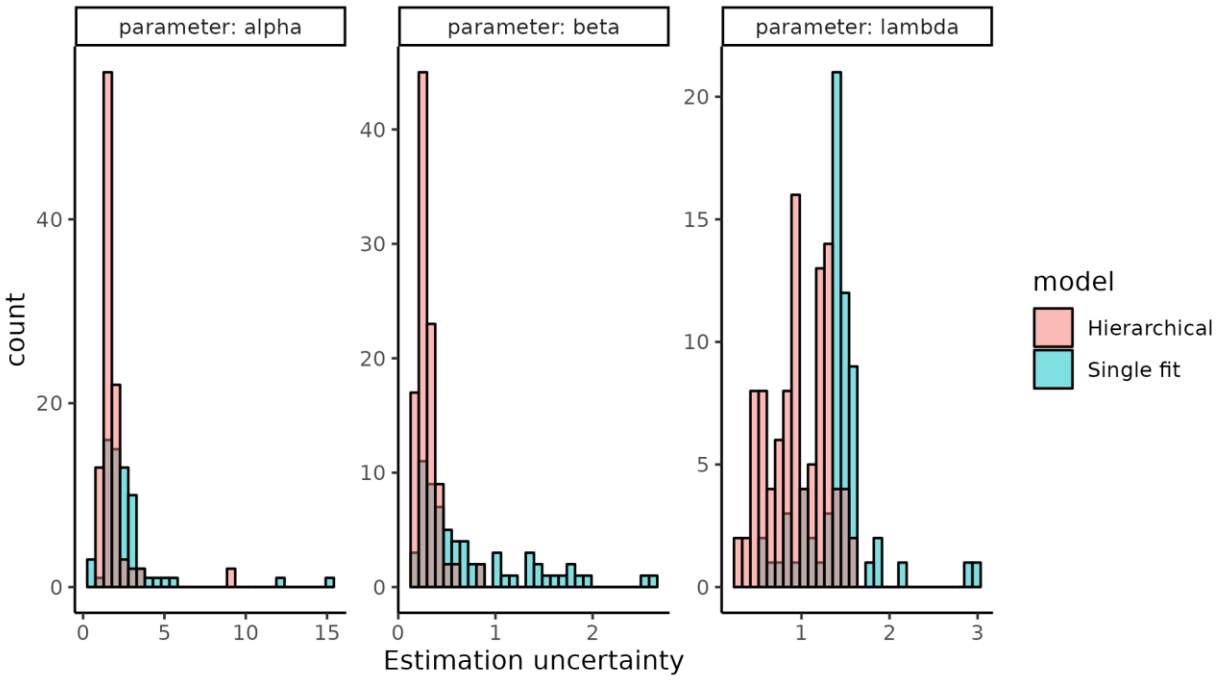


Figure 7 Estimation uncertainty for each parameter for the single and hierarchical fit models Each panel represents one of the three parameters of the psychometric function with the estimated uncertainty (95% credibility interval) depicted as histograms. The color of the histogram shows whether the model was fit using the single fit or hierarchical model.

Uncertainty minimization

Adaptive design optimization

An import consideration for parameter recovery, is the design of the experiment that the agent goes through. Referring back to figure 4, providing stimulus values in the far ends of the psychometric functions, i.e. in the ranges of [-50 ; -25] and [25 ; 50] will in most cases, give limited information on the shape of the psychometric. This could entail that the threshold and slope could be more informed by sampling stimulus values in the [-25 ; 25] range, see supplementary note 2 for discussion on the lapse rate. Therefore, selecting the input stimuli, in this interval could be assumed to be better for decreasing the estimation uncertainty in the two parameters, compared to randomly or uniformly exploring the input space. One might even go a step further; instead of selecting inputs that are more appropriate based on the mean of the population, each experiment could be individualized to each subject.

This practice of individualizing the experiment is called adaptive design optimization (ADO). The concept revolves around selecting inputs that are optimal given a specific criterion (Prins, 2013; Watson, 2017). Many of these criteria exist, including minimizing entropy, minimizing the posterior variance, or maximizing mutual information. What they all have in common is that they decrease estimation uncertainty, either of all or certain parameters. One of the main challenges of utilizing ADO is that the experiment has to be updated and individualized on a trial-by-trial basis. In the extreme this would require the algorithm, to run in tandem with the experiment. This puts significant constraints on the computation time of the algorithm. This issue has been partly solved in the existing packages by calculating a grid, of a particular resolution, of parameter values before the initialization of the experiment. This solution puts the heavy computation time before the experiment, ensuring that when the experiment is run, only a single look up is needed to provide the next stimulus value on each trial. This approach works great for experiments where each trial is independent of the next, like in a psychophysical experiment. However, if trials were mutually dependent as in a learning experiment, then the algorithm would need to calculate all possible lines of stimuli and responses, up until a certain point. This dependent structure would therefore become a daunting task, due to the combinatorial complexity. I will describe how an ADO can be implemented, utilizing the single-subject model, which was used for the single-subject parameter recovery. The goal of demonstrating how easily such an ADO can be

implemented is to show and examine the flexibility in the cognitive modeling framework. The advantages of being able to write such a custom algorithm is two-fold. First, if the model can be written to invert observed data to parameter values (i.e. fit a model to data), then it can also be used to simulate stimulus values, this therefore increases flexibility. Secondly, as this approach is not "optimal" for stimulus selection, the method can be extended to mutually dependent experiments. Building such an ADO can be done using variational inference algorithms. These algorithms can quickly estimate an approximate posterior distribution of the parameters of interest, here the pathfinder algorithm implemented in Rstan is used (Zhang et al., 2022). This algorithm locates normal approximations to the targeted density of the posterior distribution with its quasi-Newton optimization. Using this approximate normal, the pathfinder algorithm draws samples from it to provide approximate posterior samples. The rationale behind this approach to ADO is to iteratively fit the model as responses from the participant is collected. The parameter estimates from the model are then updated, and a new stimulus value is then selected based on these estimates, together with the knowledge of which stimulus values are the most informative for parameter values. For a full description of how the pathfinder algorithm was implemented, see supplementary note 3.

Figure 8 shows how the posterior distribution of the three parameters of the PF varies as a function of trials. This is visualized by using the previously uniform selection of stimulus values and the implemented pathfinder approach. As can be seen, both approaches make the parameters converge towards the real simulated values (black line), with increasing trials. However, the speed towards convergence is quite different, especially for the two parameters that the pathfinder algorithm is optimizing for, alpha and beta. After just 20 trials, using the pathfinder optimization, these parameters have found the simulated parameter value and is decreasing their estimation uncertainty (95% credibility interval). In contrast, even after 50 trials the uniform approach still has a bias in the estimation (the individual points are not on the black line), but also a substantial estimation uncertainty associated with it. For completeness, a PSI-algorithm was also used to compare the feasibility of this new approach due to the high constraint on computation time (Kontsevich & Tyler, 1999). Interestingly, the pathfinder algorithm completed the 50 trials in 14 seconds, whereas the PSI algorithm took 30 seconds. This highlights the feasibility of this approach, as experimental designs must be run relatively quickly, in order to keep the attention of the subject (Kwon et al., 2023).

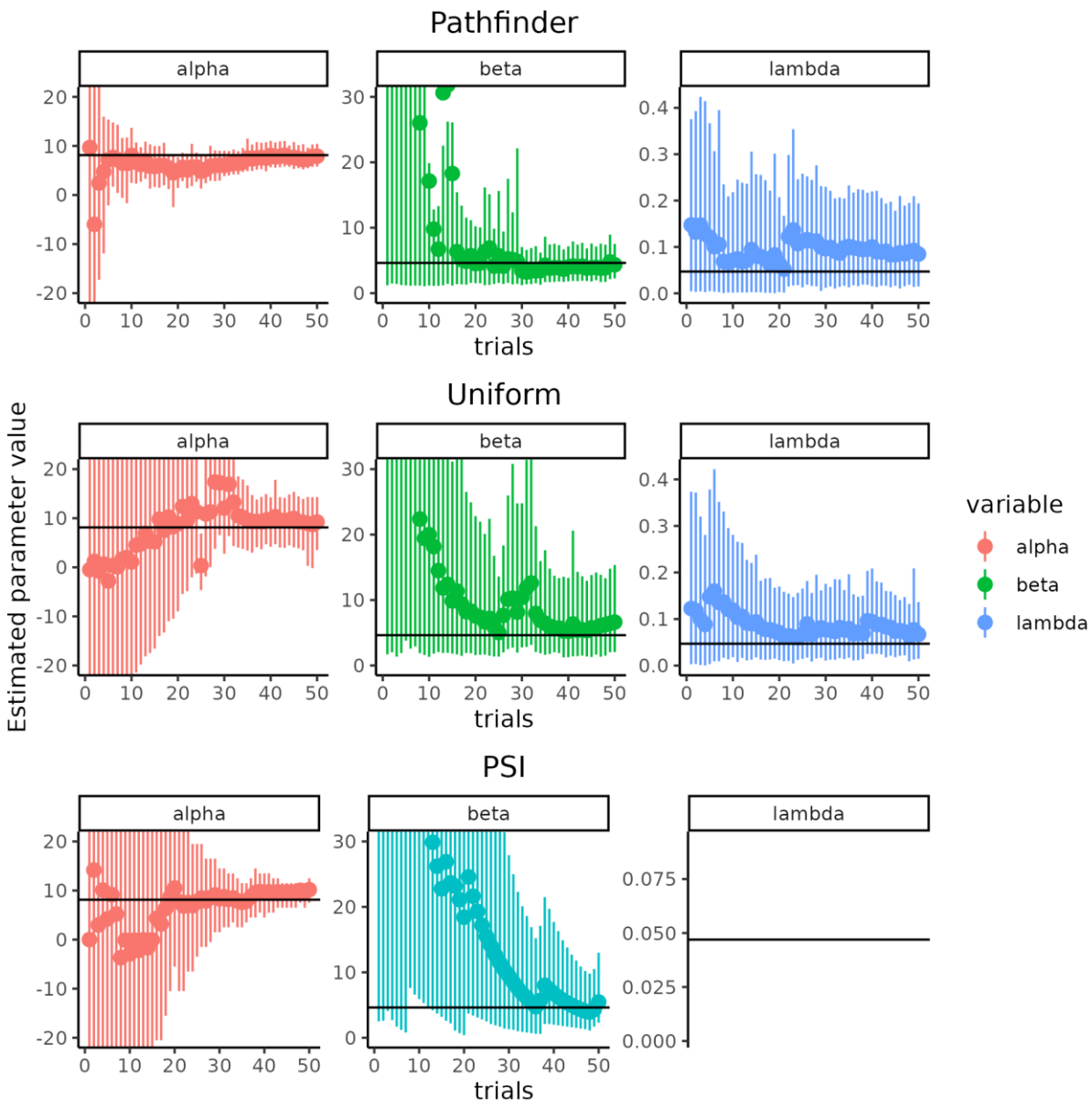


Figure 8 comparison of algorithms to obtain stimulus values of the psychometric function. Columns and colors display the three parameters of the psychometric function, while rows depict the (adaptive) algorithm used to obtain the parameter estimates. The figure displays how the Pathfinder algorithm quickly converges to the simulated value (black line) for the three parameters alpha, beta, and the lapse rate. This contrasts with the two other methods that take more trials to converge.

To show the differences more formally in the ways of selecting stimulus values across a range of trial numbers, the algorithms were run 100 times. This was done for trials ranging from 20 to 100 in a sequence of 10 trials (using the same range of parameter values, as depicted in table 1). To ensure fair comparison, each algorithm was only used to generate the stimulus sequence. This meant that data-sets were refitted using the same single fit Bayesian model used for the single fit parameter recovery, ensuring the same priors for all models. The inputs for the following analysis were, therefore, the posterior distribution of these refitted parameter values. For complete details on the fitting and optimization strategy, see supplementary note 3 and 4, including prior initialization for PSI and Pathfinder. Figure 9 shows the results of this simulation, with the top panel showing the bias, i.e. the difference between the estimated and simulated parameter values and the bottom panel the uncertainty in the estimated parameter value. Interestingly, the PSI-ADO performs the worst both in terms of bias in the slope (beta) and lapse rate (Lambda) parameter, and especially in the estimation uncertainty for all parameters. The main difference between the uniform and Pathfinder approach appears in the estimation uncertainty, especially in the threshold (alpha), where estimation uncertainty is significantly lower for all trial numbers.

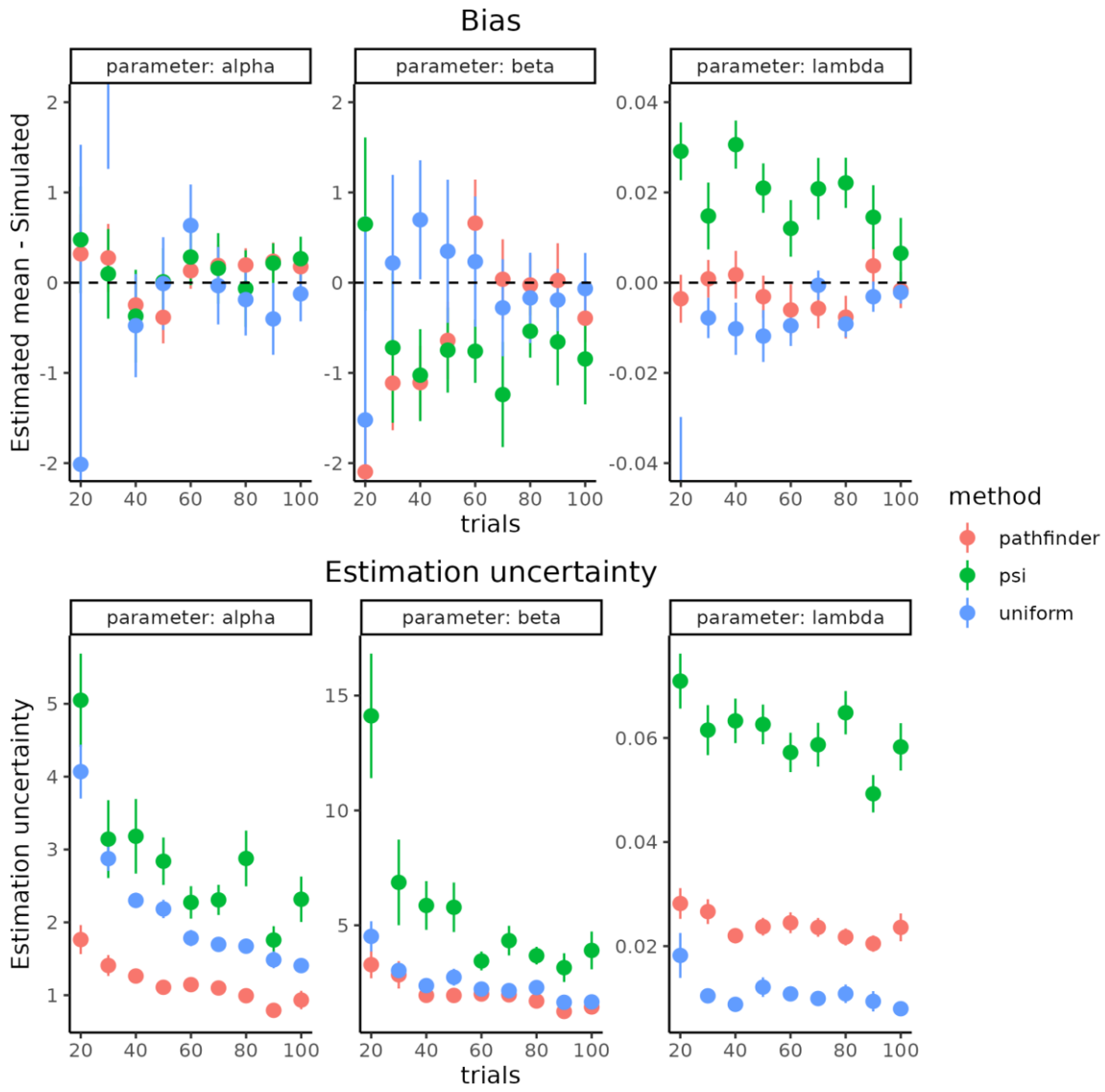


Figure 9. shows how the estimation bias (upper row) and uncertainty (lower row) changes according to the number of trials (x-axis) and parameter value estimated with the different methods of selecting the input stimulus (colors).

Having tested and compared the pathfinder algorithm, one can now examine three other focal points of minimizing estimation uncertainty, namely subjects, trials, and the influence on the mean simulated slope value. The last point is less obvious than the others but stems from the fact that increasing the slope (decreasing the steepness) of the PF will make it harder to estimate the parameters of the function. This means that estimation uncertainty is increased if other factors are held constant, the reason for this will become clear below. The number of subjects might also

influence estimation uncertainty, as the partial pooling effect of the hierarchical model will be stronger with more subjects. To investigate these three focal points, trials ranging from 20 to 200 in increments of 20, subjects between 10, 30 and 50 and lastly mean slope values of 1,2 and 3 in the unconstrained space are simulated. All other parameter values being identical to table 1. To guard against simulations that are not representative, due to either bad convergences in the ADO or in the fitting procedure, each combination was run five times. Figure 10 displays the result of this parameter recovery, across trials and group mean slope levels (i.e. simulated beta values). Figure 10 only displays the correlation approach with the inclusion of estimation uncertainty, in the upper panel, and the developed ICC_2 in the lower panel. For the two other metrics i.e. the correlation without proper uncertainty propagation and the ICC_1, see supplementary figure 3. Due to the limited influence of subjects, these have been aggregated in Figure 10. Supplementary figures 4 and 5 display the individual subject simulations separately.

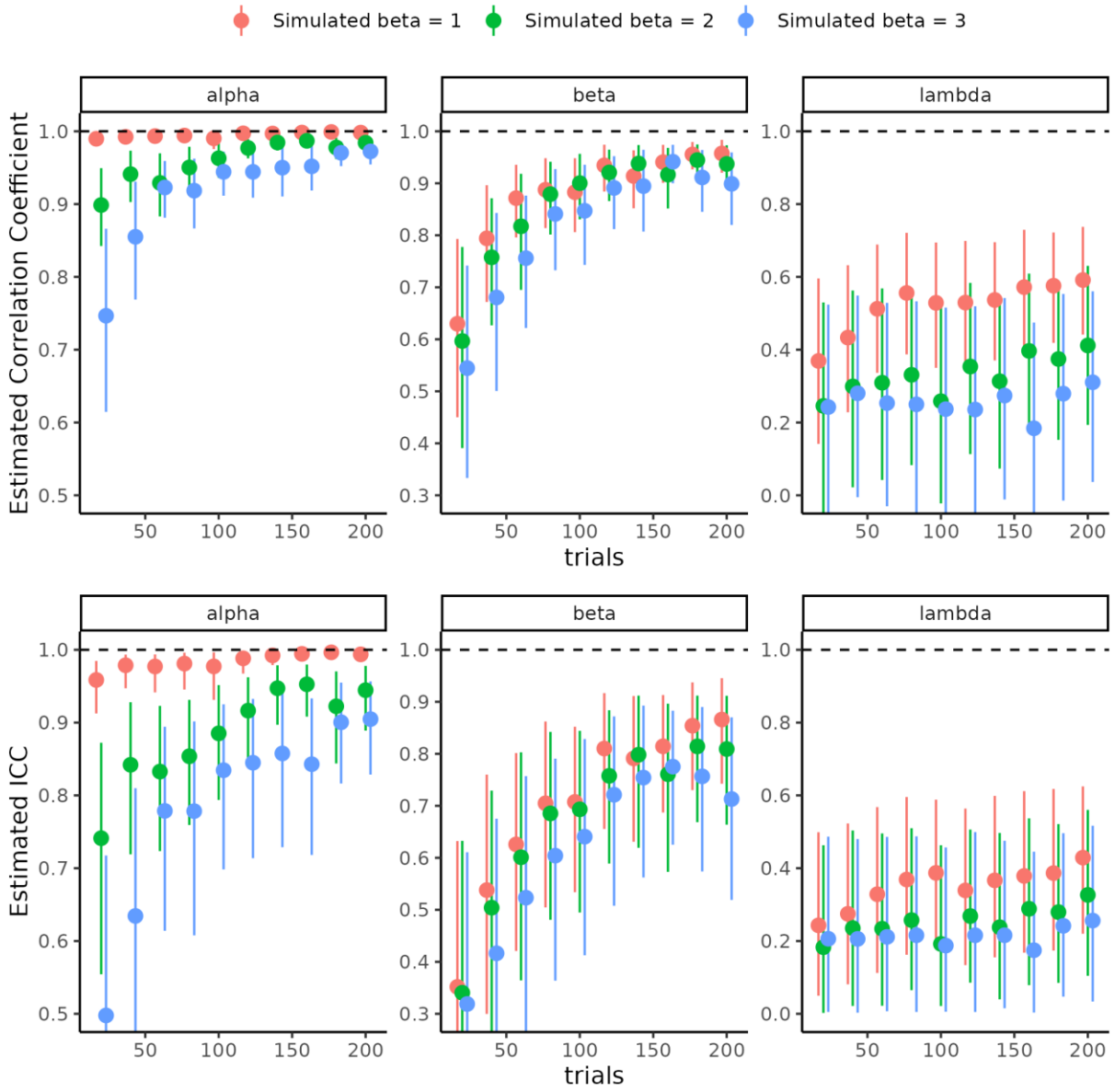


Figure 10 comparison of parameter recovery metrics, across trials and simulated steepness of psychometric function. The first row depicts how the correlation coefficient between simulated and estimated means change as a function of trials (x-axis) and the simulated mean slope (color) for each parameter of the psychometric function (columns). The bottom row shows how the estimate of ICC_2 changes based on the same metrics as the correlation coefficient. Note that the correlation coeffecient has been uncertainty propergated using bootstrapping.

From Figure 10, the main differences between the two approaches are that the ICC metric is generally lower than the correlation approach. Both approaches do asymptotically move towards one, with increasing trials and/or simulated mean slopes of the psychometric function. One way to

highlight the difference, and significance of this difference, is to plot the pairwise scatter plot of simulated vs recovered parameter estimates. These pairwise scatter plots are what both metrics in Figure 10 attempt to describe. Picking the instances where the difference between the correlation and ICC approach is the greatest will give insight to which metric might be more suitable. Figure 11 shows the pairwise scatter plot of the threshold (alpha) in three selected trials (40, 120 and 200) for both steep and shallow slopes (means of 1 and 3, respectively for beta). These instances were chosen, because the CC and ICC were similar for the steep slopes, but remarkably different with shallower slopes. Figure 11 shows why there is such a difference between the two metrics. The ICC metric is penalized considerably more by the increased estimation uncertainty and the deviation from the identity line. This is especially evident in the threshold when shallower mean slopes are used. This observation indicates that the ICC metric is more sensitive to uncertainty, compared to the CC. The same reasons apply for the difference in the slope estimate itself, and pairwise scatter plots can be found in supplementary figure 6. Lastly, both approaches suggest that the lapse rate is below the two other metrics, without much improvement with increasing trials, but with the ICC being more conservative.

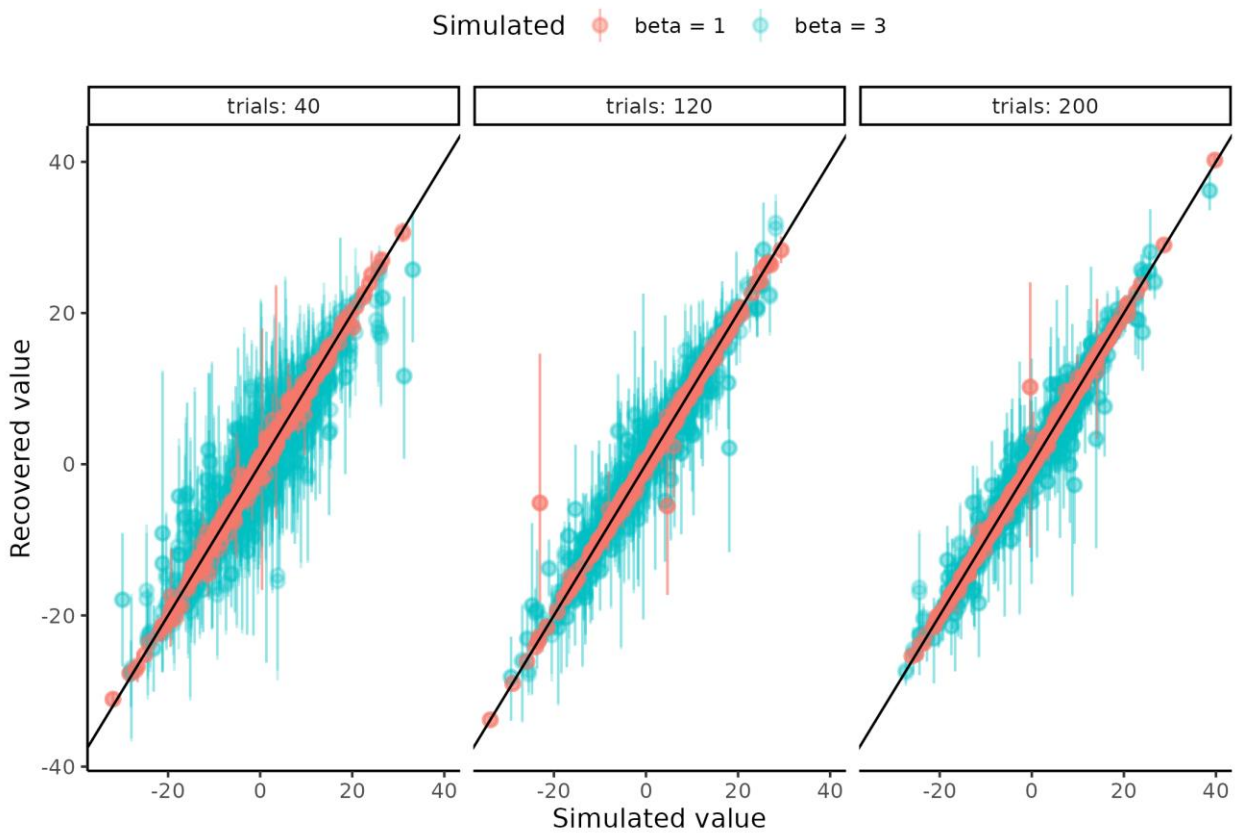


Figure 11. Showing the pairwise scatter plots of simulated vs recovered threshold (alpha) parameters when the simulated slope (beta) value is low (beta = 1) and high (beta = 3) for trials (columns).

As conveyed by the pairwise scatter plots in figure 11, the conservative ICC metric captures the fact that estimation uncertainty is a source of variability, which can be reduced, even when the CC suggests a close to perfect fit. This aligns with the behavior one would like when trying to understand and validate their model. Furthermore, the values of the ICC have a natural interpretation, as the ratio of between subject variance to total uncertainty, whereas for the CC the interpretation is not straightforward. This means that an ICC value of 0.8 indicates that 80% of the variance in the model is governed by the between subject level variance, and therefore only 20% is in the estimation or test -retest uncertainty. The ICC could be further decomposed into these constituent parts to explain what is deriving these last 20%, see supplementary figure 3. This straightforward and nuanced interpretation is not present for the CC, especially because of the arguments laid forth in the "Current problems with internal recoverability of models" section. Another import considerations, sometimes neglected in parameter recovery, is the hierarchical

structure, as mentioned in the previous section (Harrison et al., 2021; Hess et al., 2024; Hübner & Pelzer, 2020). Lastly, this approach highlights that parameter values, in a cognitive model, do not necessarily have to improve with increasing trials. This is the case for the lapse rate (lambda parameter) in this particular PF and could perhaps have been improved if the ADO algorithm was built for estimating this parameter. Nevertheless, mindlessly increasing the number of trials to hopefully decrease estimation uncertainty on a parameter, should only be done after having conducted such an analysis. This would be to ensure that resources are not wasted trying to decrease estimation uncertainty on a parameter, to a degree that is not possible, even in principle.

Increasing information in cognitive models

The previous section highlighted how the number of trials and the group level slope, but not the number of subjects, could influence the parameter recovery metric. In this section, it will be described how using data and/or information about the underlying experiment, can reduce estimation uncertainty further. Here, the incorporation of reaction times of the agents' responses are going to be used, these will serve as sources of information about the underlying PF. The focus on the reaction times is twofold; first they have a long and rigorous history in the cognitive science literature, but more importantly, are present in many experiments conducted today (Hess et al., 2024; Legrand et al., 2022; MacLeod & Dunbar, 1988; Pirolli & Anderson, 1985; Sternberg, 1969).

To incorporate the reactions times into the current formulation of the generative structure of the task, it is helpful to think of the output of PF as a probability of responding 1. Thus, in either end of the tail of the PF, the certainty with which you respond is the highest, and the midpoint between the extremes (the threshold) is the most uncertain. This descriptive formulation is what the variance of the Bernoulli distribution describes, which is the distribution that converts the probabilities from the PF to binary choices.

$$Var\big(Bern(p_t)\big) = p_t \cdot (1 - p_t)$$

Here $Var\big(Bern(p_t)\big)$ is the variance of the Bernoulli distribution at p_t, which is the probability of responding 1 at trial t. Using this information, together with the assumption that participants will respond slower when more uncertain and faster when certain, one can model the reaction times as a linear combination of this Bernoulli variance. This linear combination is thus formalized as an

intercept, to account for the individual differences, and a slope that scales the influences of the uncertainty to the variances of the underlying PF. Mathematically this would entail.

$$RT_t \sim intercept + \beta_{RT} * Var\big(Bern(p_t)\big)$$

where RT is the reaction time at trial t, intercept represents the intercept and β_{RT} represents the degree to which the uncertainty from the psychometric function influences the reaction times. Figure 12 shows a visualization of this mapping.

To stochastically model the reaction times with this formulation, a probability density function is needed to account for the noise in reaction times observed. Due to the non-negative nature of reactions times, and physical constraints of information processing (i.e. a delay from the time the stimulus is presented to which it reaches the brain of the agent), a sensible choice of this probability density function would be the shifted log normal distribution. This introduces two more variables, a non decision time (τ) and a standard deviation (σ) for the log normal distribution (Jain et al., 2015; Ranger et al., 2020). This formulation of the reactions times follows the mathematical relationship described below, where the crucial link between the psychometric function and the reaction times is the Bernoulli variance.

$$RT_t \sim LogNormal(intercept + \beta_{RT} * Var(Bern(p_t), \sigma) + \tau$$

To show how incorporation of these reaction times could help with recovery of the parameters, agents with the parameter values displayed in table 2 were simulated.

Table 2: Parameter distributions for reaction time simulations. Parameter distributions for the simulated agents and the transformations for each of the parameters when including the reaction times in the psychometric function.

Parameter	Mean	Sd	Transformation
Alpha	0	10.0	x
Beta	[1 ; 3]	0.6	Log(x)
Lambda	-4	2.0	$Logit^{-1}$(x) / 2
Intercept	-2	0.5	X
BetaRT	[1 ; 1.5]	0.3	Log(x)
Sigma	-1	0.5	Log(x)
Non-decision-time	-1	0.5	$Logit^{-1}$(x) * minRT

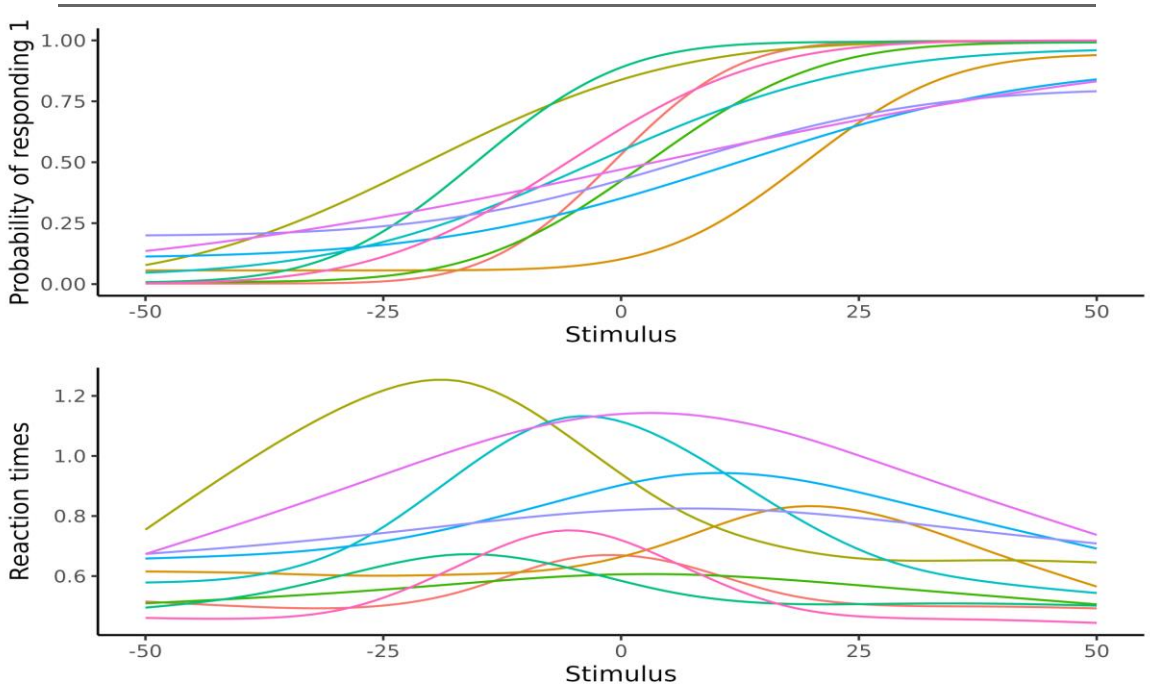


Figure 12. Visualization of the psychometric function with Reaction times. The upper panel depicts 10 psychometric functions where parameters were drawn from table 2 (Beta = 3 and BetaRT = 1.5). Lower Panel depicts the assumed relationship between the stimulus value (x) and the reaction times (y), which as can be seen is dependent on the shape of the psychometric function in the upper panel. The reaction time functions peak around the psychometric threshold and tapers off when the psychometric function asymptotes at 0 or 1.

To understand the influence of the size of coupling between the PF and the reaction times (β_{RT}), this parameter was simulated with either a high or low group mean, 1.5 and 1 respectively. The steepness of the slope of the PF also varied between high and low, 1 and 3 respectively. These values were used after having simulated and visualized the implication of them. This can be seen in figure 12, where ten simulated subjects are visualized. The figure clearly shows the relationship between the PF and the reaction time function. At high stimulus values, i.e. the most extreme x-values, the reaction times are fast, and the psychometric function is approaching 0 or 1. As the PF increases from very low stimulus values (the left side), the reactions times increase up until the threshold for that agent is reached, and then the reaction times decreases again.

Next, utilizing this model, a parameter recovery analysis can be conducted that investigates the influence of these reaction times on the recovery of the parameters. Here only the ICC_2 is depicted for the 8 combinations of slope, size of the RT coupling, and inclusion of reaction time is depicted. Similar results were obtained by using the CC, which can be seen in supplementary figure 7. Figure 13 displays the difference in parameter recovery between inclusion of reaction times in the modeling, on the 3 parameters of the psychometric function. The plot highlights increased ICC_2 values for the two parameters of particular interest, i.e. the threshold (alpha) and the slope (beta). This difference is particularly present in the slope parameter for both slope conditions (i.e. steep and shallow simulated mean slopes i.e. 1 and 3), but also in the shallow simulated slope (beta = 3), on the threshold. In these conditions the ICC_2 metric has not reached its asymptote of 1, as is the case in the steep slope simulation on the threshold. This means that inclusion of the reaction time can reduce estimation uncertainty in the slope and threshold parameter.

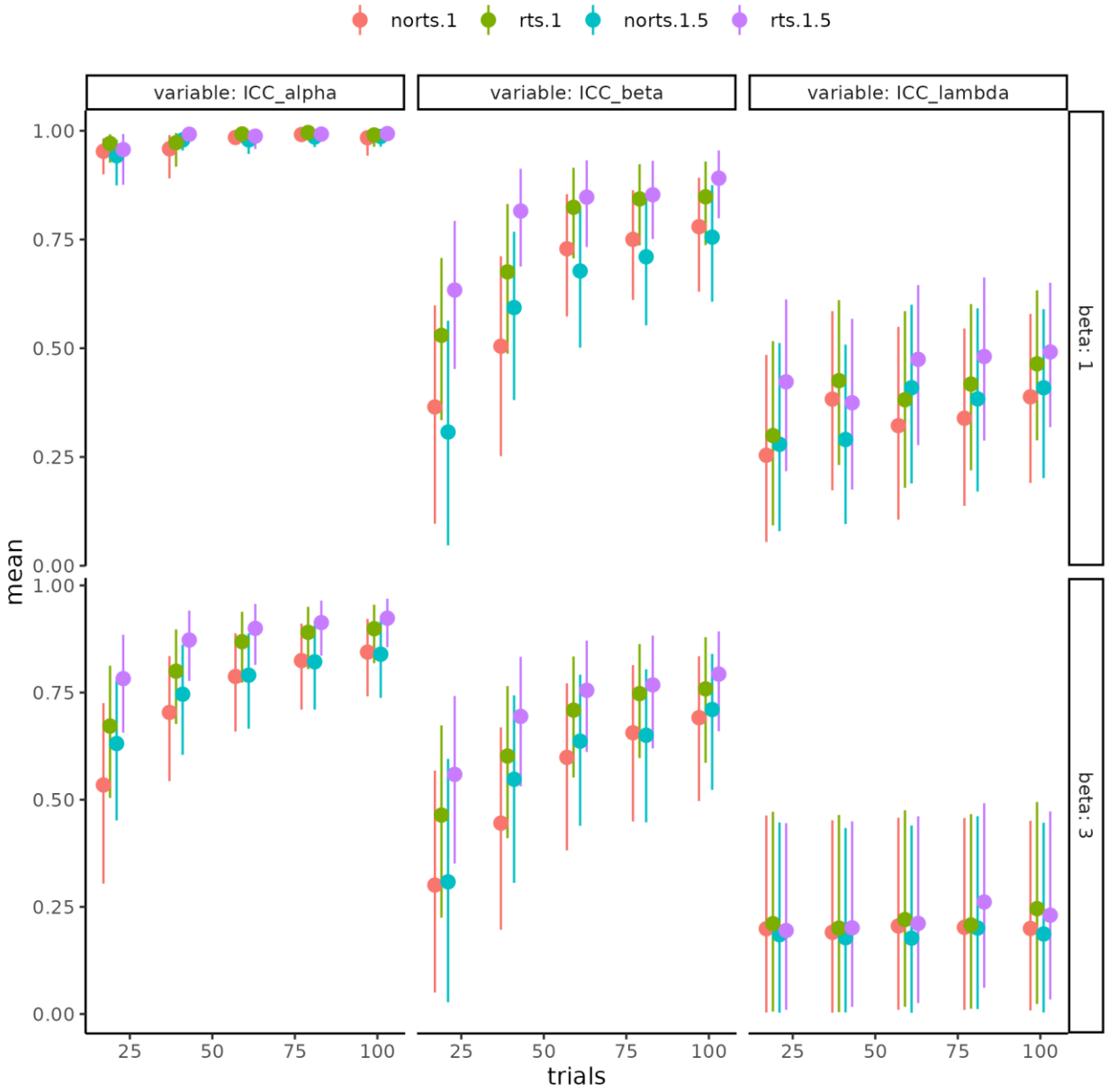


Figure 13. Parameter recovery of the psychometric function for the Intra class correlation for each parameter (columns), in each combination of including and not including reaction times and its size (color), and the simulated mean slope (rows) for differing number of trials x-axis. Stronger coupling is associated with greater intra class correlation values for both threshold slopes, with a strength dependent association i.e. the higher the coupling strength the more improve recovery when comparing to not utilizing reaction times.

Experimental data

Having investigated how the PF, both in terms of the interaction between parameter values, but also in terms of estimation uncertainty of the parameters are influenced by various factors. The thesis now turns to a re-analysis of already published data. The goal with this re-analysis is 2-fold. First, it reiterates the fact that making assumptions about the structure of the data can make big differences in the parameter estimates and their uncertainties. Second, it will serve as a starting point for understanding the utility of the internal model validity, as a metrics to gauge how trials and subjects interact on the statistical power of a model to reject a hypothesis, which will be investigated in the last section of the thesis. This final aspect of testing hypotheses will tie together how the validity steps above can help determine the ability of a particular model to conduct hypothesis testing.

Heart rate discrimination task

The Heart rate discrimination task (HRD) as introduced in Legrand et al. (2022), is an interoceptive task, entailing that participants were instructed to attend to their internal bodily states. The study recruited 223 participants, who completed the task twice, within 6 weeks between visits. HRD task has participants internalize their own heart rate for 5 seconds, meanwhile the participant's heart rate is monitored and calculated in real time. Subsequently, participants are exposed to five auditory tones with a given frequency (not the internal frequency of the tone, but the frequency of how fast the tones is presented), that is either faster or slower than their own objective heart rate. The amount this auditory tone's frequency is faster or slower, is determined by the PSI ADO algorithm introduced in the "Adaptive design optimizing" section. This means that the stimulus value for the PF for this experiment is the difference between the external tone's frequency and the observed heart rate of the participant. The binary responses are therefore either faster or slower, with faster referring to the belief that the individual heart rate was faster than the tone's frequency. For instance, one might have a heart rate of 50 beats per minute (BPM) and then hear tones in a frequency of 40 BPM, they would then be asked to respond whether this 40 BPM tone is slower or faster than their own heart rate. The authors of the experiment ran a single participant level model of each subject, in each session, and then correlated the slope and threshold of the PF. They found a medium correlation between the threshold $r = 0.5$ $p < .001$ between sessions, while the correlation for the slope was negligible $r = 0.1$, $p = .15$ (Legrand et al., 2022). This particularly

low reliability estimate of the slope of the PF, entails that this parameter is a state and not a trait of a particular individual, at least over the 6-week time span investigated. The next section investigates how these reliability estimates might change given different assumptions, on the structure of the data, but also by employing different models by incorporating additional information in terms of reaction times and confidence ratings.

The models

This section describes the models fit to the test-retest data-set described above. These models sort to examine the influence of different assumptions on the correlation between session one and two of the PF parameters. The baseline model is the single fit model and is going to be the same as fitted by the authors. This entails estimating each individual for each of the sessions separately without a lapse rate (i.e. a two parameter PF). Subsequently, a CC between the estimates from session one and two is calculated. Adding and propagating the uncertainty of these estimates will serve as the next model. Next, the same model as above with a lapse rate with be tested, in order to understand the influence of this parameter in this particular data-set. Two types of hierarchical models are going to be fit. The first is a single layer hierarchical model, amounting to modeling the two sessions from the same multivariate normal distribution with priors for each session. This model directly captures the correlation between sessions, as it is included in the variance - covariance matrix of the multivariate normal distribution. This model amounts to the model displayed in figure 3, with the participant level distributions removed. The last type of model is the nested hierarchical model (figure 3). This model assumes that all subjects have a mean level parameter which are drawn from the same population level multivariate normal distribution. Each parameter for each session is then drawn from a subject level distribution. For this last model the ICC is the statistical metric estimated by the model itself (i.e. what has been described as ICC_1), and the correlation will be calculated afterwards. In addition to examining the influence of the assumed data structure, in the fitted models, reaction times for each trial is also going to be included. This will be done in the same vein as described in the section about increasing information in cognitive models. Finally, a full model is going to be fitted, which will incorporate continuous confidence rating available in the data-set. This full model will not only incorporate the reaction times on a trial-by-trial basis, but also these confidence ratings for each trial. Confidence ratings were included in the task of the original experiment to examine the participants' interoceptive metacognitive abilities. Here these confidence ratings are used to

inform the parameters of the underlying psychometric function, similarly to the reaction times. The confidence ratings are going to be modeled in close resemblance to the reaction times, just inverted. This inversion is because at the threshold of the psychometric function the uncertainty about the stimulus representation is the highest, and therefore reaction times should be their highest, but confidence should be at the lowest. Another difference between the reaction times and the confidence ratings is their range of possible values, and therefore the probability density function used to describe them. The confidence ratings in the task were bounded between 0 and 100 ranging from complete uncertainty to certainty. The beta distribution is a natural choice of probability density function for such kind of double bounded variable, as it is bounded between 0 and 1 (Geissinger et al., 2022). The problem with using beta distribution, in this case, is the edge cases of 0 and 1's which for the confidence ratings are 0 and 100, when dividing each confidence rating with 100. One approach to circumvent this issue is to model these edge cases separately, by using a zero-one-inflated beta distribution. However, this model treats these edge values as separate processes, which does not align with the experiment, because confidence ratings are meant to represent a continuous. For simplicity, the thesis therefore subtracts a small number i.e. 0.001 from the 1 rating and adds 0.001 to the 0 ratings, making it possible to use the beta distribution for the full range of confidence ratings. Admittedly, this approach of modeling the bounded ratings between 0 and 100 is tenuous, and new methods are slowly being developed see Kubinec (2023). Reaction times of the responses were at maximum 8 seconds, and were modeled by the shifted log normal distribution introduced previously. To fully understand the parameters and implications thereof readers are referred the Github, where a shiny app has been made to demonstrate the full model (Chang et al., 2022).

Results

Table 3 displays the CC between the first and second session for the threshold and slope for each model, when uncertainty was propagated using bootstrapping. For a full table of all parameters, of all models, as well as with and without uncertainty propagation, see Supplementary table 1. This table is linked to the github of the thesis, due to the size.

Table 3. Results from reanalysis of legrand (2022). Table showing the correlation between sessions of the threshold (alpha) and slope (beta) parameter of the psychometric function using different model fomulations as well as hierarchical model structures.

alpha	beta	lapse	model	structure
0.51 [0.40 ; 0.63]	-0.00 [-0.07 ; 0.06]	FALSE	Binary	Single
0.50 [0.38 ; 0.61]	0.03 [-0.06 ; 0.11]	TRUE	Binary	Single
0.52 [0.39 ; 0.63]	0.09 [-0.03 ; 0.36]	FALSE	RT	Single
0.52 [0.41 ; 0.64]	0.04 [-0.03 ; 0.14]	TRUE	RT	Single
0.51 [0.38 ; 0.62]	0.14 [-0.05 ; 0.37]	FALSE	RT+Conf	Single
0.52 [0.39 ; 0.63]	0.12 [-0.02 ; 0.28]	TRUE	RT+Conf	Single
0.51 [0.41 ; 0.59]	0.23 [-0.01 ; 0.44]	TRUE	Binary	Hierarchical
0.49 [0.40 ; 0.58]	0.25 [0.08 ; 0.42]	TRUE	RT	Hierarchical
0.49 [0.40 ; 0.58]	0.23 [0.08 ; 0.38]	TRUE	RT+Conf	Hierarchical
0.53 [0.49 ; 0.58]	0.20 [-0.03 ; 0.43]	TRUE	Binary	Nested Hierarchical
0.54 [0.51 ; 0.58]	0.27 [0.08 ; 0.47]	TRUE	RT	Nested Hierarchical
0.55 [0.53 ; 0.58]	0.21 [0.06 ; 0.37]	TRUE	RT+Conf	Nested Hierarchical

Table 3 highlights the differences in the session-by-session correlation of the slope and threshold for the PF, when additional assumptions of the hierarchical structure is assumed (structure column). Additionally, models also included the reactions times as well as the confidence ratings (model column). Generally, table 3 shows an increase in the correlation with higher assumed structure, but also with increased complexity of the type of responses modeled. Interestingly, the models with confidence ratings included performed worse than the models with only the added reaction time, perhaps indicating improper modeling of these. The main difference in session-by-session correlation between the two hierarchical models can be found in the threshold, as the nested hierarchical model outperforms the non-nested hierarchical model.

A concern of this approach of just looking the CCs, is that a model with a high session by session correlation, might not fit the data. Therefore, an examination of the model fit is crucial, in order to ensure that the nested hierarchical model also fits the data. One approach to access this would be to examine model fit using common metrics such as leave one out cross validation, information criterion etc. The difficulty with this, is that most of the models are incompatible. This incompatibility stems from the models being fit to differing amounts of subjects, in the case of hierarchical vs single fit models, and to differing amounts of dependent variables in the case of within model architectures. Another consideration for not conducting model comparison, in the models that are comparable, is that the difference between these model is in the assumption of the data, and is therefore something that should have been decided, before modeling the data. Here, all types were used in order to investigate the differences in results. Therefore, given that was known that each subject was accessed twice (and not a new participant was tested), and that the nested Hierarchical model captures this assumption, one should be inclined to choose this model, regardless of the session-by-session correlation.

Instead of directly comparing the comparable models, one might look at posterior predictive checks. These checks investigate whether the model predictions align with the data. Posterior predictive checks were performed for the most complicated models, to ensure that the models are capturing the underlying patterns, in the collected data. Posterior predictive checks, on both group level and single subject level, can be seen in supplementary Figure 8-11 together with Supplementary Note 5.

Importance of uncertainty minimization

Throughout the thesis uncertainties, from measurements to estimations to the uncertainty of these estimations over time, have been investigated, through statistical and cognitive modeling. The section on measurement uncertainty was a brief overview highlighting ways in which computational resources can be used to account for these uncertainties. Next, the section about estimation uncertainty showed how different approaches can be utilized, from smart design of the experiment, to including additional information present, to decrease this kind of uncertainty. In the last section reliability of estimates were examined using experimental data and how the approach of adding additional information to the analysis can increase the reliability of the test re-test reliability.

To fully appreciate and explore how these uncertainties interact and their implications for hypothesis testing, the thesis will below conduct a power analysis, for the experiment analyzed in the previous section. In this power analysis, measurement uncertainty is assumed to be negligible. This amounts to assuming that participant's heart rate is estimated with infinite precision, consistent with the previous analysis, as the authors of the experiment did not disclose the uncertainty in these estimates.

Furthermore, only the simplest form of the PF with 3 parameters is going to be analyzed, focusing only on a difference in thresholds. The power analysis is limited in scope, to fully capture the potential of this way of conducting a power analysis. The simplicity of the model is for the power analysis to fully explore the effects of combinations of subjects and trials on statistical power, with the least amount of computational overhead, but see discussion and limitations for further elaboration of this. The following sections introduce power analyses, and how they are and can be contextualized.

Power analysis

When researchers are interested in the parameter values of their models, they often seek interest in how they differ by some manipulation. This could for instance be a pharmacological intervention, or a difference between healthy controls and patient populations. In such a scenario a key question is, how many participants and/or trials do I need to reliably detect a particular size of effect, between the two conditions? These estimates of trials and participants can, in principle, be calculated a priori to conducting the experiment, given some assumptions. This type of a priori

analyses thus tries to answer the question of, what is the probability that my results are going to be "significant", given some "real" underlying effect. Here "significant" refers to the standard frequentist approach of rejecting or failing to reject a null hypothesis, based on a significance level.

Usually, this concept of hypothesis testing is illustrated in a 2 by 2 matrix, with the real latent effect being in one dimension, and the model results in the other dimension (see table 4). The probabilities of landing in either of the four categories are usually described as functions of our statistical significance threshold (alpha / p-value), and the statistical power of our model and test (1-β). This framing of power analyses, is thus to imply that results are significant, if the p-value is less than a particular value (5%), and that the probability to detect this effect, given that it is present, is another arbitrary value, typically set at 80% (Chén et al., 2023; Dumas-Mallet et al., 2017).

Table 4. 2 by 2 confusion matrix of whether the is an underlying effect (Reality) and whether a model can correctly identify this effect or not whether its present or not.

	Reality (effect)	Reality (no effect)
Model result (significant)	1-β	α
Model result (in-significant)	β	1-α

Power analysis in practice

Our models in Cognitive Science will reject and fail to reject, different effect sizes at different rates, based on their magnitude, as well as the amount of data i.e. the number of subjects and/or number of trials. Increasing the number of subjects and/or trials serves to reduce the uncertainty in the estimated effects and thereby increasing the probability of detecting said effect.

With this understanding, the commonly depicted table above (table 4) is somewhat misleading, as the dimension of "reality" is a continuous variable of the size of the effect. Our models then have a specific probability of rejecting a hypothesis, based on the effect size observed at a particular set of subjects, trials and significance level. For example, consider a researcher wanting to detect an effect of gender on height, in the human population. Assuming an underlying

effect, the researchers observe X females and Y males and conducts a statistical analysis to determine whether he can reject the null hypothesis (there are no differences in height, in the two genders). Compare the above hypothesis to the hypothesis that there is an effect of age (late adolescents vs. adults) on height. The gender difference is generally larger than the age difference, and therefore with all else being equal (trials, subjects, statistical model etc.) this difference will be easier to detect compared to the difference in height, based on age. Therefore, in conducting power analyses observed effect sizes are simulated, (effect sizes in the data that is observed) with differing amounts of trials and subjects. The ability of the statistical model to reject these simulated experiments is then accessed. Usually, this involves counting the number of times the model achieves "significant" results, compared to non-significant results, which then represents the power of the model, at that number of trials, subjects and observed effect size. This approach accurately captures how we expect the model to behave when we fit the data to the model after obtaining it. It tells us if we observe a particular effect size, we will with a specific probability be able to reject the null hypothesis. The utility of such analysis therefore lies in being able to examine how many subjects and/or trials are needed, to obtain a statistical power of usually 80%, given that a particular effect size in the population is present. The assumed effect size in the population might be informed by previous studies and/or meta-analyses in the field. Additional assumptions are then needed, in order to approximate the distribution of effect sizes, as these statistical metrics also have uncertainty associated with them.

The power simulations conducted in this thesis will focus on a repeated measures design investigating a threshold difference, due to some intervention. Subjects, trials, and effect sizes in a variety of combinations are therefore simulated (figure 14). The choice of effect size metric was the Cohens' $d_r m$, as seen in the formula below. This particular effect size is suitable for repeated measures design because it accounts for the correlation between the two sessions of each participant, i.e. the test-retest reliability of the metric investigated.

The simulation process followed these steps. First, a set of agents were simulated from a multivariate normal distribution with two sessions, the parameters were informed by the group level parameters of the binary nested hierarchical model, presented in section about experimental data, (see supplementary table 2, for the exact values for each parameter). Next, the thresholds of the second session for the agents were increased by a random variable, drawn from the difference distribution. This difference distribution was calculated based on the two equations presented

below, where the second session variance was defined as 1.5 times the variance of the first session. To ensure a particular observed effect size, this process was repeated until an observed effect size of the desired value, was obtained within $\pm$ 0.01. This step of re-sampling for a particular effect size was mainly for visualization purposes (see Figure 14 and the accompanying text). After simulating the parameter values of each agent at each session, the agents were put through the pathfinder algorithm to obtain their trial-by-trial stimulus values. The complete trial-by-trial data-set was then fitted using a single layered hierarchical model, where the threshold was parameterized as a linear combination of an intercept, and a dummy coding of session with a difference parameter (supplementary note 6, for the full model description).

$$\mu_\delta = d_{rm} * \frac{\sqrt{Var_1 + Var_2 - 2 * \sigma_1 * \sigma_2 * \rho}}{\sqrt{2 * (1 - \rho)}}$$

$$\sigma_{dif} = \sqrt{Var_1 + Var_2 - 2 * \sigma_1 * \sigma_2 * \rho}$$

Mean and standard deviation of the difference distribution between the two sessions. Where Var_1 is the variance of session 1 and Var_2 is the variance of session 2. ρ is the correlation coefficient between the two sessions and μ_δ is the mean of the difference distribution with d_{rm} being the standardized effect size between the two sessions.

Power analysis results

Understanding the goal of conducting a power analysis, it can be difficult to choose the number of combinations of trials, subjects and observed effect sizes to explore. The space of these combinations is theoretically infinite, and practically also quite large. Therefore, conducting a complete power analysis with all combinations of trials, subject and observed effects sizes for a single model is unfeasible. However, this thesis will demonstrate that the variation in how well the PF rejects the null hypothesis, given subject and trial combinations is stable over observed effect sizes. This stability allows for possible ways to give good predictions of power, without needing to simulate the exact number of participants or trials. This procedure is therefore about simulating a set of trials, subjects and observed effect sizes and then extrapolate from these simulations.

For the current power analysis, a result is considered significant if less than 5% of the posterior difference distribution of the threshold crosses 0, analogous to setting an alpha value of 5% in a frequentist power analysis. Furthermore, 100 simulations are going to be run for each

subject, trials observed effect size combination to ensure reasonable estimates on the probability of rejecting the null hypothesis. To effectively display the raw results of the power analysis and facilitate visual comparison of the effects of trials and subjects, the beta distribution is going to be used. A beta distribution is going to be used to aggregate and propagate the uncertainty of the 100 simulations for each effect size. This is done by utilizing that the beta distribution can be parameterized with one parameter counting the number of times an event has happened, while the other parameter counting how many times this event did not happen.

Starting with what is analogous to a uniform prior, on the probability of rejecting the null hypothesis (Beta (1,1)), it is possible to update this probability density function with the amount of significant or non-significant results in the 100 simulations. This updating process will result in a probability density function, that contains the information in the 100 binary points (i.e. significant or not simulations). Figure 14 shows each trial/subject combination with points representing this prior uniform beta distribution being updated by the 100 data-points.

Several important observations are worth noting when viewing Figure 14. The shape of the points, for each trial/subject combination, very closely resembles a psychometric function, where both subjects and trials influence the steepness and the location of the function. This suggests that increasing the number of subjects, and trials to a lesser extent, has two important features. Firstly, it shifts the points towards higher power, with lower effect sizes. Additionally, it seems to increase sensitivity to the observed effect size, as evident by the slope of the curve getting steeper, with higher number of subjects. Investigating the number of trials' influence on power, there seems to be large diminishing returns. This means that the effectiveness of increasing the number of trials, to achieve higher power, is highly dependent on the number of trials itself. In practice and as shown in figure 14, increasing trials from 10 makes a big difference in the shape of the function, but the difference between high and very high i.e. 100 to 150 trials, has a lesser effect. The tendency of the function to be less affected by ever increasing trials is also present for the number of subjects, to a lesser extent.

This observation aligns with the expectation when considering the function at its extremes, in terms of trials and subjects. When subject and trial numbers approach infinity, one would expect, assuming the model has been shown to become increasingly better with increasing trials (like with the ICC metric presented previously), that the model would be able to detect even the smallest difference in groups. This would essentially mean that the function would consistently be at y = 1,

with x approaching 0 from the positive direction, and then jump to (0,0) in the (observed effect size, power) curve, as no difference would entail no power. Conversely, when no subjects or trials are present the curve should approach a flat line at y = 0, entailing no power for any amount of effect size. Essentially, the function would asymptote to a step function in the limit when x goes to 0, and subjects and trials goes to infinity. This can mathematically be written as:

$$\lim_{(s,t)\to\infty}\left(\Psi(x,\alpha,\beta,s,t)=\begin{cases}0 & \text{if } x\leq 0\\ 1 & \text{if } x>0\end{cases}\right)$$

where s and t are the number of subjects and trials respectively, and x, α and β are the observed effectsize, the threshold of the psychometric function and the slope of the psychometric function, respectively.

These observations will be used in the next section to extrapolate the results from figure 14. This will enable the possibility of constructing a model that maps trials, subjects, and effect sizes to statistical power.

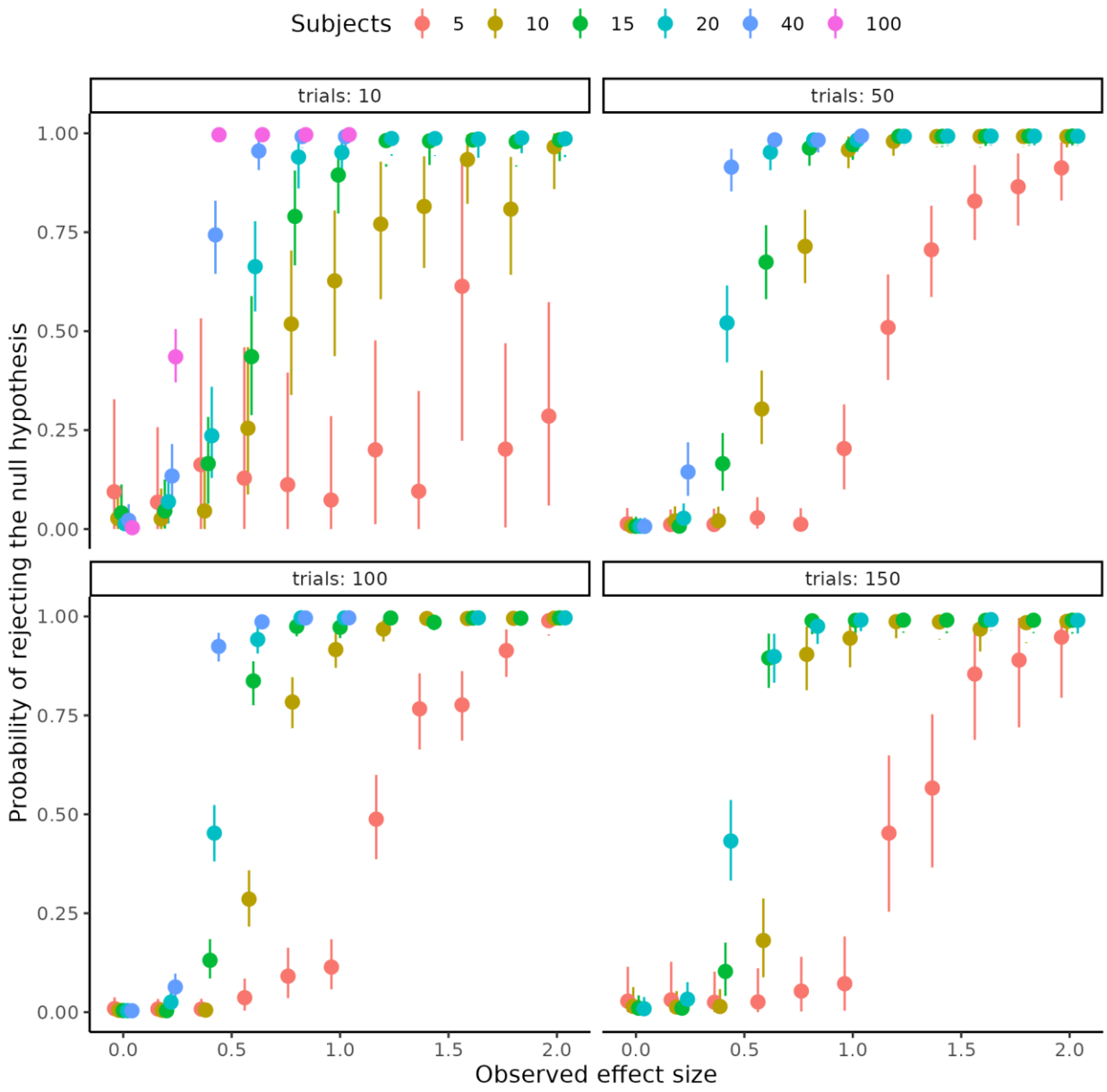


Figure 14 depicts power as a function of observed effect sizes in different combinations of trials and subjects. The x-axis represents the observed (& simulated) effect size with the y-axis depicting the statistical power of the model, i.e., the proportion of rejected null hypothesis to failed rejections.

Modeling of power analysis

Using the information from above, one needs to investigate the latent psychometric function describing the relationship between subjects, trials, and effect sizes. Ideally, this psychometric function enforces the curve going through the origin, as an observed effect size of 0 should always

entail no power. Next the parameters of this psychometric functions, i.e. the threshold and slope, need to be parameterized by the number of trials and subjects. This parameterization of trials and subjects should ensure that the function moves towards a step function right as x becomes greater than 0, when subjects and trials approach infinity.

Before fitting the general case, used for extrapolation, but also ensuring that a psychometric function is well fitting function to the problem, each set of trials and subject combinations are fit independently to the parameters of the psychometric function. This involves estimating the threshold and slope of the psychometric function for each trial and subject combination. This steps makes it possible to ensure that fitted functions pass through the points, depicted in figure 14, which will increase confidence in the following type of modeling.

Several types of psychometric functions might be used for this type of analysis, where the goal is out of sample predictability and/or extrapolation. This would mean that the best model could be selected based on leave one out cross validation. The ideal model is the model that can best describe new data, as we want to use the function for prediction on not already simulated data. This is because the overall goal with this power analysis is to use the quite sparsely simulated space of trials, subjects, and effect sizes, depicted in figure 14, to inform a model that can predict outside the realms, which it has been tested on. Therefore, these models were compared using the Pareto smoothed importance sampling leave one out cross validation (Vehtari et al., 2017, 2024; Yao et al., 2018).

Three types of psychometric functions were fit, the cumulative normal, the cumulative logistic and the cumulative Weibull function. The main differences between the normal and logistic function are that the logistic function has heavier tails than the normal allowing for more disperse observations. The difference between the Weibull and the two other functions is that the Weibull function is forced through the origin, resulting in a distinct shape compared to the other two functions.

The choice of the cumulative normal or logistic function does not necessarily violate the assumptions laid out above. This is due to the way that the parameters are going to be dependent on the trials and subjects. This can be understood if one considers an asymptote at 0, for the slope and threshold (i.e. a step function also for the cumulative normal and logistic function), when trials and subjects move to infinity. This exactly aligns with the observation from above, that the psychometric function approaches a step-function (as the slope gets closer to 0) and that the

location of this step function approaches x = 0, but never reaches it with increasing trials and subjects.

The results of the preliminary independent analysis on trials and subjects can be seen in figure 15, where the independently fit logistic psychometric functions are overlaid on the observed data-points from figure 14. The figure highlights a good fit for most of the trials and subject combinations (i.e. functions passing through the points).

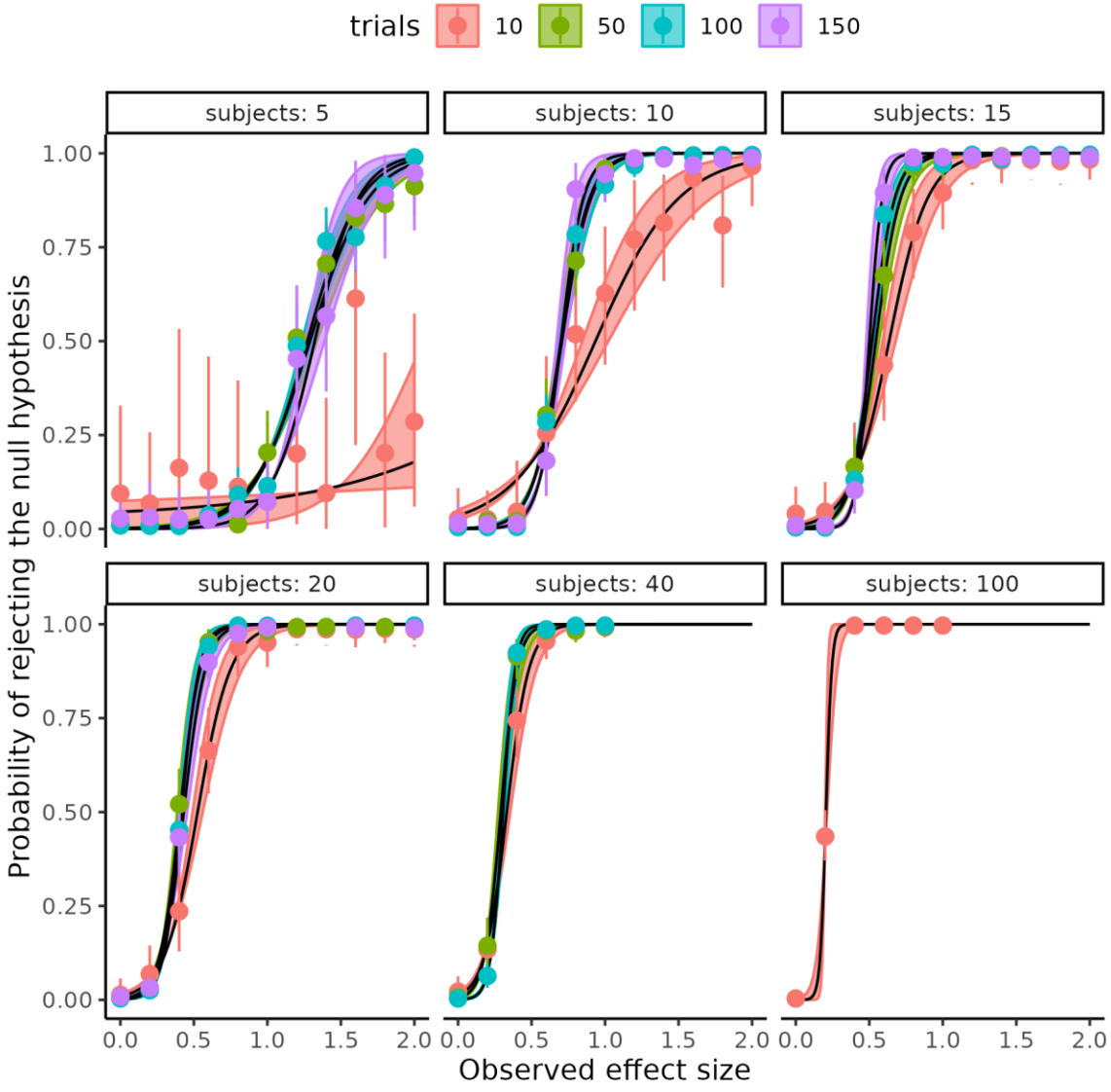


Figure 15 depicts power as a function of observed effect sizes in different combinations of trials and subjects. Lines and shaded area represent the mean and 95 % credibility interval of the **independently** fit logistic psychometric function to each trial by subject combination.

Continuous mapping of the power analysis

Moving to the continuous mapping of subjects and trials, to the psychometric function's parameters. This mapping needs to be defined as a function that relates subjects and trials, to the slope and threshold of the psychometric function. As argued above, the steepness increases, and the threshold moves to the left with increasing trials and subjects following a pattern of diminishing returns. A first choice of this mapping function could therefore be to model the two parameters as exponentially decreasing by trials, subjects, together with their interaction. An exponentially decreasing function in the complete general case would mean the following relationship.

$$\Theta = \beta_0 * exp(-\beta * X) + \alpha_{asym}$$

Here Θ represents the parameters of the psychometric function, where α_{asym} denotes the parameter value when the number of trials and subjects approach infinity. β is vector of parameters determining the steepness of the exponential decrease from the co-variates in the matrix X, here trials subjects and their interaction. The parameter β_0 serve, together with α_{asym}, as the value of the parameter when trials and subjects are 0.

Another formulation of the dependency might be a power law equation, as shown below.

$$\Theta = \beta_0 * X^{\beta}$$

Both approaches can produce the observed behavior, and difference in the two formulations depends on the underlying relationship between the parameters and the matrix X (i.e., trials and subjects and perhaps their interaction). The exponential equation assumes that as trials and subjects increase by a fixed amount, the parameters will decrease by a percentage. The power law on the other hand assumes that as trials and subjects increase by a percentage, the parameters will decrease by a percentage.

Several ways of investigating which of these two approaches results in the better fit. Firstly, plotting the parameters of the independent fits (figure 15) vs trials and/or subjects, in two different coordinate systems, either (log(y), x) or (log(y), log(x)). Which of these coordinate systems produces the best-looking linear line would be the best candidate. Supplementary Figure 12 displays the three function's parameters fitted independently on each of the two coordinate scales. Using this approach no obvious differences were found.

Another approach involves fitting both types of models, and then comparing them on leave one out cross validation as described above. However, this approach revealed problems with 15, 25 and 3 % of observations for the normal, Weibull and logistic functions, respectively. These

percentages were accessed using the pareto k diagnostic value, which was above 1 for these percentages of data-points. Essentially, this renders the comparison meaningless (Vehtari et al., 2024).

Investigating these functions, the logistic cumulative function produced the least amount of problems, with pareto k values. This function was therefore used when fitting trials and subjects as continuously informing the parameters of the latent psychometric function. The first model was the exponentially decreasing function. Four other models were fitted, with different parameterizing of the power law equation. These four models had different approaches to modeling trials and subjects and their interaction, as there is no straightforward way of combining X and β. The first power law was an additive model, with the following parameterization.

$$\beta_0 \cdot X^{\beta} = \beta_{01} + s^{\beta_1} + t^{\beta_2} + (t \cdot s)^{\beta_3}$$

The second power law, with a combination of additive and multiplicative operations:

$$\beta_0 \cdot X^{\beta} = \beta_{01} * \left(s^{\beta_1} + t^{\beta_2} + (t \cdot s)^{\beta_3}\right)$$

The third power law was a multiplicative model without an interaction.

$$\beta_0 \cdot X^{\beta} = \beta_{01} \cdot s^{\beta_1} \cdot t^{\beta_2}$$

The last power law was the multiplicative model with an interaction, but defined as the sum of subjects and trials as the normal interaction of multiplying trials and subjects would lead to a similar model, of the model without an interaction.

$$\beta_0 \cdot X^{\beta} = \beta_{01} \cdot s^{\beta_1} \cdot t^{\beta_2} \cdot (t + s)^{\beta_3}$$

Comparing these five models, with leave one out cross validation showed that the best model was the last power law model, but closely followed by the second power law model, which can be seen in table 5. Importantly for these reported models, the diagnostic values were all below 0.7.

Table 5. Model comparison of the power analysis models, using Pareto smoothed importance sampling leave one out cross validation. Expected log predictive density (elpd) difference and standard error between models is depicted in the second and third column and the absolute ratio between these in the fourth column. The Higher the elpd-ratio the bigger the scaled difference (scaled by the uncertainty) between the models and the more confident one might be that one model outperforms the other.

models	elpd_diff	se_diff	elpd_ratio
logs_power	0.00	0.00	
addititve_multipli cative	-10.06	4.65	2.16
logs_power_noint	-35.91	8.87	4.05
logs_expo	-85.86	15.74	5.45
additive	-820.02	37.48	21.88

Table 5 indicates that as the trials and subjects increase by a percentage, the parameters of the psychometric decrease by a percentage, as the top two models, which are both variations of the power law. To verify that the tested models capture the underlying simulation, Figure 16 displays the winning model superimposed on the data, with 95 credibility intervals of the mean. As seen, this closely resembles the individual independent fits, with the most drastic deviation in the 5 subjects and 10 trials condition.

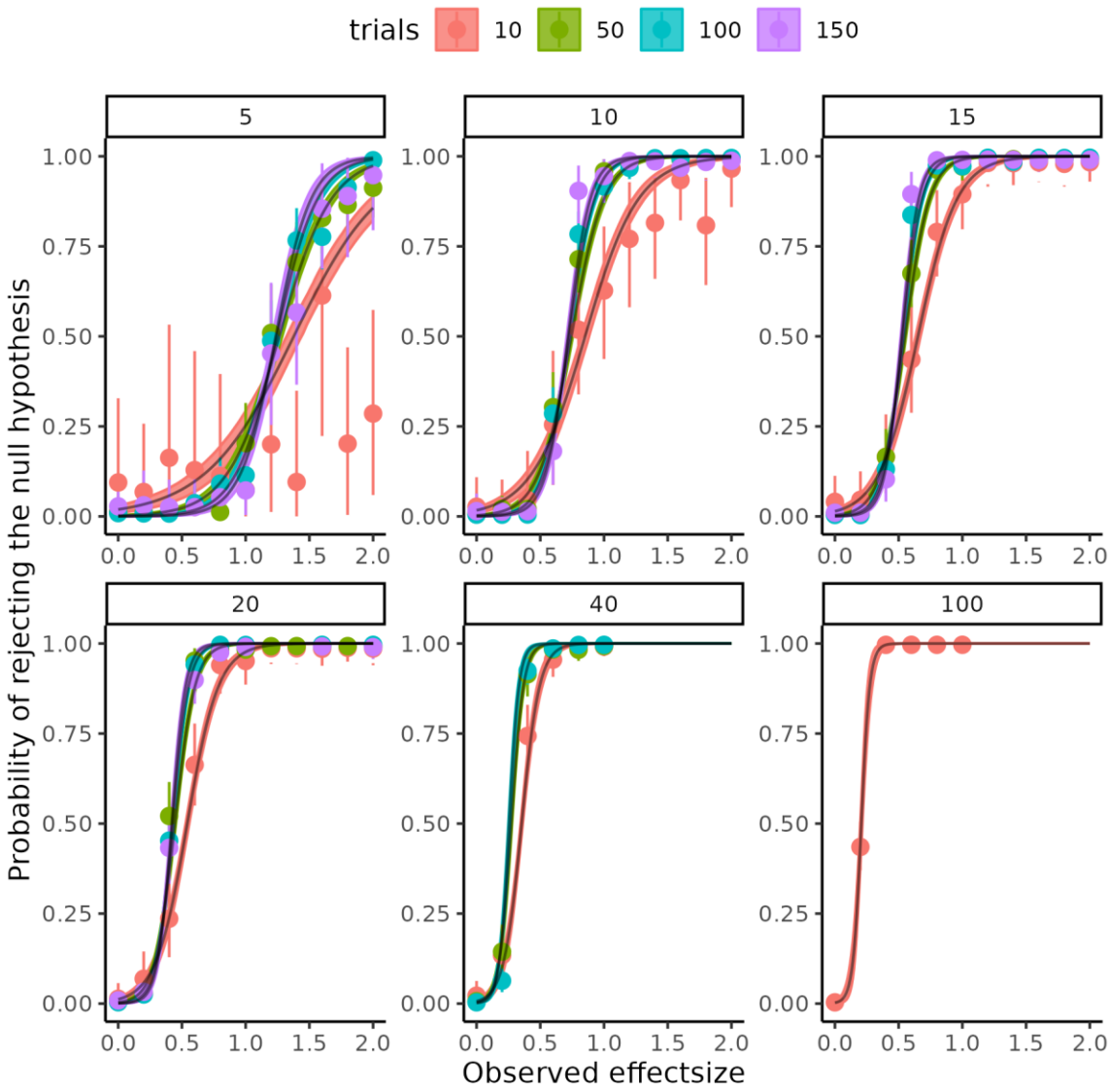


Figure 16 depicts power as a function of observed effect sizes in different combinations of trials and subjects. Lines being the **dependently** fit logistic psychometric functions to each trial by subject combination. With the shaded area being the 95-credibility interval.

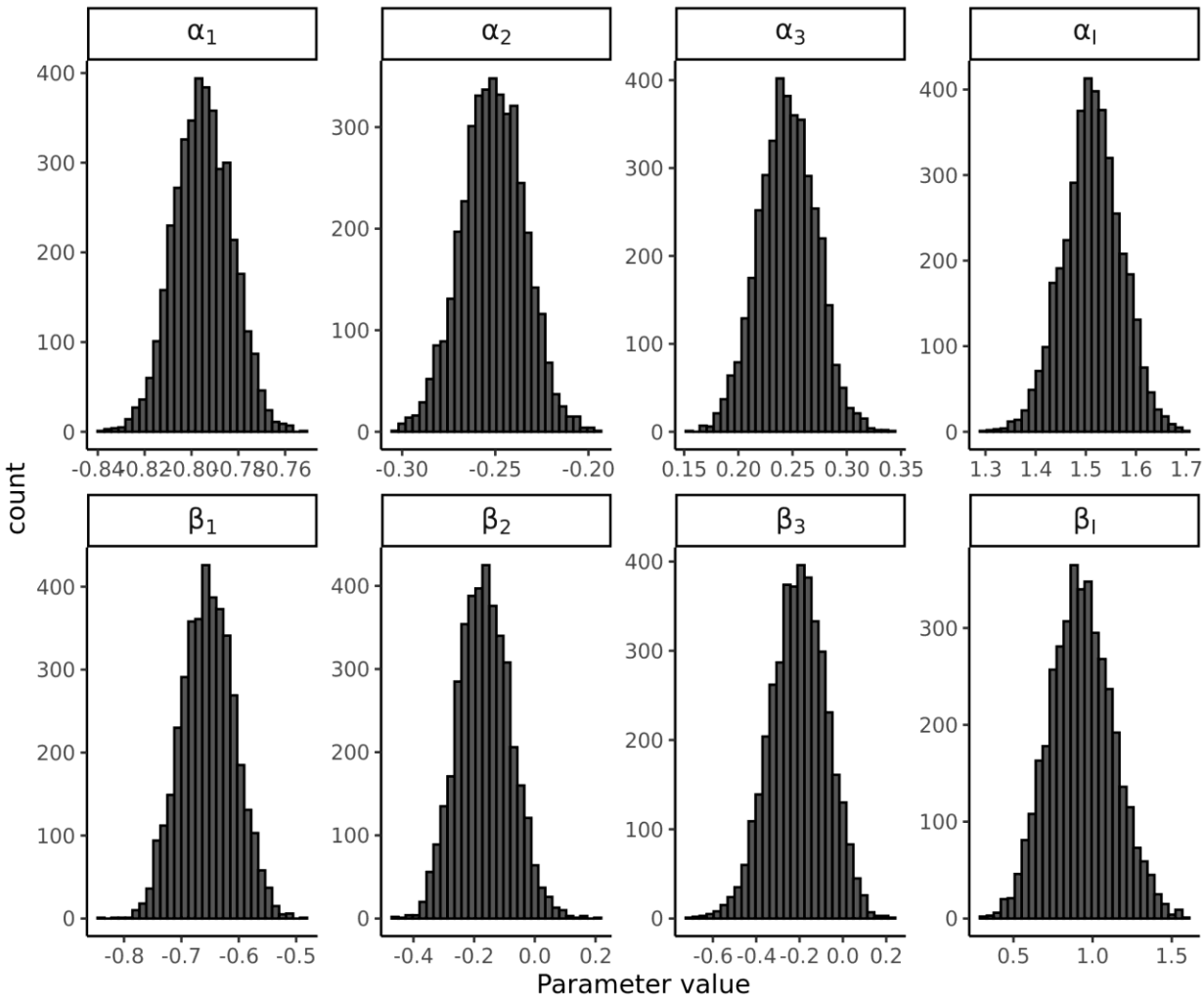


Figure 17. Marginal posterior distributions for the winning model's parameters

The marginal posterior distributions of the parameters of the winning model are displayed in figure 17. This means that the best the underlying function transforming trials, subjects and observed effect sizes into probabilities of rejecting the null hypothesis of no difference in threshold, follows:

$$\Psi(d_{\text{obs}}, \alpha, \beta \mid t, s) = \frac{1}{1 + \exp\left(-\frac{1}{\beta(t,s)} \cdot \left(d_{\text{obs}} - \alpha(t,s)\right)\right)}$$

Where

$$\beta(t,s) = \beta_I \cdot s^{\beta_1} \cdot t^{\beta_2} \cdot (t+s)^{\beta_3}$$

$$\alpha(t,s) = \alpha_I \cdot s^{\alpha_1} \cdot t^{\alpha_2} \cdot (t+s)^{\alpha_3}$$

Each of these parameters are given by the distributions depicted above (figure 17).

Utility of the power analysis

As alluded to in the initial section of the power analysis, the work presented here should be able to help an independent researcher. It should do this by helping to determine the probability of rejecting a particular observed effect size, given trials and subjects, even outside the realm of simulations presented here. However, for a researcher to utilize this function to calculate the probability of rejecting a null hypothesis, given a particular effect size in the population, further assumptions need to be made. This is because the effect size when conducting an experiment is not a fixed quantity.

In practice, this means that when conducting an experiment, we observe an effect size that is assumed to be drawn from a latent effect size distribution in the population. Mathematically, this means that the observed effect size that in an experiment is a random variable. The mean and standard deviation of this random variable is given analytically by Cohen, but could also be derived from bootstrapping (Goulet-Pelletier & Cousineau, 2018; Hedges & Olkin, 2014; Lakens, 2013). Below are the equations for the mean and standard deviation of the effect size measure used in the power analysis.

$$\mu_{d_{rm}} = \mu_{\delta} * \frac{\sqrt{2 * (1 - \rho)}}{\sqrt{Var_1 + Var_2 - 2 * \sigma_1 * \sigma_2 * \rho}}$$

$$\sigma_{d_{\mathrm{rm}}} = \sqrt{\frac{1}{n} + \frac{\mu_{d_{rm}}^2}{2 \cdot n}}$$

The equation for the mean effect size $\mu_{d_{rm}}$ is mathematically identical to the definition shown in the "Power analysis" section. The standard deviation of the random variable $\sigma_{d_{\mathrm{rm}}}$, is defined as a function of the number of subjects n and the size of the effect itself $\mu_{d_{rm}}$. Assuming that the effect size is normally distributed:

$$d_{obs} \sim N\left(\mu_{d_{rm}}, \sigma_{d_{rm}}\right)$$

The probability of rejecting (R) this sampled effect size (d_{obs}) is given by the function that was obtained above.

$$P(R \mid d_{obs}) = \Psi(d_{\mathrm{obs}}, \alpha, \beta, t, s)$$

Ideally the probability of observing a particular effect size AND reject the null hypothesis given this observed effect size is seeked. Probability theory, particularly conditional probabilities, gives us the relationship between these quantities.

$$P(R \mid d_{obs}) = \frac{P(R \cap d_{\text{obs}})}{P(d_{\text{obs}})}$$

Here $P(R \cap d_{\text{obs}})$ represents the probability that we are interested in, i.e., rejecting AND observing a particular effect size.

$$P(R \cap d_{\text{obs}}) = P(R \mid d_{obs}) \cdot P(d_{\text{obs}})$$

Integrating over all possible values of the effect size is necessary to integrating out the effect size, i.e., marginalizing.

$$P(R) = \int_{-\infty}^{\infty} P\,(R \mid d_{\text{obs}}) \cdot P(d_{\text{obs}})\, d(d_{\text{obs}})$$

Which becomes:

$$P(R) = \int_{-\infty}^{\infty} \Psi\,(d_{\text{obs}}, \alpha, \beta, t, s) \cdot N\left(\mu_{d_{rm}}, \sigma_{d_{rm}}\right) d(d_{\text{obs}})$$

Instead of trying to analytically solve this integral, one can leverage computational resources to approximate it. This can be done by taking draws of the normal distribution of the observed effect size, and then applying them through $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$. The result of such calculation will give draws from a probability distribution of rejecting the null hypothesis, i.e., P(R). The last step is to calculate the proportion of rejected null hypotheses ($p < 0.05$), to the total number of draws. Which would entail the power of the study, assuming the mean difference and variance in the two sessions.

Sampling variability of the effect size.

The above high-level explanation of calculating power for an experiment might be quite difficult to understand, and therefore implement for independent researchers. To make this more accessible, I will demonstrate below how this can be done using the concepts described above. The next section will provide a practical understanding of the parts that should go into a power analysis and how different factors will influence power.

Firstly, I'll examine and show the influence, and need, for the sampling distribution of the observed effect sizes. To demonstrate this, it is assumed that the group mean difference of the threshold in the psychometric function is -5, and the variance in the second session is 1.5 times the variance of the first session. These assumptions entail that the intervention increases the variation in the threshold, but that there is a clear effect of the intervention of the threshold. The assumptions for the choice of mean difference and difference in variance can be visualized by repeated sampling from a multivariate normal distribution with the following parameterization:

$$\binom{x_i}{y_i} \sim \mathcal{N}\left(\binom{\mu_1}{\mu_2}, \begin{bmatrix} \sigma_1^2 & \sigma_1 \cdot \sigma_2 \cdot \rho_{12} \\ \sigma_1 \cdot \sigma_2 \cdot \rho_{21} & \sigma_2^2 \end{bmatrix}\right)$$

Here μ_1 and σ_1 are given by the re-analysis of the experimental data and were -8, 8 respectively. Given the assumptions above, μ_2 and σ_2 become -3, 10. The number of subjects is then varied by only drawing a particular number of random variables from this multivariate normal ($s \in (10,40,70,100)$). To also investigate the effect of the correlation coefficient ρ on the distribution of effect sizes i.e. $p(d_{obs})$, this parameter is also varied ($\rho \in (0,0.3,0.6,0.9)$).

The results of this simulation can be seen in figure 18. Here it is shown that both the sample size i.e. subjects, but also the correlation coefficient between the sessions is important for the variances of the observed effect size.

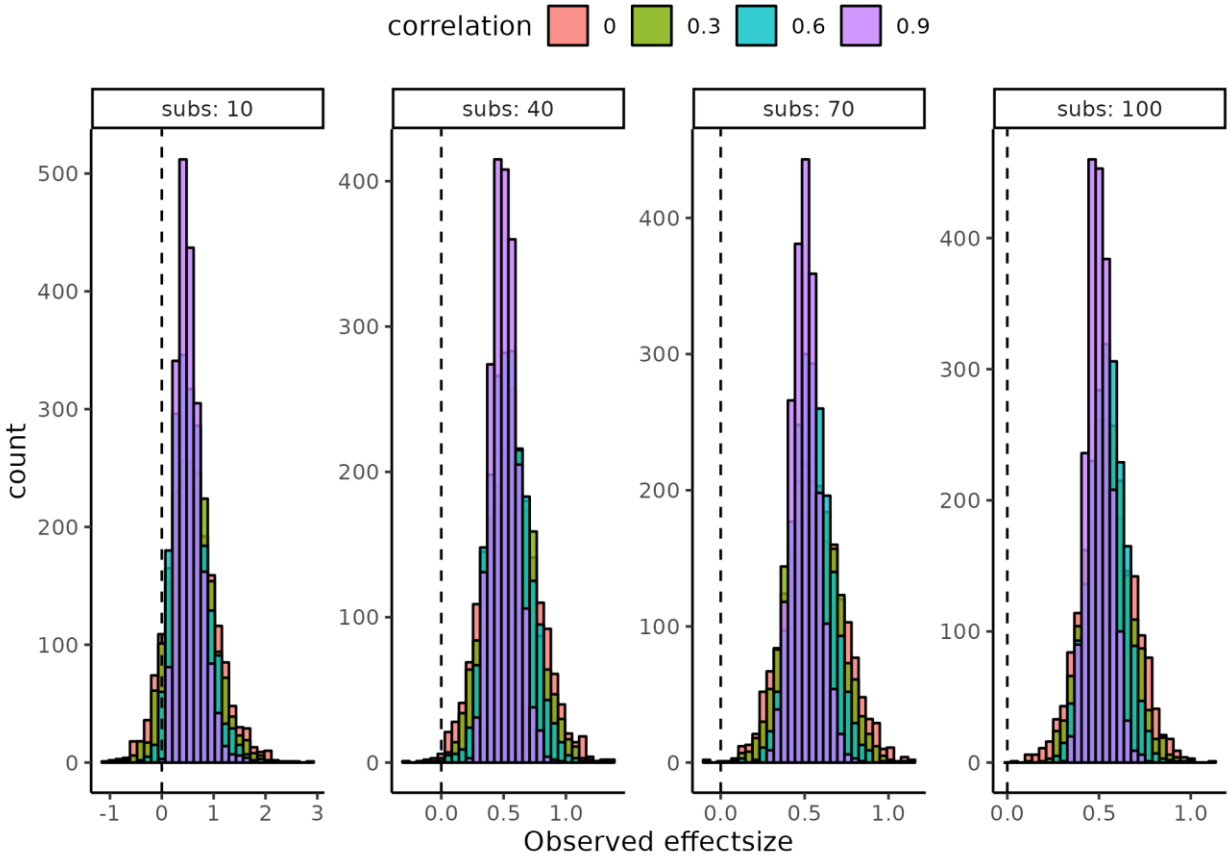


Figure 18. Sampling distributions of effect sizes across subjects (facets) and session by session correlations (colors)

Now, it is possible to visualize how these observed effect size distributions fit into the probability of rejecting the null hypothesis, i.e., $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$. Note, the observed effect size

distributions above are not dependent on the number of trials, in the experiment. The function derived above from the continuous power analysis incorporates this information, together with other factors, that might change the power of the experiment, given an observed effectsize distribution. As shown above, the implications of the function $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$ can be visualized as psychometric functions in a (d_{obs} , $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$) coordinate system with trials and subjects being fixed. Another more informative way to investigate varying number trials and subjects, is to visualize these implications in a 3-dimensional grid of (Subjects, d_{obs} , $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$) with facets being a particular set of trials. This visualization can also serve the purpose of projecting the above distributions (figure 18), unto the space of $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$.

Figure 19 displays the projection of the histograms from Figure 18 as ellipse, where the vertical width of the ellipse (the major axis) is given by the 95% Highest density interval of the histograms and the horizontal width (the minor axis) is for visualization purposes. In Figure 19, it is shown how the correlation coefficient, the number of subjects as well as the number of trials, affect the power (the background color), given an observed effect size.

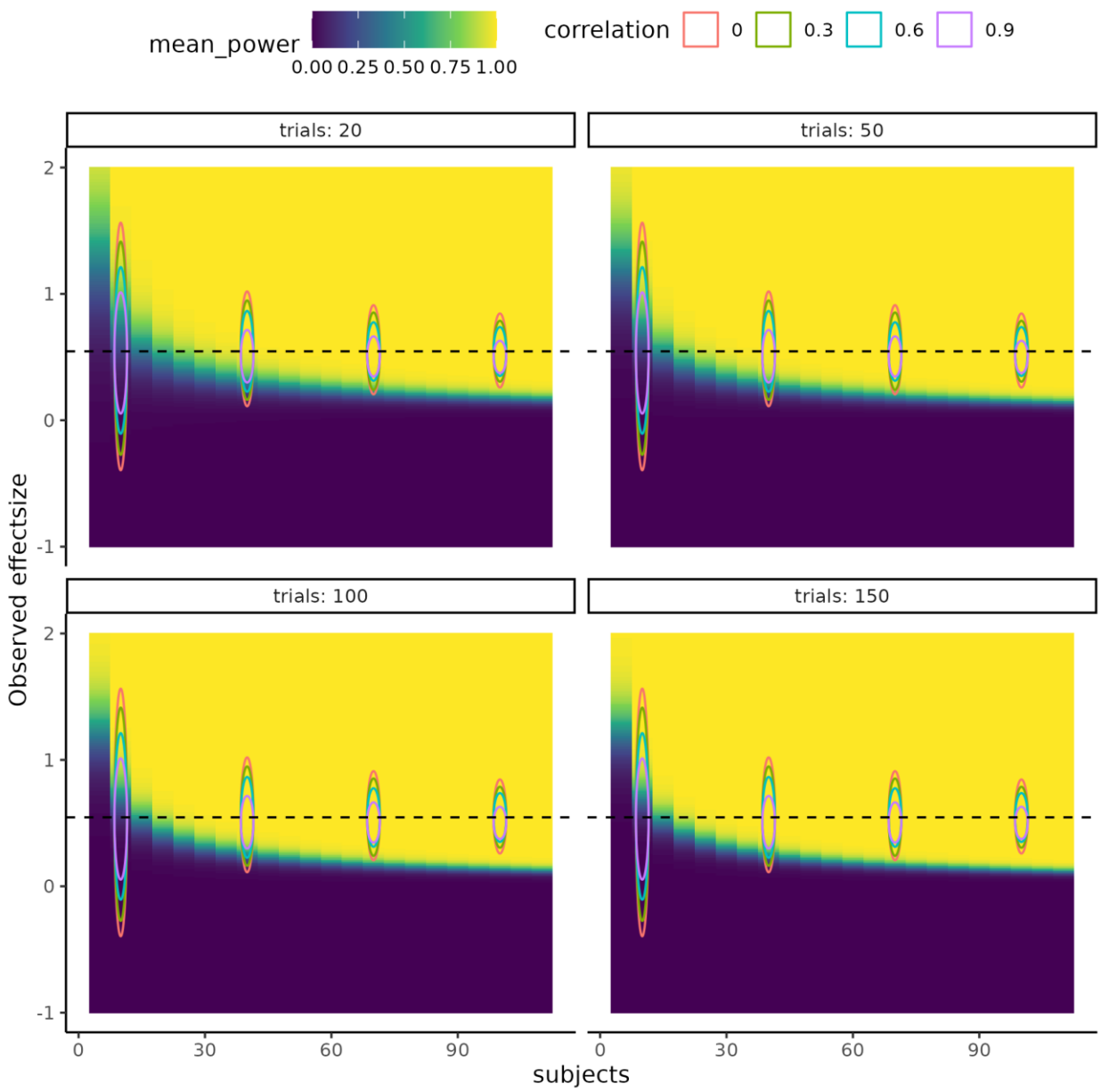


Figure 19. Visualization of how the power of a particular study, where the mean difference between threshold is assumed to be 5 and the variance of the difference being 12. The power of the experiment (background), is informed by the session-by-session correlation of the parameters, displayed as the color of the sampling distribution of the observed effect sizes (ellipses). Furthermore, the Figure shows how the number of trials and subjects include the power of the study. Subjects increase the power of the study while also decreasing the sampling variability (width of ellipses), whereas trials only increase the power of the study. Trials effect on power can be seen by investigating the upper left corner of the plots in the four facets, here one can observe that the background color turns more yellow as trials increase.

Practical implementation of the power analysis

Now, considering a practical example of a researcher wanting to conduct a power analysis, utilizing the simulation and modeling from above. Two assumptions have been made: either a mean effect size or a mean difference of the intervention is assumed, and the variance introduced by the intervention. Below, I investigate a mean difference of the intervention of 4 in threshold, and that the intervention does not increase variability, i.e., the variance in both groups is assumed equal. To fully appreciate the power of this approach, one could even imagine sampling these values as random variables, and not as point estimates (Not done here). Using the effect size equations above, it is possible to derive the mean difference and therefore simulate observed effect sizes which are then put into $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$ and the probability of rejecting that draw is calculated. This process is repeated over the 4000 draws of the posterior distribution of the parameters of $\Psi(d_{\text{obs}}, \alpha, \beta, t, s)$. Lastly, calculating the ratio of rejected to failed rejected null hypotheses gives the estimate of statistical power. In the case of not including the sampling varability (prior probability) of the observed effect size, the effect size estimate is just repeatedly entered as 0.5 (calculated from the assumptions of a mean difference of 4 and no varability change in the intervention). This idea of not including the sampling varability, is equivalent to investigating a minimum effect size of 0.5 being deeemed the minimum effect size of interest.

Figure 20 depicts a grid of subjects X trial that spans the space of power to reject the null hypothesis, here the observed effect size has been "integrated" out. This integration was done with either a constant of 0.5 (left column) or with a normal distribution with a mean of 0.5 and a variance of $\frac{1}{n} + \frac{0.5^2}{2 \cdot n}$, with n being the number of subjects. As a reference frame the red dashed line in figure 20, at subjects = 25, depicts the results from plugging the same assumptions, here $\mu_1 = -8$ $\sigma_1 = 8$ $\mu_2 = -4$ and $\sigma_2 = 8$ and $\rho = 0.54$, into the statistical software tool G*power (Faul et al., 2007).

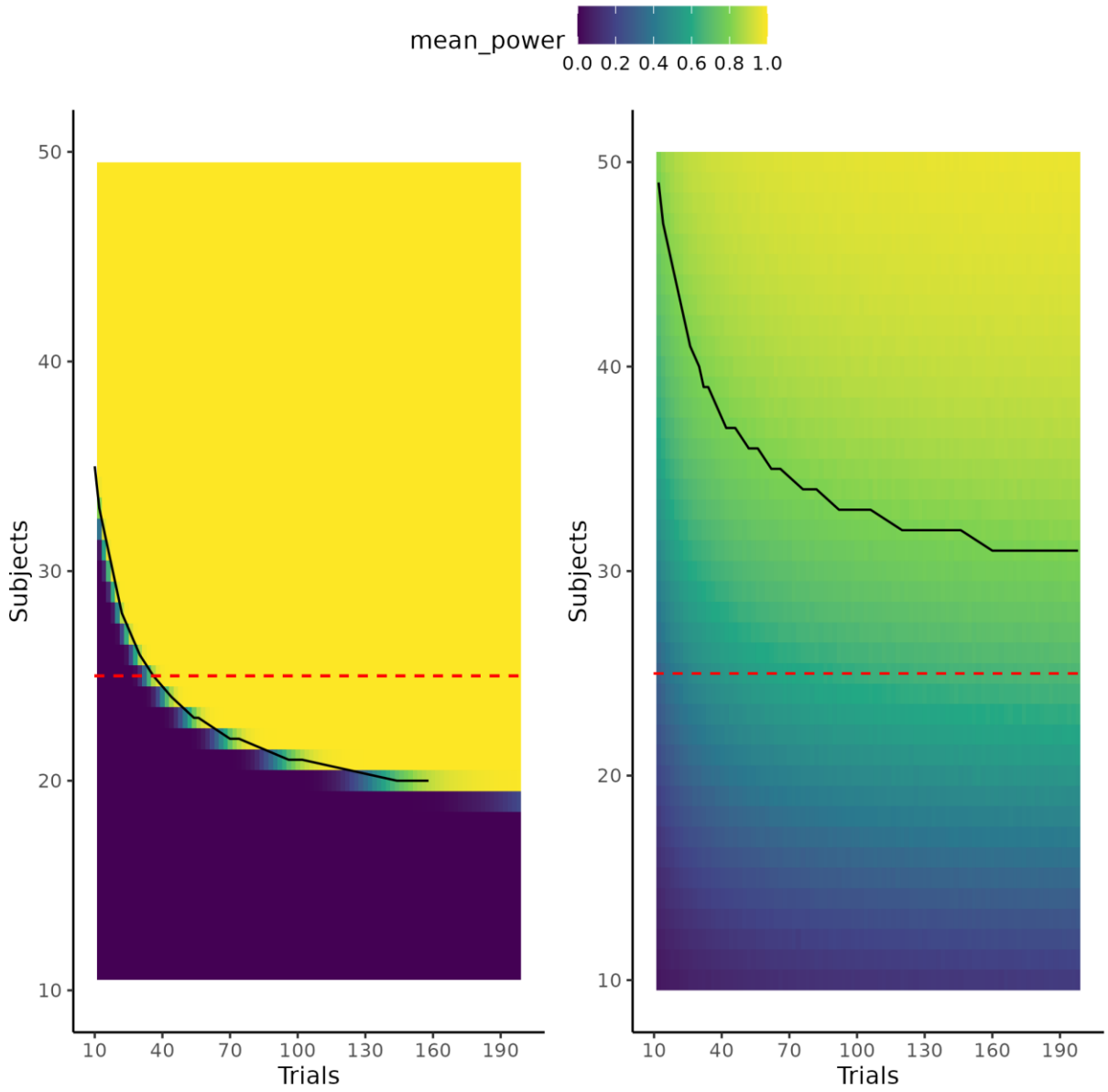


Figure 20. Displays the grid of subjects X trials to obtain a particular level of statistical power (background) given a group level difference of 4 and equal variance. Left column displays an analysis omitting sampling variability (a minimum effect interesting effect size calculation), with the right column including sampling variability.

Discussion

The thesis has investigated improvements in uncertainty handling in the field of cognitive science, particularly in the developing field of cognitive modeling. This was done with the use of simulations, highlighting that a deep mathematical understanding is not necessary to understand and/or do calculations with uncertainties. The thesis outlined three types of uncertainty; measurement uncertainty being the lowest level, often overlooked or disregarded in the field, despite of the unpredictable influence it can have on the resulting statistical metrics. Researchers should firstly be aware of measurement uncertainty and examine the extent to which it can be safely ignored in their statistical models. Even in measures like reaction times, commonly used in Cognitive Science (MacLeod & Dunbar, 1988; Pirolli & Anderson, 1985; Sternberg, 1969), measurement uncertainties are present, and depend on the soft and hard-ware the experiment (Crocetta & Andrade, 2015; Holden et al., 2019; Ohyanagi & Sengoku, 2010).

Estimation uncertainty, introduced as the uncertainty associated with doing computations, is often displayed as the standard error of statistical metrics. The main focus of the thesis was to investigate this type of uncertainty in the field of cognitive modeling and revise some of the statistical metrics used to validate a particular model. This was demonstrated using a psychometric function, that maps stimulus values to probabilities by three parameters the threshold (α), slope (β) and (λ). It was argued that the statistical metric commonly used, the correlation coefficient, between simulated and recovered parameters values was not a sensible metric to determine the extent of internal model validity (Schurr et al., 2024). Two important aspects of the correlation coefficient made it insensible for internal model validity. Firstly, the decision of choosing what size of correlation coefficient should be deemed enough, is not straightforward, because the interpretation of the correlation coefficient itself in the regard of model validation is not straightforward. This is particularly true when highlighting that the correlation coefficient is invariant of a linear transformation. Secondly, it was shown that in instances where the simulated and recovered parameter values did show good dependency, the correlation coefficient rapidly approached an asymptote at 1. This occurred when more information could be gained by increasing the number of trials, demonstrating its limited inclusion of the estimation uncertainty. The thesis therefore suggested using a variant of the intra class correlation coefficient (ICC) as the statistical metric for examining internal model validity, as recently proposed in the literature (Schurr et al., 2024).

It was shown that the ICC metric was more sensitive to estimation uncertainty in the parameters, with a sensible interpretation of the ratio between desirable and undesirable uncertainty. With this new metric, the thesis explored ways to decrease the undesirable uncertainty and thereby increase the ICC metric. Two ways were investigated, revolving around either incorporating smart experimental designs that are optimized for each individual or incorporating reaction times into the cognitive model. These methods are neither mutually exclusive or incompatible and could be implemented in experiments to decrease estimation uncertainty, in the parameters of the psychometric function.

The second approach of jointly modeling several dependent variables and their interactions, has been incorporated in the cognitive modeling literature for quite a while, however is now slowly re-gaining traction (Hess et al., 2024; Pedersen et al., 2017; Stone, 2014). What these two methods have in common is that they do not increase the efficiency of the study by increasing trials, which is usually is the default for decreasing estimation uncertainty of subject level parameters.

The obvious problem with increasing the number of trials is resource costs. This is both in terms of money, but also in the time spent for the participant and the experimenter. From an ethical perspective, this is especially true of the time investment from the participants' side, and particularly when patient populations are investigated. However, the most problematic aspect of mindlessly increasing the number of trials, becomes more obvious when we carefully consider what we are studying. In Cognitive Science, we are studying a complex system that has its own goals, desires, and motivations, and it is not trivial to know how this participant will behave if the task is twice as long. Will the participant employ a different strategy, knowing that the experiment is going to take longer, or will they halfway through the experiment employ a different strategy, due to boredom. Even if participants keep the same underlying cognitive strategy, that we are trying to model, then one would still expect that attentional lapses increase and overall engagement in the task to decrease. This would entail that each additional trial, perhaps after a certain point, would be less informative. The thesis went on to investigate the last type of uncertainty, test-retest uncertainty, by re-examining a data-set from a test-retest reliability study. Here it was shown that a re-analysis of the original data could achieve better test-retest reliability. This was done by incorporating knowledge about the structure of how the data was gathered, together with incorporating information already represented in the data i.e. reaction times.

The re-analysis was then used to exemplify of how a power analyses of a cognitive models can be conducted. This was achieved by simulating and then fitting the cognitive model to different observed effect sizes in combinations of different trials and subject. This approach allowed for modelling of the latent power curve, relating observed effect size, trials, and subjects to the probability of rejecting a null hypothesis, in an experiment. Using posterior predictive checks and leave one out cross validation, a particular power law related the parameters of the power curve, to subjects and trials with good predictive abilities. With this analysis it was shown that the number of trials in an experiment, can be added to a power analysis, which is not standard practice in widely used statistical software tools like G*power (AARTS et al., 2015; Faul et al., 2007; Ioannidis, 2005). This power analysis showed that G*power's estimation of sample size for the purposed test, was more liberal requiring 25 subjects, whereas the full uncertainty propagated power analysis, based on simulations from the fitted model, suggested approximately 30 subjects. Crucially, this was only the case if the number of trials were larger than at least 150.

The rest of the discussion of the current thesis will revolve around the implications of improperly accounting for uncertainties in science and how this might be a contributing factor to the replication crisis.

Power analyses, certainty, and replication crisis.

In recent years, some scientific fields especially psychology, social science and medicine, have been under scrutiny due to a lack of and failure of replication of previous studies (Forbes et al., 2023; Wiggins & Christopherson, 2019). Many contributing factors have been identified, such as publication bias and questionable research practices (QRP). These QRP, involves p-hacking (conducting statistical analyses until significant) or HARKing (hypothesizing after the results are known) (Head et al., 2015; Kerr, 1998).

A quite paradoxical aspect of the replication crisis is the use of power analyses, which are advised as a means to increase replicability. The argumentation of conducting power analyses before data collection, is that many studies in social science generally have low to very low power, to detect a small to medium effect size (Felix Singleton & Fidler, 2023). Power analyses are therefore promoted, to ensure sufficient statistical power to detect the size of the effect of interest. The argument is sound if the analysis of statistical power is accurate or accurate enough. What this thesis has highlighted is that the use of very popular tools like G*power, for conducting these types of power analyses, underestimate the number of subjects needed, by not including the effect of

trials in the estimation process. Therefore, the assumption from above might be misleading and problems might arise where researchers have too much confidence in their experiment, due to having conducted a power analysis, then is actually justified, hence the paradoxical aspect. This mimics the false sense of certainty on measurements, that are assumed by this popular software or measurements in Cognitive Science in general. Therefore, instead of increasing replicability and certainty in the effects observed, utilizing these tools might paradoxically decrease them, as researchers might be tricked into conducting less powered studies, due to the recommendations of the software.

Ways of combating the replication crisis.

A significant number of scientists have suggested to move the arbitrary statistical significance threshold from 0.05 to 0.005, to combat the replication crisis (Benjamin et al., 2018). Interestingly, lowering of the statistical threshold for significance would, in principle, lead to the conclusions drawn from this thesis of including and propagating uncertainty. The comparability of these two approaches depends on the structure and uncertainty of the data. However, in most cases including and propagating uncertainties, would have the effect of lowering the resulting statistic and therefore increasing the resulting p-value. These two approaches, i.e. increase the statistical significance threshold or properly propagating uncertainty, have very different reasons, even though they share the same goal. Lowering of the significance threshold would be a means to an end, instead of addressing the underlying problems, which the authors also do acknowledge (Benjamin et al., 2018).

Another interesting idea that coincides with the general theme of the thesis, and to combat the replication crisis, is that of preregistration, registered reports, and blind analyses (Chambers & Tzavella, 2022; Evans et al., 2023; Klein & Roodman, 2005; MacCoun & Perlmutter, 2015). The common theme of these interventions is that they acknowledge the subjectivity not only the data collection, but also in the data analysis pipeline. This subjectivity is both what introduces biases, but also what drives novel ideas, meaning that it becomes a trade-off between exploration and exploitation. This trade-off needs to be addressed, to partly guard against unwanted subjectivity. The interventions guard against this unwanted subjectivity by having the analysis pipeline either fixed before data collection or scrambling the data such that the results of the analyses cannot be known when producing the analysis pipeline. The rigorous checking, testing, and validating of

cognitive models shown here is not at stake with these interventions, but instead facilitates them. This is because most of the checking, testing and validation should be done on simulations.

However, there are still considerations when analyzing experimental data, especially on the model convergence side, where in or excluding covariates or reparameterization of the models might be necessary. In this regard, the blind analysis intervention might be a valuable insight from physics. Here the experimental data is scrambled in various ways, such that models and analysis pipelines can be done on data that resembles the collected data, but without being able to know the results before the data is un-blinded (Klein & Roodman, 2005; MacCoun & Perlmutter, 2015). Decisions are therefore made on scientific justifications, instead of on completely subjective criteria that could make the experimental results fit a research paradigm or perhaps even worse, produce significant results, where none are present. While the distinction between decisions based on scientific justification and subjective nonsense may be fuzzy and narrow, interventions, like those described, can help mitigate unwanted incentives such as publishing pressure and the temptation to fit results to a particular research paradigm or hypothesis (Quaia & Vernuccio, 2022). This approach could give rise to more rigorous methods and analysis pipelines, as it hinders arbitrarily stopping the development of the pipelines, when the results fit the preconceived notions of a scientific paradigm. Instead, it forces researchers to stop only when they are satisfied with the assumptions and implementations made. This process might also help researchers understand the uncertainty that is associated with many of the methods or practices commonly used in the literature.

Why and how computational tools are becoming vital in science.

Cognitive or even computational modeling could serve as a fresh start needed in the sciences that have been troubled by the replication crisis. The more sophisticated models embedded in these frameworks might be the steppingstone to engage in more theoretically driven analyses, hopefully reducing the number of non-reproducible studies. However, for this movement to succeed, it is essential that rigorous metrics are enforced to assess the model's internal validity. Models lacking any type of internal validity or identify-ability can therefore be discarded from the beginning. There may be instances where mathematical formulation of theories are developed, but that in practice the formulation is computationally intractable. It would be a shame to spend years investigating these kinds of models, and their implications in a field of research, only to discover that the model is intractable in practice (Ho & Griffiths, 2021; McClelland, 2009; Zuidema et al.,

2020). One might think that a deeper mathematical understanding is a necessity for understanding and building these more complicated models. However, what this thesis has argued is that this is not necessarily the case, simulations allow researchers to observe the implications of their assumptions. Another argument for why more sophisticated models are not necessarily off the shelf due to high level mathematically understanding is the increase in adaption of sophisticated hierarchical models which are mathematically much more complex than single level models, yet they have been widely adopted in the literature (Dedrick et al., 2009).

This is not to say that a better understanding of the machinery and mathematics itself would not be beneficial, but perhaps no longer a necessity. This would therefore also imply that the way that statistical methods and tools are taught might need to change. In fields where mathematical methods are not commonplace, students and researchers could be taught statistical methods with the use of coding and simulation examples, instead of flowcharts for which statistical analysis to conduct when. This would involve providing individuals with the tools for understanding and reflecting on these statistical models and their assumptions. This is like how a good scientific program does not merely teach students the right theories or hypotheses, but rather teaches them to think in a scientific way, such that the individual can decide and test these themselves. In this context the tools for understanding, reflecting, and experimenting with statistical models and concepts, would be programming experience in statistics. This would allow the researcher to more concretely grasp the assumptions that are being made but would also provide the tools for examining what happens when they are broken. Moreover, this framework would also necessitates a more generative approach to modeling, making the researcher more closely engaged in the statistical process of analyzing the data, instead of just picking an off the shelf model from a flowchart (Velarde Camaqui & Díaz Méndez, 2023).

Standing on the shoulders of giants

All of the models used in the current paper were fitted using Stan with the cmdstanr interface, which uses full Bayesian statistical inference with Markov chain monte Carlo sampling (Gabry et al., 2024). As described in the section about modeling definitions, fitting, and building models in this framework is extremely flexible. An additional benefit of this framework is that the code for simulating the generative process is close to identical in nature, to the code that specifies the model. This similarity makes it easy for users with a generative framework to code up these types of models. The additional benefits to using Stan and its Hamiltonian Monte Carlo (HMC) algorithm

is that when issues arise the algorithm will complain. This helps reduce the risk for erroneous inference, due to the sampling algorithm or typos in the code (Vehtari et al., 2021).

The thesis used Bayesian inference and Stan, due to its flexibility in model formulation, rather than the inherent differences between Bayesian and frequentist statistics. However, Bayesian inference does allow for a more optimistic way to interpret the replication crisis discussion above. Instead of starting each experimental analysis from the perspective that nothing or very little is known about the parameters of interest, perhaps incorporating information from previous studies would be beneficial. This is what the priors in the Bayesian inference scheme identifies. This is in essences what science is about, a hierarchical organization of knowledge, where each step rests on the step below, i.e. on auxiliary assumptions as put by the Duhem–Quine thesis (Ariew, 1984). Here priors can be thought of as in the top level of the hierarchy, that then informs the lower-level implications, but that the strength and location of these priors are informed by lower levels, i.e., empirical evidence. This view on science also matches that of uncertainties, as these are also hierarchically organized. So, in the same way that the result of a scientific theory is only as strong as its auxiliary assumptions; the strength of an analysis, that builds on a theory, is also only as strong as the (un)certainty of the data.

What the Bayesian inference allows is that prior information from similar studies can be used in modelling, allowing researchers to not start their scientific studies from scratch, but pick up where others left off. This would essentially entail, instead of collecting a larger number of subjects to achieve the desired statistical power, this could be done by two independent laboratories. Here we can think that the second laboratory uses the information provided by the first in their priors. This essentially is already what is being done when conducting meta-analyses. This approach incentives publications of all types, as the findings of one researcher would serve as a stepping stones for the next, making the problem of publication bias, where null findings are unpublished, less incentivized (Laitin et al., 2021).

Limitations

One of the main focal points of the thesis was investigating the correlation coefficient as a statistical metric for internal model validation of cognitive models. It was shown that a modified version of the intra class correlation was more sensible as a statistical metric for model validation. However, due to limitations on time and computational resources, no analysis was conducted, comparing these statistical metrics to the power analysis displayed in Figure 20. Future studies

should investigate the link between how these metrics behave and compare them to the power analysis conducted. A clear link between these quantities would make the need for conducting a power analysis superfluous, as one could imagine that the information for a power analysis could be contained in the validation analysis. A thorough investigation of this link would mean that the somewhat arbitrary choice of trials, when designing an experiment, would no longer be arbitrary. It would instead be informed by how estimation uncertainty in the parameters of interest, change based on the number of trials (Miller, 2024).

Another limitation of the current study is the limited power analysis conducted. Ideally, the thesis would have investigated other parameters of the psychometric function. A particular interest would be on the slope of the psychometric function, as it was shown that changes to this parameter changes the estimation uncertainty of all the other parameters. One might suspect that an intervention that increases the steepness of the slope, would also make it easier to detect a change in the threshold. As both the correlation coefficient and ICC metric showed that with increased steepness of the function, less estimation uncertainty was present in the threshold. This highlights how the parameters of the model interact, which can be accounted for by performing these simulations. Therefore, future investigations should expand upon this power analysis to include other parameters, especially the slope of the psychometric function.

Future studies should also investigate how incorporating the reaction times into the power analysis would change the statistical power function. This research would not only help elucidate the question posed above, about the relationship between the internal model validation metric, trials and power, but could also give an estimate of the increased efficiency pf incorporating information already present in most experiments. The reasoning for only conducting the single power analysis on the threshold in the current thesis highlights one of the main hurdles of the framework purposed: computational resources. Fitting models using HMC and Bayesian inference is both time and computational resource intensive, compared to frequentist inference in packages such as lme4, lmertest or GAMLSS implemented in R (Bates et al., 2015; Kuznetsova et al., 2017; Stasinopoulos & Rigby, 2008). This additional invested time for doing computation can partly be negated with an access to bigger machines. Here parallelization of the computational burden, especially when several chains are needed to ensure convergence, is essential. Fortunately, the access to bigger machines, both privately but also on an institutional level, is something that is growing in accessibility and already available to many universities or research centers. This

increase in computational availability has also been correlated with research competitiveness (Apon et al., 2010).

References

AARTS, A. A., al, et, & LIN, S. C. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–950. https://doi.org/10.1126/science.aac4716

Apon, A., Ahalt, S., Dantuluri, V., Gurdgiev, C., Limayem, M., Ngo, L., & Stealey, M. (2010). High Performance Computing Instrumentation and Research Productivity in US Universities. *JITI Journal of Information Technology Impact*, *10*, 87–98.

Ariew, R. (1984). The Duhem Thesis. *The British Journal for the Philosophy of Science*, *35*(4), 313–325. https://www.jstor.org/stable/687336

Bahadori, M., Soltani, M., Soleimani, M., & Bahadori, M. (2023). *Statistical Modeling in Healthcare: Shaping the Future of Medical Research and Healthcare Delivery* (pp. 431–446). https://doi.org/10.4018/979-8-3693-0876-9.ch025

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1350–1365. https://doi.org/10.1098/rstb.2011.0420

Baldi Antognini, A., Frieri, R., & Zagoraiou, M. (2023). New insights into adaptive enrichment designs. *Statistical Papers*, *64*(4), 1305–1328. https://doi.org/10.1007/s00362-023-01433-0

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berker, A. O. de, Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*(1), 10996. https://doi.org/10.1038/ncomms10996

Boekel, M. A. J. S. van. (2021). To pool or not to pool: That is the question in microbial kinetics. *International Journal of Food Microbiology*, *354*, 109283. https://doi.org/10.1016/j.ijfoodmicro.2021.109283

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022). *Shiny: Web application framework for r*. https://shiny.rstudio.com/

Chén, O. Y., Bodelet, J. S., Saraiva, R. G., Phan, H., Di, J., Nagels, G., Schwantje, T., Cao, H., Gou, J., Reinen, J. M., Xiong, B., Zhi, B., Wang, X., & Vos, M. de. (2023). The roles, challenges, and merits of the p value. *Patterns*, *4*(12), 100878. https://doi.org/10.1016/j.patter.2023.100878

Coates, D. R., & Chung, S. T. L. (2014). Changes across the psychometric function following perceptual learning of an RSVP reading task. *Frontiers in Psychology*, *5*, 1434. https://doi.org/10.3389/fpsyg.2014.01434

Courtin, A. S., Delvaux, A., Dufour, A., & Mouraux, A. (2023). Spatial summation of cold and warm detection: Evidence for increased precision when brisk stimuli are delivered over larger area. *Neuroscience Letters*, *797*, 137050. https://doi.org/10.1016/j.neulet.2023.137050

Crocetta, T., & Andrade, A. (2015). THE PROBLEM OF MEASURING REACTION TIME USING SOFTWARE AND HARDWARE: A SYSTEMATIC REVIEW. *Revista de Psicologia Del Deporte*, *24*, 341–349.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and

Applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, *4*(2), 160254. https://doi.org/10.1098/rsos.160254

Durstewitz, D., Koppe, G., & Toutounji, H. (2016). Computational models as statistical tools. *Current Opinion in Behavioral Sciences*, *11*, 93–99. https://doi.org/10.1016/j.cobeha.2016.07.004

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593

Evans, T. R., Branney, P., Clements, A., & Hatton, E. (2023). Improving evidence-based practice through preregistration of applied research: Barriers and recommendations. *Accountability in Research*, *30*(2), 88–108. https://doi.org/10.1080/08989621.2021.1969233

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Felix Singleton, T., & Fidler, F. (2023). *The statistical power of psychology research: A systematic review and meta-analysis*. https://osf.io/yr8st/download

Forbes, H. J., Travers, J. C., & Johnson, J. V. (2023). Chapter 11 - Supporting the replication of your research. In D. J. Cox, N. Y. Syed, M. T. Brodhead, & S. P. Quigley (Eds.), *Research Ethics in Behavior Analysis* (pp. 237–262). Academic Press. https://doi.org/10.1016/B978-0-323-90969-3.00003-7

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *Cmdstanr: R interface to CmdStan*. https://mc-stan.org/cmdstanr/

Geissinger, E. A., Khoo, C. L. L., Richmond, I. C., Faulkner, S. J. M., & Schneider, D. C. (2022). A case for beta regression in the natural sciences. *Ecosphere*, *13*(2), e3940. https://doi.org/10.1002/ecs2.3940

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Gomes, D. G. E. (2022). Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, *10*, e12794. https://doi.org/10.7717/peerj.12794

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Harrison, O. K., Köchli, L., Marino, S., Luechinger, R., Hennel, F., Brand, K., Hess, A. J., Frässle, S., Iglesias, S., Vinckier, F., Petzschner, F. H., Harrison, S. J., & Stephan, K. E. (2021). Interoception of breathing and its relationship with anxiety. *Neuron*, *109*(24), 4080–4093.e8. https://doi.org/10.1016/j.neuron.2021.09.045

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., & Olkin, I. (2014). *Statistical Methods for Meta-Analysis*. Academic Press.

Hess, A. J., Iglesias, S., Köchli, L., Marino, S., Müller-Schrader, M., Rigoux, L., Mathys, C., Harrison, O. K., Heinzle, J., Frässle, S., & Stephan, K. E. (2024). *Bayesian Workflow for Generative Modeling in Computational Psychiatry*. bioRxiv. https://doi.org/10.1101/2024.02.19.581001

Ho, M. K., & Griffiths, T. L. (2021). *Cognitive science as a source of forward and inverse models of human decisions for robotics and control*. arXiv. https://doi.org/10.48550/arXiv.2109.00127

Holden, J., Francisco, E., Lensch, R., Tommerdahl, A., Kirsch, B., Zai, L., Dennis, R., & Tommerdahl, M. (2019). *Accuracy of different modalities of reaction time testing: Implications for online cognitive assessment tools*. bioRxiv. https://doi.org/10.1101/726364

Hübner, R., & Pelzer, T. (2020). Improving parameter recovery for conflict drift-diffusion models. *Behavior Research Methods*, *52*(5), 1848–1866. https://doi.org/10.3758/s13428-020-01366-8

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ivanova, A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., & Isik, L. (2022). *Beyond linear regression: Mapping models in cognitive neuroscience should align with research goals*. https://doi.org/10.48550/arXiv.2208.10668

Jain, A., Bansal, R., Kumar, A., & Singh, K. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research*, *5*(2), 124–127. https://doi.org/10.4103/2229-516X.157168

Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. *Journal of Development Economics*, *161*, 102992. https://doi.org/10.1016/j.jdeveco.2022.102992

Johannes, W. J., & Smilde, P. L. (2009). Fundamentals of Gravity, Elements of Potential Theory. In W. J. Johannes & P. L. Smilde (Eds.), *Gravity Interpretation: Fundamentals and Application of Gravity Inversion and Geological Interpretation* (pp. 23–111). Springer. https://doi.org/10.1007/978-3-540-85329-9_2

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein, J. R., & Roodman, A. (2005). BLIND ANALYSIS IN NUCLEAR AND PARTICLE PHYSICS. *Annual Review of Nuclear and Particle Science*, *55*(1), 141–163. https://doi.org/10.1146/annurev.nucl.55.090704.151521

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. https://doi.org/10.1016/s0042-6989(98)00285-5

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Kubinec, R. (2023). Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds. *Political Analysis*, *31*(4), 519–536. https://doi.org/10.1017/pan.2022.20

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Kwon, M., Lee, S. H., & Ahn, W.-Y. (2023). Adaptive Design Optimization as a Promising Tool for Reliable and Efficient Computational Fingerprinting. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 798–804. https://doi.org/10.1016/j.bpsc.2022.12.003

Laitin, D. D., Miguel, E., Alrababa'h, A., Bogdanoski, A., Grant, S., Hoeberling, K., Hyunjung Mo, C., Moore, D. A., Vazire, S., Weinstein, J., & Williamson, S. (2021). Reporting all results efficiently: A RARE proposal to open up the file drawer. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(52), e2106178118. https://doi.org/10.1073/pnas.2106178118

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Legrand, N., Nikolova, N., Correa, C., Brændholt, M., Stuckert, A., Kildahl, N., Vejlø, M., Fardo, F., & Allen, M. (2022). The heart rate discrimination task: A psychophysical method to

estimate the accuracy and precision of interoceptive beliefs. *Biological Psychology*, *168*, 108239. https://doi.org/10.1016/j.biopsycho.2021.108239

Lithfous, S., Després, O., Pebayle, T., Casadio, C., & Dufour, A. (2020). Accurate Determination of the Cold Detection Threshold with High-Speed Cooling of the Skin. *Pain Medicine (Malden, Mass.)*, *21*(12), 3428–3436. https://doi.org/10.1093/pm/pnaa246

Luijcks, R., Vossen, C. J., Hermens, H. J., Os, J. van, & Lousberg, R. (2015). The Influence of Perceived Stress on Cortical Reactivity: A Proof-Of-Principle Study. *PLoS ONE*, *10*(6), e0129220. https://doi.org/10.1371/journal.pone.0129220

Ma, A. C., Cameron, A. D., & Wiener, M. (2024). Memorability shapes perceived time (and vice versa). *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-024-01863-2

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189. https://doi.org/10.1038/526187a

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126–135. https://doi.org/10.1037/0278-7393.14.1.126

Makowski, D., Wiernik, B. M., Patil, I., Lüdecke, D., & Ben-Shachar, M. S. (2022). *correlation: Methods for correlation analysis*. https://CRAN.R-project.org/package=correlation

Makowski, D., Wiernik, B. M., Patil, I., Lüdecke, D., & Ben-Shachar, M. S. (2023). *Correlation: Methods for correlation analysis*. https://easystats.github.io/correlation/

Maravelakis, P. (2019). The use of statistics in social sciences. *Journal of Humanities and Applied Social Sciences*, *1*(2), 87–97. https://doi.org/10.1108/JHASS-08-2019-0038

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*. https://doi.org/10.3389/fnhum.2014.00825

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00039

McClelland, J. L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

Meier, M., Martarelli, C. S., & Wolff, W. (2024). Is boredom a source of noise and/or a confound in behavioral science research? *Humanities and Social Sciences Communications*, *11*(1), 1–8. https://doi.org/10.1057/s41599-024-02851-7

Miller, J. (2024). How Many Participants? How Many Trials? Maximizing the Power of Reaction Time Studies. *Behavior Research Methods*, *56*(3), 2398–2421. https://doi.org/10.3758/s13428-023-02155-9

Ohyanagi, T., & Sengoku, Y. (2010). A solution for measuring accurate reaction time to visual stimuli realized with a programmable microcontroller. *Behavior Research Methods*, *42*(1), 242–253. https://doi.org/10.3758/BRM.42.1.242

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234–1251. https://doi.org/10.3758/s13423-016-1199-y

Pirolli, P. L., & Anderson, J. R. (1985). The role of practice in fact retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 136–153. https://doi.org/10.1037/0278-7393.11.1.136

Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13*(7), 3. https://doi.org/10.1167/13.7.3

Quaia, E., & Vernuccio, F. (2022). Finding a Good Balance between Pressure to Publish and Scientific Integrity and How to Overcome Temptation of Scientific Misconduct. *Tomography*, *8*(4), 1851–1853. https://doi.org/10.3390/tomography8040155

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ranger, J., Kuhn, J. T., & Ortner, T. M. (2020). Modeling Responses and Response Times in Tests With the Hierarchical Model and the Three-Parameter Lognormal Distribution. *Educational and Psychological Measurement*, *80*(6), 1059–1089. https://doi.org/10.1177/0013164420908916

Saccenti, E., Hendriks, M. H. W. B., & Smilde, A. K. (2020). Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Scientific Reports*, *10*(1), 438. https://doi.org/10.1038/s41598-019-57247-4

Schurr, R., Reznik, D., Hillman, H., Bhui, R., & Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-024-01814-x

Stasinopoulos, D. M., & Rigby, R. A. (2008). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, *23*, 1–46. https://doi.org/10.18637/jss.v023.i07

Sternberg, S. (1969). Memory-Scanning: Mental Processes Revealed by Reaction-Time Experiments. *American Scientist*, *57*(4), 421–457. https://www.jstor.org/stable/27828738

Stone, J. V. (2014). Using reaction times and binary responses to estimate psychophysical performance: An information theoretic analysis. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00035

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). *Pareto Smoothed Importance Sampling*. arXiv. https://doi.org/10.48550/arXiv.1507.02646

Velarde Camaqui, D., & Díaz Méndez, R. E. (2023). *Flowchart for choosing inferential statistical test*. https://doi.org/10.21125/iceri.2023.2393

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10. https://doi.org/10.1167/17.3.10

White, C. N., Servant, M., & Logan, G. D. (2018). Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychonomic Bulletin & Review*, *25*(1), 286–301. https://doi.org/10.3758/s13423-017-1271-2

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, *14*(4), 1261–1295. https://doi.org/10.1214/aos/1176350142

Wu, J., Tong, H., Liu, Z., Tao, J., Chen, L., Chan, C. C. H., & Lee, T. M. C. (2021). Neurobiological effects of perceived stress are different between adolescents and middle-aged adults. *Brain Imaging and Behavior*, *15*(2), 846–854. https://doi.org/10.1007/s11682-020-00294-7

Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2021). ADOpy: A python package for adaptive design optimization. *Behavior Research Methods*, *53*(2), 874–897. https://doi.org/10.3758/s13428-020-01386-4

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

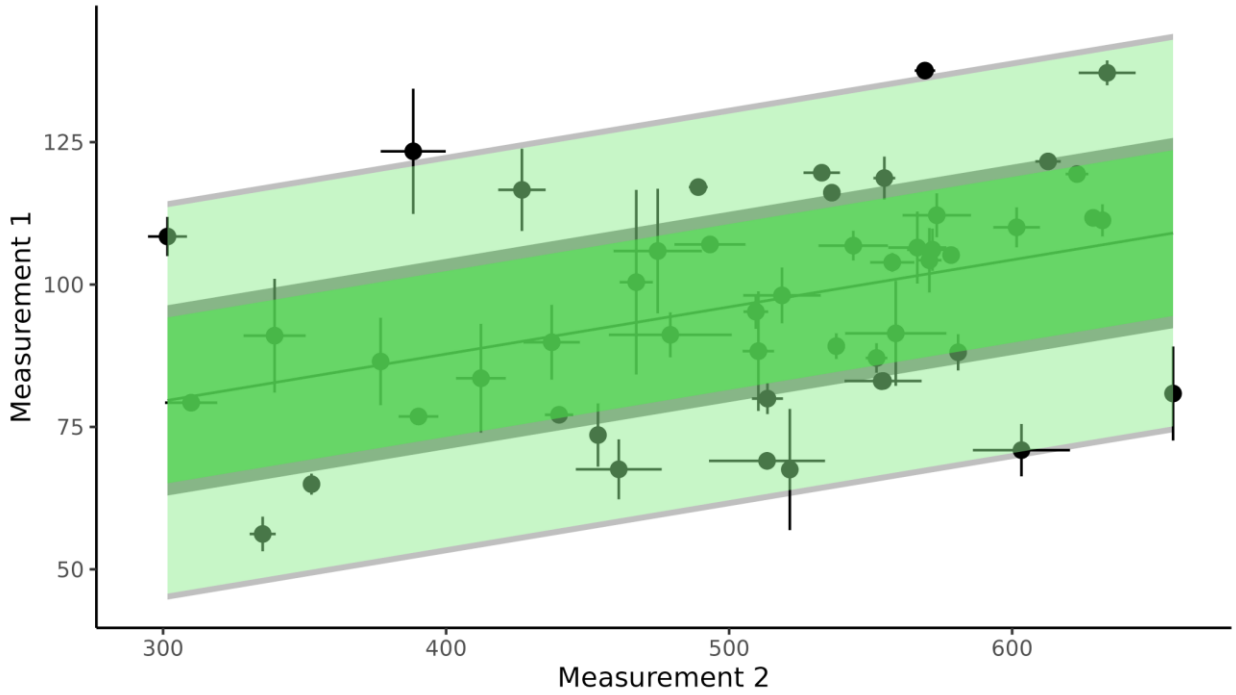
Zhang, L., Carpenter, B., Gelman, A., & Vehtari, A. (2022). *Pathfinder: Parallel quasi-Newton variational inference*. arXiv. http://arxiv.org/abs/2108.03782

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., & Gentner, T. Q. (2020). Five Ways in Which Computational Modeling Can Help Advance Cognitive Science: Lessons From Artificial Grammar Learning. *Topics in Cognitive Science*, *12*(3), 925–941. https://doi.org/10.1111/tops.12474

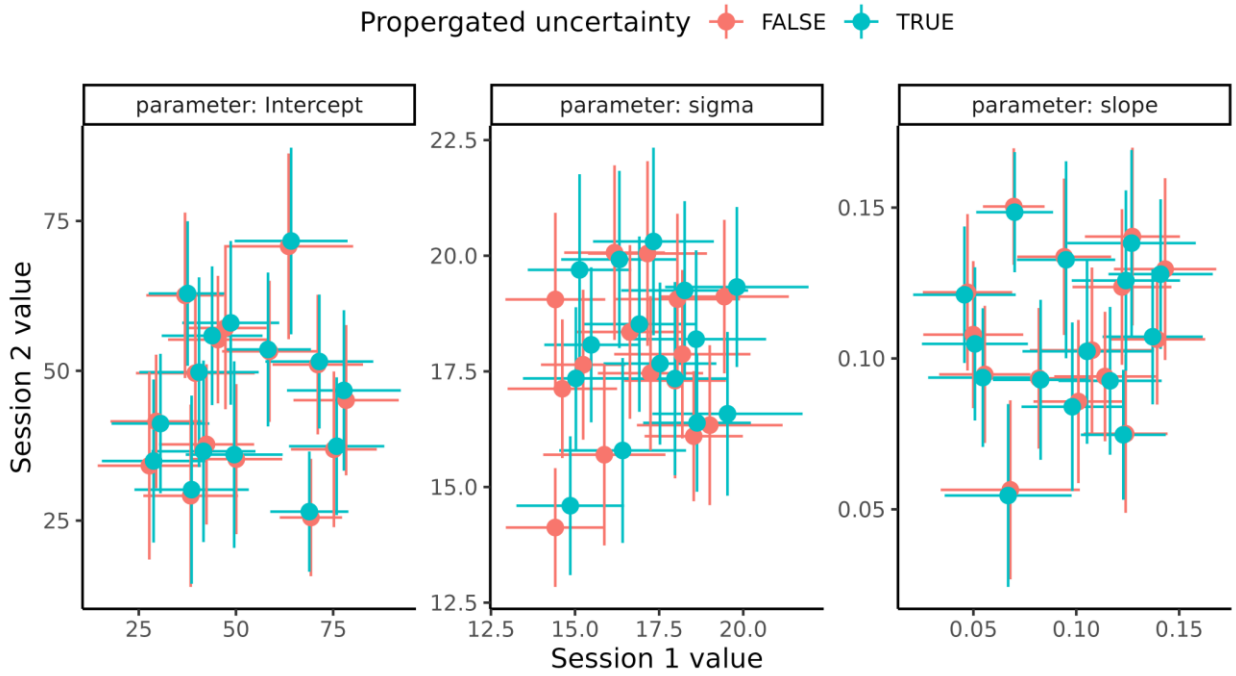
Supplementary material

Supplementary analysis 1

As mentioned in the main text, introducing measurement uncertainty or any type of uncertainty into statistical metrics and easily be done using bootstrapping / re-sampling. Here I will demonstrate an example from a simple linear regression analysis with 3 parameters an intercept, a slope and a residual variance. To do this we consider an idealized example of a fictional researcher wanting to understand the relationship between reaction times and stress while incorporating all types of uncertainty. To do this the researcher conducts an experiment where participants are measured several times under different conditions to introduce stress. In this example both of these measures have associated uncertainty, see the individual data points in the Supplementary analysis Figure 1. The relationship between reaction times and stress is determined by the slope of the regression line depicted in the Supplementary analysis Figure 1. Estimation uncertainty can thus be thought of as the uncertainty in the parameter estimates achieved by fitting a linear model to the data. See the linear model in Supplementary analysis Figure 1. Finally, test re-test uncertainty can be thought as when the researcher's study on reaction times and stress is tested twice on different days to understand how stable the relationship is over time. As the relationship is measured by the parameters of the model the stability of the relationship is measured by the stability of the parameters. One might imagine that the amount of sleep acquired before the experimental day could influence both measures of the task i.e. reaction time and susceptibility to stress and perhaps even their relationship. Supplementary analysis Figure 2 displays how the parameter estimates of the same model as presented in Supplementary analysis Figure 1 with and without accounting for uncertainty propagation change based on the propagation of uncertainty. As can be seen from Supplementary Analysis Figure 1 accounting for the measurement uncertainty does not change much the prediction made by the model, however when propagating these extra uncertainties into the next analysis of the parameters i.e. from session to session in Supplementary analysis Figure 2 the change in results become more pronounced. The main effect for the current linear model is that the residual variance (sigma) and the intercept is underestimated without error propagation and the slope parameter is overestimated.



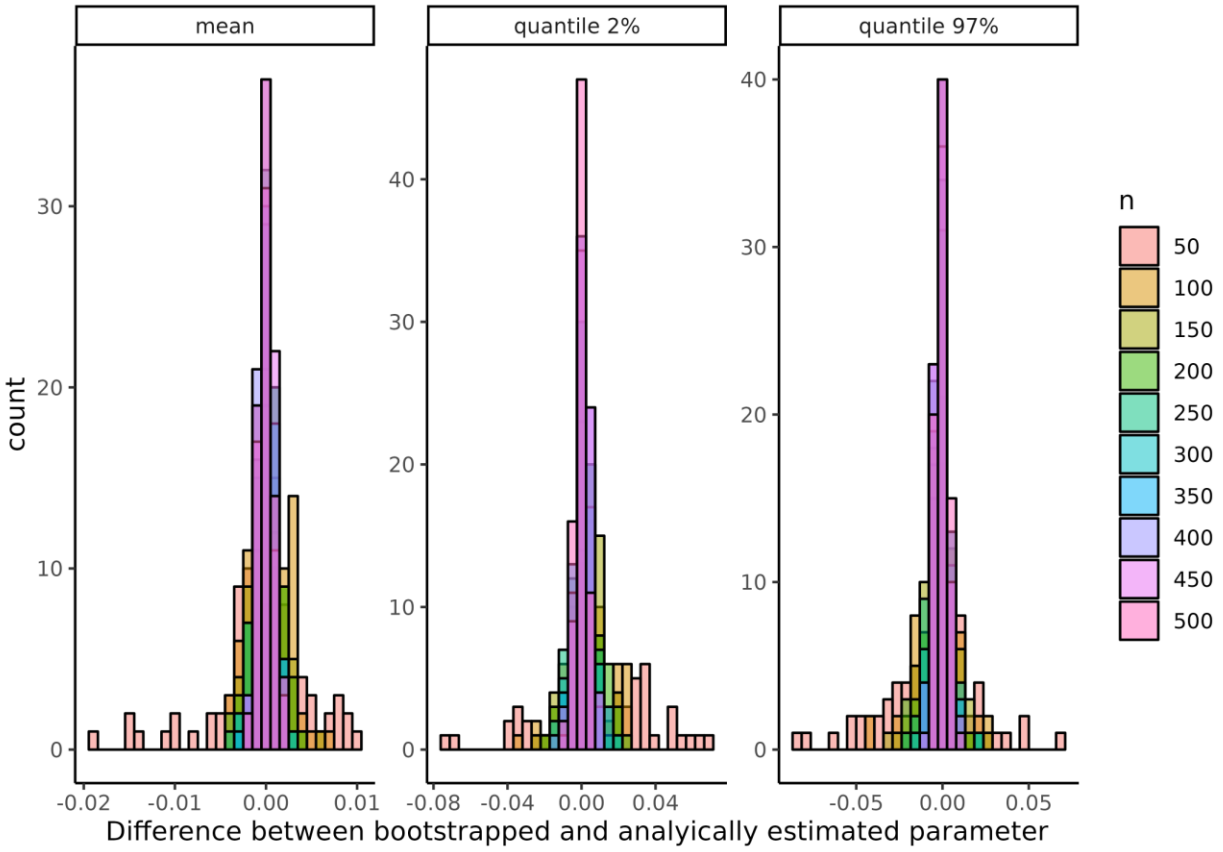
Supplementary analysis Figure 1: Measurement and Estimation uncertainty. The figure displays a linear regression between two measurements of for instance reaction time and stress with measurement uncertainty depicted as vertical and horizontal error bars on individual points. The mean of the regression line with and without propagated uncertainty is highlighted in grey and dark green respectively. Lastly a prediction interval is depicted as the shaded area around the mean of the regression line with and without propagated uncertainty again in grey and green respectively. The difference between the green and grey lines are therefore the difference between accounting for measurement uncertainty and not accounting for it.



Supplementary analysis Figure 2, Test-retest uncertainty. Displays the results of fitting the linear regression in Figure 1 twice, with and without accounting for measurement uncertainty. Each facet represents one of the three parameters of the linear model, the intercept the residual uncertainty (sigma) and the slope respectively from left to right. Colors represented weather the measurement uncertainty was propagated or not.

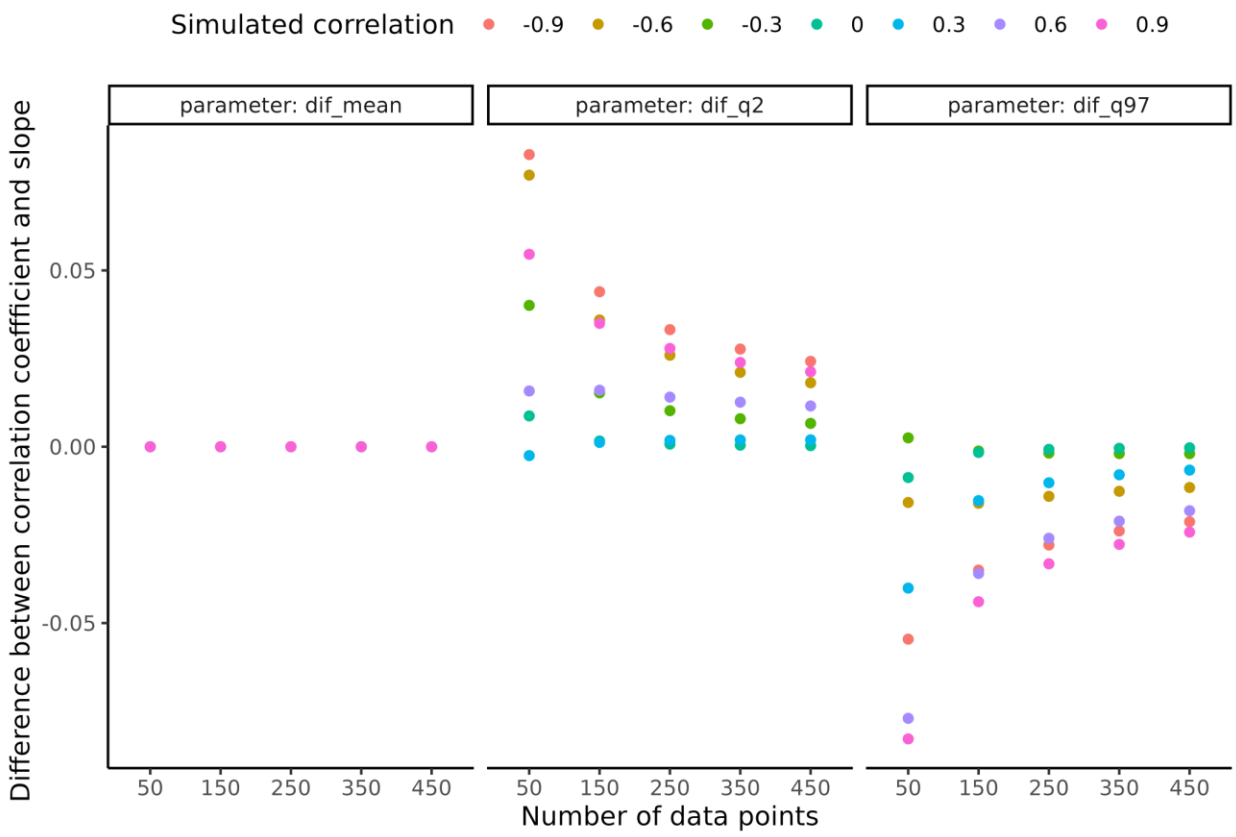
Supplementary Figures

Supplementary figure 1



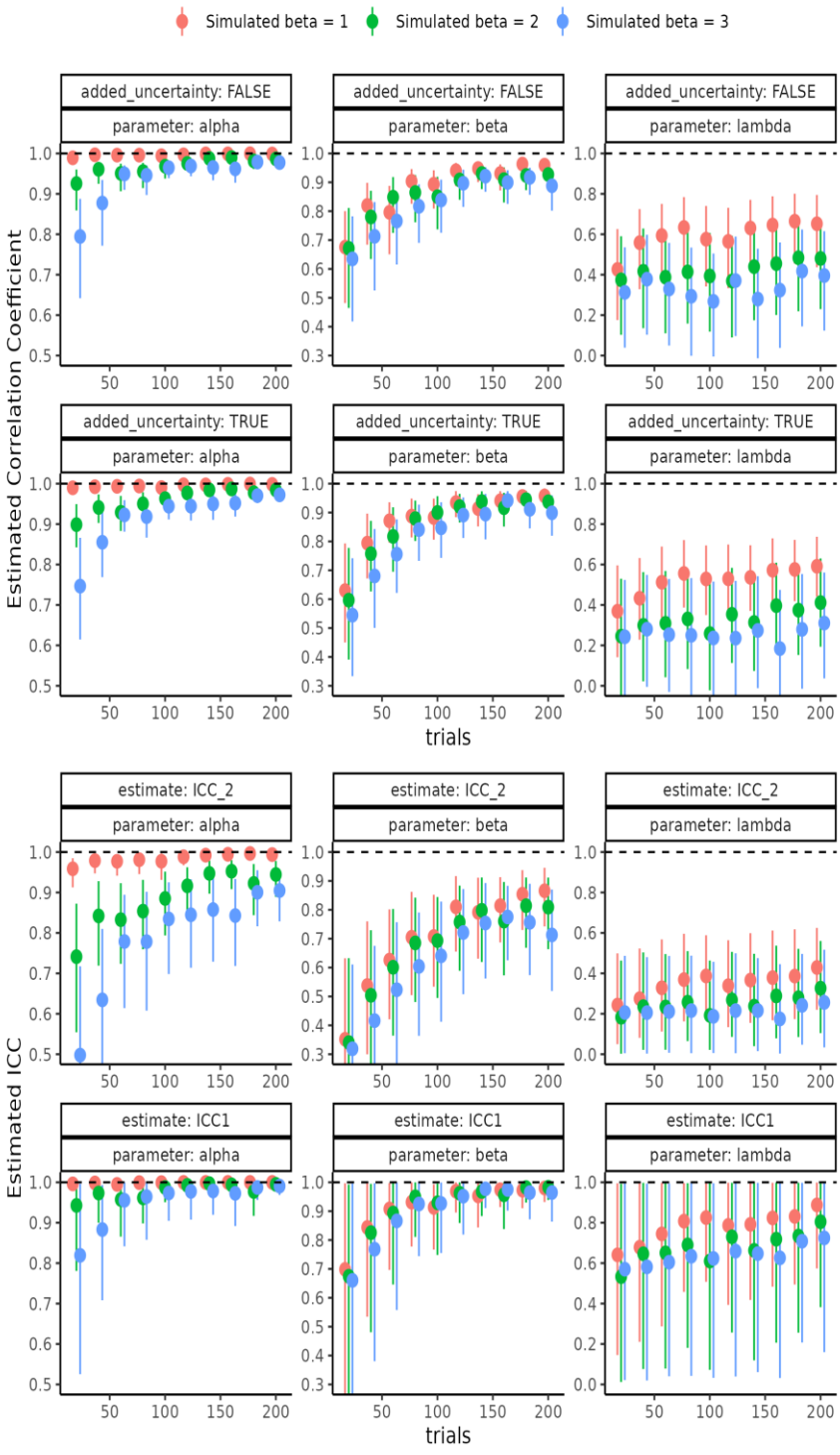
Supplementary figure 1 Comparison of analytical and bootstrapping correlation coefficient. Histograms of the difference between different metrics of the correlation coefficient when bootstrapping and analytically calculating the correlation coefficient. Facets show the different used metrics when evaluating the correlation coefficient i.e. the mean, the 2% quantile and the 97% quantile. colors represent the sample size, i.e. the number of datapoints the simulated correlation coefficient was based on.

Supplementary figure 2



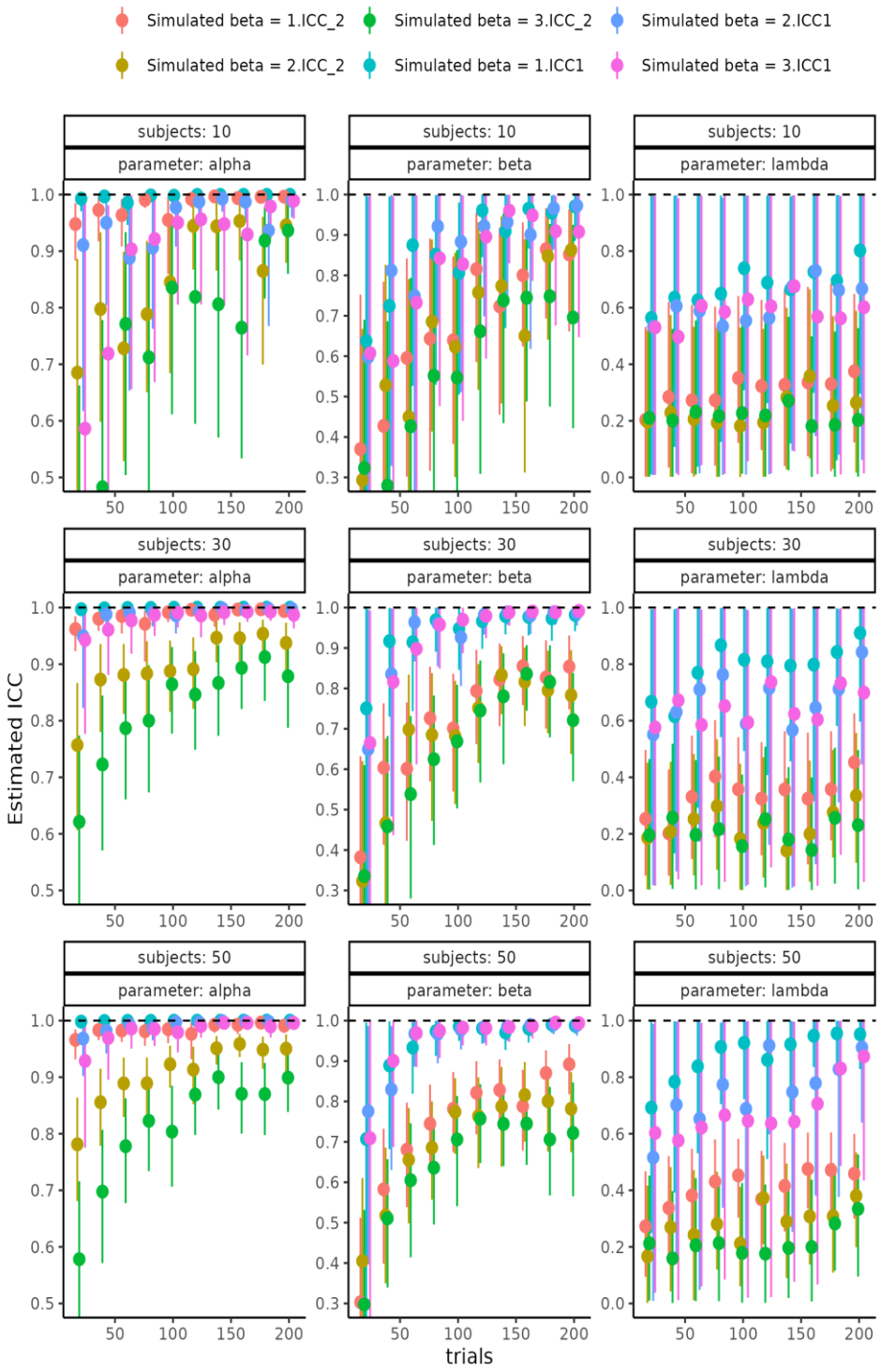
Supplementary figure 2. Comparison of analytical and bootstrapping correlation coefficient. Points depicting the differences in estimated correlation coefficient using bootstrapping and the analytical solution for thee measures of the correlation distribution i.e. mean (dif_mean), 2% quartile (dif_q) and the 97% quartile (dis_q97). X-axis represents the number of data points simulated with the colors depicting the simulated size of the correlation coefficient.

Supplementary figure 3



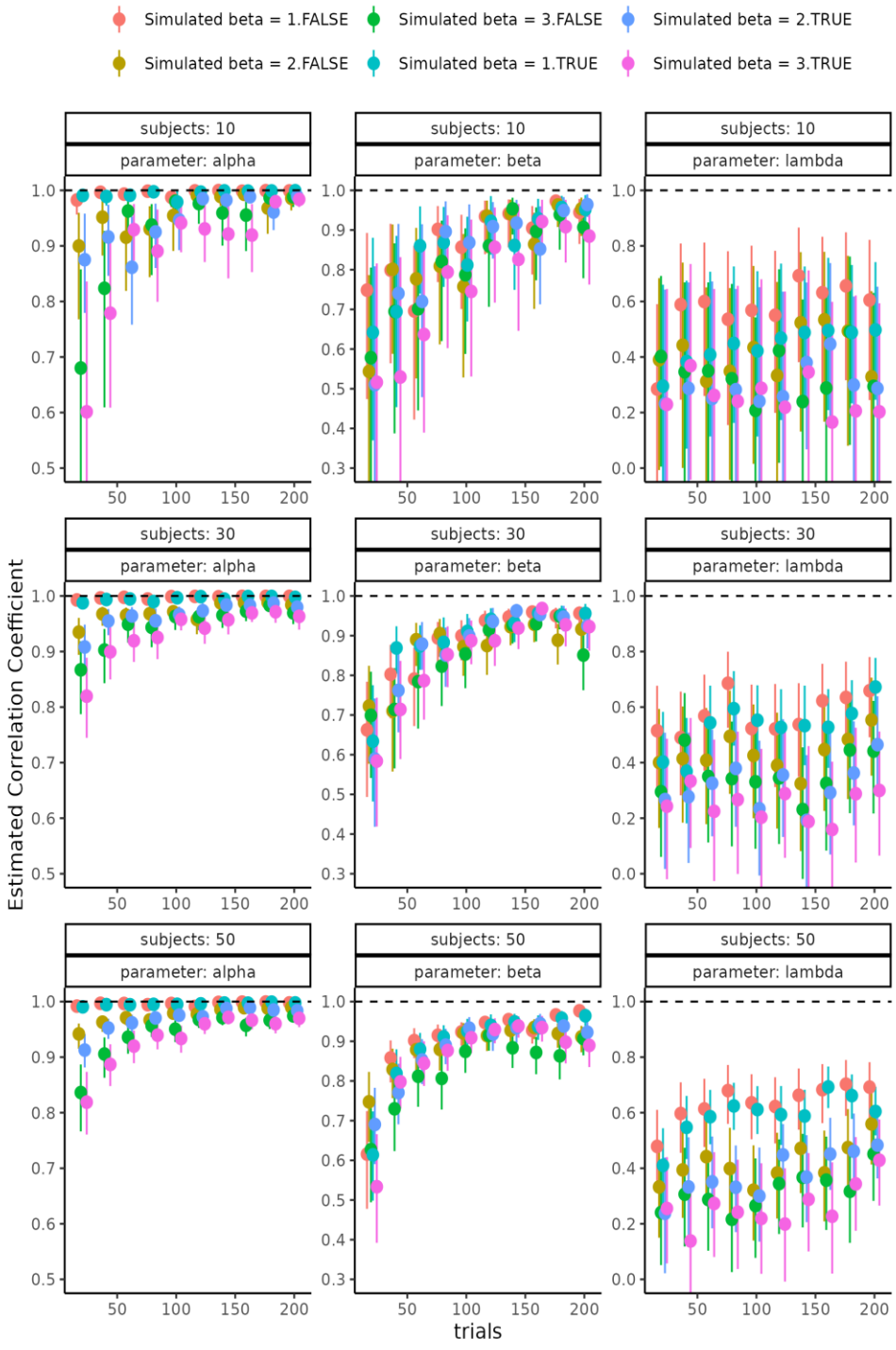
Supplementary figure 3 Parameter recovery of the psychometric function for all four parameter recovery metrics investigated. That is the correlation coefficient with and without uncertainty propergation and ICC1, ICC2.

Supplementary figure 4



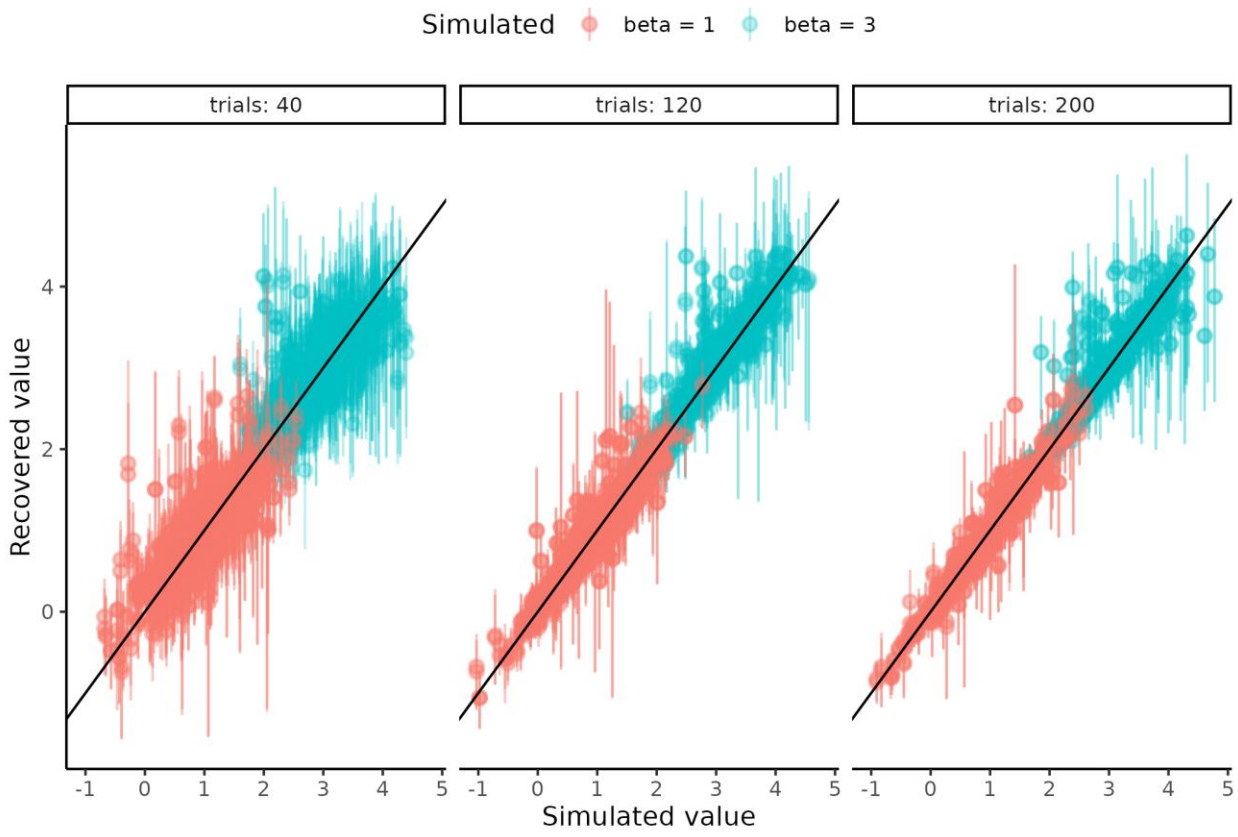
Supplementary figure 4. Parameter recovery of the psychometric function using the intra class correlation coefficient. Identical to supplementary figure 3, but only for the intra class correlation coefficient, but stratified by number of subjects.

Supplementary figure 5



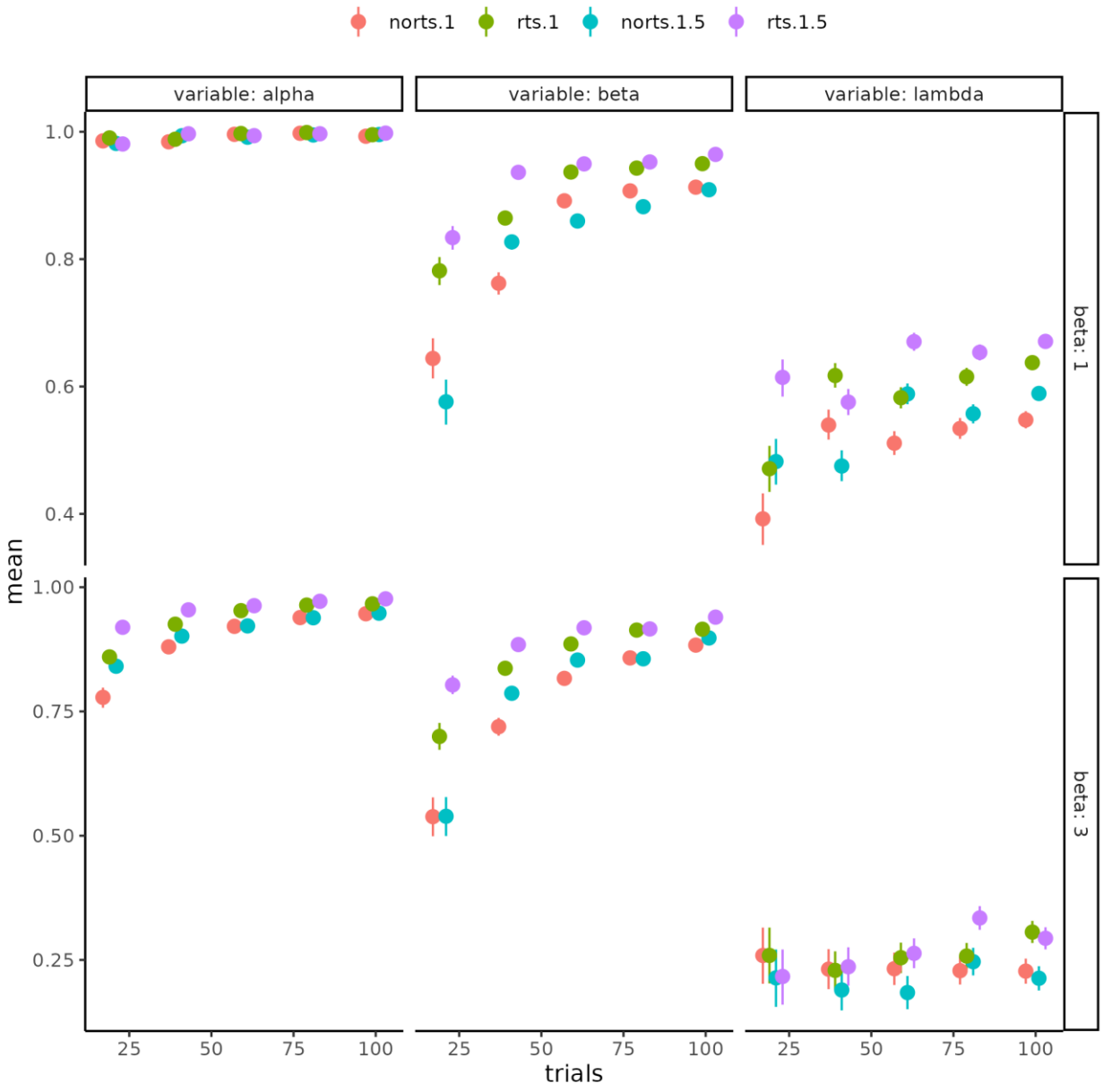
Supplementary figure 5. Parameter recovery of the psychometric function for the Correlation coefficient metrics. Identical to Supplementary figure 4, however instead of investigating the ICC here the correlation coefficient is depicted. True and False in the colors highlight whether uncertainty has been propagated.

Supplementary figure 6



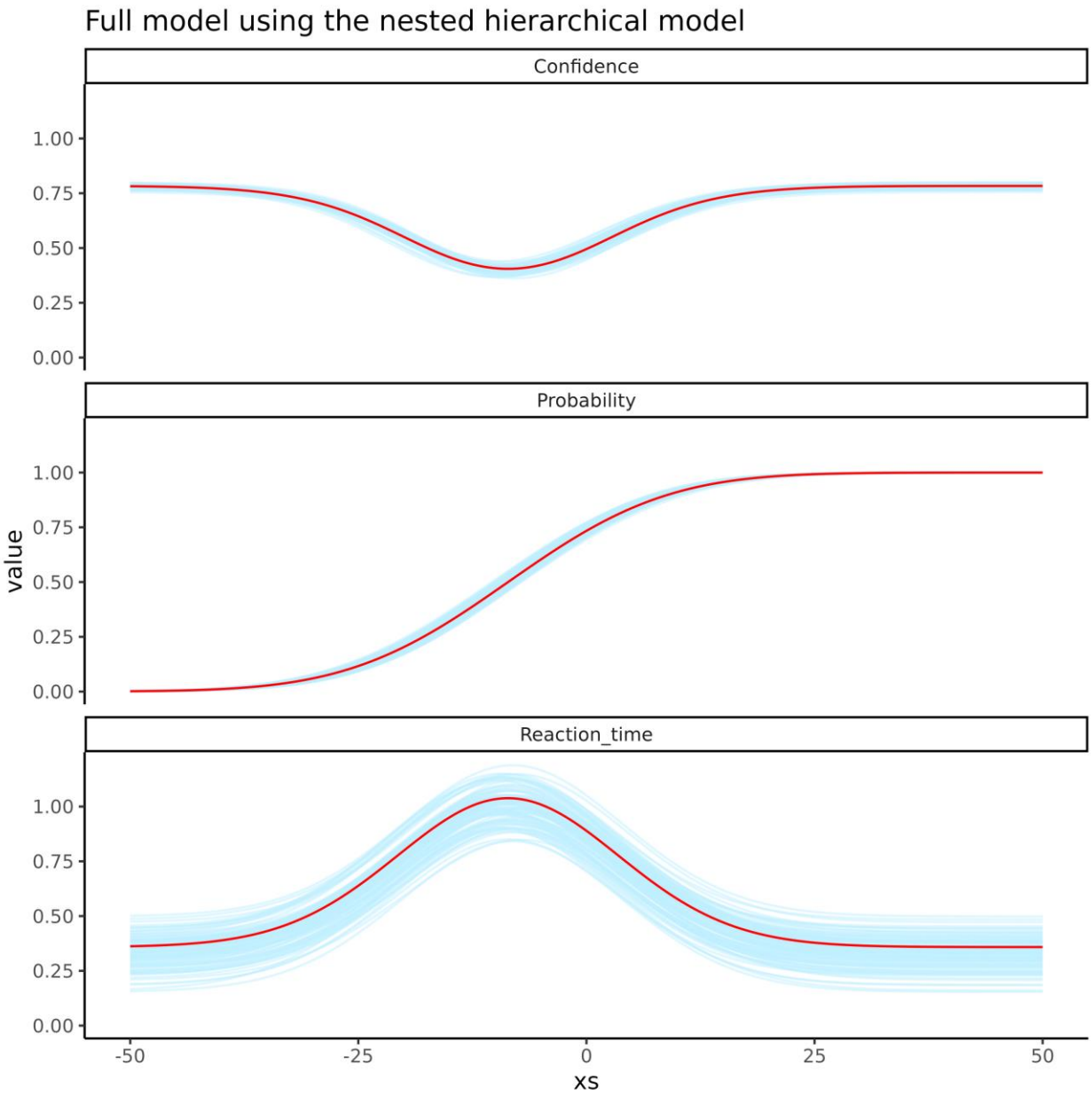
Supplementary figure 6. Pairwise scatter plot of simulated vs estimated slope values for steep slopes (beta = 1 (red)) and shallow slopes (beta = 3 (blue)). Figure shows how increases in trials (left to right) increases the degree to which points all close to the identity line (black line).

Supplementary figure 7



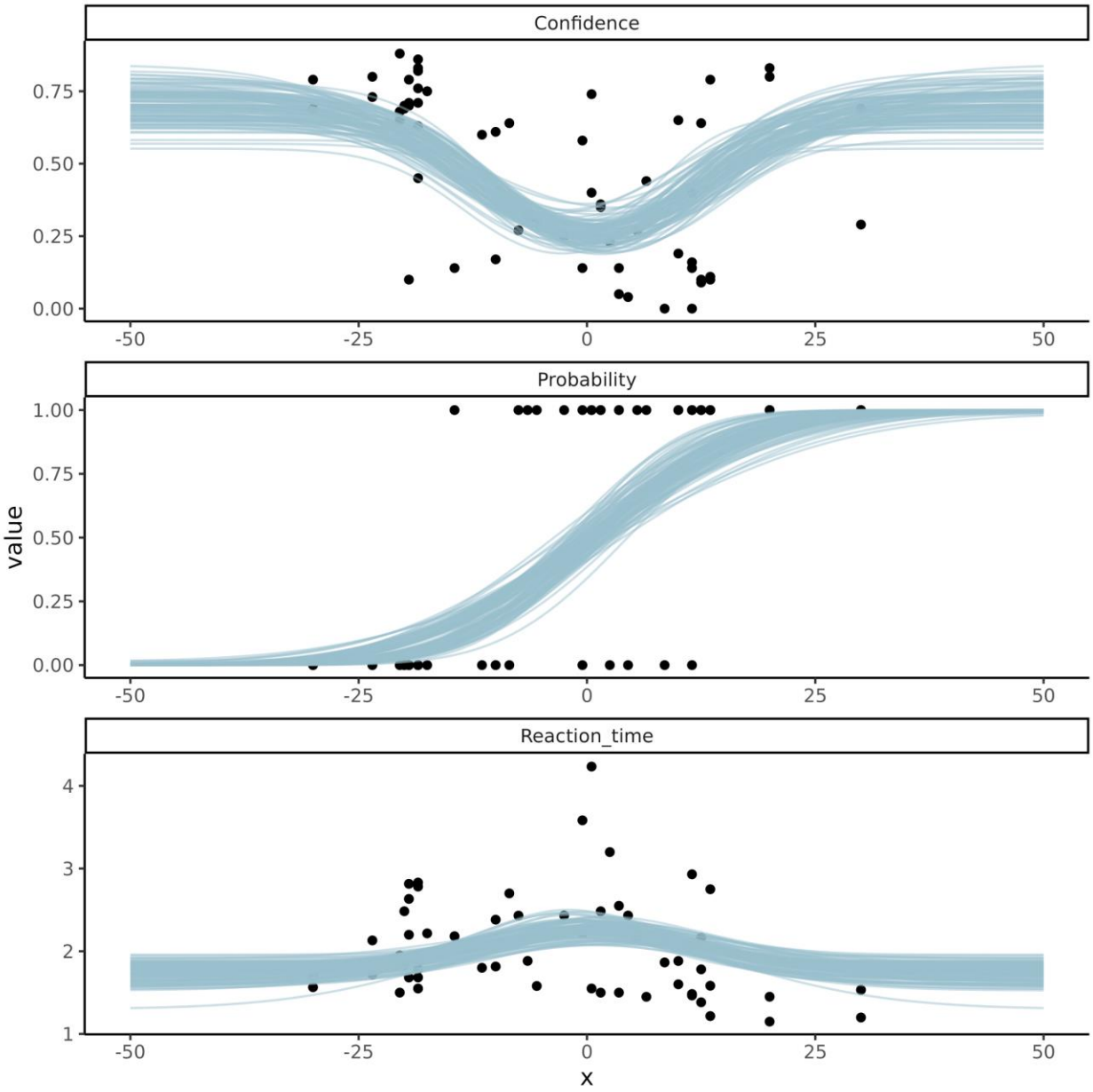
Supplementary figure 7. Parameter recovery of the psychometric function for the correlation coefficient (y-axis) for each parameter (columns) in each combination of including and not including rts and its size (color) and the simulated mean slope (rows) for differing number of trials x-axis.

Supplementary figure 8



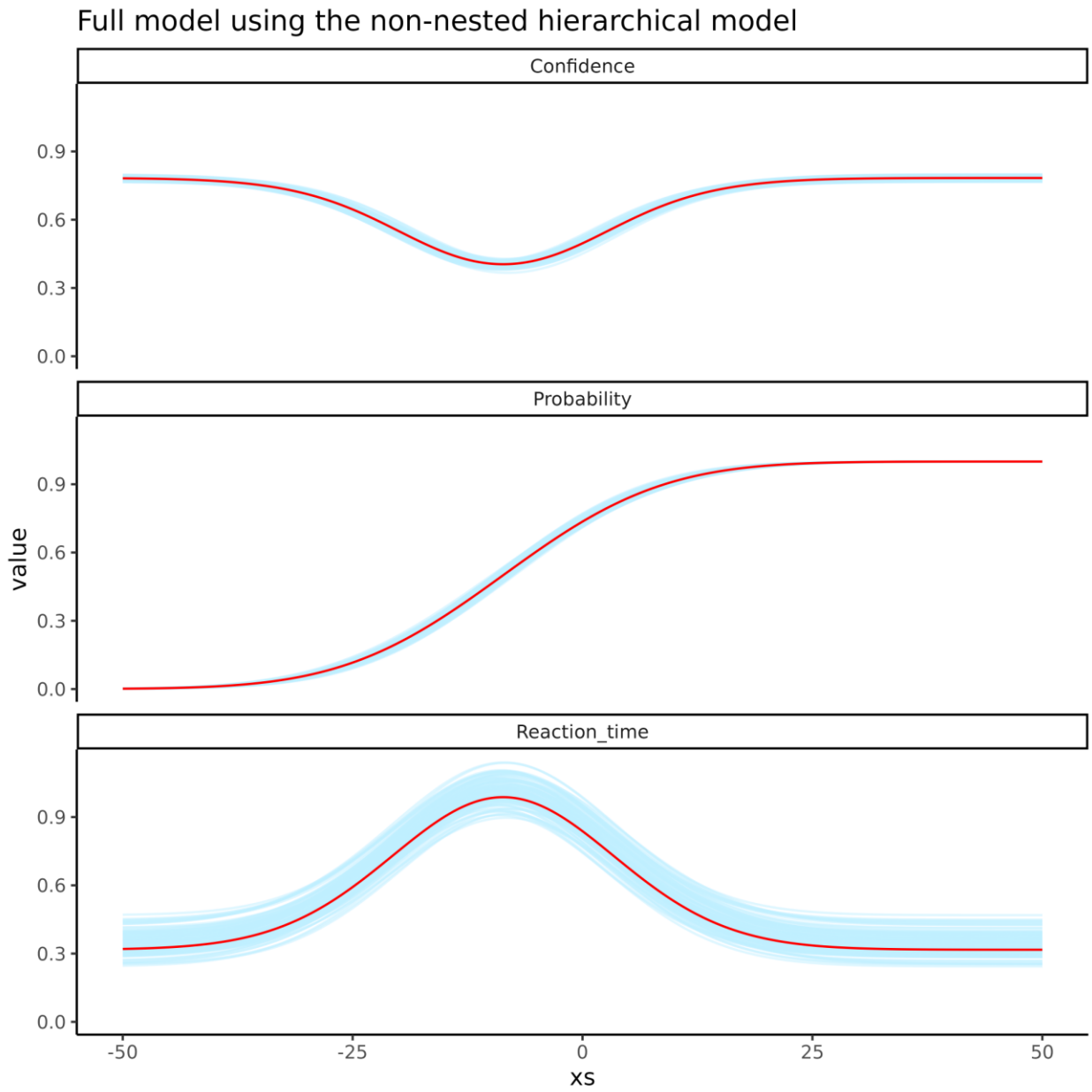
Supplementary figure 8. Group level posterior predictive checks of the 3 types of responses, Confidence, binary (faster or slower) and reaction time on the binary response for the Nested Hierarchical model. Facets represent the 3 types of responses, 0-1 Confidence ratings, 0 or 1 binary responses of (faster or slower) and the reaction time for these binary responses. Red line depicts the mean of the posterior and the blue lines represents 100 posterior draws.

Supplementary figure 9



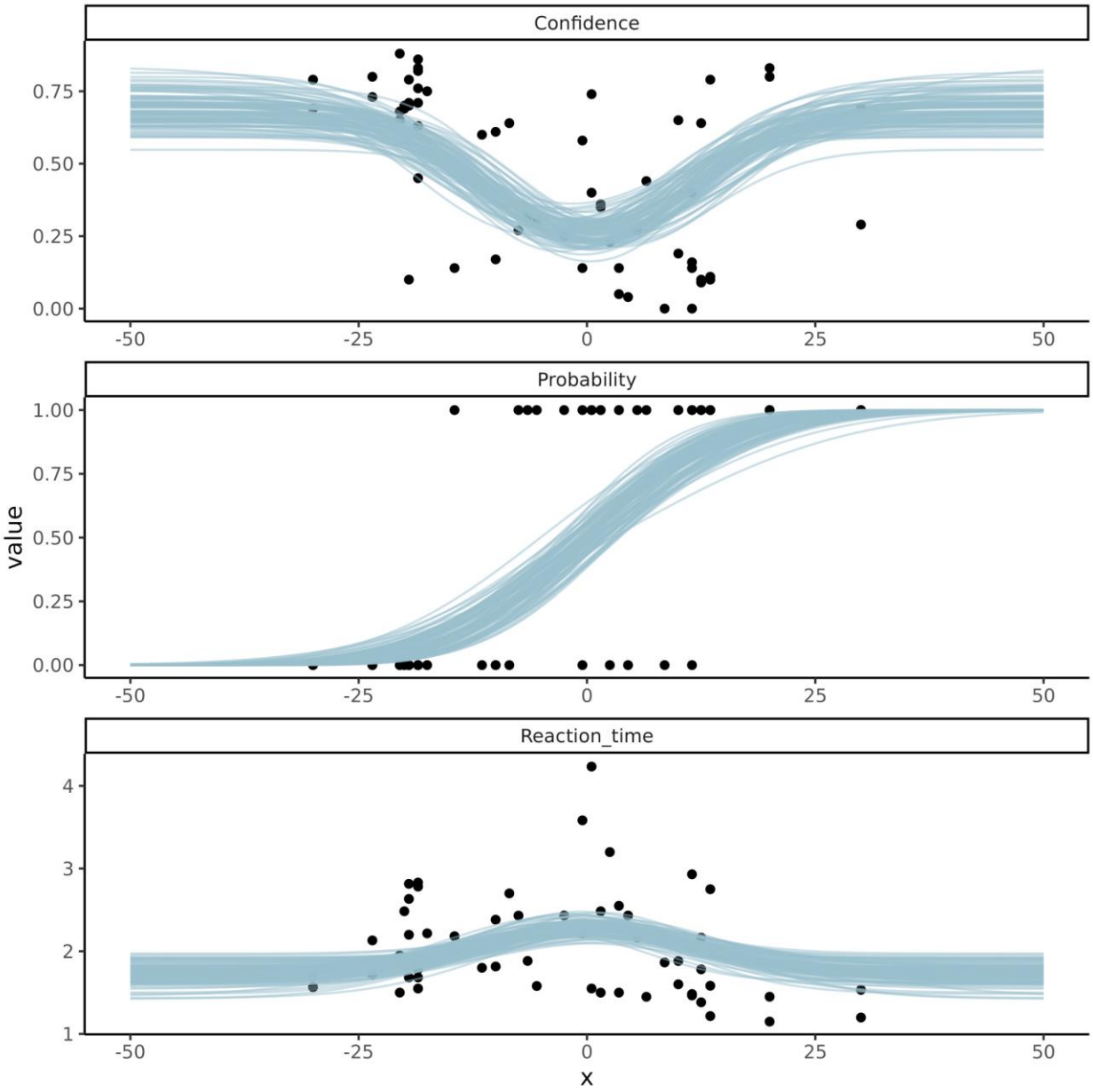
Supplementary figure 9. Single subject level posterior predictive checks of the 3 types of responses, Confidence, binary (faster or slower) and reaction time on the binary response for the Nested Hierarchical model. Facets represent the 3 types of responses, 0-1 Confidence ratings, 0 or 1 binary responses of (faster or slower) and the reaction time for these binary responses. Red line depicts the mean of the posterior and the blue lines represent 100 posterior draws. Black points represent the actual responses of the participant.

Supplementary figure 10



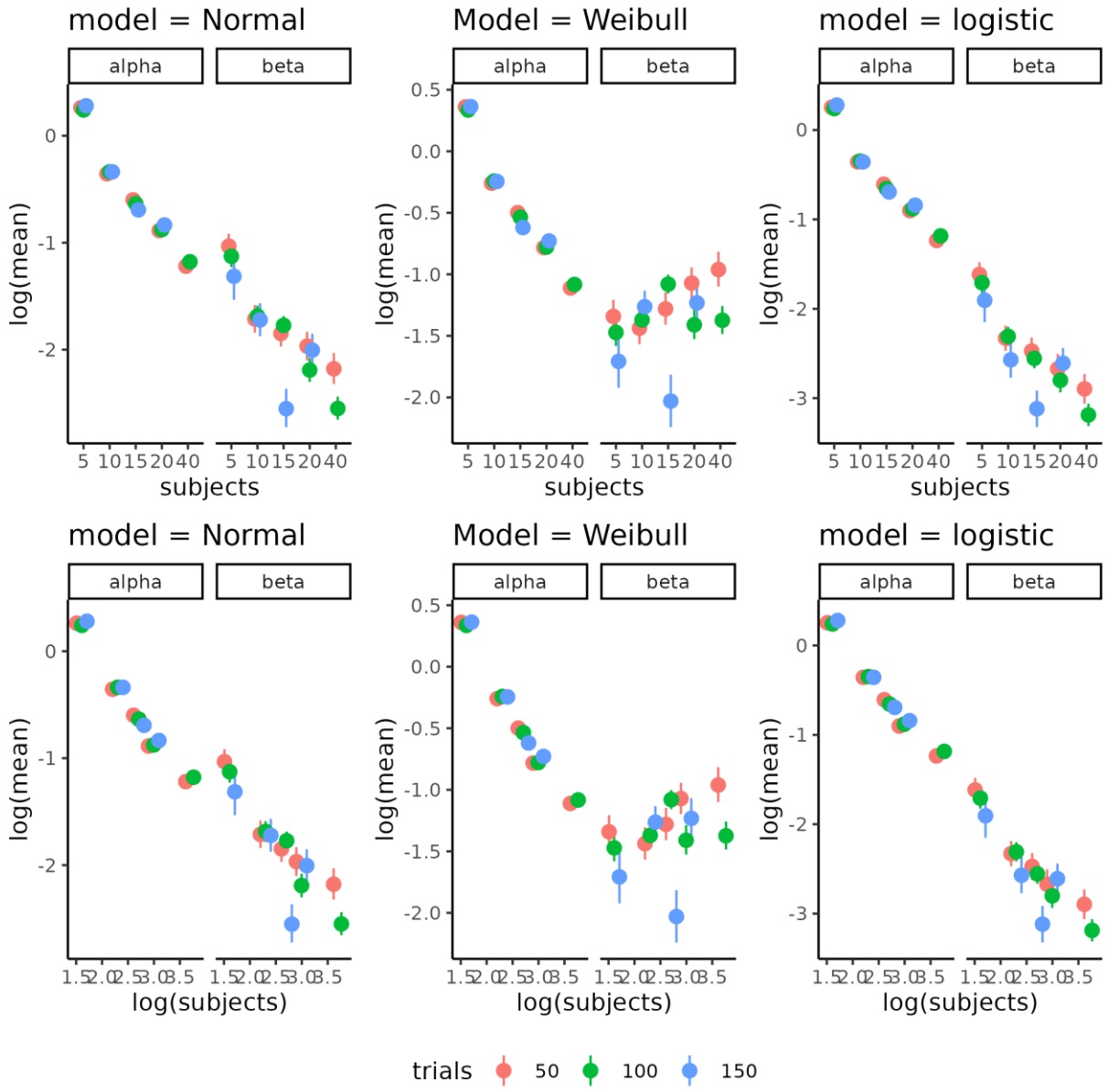
Supplementary figure 10. Group level posterior predictive checks of the 3 types of responses, Confidence, binary (faster or slower) and reaction time on the binary response for the non-nested Hierarchical model. Facets represent the 3 types of responses, 0-1 Confidence ratings, 0 or 1 binary responses of (faster or slower) and the reaction time for these binary responses. Red line depicts the mean of the posterior and the blue lines represent 100 posterior draws.

Supplementary figure 11



Supplementary figure 11. subject level posterior predictive checks of the 3 types of responses, Confidence, binary (faster or slower) and reaction time on the binary response for the non-nested Hierarchical model. Facets represent the 3 types of responses, 0-1 Confidence ratings, 0 or 1 binary responses of (faster or slower) and the reaction time for these binary responses. Red line depicts the mean of the posterior and the blue lines represent 100 posterior draws. Black points represent the actual responses of this participant.

Supplementary figure 12



Supplementary Figure 12. Parameters estimates of the individually fit psychometric functions (columns) on trials and subjects. The top row depicts a (log(y),x) coordinate system whereas the bottom row depicts a (log(y), log(x)) coordinate system. A straight line relationship between subjects (trials) and the log of the parameter value (top row) would indicate exponential relationship whereas a straight line in the log; log coordinate system would imply a power law relationship in the native (x,y) space.

Supplementary tables

Supplementary table 1

see Supplementary table 1 which is linked to the github of the thesis. As the table is too big to show on a single page.

Supplementary table 2

Table supplementary 2: Group mean parameter distributions of Binary test re-test re-analysis. Summary statistics of the group mean parameters of the Binary hierarchical model, used as the baseline parameters for the power analysis.

variable	mean	q5	q95
mu_threshold	-8.35	-9.50	-7.17
mu_slope	2.26	2.21	2.31
mu_lapse	-4.93	-5.93	-4.15
std_threshold_between	7.92	6.86	8.98
std_slope_between	0.14	0.03	0.23
std_lapse_between	0.48	0.04	1.19
std_threshold_within	7.36	6.70	8.08
std_slope_within	0.34	0.26	0.41
std_lapse_within	1.99	1.46	2.61

Supplementary Notes:

Supplementary note 1: Priors

In the section for Standard parameter recovery a single fit psychometric function was used to demonstrate the standard parameter recovery approach to internal validity of the psychometric function. The priors for that model were as follows for the threshold, slope and lapse rate respectively.

$$\alpha \sim N(0,20)$$
$$\beta \sim N(0,2)$$
$$\lambda \sim N(-4,2)$$

Next the priors for the nested hierarchical model was as follows:

$$\mu_{\alpha} \sim N(0,10)$$
$$\sigma_{\alpha(within)} \sim N(0,10)$$
$$\sigma_{\alpha(between)} \sim N(0,10)$$
$$\mu_{\beta} \sim N(0,3)$$
$$\sigma_{\beta(within)} \sim N(0,3)$$
$$\sigma_{\beta(between)} \sim N(0,3)$$
$$\mu_{\lambda} \sim N(-4,2)$$
$$\sigma_{\lambda(within)} \sim N(0,3)$$
$$\sigma_{\lambda(between)} \sim N(0,3)$$
$$\rho_{between} \sim LKJ(2)$$
$$\rho_{within} \sim LKJ(2)$$

Supplementary note 2: Lapse rate explanation

As mentioned in the main text, the lapse rate can be quite difficult to estimate if the proportion of lapses are low. In the main text it is argued that lapse rate of approximately 1% is difficult to estimate as this would on average in the simulation with 100 trials per subject amount to 1 trial where the subject would have a lapse. However, from the model's perspective, these lapses are not created equally, which makes it even more difficult to estimate this parameter. This is because of the difference between having a lapse when the stimulus value is extreme in either end (high or low) or having a lapse when the stimulus value is close to the simulated threshold. A lapse close

to the simulated threshold would from the model's perspective does not interfere with the estimation of the parameters as this response could just be due to the stochastic nature of the task (i.e. the y-axis of the Psychometric function is the probability of responding 1). Therefore, only lapses at the extreme ends would inform the model about the underlying probability of having a lapse.

Supplementary note 3: Pathfinder explanation

The pathfinder algorithm was used for optimization of the trial-by-trial simulations. The implementation was as follows. Firstly a randomly drawn stimulus value in the domain of [-50 ; 50] was drawn. After the representation of this stimulus value a response is collected based on the simulated parameters. The stimulus value together with the responses is then fit to pathfinder which uses the priors of the model with this observation to update the parameters. The next stimulus is then selected by taking the posterior mean of the threshold. This was done for the first 5 trials to get reasonable estimates of the parameters of the model. To properly explore the width and size of the psychometric function the stimulus values were after the first 5 trials selected based on a single draw from the posterior threshold and slope. This meant extracting a draw of the threshold and then randomly either adding the draw of the slope or subtracting it. For the full code see. The priors for this model are identical to those found in the single subject psychometric function (see supplementary note 1)

Supplementary note 4: Comparison of ADO algorithms

Comparison between the 3 tested models rested on having each algorithm simulate stimulus values based on simulated parameter values. These simulated stimulus values were then used to obtain responses of the agents (again using the simulated parameter values). Each set of stimuli and responses for each algorithm were then fit using the same Bayesian model (see supplementary note 3 and 1) and the posterior distributions were computed for each of the three parameters values (i.e. threshold α, slope β and lapse rate λ).

Supplementary note 5: Posterior predictive checks

Supplementary figures 8 and 10 display the posterior predictive check for the two hierarchical models with the three types of responses i.e. binary, reaction time and confidence ratings. In these plots the group means of the parameters are depicted giving an indication of the overall structure of the behavioral responses of the participants.

Supplementary figures 9 and 11 depicts instead of the group level estimates a particular participant at a particular session with overlaid data. Interested readers are referred to the github to explore other participants than depicted here.

Note that for the non-nested hierarchical model the group level estimates are the mean of the two sessions whereas this is directly estimated in the nested hierarchical model.

Supplementary note 6: Power analysis model description

The full stan code for the model used in the power analysis, see the github. The model assumes that all three parameters of the psychometric function is drawn from a multivariate normal distribution with group means and between subjects' variances, furthermore two group differences are also drawn from this multivariate normal distribution that calculates the difference between slopes and thresholds between sessions. The variance co-variance matrix was decomposed using the Cholesky-decomposition and the LKJ-prior was used for the correlation coefficients between parameters which was set to $\eta = 2$ i.e. a quite wide prior on the correlation but with less mass for more extreme correlations.