Uncertainty in Thermosensory Expectations Enhances an Illusion of Pain

Jesper Fischer Ehmsen,

# Abstract

Trusting papers introducing models without formally testing them is like trusting a mathematician to measure a 1m stick because of a phd in mathematics or physics instead of having him measure the stick and compare it to the 1m stick.

**Keywords:** Uncertainty, Cognitive modeling

# Introduction

Imagine measuring the length of an A4 paper with a ruler and then using a well validated questionnaire to determine your anxiety level. Which of these two measures, i.e. the length of the A4 paper or your anxiety level, would have the greatest faith in? In this thesis I will investigate shortcomings in uncertainty handling in cognitive science literature while also providing ways to account for these uncertainties when statistical or computational models are applied. To do this I will draw examples from the study of physical systems as uncertainty quantification is more straightforward and provides an entry point to understanding why and how cognitive science field deals with uncertainty. Lastly using commonly known models in the cognitive science literature like the psychometric function I will highlight opportunities of optimizing task design and modeling choices, not only for a more holistic view of cognition and the humans that we are studying but also to minimize the uncertainty associated with our statistical models.

## *Uncertainties in science and examples from physics*

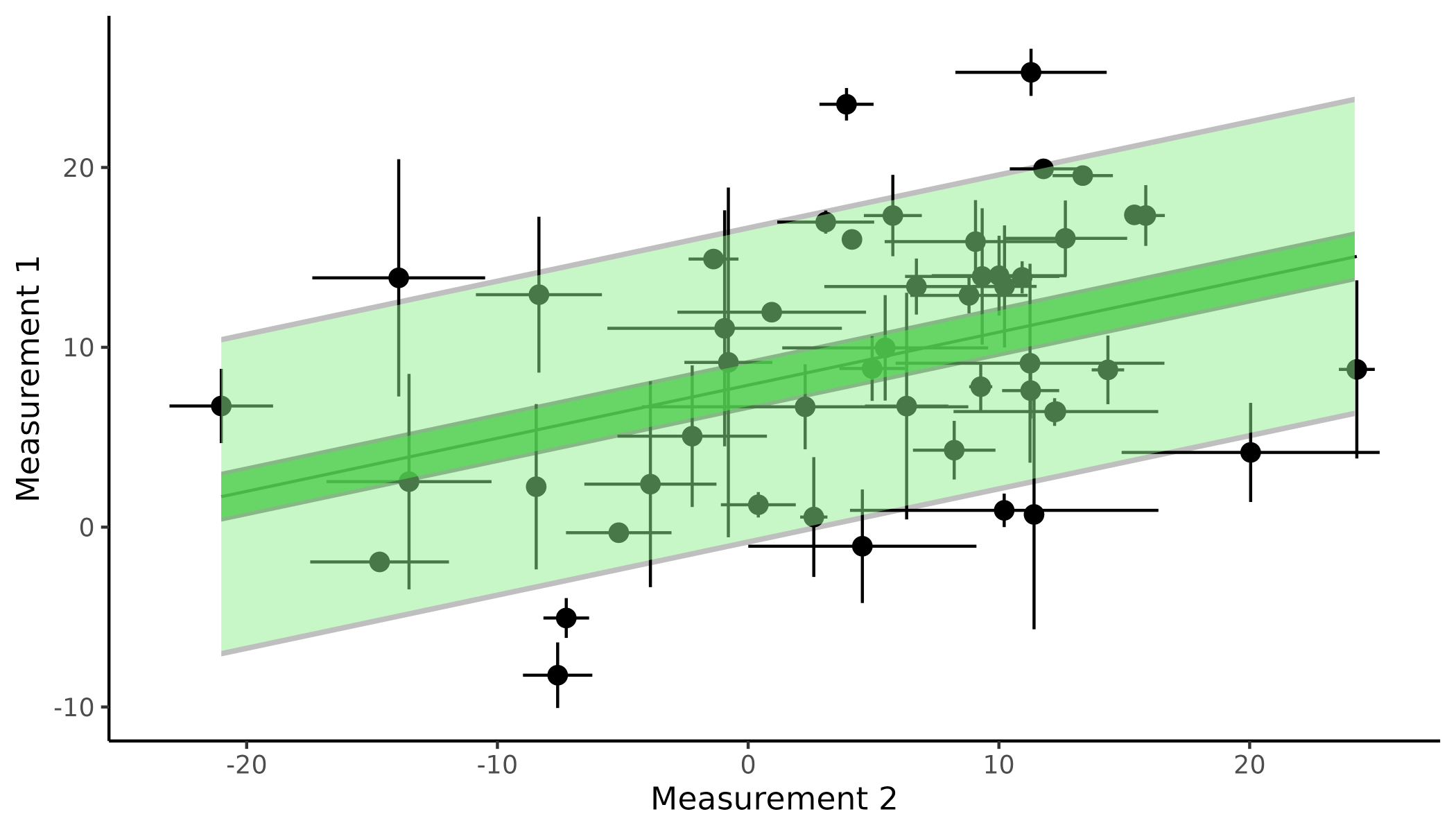
Science is a systematic way to organize knowledge in hierarchies, with a foundation in testable hypotheses. Knowledge can be hard to define, however for this thesis it will suffice to define knowledge as the idea of knowing when something will or would happen. For instance, a dropped cup will fall towards the ground. This is to say that knowledge is the relationships that we believe to be true with differing amounts of certainty. The reality is that even though we might say that we are completely sure, i.e. know, that the cup will fall towards the ground this is an assumption that is true most of the time but given that the natural world is bounded on probabilities, complete certainty is unwarranted. This probabilistic nature can at one level be described by quantum mechanics, however from a less fundamental level the cup might not fall towards the ground or shatter because of some unseen events (example?). The interest here is not always in unseen events per say but instead in the predictability and (un)certainty of the expected. Taking the falling cup as an example, we would normally not be interested in the probability that the cup will hit the ground and shatter, but instead in the acceleration of the cup and the uncertainty in this estimate. What scientists have shown is that objects dropped on earth will accelerate towards the ground with an acceleration of 9.82m/s^2 (REF) (depending on where you are on earth). However, this number does not mean anything without an estimate of the uncertainty, while also accounting for the assumptions that are entailed with these numbers. The first proposition is well studied and is that we are X % sure that it will be in the range between Y and Z. The second proposition is also quite well studied as we know that the density of the medium that the cup is dropped in is quite important and the shape and weight of the cup if dropped outside a vacuum. There are X main points of the example which this thesis will explore, firstly uncertainties are organized in hierarchies as knowledge and are just as important as beliefs as without one, the other is meaningless. Secondly, taking these uncertainties seriously and herein estimating and propagating them should not be a choice or something that can be avoided, but a necessity of all scientific endeavors, where they can be quantified or at least be approximated.

Taking its outset in the published literature, the thesis will establish issues regarding uncertainty handling and propagation and use simulations to highlight the problems with neglecting a proper account of uncertainty in our statistical models. After highlighting these potential issues, the thesis will provide possible ways of dealing with the shortcomings by means of simulations. The goal of this thesis is to illuminate the often-overlooked uncertainties in the data collected on human behavior and cognition while providing ways of dealing with it, such that the uncertainty reported in the published literature more accurately reflects the uncertainty we should have in the results. The thesis will argue that accounting for uncertainties is more important than ever, especially in research of complex systems such as humans as computational resources have made it possible to easily develop more sophisticated analyses and models that have dependencies on lower-level analyses, which makes the need for proper uncertainty handling even more imperative. Furthermore, these computational resources do allow for uncertainty propagation without understanding the underlying mathematics, making it accessible to most researchers with some coding experience. To effectively communicate 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 in cognitive science and examine which parts of the literature might be more susceptible to overconfident findings by neglecting uncertainty propagation. Next the thesis will be investigating different types of models used in cognitive science and examine how best to account for uncertainties that are commonly neglected.

## *Levels of uncertainty and uncertainty propagation*

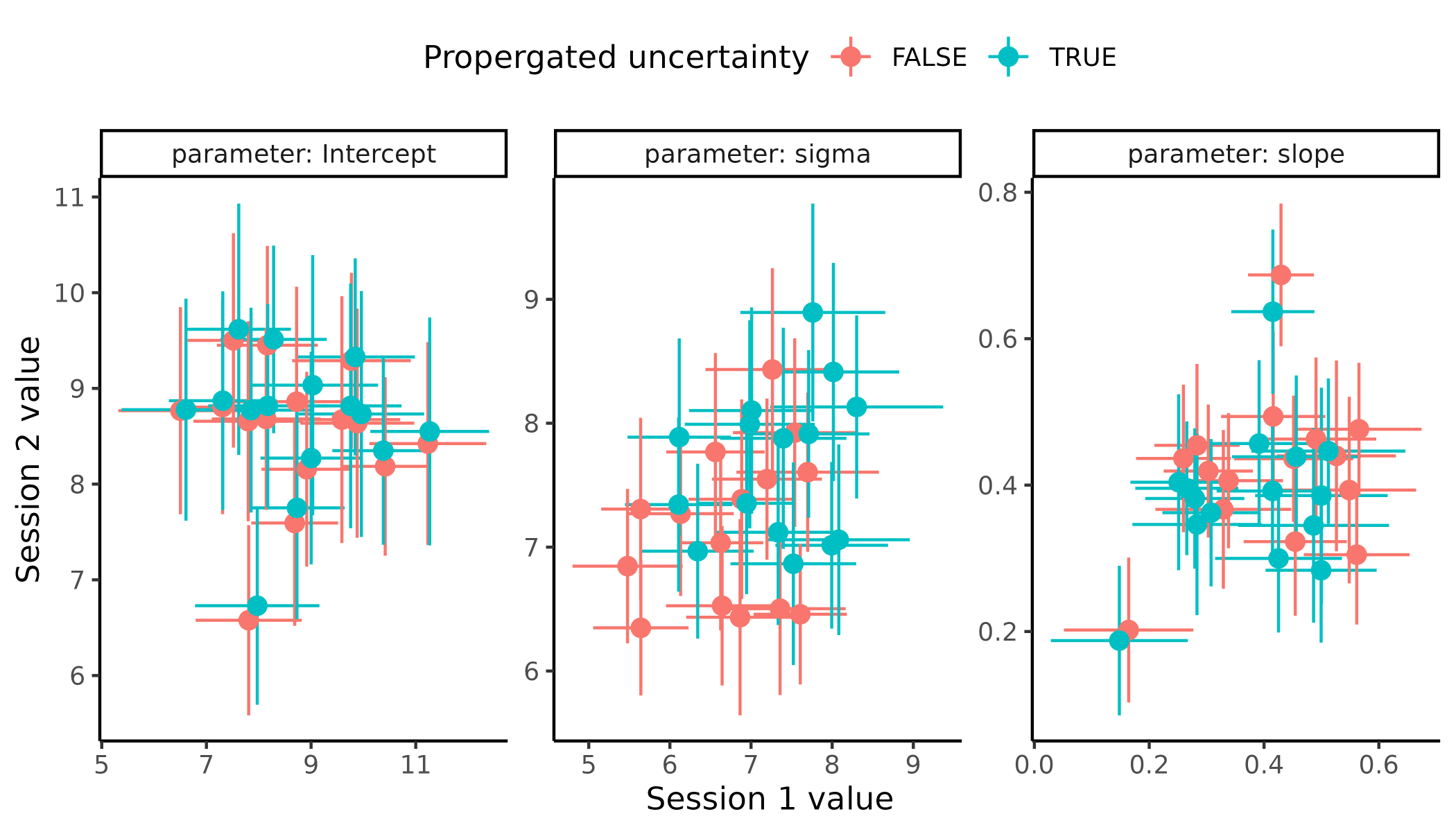
I will here broadly define 3 different types of uncertainty, measurement, estimation, and test-retest reliability uncertainty see figure 1 for a visualization. These definitions are not exhaustive and will be centered around how experimental studies in cognitive science are conducted, from data collection to data analysis. The first aspect of uncertainty is to acknowledge that uncertainties can be defined in hierarchies as uncertainty propagates these hierarchies. This uncertainty propagation means that as you do calculations based on measures with uncertainty, the uncertainty propagates to the results of the calculations. In this thesis I will be using simulations to show how uncertainty propagation can be understood and handled with a need for rigorous mathematical proofs for a more mathematical treatment see (REFS).

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 the falling cup to reach the ground. This level of uncertainty is often neglected in cognitive science when applying statistical models, because they are thought to be minuscule as in the case of reaction time tasks, which may or may not be true given the experiment setup ([Crocetta & Andrade, 2015](#ref-crocetta_problem_2015); [Holden et al., 2019](#ref-holden_accuracy_2019); [Ohyanagi & Sengoku, 2010](#ref-ohyanagi_solution_2010)). This is not to say that cognitive scientists do not care about them, as moving towards more sophisticated measurement methods is an ongoing endeavor. For instance, moving from lower to higher magnetic fields or more electrodes when using functional magnetic resonance imaging or electroencephalography respectively ([Glover, 2011](#ref-glover_overview_2011)). Minimizing this kind of uncertainty most often revolves around getting better tools to measure the variable(s) of interest. However, there might also be other avenues where a more explicit quantification of measurement uncertainty might be appropriate. One of the main instances coming to mind is the use of questionnaires, that try to quantify how mental health conditions such as depression or anxiety correlate with cognitive measures or parameters. Many of the main questionnaires used to assess depression, anxiety, stress etc. use several questions that are then added together to give a score of the mental health condition without a quantification of the uncertainty ([Cohen, 1994](#ref-cohen_perceived_1994); [Johnson et al., 2019](#ref-johnson_psychometric_2019); [Kroenke et al., 2001](#ref-kroenke_phq-9_2001); [Xiao et al., 2023](#ref-xiao_psychometric_2023)).



**Figure 1 Measurement and Estimation uncertainty;** displays a linear regression between two measurements 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 next level of uncertainty is when a particular model is fit to some data, or more broadly when calculations are done on data with uncertainty. In cognitive science we achieve parameter estimates from our collected data and these parameter estimates have uncertainty associated with them, this uncertainty will be referred to as estimation uncertainty. Estimation uncertainty is most often quantified by the statistical model be that the standard error of a regression coefficient or the width of a posterior distribution of a parameter in a Bayesian framework. Minimizing this estimation uncertainty is what most scientists care about, as inevitably most cognitive science experiments revolve around null hypothesis testing, which in most cases will involve testing whether the parameter estimate includes a particular value mostly, 0. To minimize this type of uncertainty the standard approach is to get more data, given they are from the same population and behave similarly. In cognitive science this might include achieving more trials or subjects to get a more precise estimate of interest i.e. minimizing estimation uncertainty. In cognitive science the minimizing of estimation uncertainty is however not free or free of uncertainty itself. Firstly, both approaches, increasing trials and subjects, utilize more resources, but more importantly increasing the number of trials in a cognitive task might even increase the estimation uncertainty. This can happen for several reasons, but boredom, habituation, fatigue and lack of engagement can become big contributors when experimental tasks become very long ([Jeong et al., 2023](#ref-jeong_exhaustive_2023); [Meier et al., 2024](#ref-meier_is_2024)). Other contributors might be that subjects switch between cognitive strategies and if not properly accounted for in the analysis might be interpreted as additional noise. Next increasing the number of subjects included in a study will many times decrease estimation uncertainty on 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 reduce the number of resources used. However, there are many times also other ways to minimize estimation uncertainty ([Baldi Antognini et al., 2023](#ref-baldi_antognini_new_2023); [Stone, 2014](#ref-stone_using_2014)). Changing the task design such that responses will give more information for the parameter values when estimated. This optimization strategy involves individualizing the task design such that each presented stimulus is the most informative. This task design optimization is frequently used in psychophysical experiments where adaptive algorithms are used to select the upcoming stimuli such that it minimizes the uncertainty in the estimated parameter values of the psychometric function. Example algorithms are PSI, QUEST ([Prins, 2013](#ref-prins_psi-marginal_2013); [Watson, 2017](#ref-watson_quest_2017); [Yang et al., 2021](#ref-yang_adopy_2021)). The next level of uncertainty stems from the fact that these parameter estimates will vary over time, as humans vary over time. This type of uncertainty will be referred to as test-retest uncertainty To fully appreciate the relationship between these types of uncertainty and why they are hierarchically organized, one might think of a researcher that want to conduct an experiment on how total time spent awake (sleep deprivation) influences reaction times in an experimental paradigm. Now imagine that the total time spent awake is measured with uncertainty either because of the participants’ accuracy in estimating when they fell asleep the previous day or in the device that should have measured it, this is one type of measurement uncertainty in the experiment. Now each reaction time of the participant also have measurement uncertainty, but more importantly because the reaction times are measured several times the mean reaction time is going to have estimation uncertainty associated with it, here the standard error of the mean. Lastly outside the current experiment of the researcher’s reaction times fluctuate even in the same person from day to day highlighting test-retest uncertainty.



**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 and the slope respectively from left to right. Colors represented weather the measurement uncertainty was proporgated or not.

The main message here is that to get reliable estimates and, in the end, to make reliable inference one needs to account for all these 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 won’t matter if you cannot estimate it or measure it reliably in the first place.

## *Investigating measurement uncertainty in questionnaires*

To keep a consistent theme, I will throughout the thesis be demonstrating how computational resources have made the need for analytical solutions involving tedious assumptions sometimes irrelevant. This is highly relevant as closed-form-problems where an analytical solution is known or even attainable are becoming less and less frequent with the surge in popularity of more and more complex models, see section about modeling definitions for further elaboration.

To demonstrate how and by how much adding uncertainty into questionnaires will change interpretations, I will here focus on the perceived stress scale (PSS) ([**cohen\_perceived\_nodate?**](#ref-cohen_perceived_nodate)). Which is a questionnaire of 10 items that are scored on a 5-point Likert scale from 0 (never) to 4 (very often), with 1 (almost never) 2 (sometime) 3 (fairly often). Trying to avoid response bias and random selection of questions, the last 5 items’ scores are reversed and then the sum of the scores is calculated to give a perceived stress score with defined intervals of scores of 0-13 14-26 27-40, indicating low stress, medium stress and high perceived stress respectively. This questionnaire has been used in various papers ([Berker et al., 2016](#ref-de_berker_computations_2016); [Luijcks et al., 2015](#ref-luijcks_influence_2015)),indicating widespread acceptance across scientific fields ([Kuiper et al., 1986](#ref-kuiper_global_1986); [Treadway et al., 2013](#ref-treadway_perceived_2013)). In this section I will demonstrate how we can use computational resources to both understand and explore how adding measurement uncertainty to this particular questionnaire will change the interpretation of what many papers in the cognitive science literature do, that is correlating the questionnaire score with other variables such as behavioral measures like mean reaction times, computational parameters like learning rates or structural properties of the brain like the myelination or grey matter volume in a region of interest ([Wu et al., 2021](#ref-wu_neurobiological_2021)).

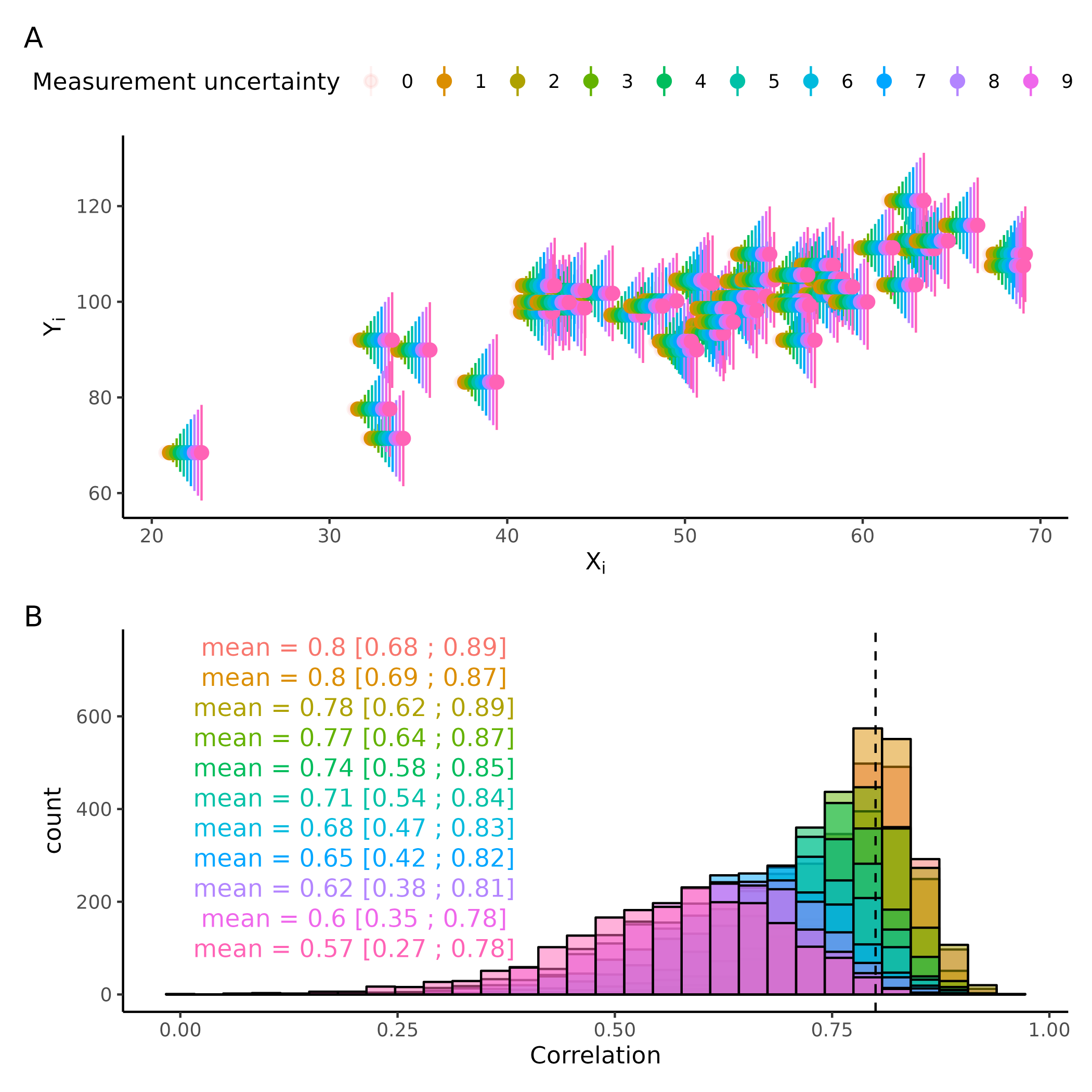
To make this section tangible I will start by showing how uncertainty estimates of statistics (here persons correlation coefficient) can be obtained using computations instead of analytically solutions. The way to do this is to resample the collected data with replacement i.e. bootstrapping the data and then recalculate the test statistic of interest. Iterating this process gives a distribution of test statistics which with enough iterations will contain the same information as the analytical solution. This for the simplest case of recalculating the correlation coefficient and its uncertainty might seem somewhat tedious compared to taking the direct analytical solution, as this is already implemented in most statistical softwares and packages, however once setup and understood this approach allows for adding not only measurement uncertainty, but a more general way of thinking about the uncertainty of statistical metrics. One of the advantages of having an analytical solution to this simple case of recalculating the uncertainty of the correlation coefficient is to ensure that the code and scripts are properly set up. This therefore serves as a validation step before exploring territories where analytical solutions are scarce or nonexistent.

The first step here is therefore to show that the two approaches of simulating and analytically estimating the uncertainty of the correlation coefficient is identical across different ranges of correlations and sample sizes. To do this, I’ve simulated data from a multivariate normal distribution with the following parameters.

Where

The multinormal distribution (Eq. 1) produces random variables with a means a standard deviation and crucially with a correlation coefficient between all random variables , here the subscript indicates that there can be x random variables been sampled together. This distribution is perfect for understanding how the correlation coefficient changes as it’s a parameter of the distribution. Now demonstrating that using bootstrapping and the analytical solution implemented in the “stats” package in r are identical, I simulate correlation coefficients ranging from -0.9 to 0.9 in increments of 0.1 with the total number of samples per random variable being between 50 and 500 in increments of 50. Supplementary Figure XX shows how the mean correlation coefficient changes based on the number of simulations and the approach used to obtain it as well as how the uncertainty of that estimate changes due to the same factors. Having shown that the two approaches are identical we can add measurement uncertainty to each observation. This has again been analytically solved and solutions exist to calculate the correlation coefficient under these circumstances (REF). To add measurement uncertainty to the measurements we can instead of randomly resampling pairs of data points from the original data, as done for the simplest case above, one resamples these pairs as means of an error distribution where the uncertainty of this distribution is the measurement uncertainty. A mindless choice of error distribution would be the normal distribution which would reflect the fact that the directionality of the uncertainty is assumed to be bidirectional i.e. with no preferred direction. Of note here is that one might sample the original data from other error distributions for instance if values are strictly positive or bounded in other ways 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 questionnaire values. For the first demonstration of adding measurement uncertainties to observed data. We simulate normally distributed noise which means simulating new “observed values” from a normal distribution with a mean of the observed observation and a standard deviation or simply adding noise by adding a normally distributed random variable with mean 0 and standard deviation which serves as the degree of measurement uncertainty, these two implementations are identical (see supplementary XXX). This can be seen in figure X, here uncertainty is added to firstly just the x values (A) then to the y values (B) and lastly to both (C). As is clear the estimated correlation coefficient using bootstrapping is being attenuated with increasing measurement uncertainty.

Mostly when cognitive scientists are displaying such a scatter plot in their results no error bars are present even though both x and y values are associated with error, adding to an illusion of certainty in the measures and the presented statistics such as the correlation coefficient might be inflated.



**Figure 3 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 coeficient without uncertainty.

Moving to the example of the PSS questionnaire, one might borrow ideas from physics of adding a small uncertainty to each of the 10 questions of the questionnaire and then propagate the uncertainty using error propagation, the same way physicists would add 0.5mm uncertainty to a measurement of length using a ruler. If we assume that the measurements / question entries are independent, this becomes one of the simplest cases of error propagation and the additional uncertainty from each added question to the questionnaire score will add the following uncertainty to the overall score:

Where is the uncertainty on the overall score and , … are the uncertainties of the individual questions. Assuming the uncertainty on each question is identical, then this expression simplifies to:

Where n is the number of questions in the questionnaire and is the uncertainty associated with each question and being the uncertainty associated with the PSS. A couple of key things should be noted, the independence of questionnaire scores is surely unjustified as we expect the first 4 questions to be highly positively correlated and these 4 being negatively correlated with the last 4, as these have been reversed in meaning. Next the error distribution for each question and therefore also the resulting questionnaire score cannot have a normal distribution as the error distribution because we know that the minimum and maximum score of each question cannot exceed 0 or 4 and the maximum score is bounded between 0 and 40. One might therefore first try a truncated norm distribution between 0 and 4 as the error distribution for each question with the standard deviation of this distribution being the amount of error we assume to be present for each question.

## *Modeling definitions*

This thesis will revolve around building, refining, testing, and designing models of cognition. To do this cognitive modelling will be deployed. Here cognitive modelling is meant as an intermediate level in a hierarchy of computational models on top, and statistical models in the bottom. The distinction between these concepts can be found in their flexibility, assumptions, and scope of investigation. It should be noted that all these types of models have many things in common such as being mathematical representations of a data generating process and that these are working definitions with fuzzy boundaries.

*Statistical models* are the models primarily used in medical, social, and educational sciences, these models mostly consist of linear and generalized linear (mixed) models. What these models have in common is that they are linear combinations of independent variables which are sometimes transformed (making them generalized) to a particular domain such that this linear combination maps to a dependent variable. The mathematical representation of such models are as follows:

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, which makes predictions on probabilities and strictly positive values like reaction times possible. is a vector of regression coefficients of P predictors which gets estimated, X Is a matrix of independent variables of size [N, P]. Lastly is a vector of N elements containing the errors of the model predictions on the dependent variables. The benefit of these regression models is that maximum likelihood estimators are available meaning that parameters estimates can be calculated using a frequentists statistical framework, making the estimation process fast and efficient. However, the downfall of these models 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 is highly nonlinear in many ways ([Ivanova et al., 2022](#ref-ivanova_beyond_2022)).

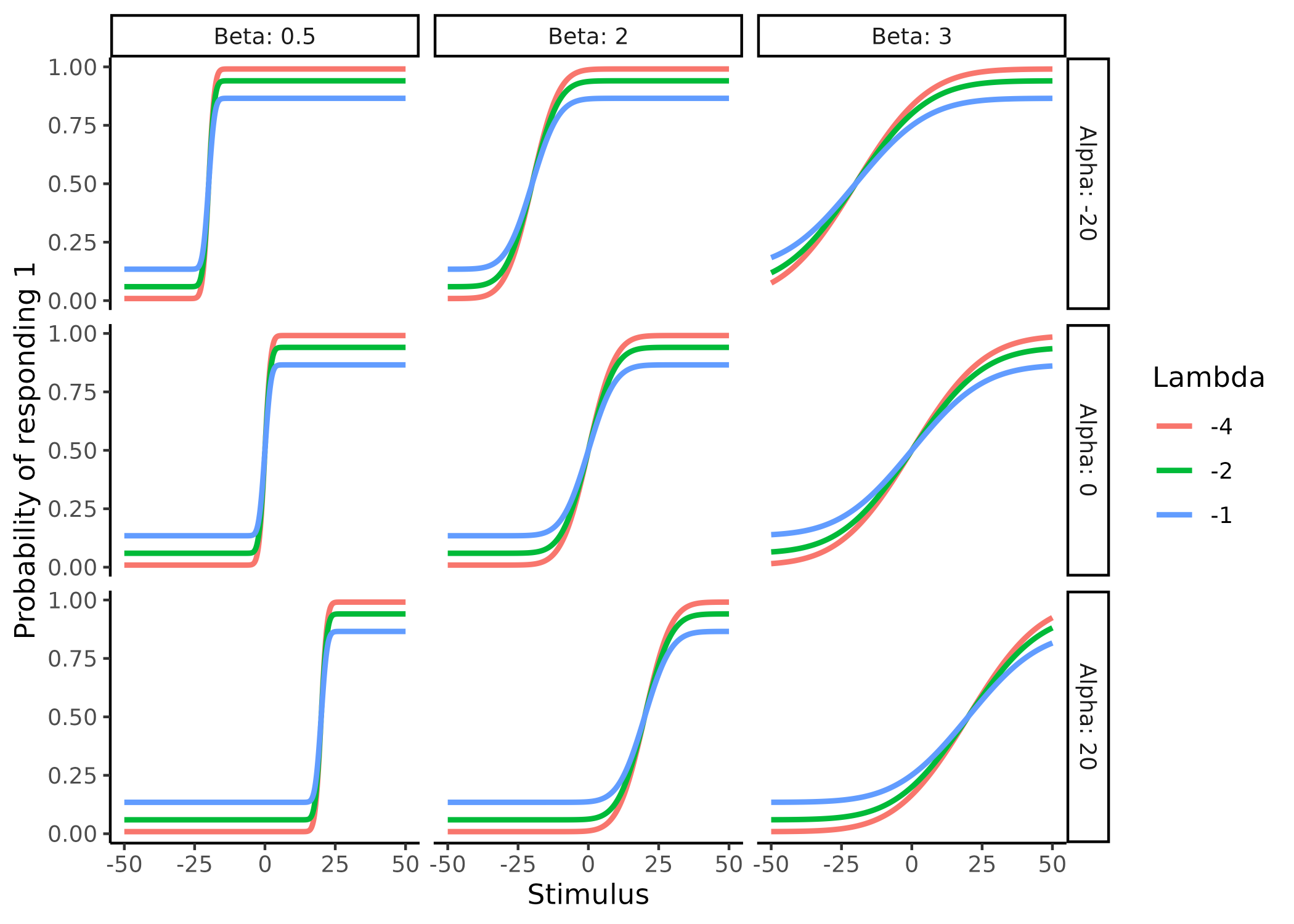
*Cognitive models* are models that are meant to resemble the generative processes of human behavior more closely. These models are generally more theoretically driven as the constraint of linear combinations is avoided, by employing different optimization schemes that use sampling algorithms to obtain results. In many cases cognitive models are estimated in a Bayesian framework due to the flexibility of specifying models. The main advantage of these models 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 with this vague definition, however many times these arbitrary definitions do add value in communicating what general framework we are working in and thereby what methods are used.

## *Model descriptions*

For the sake of this thesis the psychometric function (PF), will be investigated as this has been a stable corner stone in the cognitive science literature across different subfields (REFS). The psychometric function is a continuous function that maps real or positive inputs into probabilities, i.e. the domain is whereas the range is . In most cases the PF used is like a logistic regression in statistical modeling and is commonly used in perceptual research where the inputs are stimulus intensities, and the probabilities are then converted into binary forced choices through a Bernoulli or binomial distribution. The mapping of inputs to probabilities is usually done through a cumulative density function such as the cumulative logistic or normal distribution, which amounts to conducting a logistic or probit regression in the statistical framework. The main difference between the statistical and cognitive framework of the PF is the number of parameters. The least number of parameters used to describe the PF is 2 the threshold and the slope (,). These two parameters describe the center of the curve, with 𝛼 being the intensity of the stimulus at probability 0.5 and 𝛽 being the steepness of the function around this value. In the cognitive modeling framework one or two more parameters are typically introduced the lapse and guess rates (, ). These two parameters together handle the tails of the psychometric functions and essentially makes the probability in the two ends of the psychometric no deterministic i.e. the upper and lower bounds become 𝛾 and 𝜆 instead of 0 and 1. These parameters help with fitting the PF to data where sometimes attentional slips or wrong button presses happen and it can 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 (REF). This also makes intuitive sense as the function cannot predict deterministic (i.e. probabilities of 0 or 1) if there are 0 responses at a high stimulus level which was caused by a lapse. Figure XX depicts how all these parameters change the shape of the PF as well as how including the lapse and guess rates help estimate the slope of the PF. For the sake of this thesis, I’ll be using the cumulative normal distribution to map stimulus values to probabilities with a single lapse rate. This single 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. This mathematical formulation of the function is as follows:



**Figure 4 Psychometric parameters.** Displays how the parameters alpha (), beta () and lambda () of the psychometric fucntion changes 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 asympotes in extreme stimulus (x) values.

## *Model validation.*

This section will highlight the main ways in which computational models in the literature are being tested and validated and takes offset in the seminal paper from Wilson and Collins ([Wilson & Collins, 2019](#ref-wilson_ten_2019)) describing 10 simple rules of computational modeling, which is commonly cited when validation of computational models is described in experimental papers.

When conducting statistical analyses using statistical or cognitive modeling we estimate parameters. Mostly these parameters are what we are interested in, maybe a slope from a regression model, the threshold of a psychometric function or a cognitive interpretable parameter in a cognitive model like the learning rate of a reinforcement learning model. These parameters are important to scientists as these are the things we usually make our inference on. There are at least three main challenges in this scientific endeavor. 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 recoverability)? And lastly how do we know that we can distinguish between competing models (external recoverability) The answer to these challenges must be found in simulations when our models become more and more complex. This simulation practice revolves around selecting an appropriate range of parameter and using these to simulate behavioral data from our models and then refitting the behavioral data to then see how well the model approximates the simulated parameter values. It should therefore come as no surprise that ensuring that in these simulations, we would like our models to perform well, such that we can have faith in them when the real underlying process is unknown, i.e. analyzing real world data. An appropriate range of parameter values for a particular model can be difficult to select as is exactly the problem of identifiability. However, 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 psychometric function this amounts to ensuring that the slope is strictly positive as this ensures that increasing levels of stimuli (when x- increases) will produce greater probabilities of responding 1, but also ensure that the standard deviation of the underlying probability density function is strictly positive. The lapse rate of the psychometric will be constrained between 0 and 0.5 to again ensure that the shape of the psychometric as values 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 psychometric, as negative slope values will. Not containing the PF in this way could lead to two distinct solutions to a given problem as negative slope values and lapse rates above 0.5 would be able to produce the same mathematical transformation of stimulus values to probabilities making the solution non unique (figure Y).

From a more theoretical level an appropriate range of parameter values can be narrowed down by looking at the function of interest and investigating whether the observed behavior (given the parameter values) is physically or biologically plausible and which values we would expect are more frequent. For the PF we might expect a few of our participants to not be particularly interested in the task and therefore just respond at random, which would amount to having a lapse rate of 0.5, however this behavior is quite unlikely and expecting only few lapses in the experiment, given that it’s conducted in a quiet environment is likely. Lastly using empirical knowledge from the literature at large helps narrow the parameter space further. For the sake of argument, we’ll investigate whether we might be interested in the detection threshold for cold stimulation to the skin. Just given this information alone we can narrow down the threshold for the cold detection to being below the skin temperature of around 30-34 degrees ([Courtin et al., 2023](#ref-courtin_spatial_2023)) and -273 degrees, however common knowledge, but also the scientific literature would suggest that thresholds between 28 and 33 would capture most of the population. These same arguments would apply for the slope, but also in other domains of inquiry. This practice of investigating the assumptions of the used parameter values is closely related to those of prior predictive checks when doing Bayesian inference. Prior predictive checks serve as a check of the model, without having seen any data. This check also revolves around simulating data from just the priors of the model and then investigating whether these conform with both what is physically and theoretically plausible, but also serves as a tool to investigate that the model can capture the behavior that is expected from the given experiment.

The next challenge is about internal recoverability i.e. can our model estimate the exact parameter values that was used to simulate the behavioral data that the model estimates the parameter values on. To test and validate our models of recoverability we simulate behavioral data from pre-specified parameter values which have been deemed to be appropriate using the first step. We then feed our models with this behavioral data and investigate how well the model can estimate the latent simulated parameters. This exercise of simulating behavior and then re-estimating the parameter values from the simulated behavior is commonly known as parameter recovery and if this procedure succeeds, then the parameters are said to be recovered. The satisfactory criterion often refers to some correlation coefficient, between the estimated and simulated parameter values. 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 real world data. This is because if we had known the parameters values beforehand (i.e. simulated them) then we know that they are somewhat close to the estimated parameter values we got from fitting our model to the data. The assumption that if our model recovers the parameter values well in a simulated setting then it must also do so when fitted to real world data where the underlying parameters are unknown, This assumption is of cause not necessarily true and rests on axillary assumptions. These axillary assumptions are grounded in that the underlying generative cognitive model is the same or at least close to the same as the one used to model the data. Because the process of doing parameter recovery assumes that we know the underlying generative model, which is not the case when fitting real world data. To further elucidate this point we imagine using the 3 parameter PF described above, we find that it recovers its parameters well using simulated parameters from the same model. However, if we instead of simulating data from the same underlying model, instead simulated data where the underlying cumulative distribution was the logistic or the hyperbolic secant, we might find that our model cannot well recover the parameters. This is of cause nonsensical from the beginning, as how might our model recover parameters of another model, but many times the differences in our model space (i.e. the models that we think underlie the generative process) are similar and the parameters have similar meanings as they come from the same or similar underlying theory, meaning that they can be compared. This last 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 also 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 and being able to compare them, while ensuring that all these models are distinguishable. What is therefore commonly done is using the theoretical framework(s) to build competing models which contain different assumptions of the underlying generative process and then comparing this subset of the entire model space, as these are the models that our theories deem relevant. This highlights two important aspects, firstly our models reflect our theories and are therefore at best as good as our theories and secondly, we are surely missing the real generative model in most cases, but as our theories improve and therefore also our models the assumption is that our models will converge towards the real generative model, entailing that our theories move towards the truth. In practice what is commonly done is that models are fit to real world data and then compared on how well they can describe the data using statistical metrics such as information criteria. The problem with this approach is that how do we know that we can accurately distinguish the model space that we are testing. This challenge has been addressed using model recovery, which is the act of simulating data from the entire model space and then refitting all generative models to the entire model space. Going back 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, further we want to ensure that we can distinguish between the normal and logistic cumulative distributions which transform stimulus values into probabilities in different fashions. In this practical example the model space consists of 4 models i.e. two or three parameters for each of the two types of PFs. One would therefore simulate data from these 4 distinct models and fit them all individually to each of the 4 simulated datasets and lastly determine which of the 4 models describe the data the best in each case. The result of such model recovery is a N times N matrix with N being the number of models, the rows being which model was used for the simulation and columns being which model was used to fit the data. The entries of the matrix are commonly depicted as the probability of choosing a particular model given the data simulating model. An identity matrix therefore represents that the models are completely distinguishable and anything else would indicate that in some of the simulations the best fitting model was not the model that simulated the data.

## *Limitations of current model validation steps*

The model validation steps above should ideally serve to increase our faith in our models, their parameters, and the comparison between them. However, the metrics used to access these different types of validations are flawed. Firstly, the metrics used can be easily manipulated (faithfully or not) to show good model validation when masking the actual poor or terrible validation. This problem can thus introduce false faith in the model and overconfidence in the inference made based on it. Next the metrics used are not sensitive or specific enough to give the modeler information about where models for instance are similar or where they break down, thereby leaving valuable insights hidden. In this section I will highlight the metrics commonly used in the literature for model validation which are described in ([Wilson & Collins, 2019](#ref-wilson_ten_2019)), focusing on two of the challenges described above; internal recoverability and external recoverability.

As mentioned above internal recoverability of computational models are accessed with parameter recovery, where behavioral data is simulated from a model given a set of parameters. This behavioral data is then fitted to the model which then optimizes for the parameters given the data. What is commonly done is then estimating the correlation coefficient between the estimated and simulated parameters. In their seminal paper (REF) describes that in a perfect world the estimated and simulated parameters should be tightly correlated without any bias, and that a weak correlation could mean bugs in the code or an underpowered study. They also reiterate that plotting simulated vs estimated parameters should be done to access if ranges of parameter values are problematic and whether there might be biases. I will here argue that the correlation coefficient is an inappropriate metric and that a version of an intra class correlation (ICC) is better suited for the task. Acknowledging two important things; neither metric is perfect, and visually inspecting the simulated vs estimated parameter scatterplot is crucial. The importance in using the right metric is therefore as a precautionary step given that some literatures are starting to just report correlation coefficients without this crucial scatter plot, which I’ll argue then in some cases would make the correlation coefficient meaningless. External recoverability or model recovery is highly dependent on the range of parameter values used for simulating the behavioral data and the metric used for accessing the best fitting model, this is also stated in the paper by (REF). For this thesis I will not argue for which metric to use as with parameter recovery as this will be highly dependent on the specific case, but instead highlight some missed opportunities in this step of the model validation phase. This missed opportunity is to better understand the model space and therefore where one might put the emphasis on a particular task. (expand or not?) These precautionary steps are crucial to enforce, in the development stages of new statistical models as they will serve the basis of model validation and if not sensitive or specific enough many resources might be used in using a model that in reality cannot be properly identified. This would therefore serve as a roadblock for scientific progress as years might pass before someone realizes that the model used in the field is not behaving properly. From a philosophy of science perspective this amounts to ensuring that the axillary assumptions that our current investigations rest on, i.e. the models that we use to test our hypotheses, are valid.

## *Current problems with internal recoverability of models (parameter recovery)*

The first and perhaps biggest problem of internal recoverability of computational models Is that it is not universally done, which from a readers perspective 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 that interactions between parameters are either neglected or disregarded. This is less of a concern for individuals using an established cognitive model wanting to ensure that given their experimental design and ranges of parameters are sensible, but a big concern in the highly cited method papers describing and formalizing the models. A prime example of this is the Hierarchical Gaussian filter paper (REF) where after having laid out the equations of model 2 of the most crucial parameters of the model are held constant when performing parameter recovery. Even in much more simple models such as with the PF described above, I will show that there are tradeoffs and interchangeability between parameters. The last problem with parameter recovery is the metric used to access it. As has been suggested elsewhere, the correlational approach to parameter recovery is at best insufficient and at worst misleading. The three most obvious problems with using correlations are (REF).

Correlation coefficients are invariant to linear transformations, making two sets of variables i.e. [1,2,3] and [1,2,3] have the same correlation after transforming on one of the sets with linear transformation. y=2\*x+3 (or report as a matrix idk) Resulting in the sets [1,2,3] and [5,7,9]. This invariance to linear transformations does not make sense for parameter recovery as we want a metric that penalizes this behavior. The domain of correlations is between -1 and 1. This directionality also does not make sense given that a correlation coefficient of -1 would mean perfect parameter recovery, with a negative sign of the simulated or estimated parameter meaning that you do recover the value (or the linear transformed value) just not the sign. Ideally, we would want a metric that goes from no recovery to perfect recovery.

Lastly, the interpretation of the correlation coefficient in terms of parameter recovery is difficult. What is a sufficiently large correlation coefficient for the parameter to be said to be recovered and what types of uncertainty is causing the correlation to be less than ideal. All these issues are similar to what researchers face when wanting to estimate the stability and or test -retest reliability of different metrics over time, where the solution has been to use the intra class correlation (ICC) as the metric instead of simple correlation coefficients.

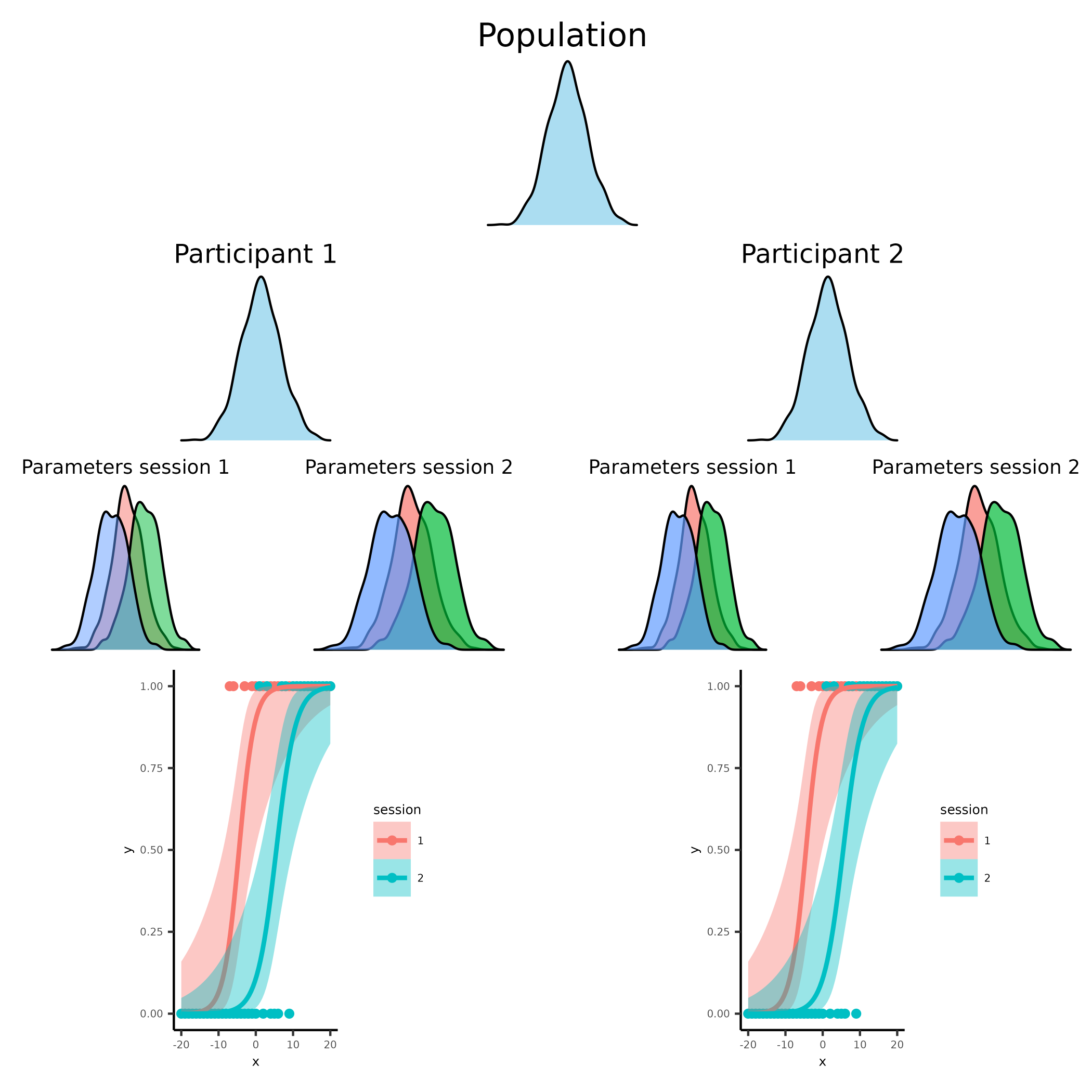
## *ICC Parameter recovery*

Given that the idea of using the ICC as a metric for parameter recovery is relatively new and to the authors knowledge has only been suggested and not been used anywhere in the literature (REF) I will here outline what the ICC is and how it can overcome some of the shortcomings of the correlation coefficient. The ICC is in its simplest form a ratio of irreducible variances (uncertainties) to the total variance in the data. In practical terms the irreducible uncertainty is the uncertainty between subjects and whereas the total uncertainty can have several parts. In order to calculate the ICC one needs a model that can properly model and account for these different types of variance and the typical approach are hierarchical models, where known structure of the data is added to the model.

Taking an example, we imagine a researcher doing a test-retest reliability study on a parameter of a cognitive model. His subjects are coming in for x sessions and doing the same cognitive task each time. We will now assume that all subjects come from the same underlying distribution of say humans (i.e. the population), this is the highest level in the hierarchy and is governed by a population mean and a population variance, i.e. the between subject variance. The next level in the hierarchy is the subject level, here each subject has their own means and variances (within subject variances), their means are drawn from the population distribution. Now for each session that the subject is in a parameter value is drawn from this subject level distribution which then governs the participants’ behavioral responses. This nested hierarchical structure is demonstrated in figure Y, as can be seen each of the levels are governed by the levels above and each of the levels has variance associated with it, where the between subject variance is the variance of the top 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 within and between subject variances.

Where is the variance between the subjects’ parameter estimates and is the within subject variance. Given that we are interested in the performance of the model we can simulate agents that have no within subject variance i.e. the same true parameter values for each session and then see how the number of subjects and or trials of the cognitive task will influence the model’s ability to pick up this association. This approach has one clear problem it does not necessarily tell us something about how well the model estimates the true parameter values for each participant at each session, as it just looks at how close each parameter is to itself between sessions. To capture this, one might use the mean squared errors (MSE) between the simulated and estimated parameter values, which serves as a residual error of the model. Including this into the ICC formulation is easy as this is just another source of variance which can be added into the denominator, highlighting the fact that the ICC is a partitioning of variance in the model. This partitioning of variance is exactly what we are interested in when building models and validating them, as this tells us where the model fails and where it might excel. Formally we add the MSE into the equation and get.

Where is the MSE. This conceptualization allows us to put parameter recovery for a model into a single value for each parameter that ranges from 0 to 1 which is going to be trial and subject level dependent, but also dependent on the simulated ranges of parameter values.

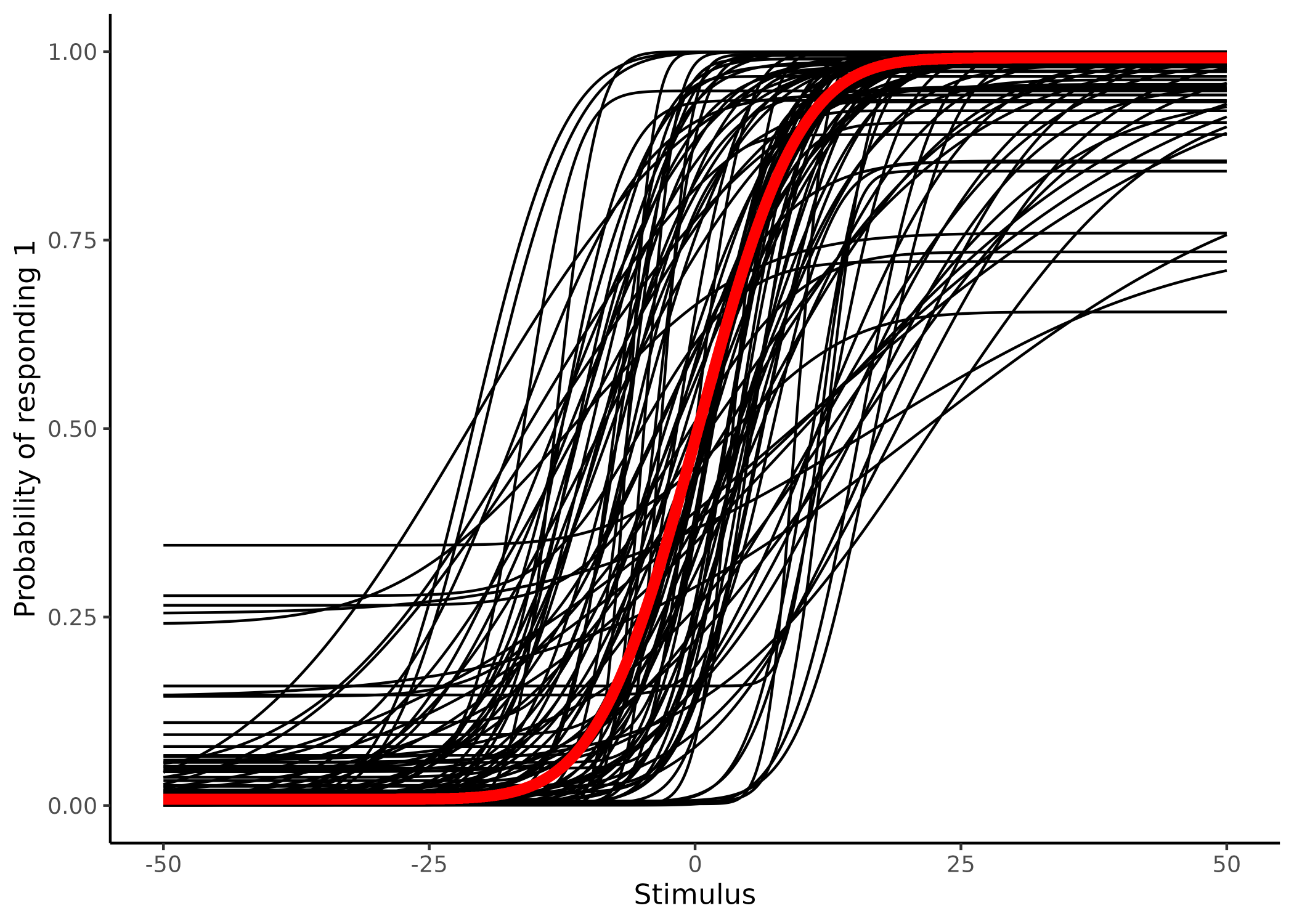


**Figure 5. Visualization of the nested hierarchical model.**

## *Standard parameter recovery.*

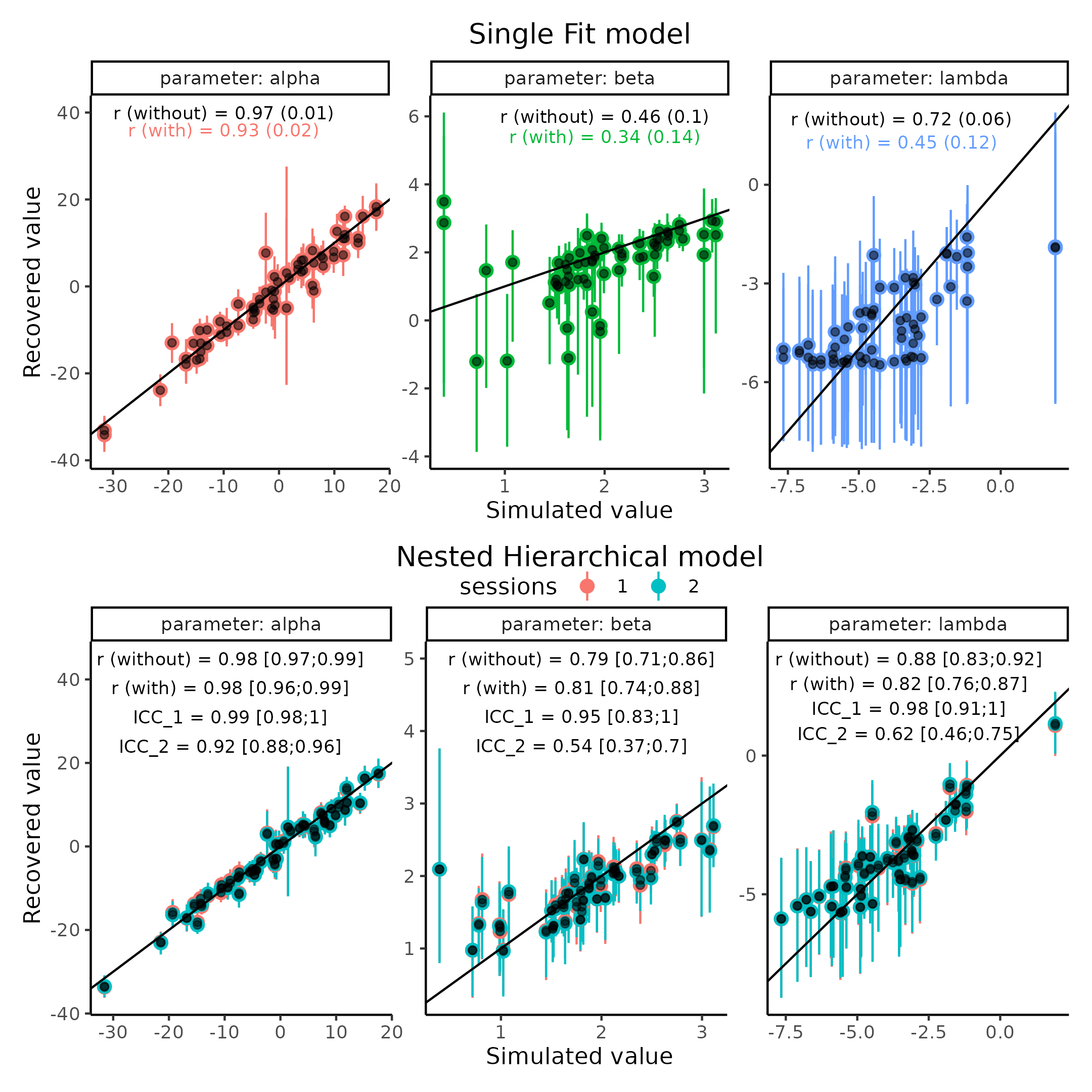
The model and task used to demonstrate this is going to be the 3 parameter PF described above which is widely used in the cognitive science literature from XX to YY to ZZ. (REFS). After having specified the model, we can simulate behavioral data from different ranges of parameters to select appropriate ranges of parameter values. Firstly, I’ll be selecting the parameter range shown in table Y and figure XXA to simulate from to show the common way of doing parameter recovery. Using the probabilistic programming language Stan and its interface with R, Rstan we can now invert the behavioral data from the simulated parameters to obtain estimates of these parameters.

| Parameter | Mean | Sd | Transformation |
| --- | --- | --- | --- |
| Alpha | 0 | 10.0 | x |
| Beta | 2 | 0.6 | Log(x) |
| Lapse | -4 | 2.0 | Logit^-1(x) / 2 |



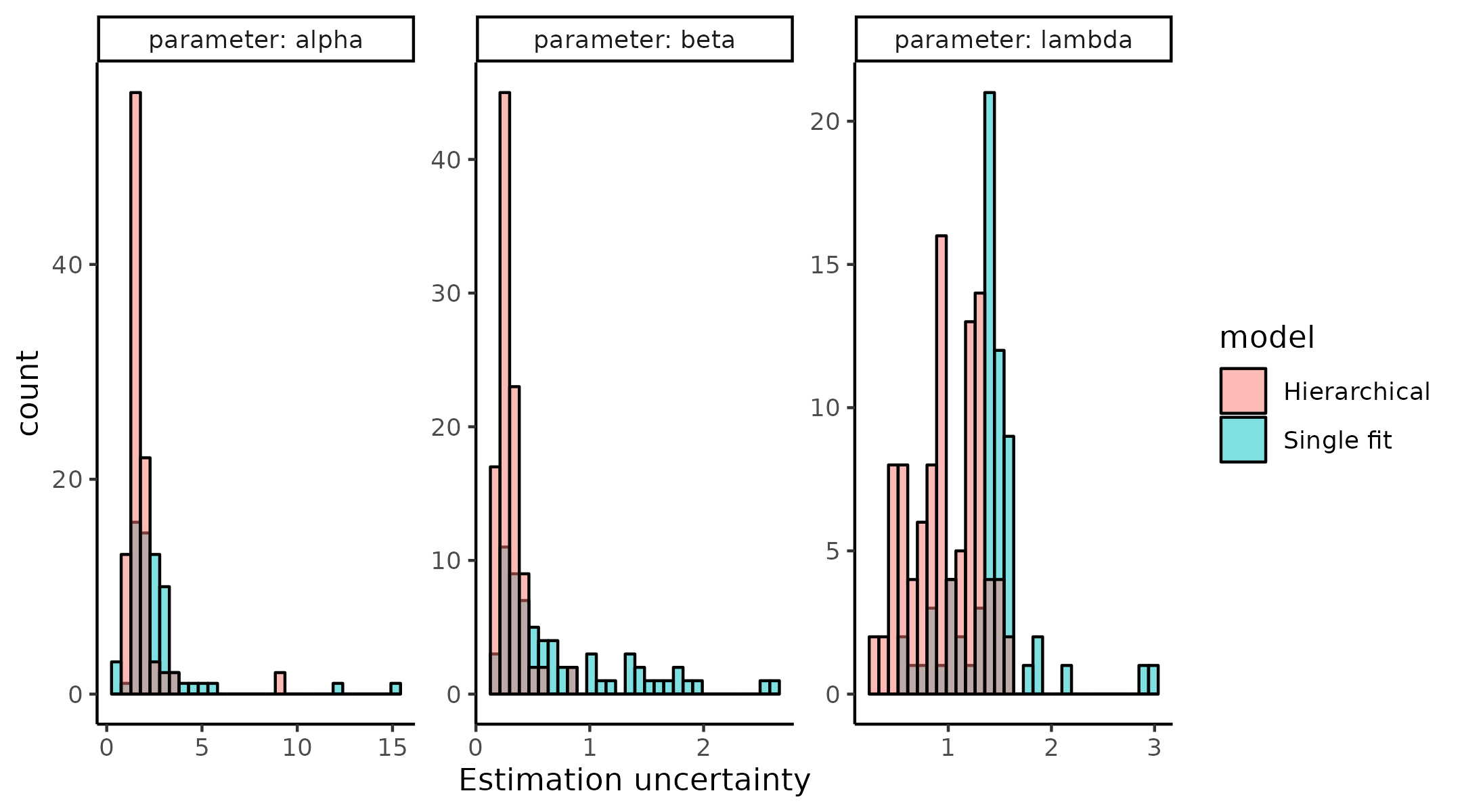
**Figure 6. Displaying 100 samples of the parameters of the psychometric function from table1.** Visualization of the implications of the simulated parameters of table1. Black lines depicting individual subjects, while the red line depicts the group mean.

For the sake of argument, I have plotted the pairwise scatter plot of estimated vs simulated parameter values in figure F this is of course with the added estimation uncertainty, which is not normally done in the literature. The simulation is done with 100 data sets.



**Figure 7. Parameter recovery for the three parameters of the psychometric function.** Scatter plot of Simulated vs recovered parameter values, with error bars displaying the 95 highest density interval for that parameter on that simulation.

From the figure adding the estimation uncertainty to the correlation coefficient paints a better picture of what is actually going on, and in this case, it reduces the correlation coefficient on all parameters. This is however not necessarily always the case, as if a couple of points fall way off the identity line with high uncertainties, they will have less weight when accessed with uncertainty compared to without. What this recovery analysis, even when including the estimation uncertainty does not account for is the hierarchical structure that is becoming standard practice in most cognitive science experiments (REF). Therefore, the model fit to the real data is different from the model tested. To test and compare the nested hierarchical parameter recovery and the ICC to the correlation approach, I first simulate agents according to table 1, however this time only simulating 50 agents but duplicating them, such that 100 sets of parameter values are again simulated. The nested hierarchical model is then fit to this set of 50 agents and the correlation with and without estimation uncertainty and the ICC as described in equation 1 and 2 now referred to as and are calculated. As an additional step each of the 100 agents generated were also fit to the single subject level model, which makes the comparison of correlation coefficients across the nested hierarchical and individual model clear.



**Figure 8 Estimation uncertainty for each parameter for both single and hierarchical fit models** (is it simulated vs reocvered or oppisite?)Each panel represents one of the three parameters of the psychometric function.

As can be seen this hierarchical fit does improve the parameter recovery, both from a visual inspection and the correlation estimates, with and without accounting for estimation uncertainty. It should here be noted that a single simulation like this wouldn’t be enough to ensure that the parameters are nicely recovered as a good example of this is the lambda parameter. The pairwise scatter plot of the nested hierarchical model seems to suggest that this parameter is quite well recovered but, if we back calculate a lambda value of -5 corresponds to a lapse rate of 1.3%, which obviously is hard to find when there are 100 trials for each subject, see supplementary material XYZ for a deeper explanation and similar runs of the same model showing way worse recovery of this exact parameter. (supplementary it is even lower than the 0.7% for a lapse of -5 as the stimulus value also needs to be far enough away from the underlying psychometric to actually matter in the estimation). Turning the attention to the ICC values, we firstly observe that on each of the 3 parameters has can upper bound at the maximum value of 1, which is somewhat confirmed looked at the scatter plot as one would assume that all the session one estimates are hidden behind the session two estimates with only a few estimates deviating slightly. The estimate distinguishes the 3 parameters the best, of the 4 different metrics, which conforms with the visual inspection of the scatter plots. An interesting observation in Figure XXY is that the difference between lambda and beta is minute in all the metrics used. In the next section I’ll show how we can do better as this is a result of quite high estimation uncertainty present in the estimates of beta.

## *We can do better!*

A couple of import bits of information have been left out in the parameter recovery analysis described above, including the priors of the Bayesian model, estimation, and handling of convergence for the model, but most importantly, what is the design of the experiment that the simulated agent goes through? Looking back at figure XXA 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 for most agents give next to no information on the shape of the psychometric and therefore the parameters we mostly care about i.e. alpha and beta, as on average the agents’ psychometric functions are monotonically increasing in the interval of [-25 ; 25]. Therefore, selecting stimuli (inputs) in this interval must be better for decreasing the estimation uncertainty in the two parameters we care about, compared to randomly or uniformly exploring the input space. We might even go a step further and instead of selecting inputs that are more appropriate for the mean of the population we could individualize each experiment to the agent or subject. This practice of individualizing the experiment of interest is called adaptive design optimization (ADO) and has quite a big literature behind it and revolves around selecting inputs that are optimal given a specific criterion. Many of these criterions exists such as minimizing entropy, minimizing the posterior variance or mutual information, but what they all have in common is that they do decrease estimation uncertainty of either all or certain parameters to a meaningful degree (REFS). In order to keep in the same theme as the rest of the thesis I will instead of utilizing the few available packages that exist for doing ADO for psychometric functions I will show how utilizing the single fit model which was built for conducting the simple single subject parameter recovery can be utilized together with the knowledge that the most informative stimuli for determining the shape of a single agents’ psychometric function is somewhere in the middle region of that psychometric. One of the main challenges of utilizing ADO is that because the experiment is updated and individualized an algorithm determining the next stimulus must run in tandem with the experiment. This puts quite a high constraint on computation time of the algorithm, this issue has partly been solved in the existing packages by before conducting the experiment mapping out a grid at a particular resolution of parameter values at a current trial and then what the optimal stimulus value to present is. This clever solution puts the heavy computation time before the experiment and ensures that when the experiment is run only a single look up is needed to provide the next stimulus value. This approach works great for psychophysical experiments or other experiments where each trial is independent of the next (maybe expand idk). However, to provide something that is more generalizable and can be continuously updated on a trial-by-trial basis other approaches might be more appropriate. Illustrating such an approach can easily be done with the same model used to fit individual subjects when using the R and Stan, the quick estimation of the posterior distribution is then done using a variational inference algorithm, particularly pathfinder (REF). Figure 5 shows how the posterior distribution of the 3 parameters of the PF varies as a function of trials in both the uniform and pathfinder approach to selecting stimulus values. As can be seen both approaches makes the parameters converge towards the real simulated values (black line), however the speed at which this happens is clearly very different, especially for the two parameters we are the most interested in i.e. alpha and beta. For these two parameters after just 20 trials of pathfinder the optimization has found the simulated parameter value and decreased the estimation uncertainty (posterior variance) to close to 0 whereas even after 50 trials the uniform approach still has a bit of 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 (REF) was also used to compare to ensure that the pathfinder algorithm was not too slow or bad.

To show the improvement more rigorously in reduced estimation uncertainty especially across a range of trial numbers, the Pathfinder, Uniform and PSI algorithms were run 100 times for trials ranging from 20 to 100 in a sequence of 10 trials, in order to make the comparison as fair as possible each of the algorithms were only used to generate the stimulus sequence, meaning that all three types were refitted using the same single fit Bayesian model. For complete details on the fitting and optimization strategy see supplementary material XX.

Using the newfound ADO we can now investigate the three remaining elephants in the room, subjects, trials and the influence on the mean simulated slope value. The last point is less obvious than the two others but stems from the fact that increasing the slope (decreasing the steepness) of the PF will make it harder to estimate, but also influence the recovery on the threshold, which will become clear below. For this purpose, I’ll simulate trials ranging from 20 to 200 in increments of 20, subjects being between 10, 30 and 50 and lastly mean slope values of 1,2 and 3 in the unconstrained space, 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 5 times. Figure XXX) shows how the correlation approach with added uncertainty to parameter recovery fairs (for the standard approach of no uncertainty see supplementary XXX). Figure YYY shows how the fairs on the same simulated datasets, (for the analysis with only the within subject variance see supplementary QQQ).

What seems to be the main difference between the two approaches is the in the lower number of trials and especially in the comparison between the high simulated slopes (lowest panel) for the threshold as both approaches seem to suggest that in high number of trials (> 100) and in steep slopes (beta >= 2) that the threshold is fully recovered. The difference is clearly in the lowest panel where the ICC approach suggest that there is still variance left unexplained, to investigate this we can plot the pairwise scatter plot of the high simulated slopes (beta = 3) and low simulated slopes (beta = 1) on different trials and subjects. Figure XYX clearly shows why there is such a difference between the two approaches, the ICC metric is much more stringent on the higher-level estimation uncertainty when the simulated slope is less steep. Turning the attention to the slope itself, there also seems to be a difference. What is present is again that the ICC metric has lower values in general and is not asymptotic at one with the configurations used here (to see the pairwise scatter plots see supplementary material XXX). Lastly both approaches suggest that the lapse rate is well below acceptable ranges, but still with the ICC being more conservative. As conveyed by the pairwise scatter plots the conservative ICC metric capture the fact that estimation uncertainty is a source of variability that can still be reduced even when the correlation coefficient (also with the estimation uncertainty propagated) might indicate perfect fit. This is exactly the behavior one would like to have when trying to understand their model as this information is much more sensitive, furthermore the values also have a natural interpretation. An ICC value of 0.8 means that 80% of the variance in the model is governed by the between subject level and only 20 % is in the estimation or test -retest uncertainty, the ICC could of cause be further decomposed into what proportion of variance of the 20% is from estimation and what is from test retest uncertainty, however for this particular model it seems like most if not all is from estimation uncertainty (see supplementary XYXX). This straightforward interpretation is not present for the correlation coefficient, especially because of the arguments laid forth in the “current problems with internal recovery” section.

Many if not all papers describing the test retest reliability of cognitive models in the literature finds that hierarchical models are better but also argue some something around the correlation coefficient, here the ICC metric is again really helpful as the estimation uncertainty is going to be soaked up somewhere in the model and that is going to go towards the within subject variance or at least increase it such that in reality the ICC might have been higher than observed but because you did not have enough trials the uncertainty around the estimate i.e. the estimation uncertainty is what is causing the low test retest reliability.

## **Can we do better?**

Now I’ve hopefully convinced that the ICC approach to parameter recovery is superior to both the standard and the uncertainty propagated correlational approach in that it better reflects our expectations given the pairwise scatter plots of simulated vs estimated parameter values. The question now becomes given our more nuanced view of parameter recovery what we can do to improve it. The obvious answer given the plots above seems to be increasing the number of trials, as the number of subjects does not seem to influence the estimates to a meaningful degree. However mindlessly increasing trials to gain a certain recovery and or statistical power can be troublesome in non-obvious ways. The obvious problems with increasing the trials number are resources costs, both in terms of money to the participants completing the experiment, the experimenter, but also the time investment. However, the most problematic aspect becomes more obvious if we take a step back and think carefully about what we are studying. We are studying a complex system that has its own goals, desires and motivations and it is not trivial to how this participant will behave if the task is double the length. Firstly, will the participant employ a different strategy knowing that the experiment is going to take X time longer, or will they halfway through employ a different strategy. Even if the participant keeps the same underlying cognitive strategy that we are trying to model, then one reasonable assumption would be that attentional lapses and engagement in the task will decreasing, making each additional trial after a certain point less “valuable”. I will here argue that in many of the cognitive science paradigms there might be no need for increasing trial counts to increase the recovery of parameters, but to utilize the data the participant has already provided in better and more sophisticated ways. For the sake of this thesis, I will look at incorporating the reaction times of the agents’ responses as sources of information about the underlying psychometric function of their binary choices. I’ll be focusing on the reaction times as these have a long and rigorous history in cognitive science literature, but more importantly are present in most experiments conducted today (REF).

<https://www.jstor.org/stable/27828738> <https://psycnet.apa.org/record/1986-00316-001> <https://psycnet.apa.org/record/1989-07414-001>

To show how these reaction times help the recovery of the parameters of interest i.e. the threshold and slope of the psychometric, I’ve chosen to simulate agents with the parameter values displayed in table 2. To understand the influence of the size of coupling between the binary responses and the reaction times I’ve chosen to simulate two different means for this coupling parameter i.e. 1 and 1.5. Again, showing and understanding what these parameter values mean we simulate the parameters and display the behavior. This can be seen in figure XXX and YYY for each of the two-coupling means for 20 simulated subjects when the psychometric slope is 3, for the visualization of what happens with a steeper slope i.e. beta = 1 see supplementary XXX:.

With these simulations we can now visualize what including the reaction times into the modeling means for the parameter recovery for the influence on the other metrics like the correlation coefficient see supplementary material XXX. Figure XXX and YYY displays the results of this analysis the first plot showing the means and 95% confidence intervals of the ICC\_1 for the stimulations for the three parameters of the PF together.

## *Real data!*

Having rigorously investigated how the psychometric function behaves and how the certainty of the parameters interacts with each other but also with the number of trials for each subject, one can now turn to real data. I’ll in this section introduce the published dataset that I will re-analysis utilizing the psychometric functions introduced above. The goal with this re-analysis is 2-fold. Firstly, it reiterates the fact that the assumptions about the structure of the data can make quite a difference in the parameter estimates and their uncertainty. Secondly, it will serve as a starting point to understand why the internal model validity using the ICC can be helpful as a metric to gauge how trials and subjects interact on the statistical power of a model to reject a hypothesis. This last aspect of testing hypotheses will tie together how these validity steps help determine the ability of a model to do what researchers are many times interested in i.e. hypothesis testing. The last point of the thesis is going revolve around conducting a thorough power analysis of the current model, utilizing the published dataset described below here I will compare the ICC metric for the model to its ability to reject hypotheses at certain trial and subject numbers. In this regard of conducting a power analysis I will again highlight where uncertainty creeps in and how we can deal with and account for these, as common practices are insufficient.

## *Heart rate discrimination task*

The article where the dataset was published is (REF) and is an interoceptive task. Here the authors collected 223 participants who came in twice to complete a heart rate discrimination (HRD) task within 6 weeks between visits. The HRD task is comprised of two distinct tasks, a comparison and an interoceptive task. Here I’ll focus on the Interceptive task where participants were asked to internalize their own heart rate for 5 seconds. Meanwhile the participant attends to their own heart rate, the heart rate is monitored and calculated in real time. Next based on the observed heart rate participants will hear five auditory tones in a 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 frequency was faster or slower was determined by the PSI procedure introduced in the Adaptive design optimizing paragraph. This means that the stimulus value for the psychometric function is the difference between the external tones frequency and the observed heart rate of the participant in the current trial and the responses are given by faster or slower with faster being coded as 1 and slower being coded as 0. This means that a participant might have a heart rate of 50beats per minute (BPM) at a particular trial and then hear tones in a frequency of 40 BPM and are asked to respond whether they think this 40BPM is slower or faster than their own heart rate. The authors of the experiment, described above, ended up running single participant level models of each subject, for each session, and then correlating the slope and threshold of the psychometric function. They found a medium correlation between the threshold r = 0.51 p < .001 between sessions and a negligible correlation r = 0.1, p = .15 for the slope. In the next section I will show how this reliability might change given different model assumptions and different models, to demonstrate that thinking hard about what model is fitted is worthwhile.

## *The models*

In this section I will describe the models that I’ll fit to this big test-retest dataset to examine the influence of the model fit on the correlation between session one and two. The single fit model is going to be the references and going to be the same as the original authors did. That is estimating each individual for each of the sessions individually without a lapse rate (i.e. a two parameter psychometric function) and then post hoc correlating the estimates between session one and two. I will add the propagated uncertainty to these estimates as they do not seem to be adjusted by the authors. Next, I’ll investigate the same model as above but adding the third lapse parameter. The hierarchical model is going to model the two sessions from the same multivariate normal distribution. This model directly models the correlation between sessions as its included in the variance - covariance matrix of the multivariate normal distribution. The last type of model is the nested hierarchical model, this model assumes that all subjects have a mean level parameter which is drawn from the same multivariate distribution, then each parameter for each session is then drawn from a subject level distribution, identical to the model presented in the ICC parameter recovery section. For this last model the ICC is the statistical metric estimated by the model itself, and the correlation will afterwards be calculated. Additionally, each of these models will be fitted using the reaction times as described in the XXXX section to investigate the influence of adding this additional information. A final full model is going to be fit utilizing even more information already available in the dataset. This model will not only incorporate the reaction times on a trial-by-trial basis, but also the confidence ratings for each trial. These confidence ratings were included in the task of the original experiment to examine the participants’ interoceptive metacognitive abilities. These confidence ratings are going to be modelled in close resemblance to the reaction times, just inverted. This inversion is because in the middle of the psychometric function the uncertainty about the stimulus representation is the highest and therefore reaction times should be their highest as well, but confidence should be at the lowest. Another difference between the reaction times and the confidence ratings is their range of possible values. Confidence ratings were bounded between 0 and 100 indicating complete uncertainty and certainty respectively. A natural likelihood function for such kind of double bounded variables is the beta distribution as its already bounded between 0 and 1. The only problem with using this likelihood function is the edge cases of 0 and 1’s which for the confidence ratings are 0 and 100. One approach is to model these edge values separately using a zero-one-inflated beta distribution. This approach, however, models these edge values as separate processes which does not make sense in this case as the confidence ratings are meant to represent a continuous measure of confidence. I will therefore here subtract 0.1 from the 100 ratings and add 0.1 to the 0 ratings making it possible to use the beta distribution for the full range of confidence ratings. This approach of modeling the bounded ratings between 0 and 100 is tenuous and new mixture methods are slowly being developed for a more holistic approach see

<https://www.cambridge.org/core/journals/political-analysis/article/abs/ordered-beta-regression-a-parsimonious-wellfitting-model-for-continuous-data-with-lower-and-upper-bounds/89F4141DA16D4FC217809B5EB45EEE83>

Reaction times of the responses were at maximum 8 seconds and can therefore still be modeled by the shifted lognormal distribution introduced above.

## *Results*

Table 3 displays the correlation coefficient between the first and second session for the threshold and slope for each model when uncertainty has been propagated using bootstrapping. For a full table of all parameters of all models as well as with and without uncertainty propagation see supplementary table XYX

| alpha | beta | lapse | model | structure |
| --- | --- | --- | --- | --- |
| 0.48 [0.34 ; 0.60] | -0.01 [-0.06 ; 0.05] | FALSE | Binary | Single |
| 0.50 [0.39 ; 0.62] | 0.01 [-0.06 ; 0.10] | TRUE | Binary | Single |
| 0.52 [0.39 ; 0.64] | -0.00 [-0.10 ; 0.06] | FALSE | RT | Single |
| 0.52 [0.39 ; 0.65] | 0.01 [-0.11 ; 0.10] | TRUE | RT | Single |
| 0.52 [0.41 ; 0.64] | 0.00 [-0.14 ; 0.17] | FALSE | RT+Conf | Single |
| 0.52 [0.40 ; 0.62] | 0.03 [-0.13 ; 0.22] | TRUE | RT+Conf | Single |
| 0.51 [0.41 ; 0.59] | 0.23 [0.01 ; 0.45] | TRUE | Binary | Hierarchical |
| 0.49 [0.39 ; 0.58] | 0.26 [0.09 ; 0.41] | TRUE | RT | Hierarchical |
| 0.48 [0.39 ; 0.57] | 0.21 [0.06 ; 0.36] | TRUE | RT+Conf | Hierarchical |
| 0.54 [0.49 ; 0.58] | 0.19 [-0.01 ; 0.42] | TRUE | Binary | Nested Hierarchical |
| 0.54 [0.51 ; 0.57] | 0.27 [0.07 ; 0.45] | TRUE | RT | Nested Hierarchical |
| 0.55 [0.53 ; 0.58] | 0.20 [0.06 ; 0.37] | TRUE | RT+Conf | Nested Hierarchical |

Table 1 clearly highlights the fact that the additional assumptions of the hierarchical models both nested and unnested increases the session-by-session correlation of the slope of the psychometric function, and that additionally including the reactions time increase the correlation even more. 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 in this regard. A concern of this approach of just looking the correlation coefficients is of cause that a model with a high session by session correlation might not fit the data the best, as the latent underlying stability i.e. correlation might be 0. One approach would therefore be to examine model fit using common metrics such as cross validation, information criterion etc. This difficulty here is that most of the models are incompatible; because they have been 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 architecture. The only models being compatible for comparison are the two hierarchical fit models with the same model architecture, resulting in a very limited comparison.

## *Power analysis*

As a last step in this model building, formulating, and testing framework we can easily and accurately inform future work. When researchers are interested in the parameter values of their models, they are many times, especially in computational psychiatry, also interested in how they differ using pharmacological interventions or between healthy controls and patient populations. The question in such a scenario is how many participants and or trials do you need to reliably detect a particular size of effect, between the two conditions? In the next section I outline the idea of power analyses and how it can be conducted within the framework of cognitive modeling. To do so I will utilize the results from the above analysis of the test retest dataset of (REF).

<https://www.sciencedirect.com/science/article/pii/S0301051121002325?via%3Dihub>

Given the due diligence of the model validation steps it will be easy for the person using the model to get an accurate estimate of what effect size can be detected with differing amounts of trials and subjects. For the current analysis I’ll be using the group level estimates from above here focusing on the simplest model for computational efficiency. Before conducting this power analysis a few details about power and power analyses should be explicitly highlighted.

## *What is Power analysis*

The idea about power analysis is simple; we want to calculate a priori to conducting our experiment, the probability that our results are going to be “significant” given that there is some “real” underlying effect. Usually this is depicted in a 2 by 2 matrix with the real latent effect being in one dimension and the model results in the other dimension. The probabilities of landing in either of these 4 categories is usually described as functions of our statistical significance threshold (alpha / p-value) and the power (1-power).

| alpha | beta | lapse | model | structure |
| --- | --- | --- | --- | --- |
| 0.48 [0.34 ; 0.60] | -0.01 [-0.06 ; 0.05] | FALSE | Binary | Single |
| 0.50 [0.39 ; 0.62] | 0.01 [-0.06 ; 0.10] | TRUE | Binary | Single |
| 0.52 [0.39 ; 0.64] | -0.00 [-0.10 ; 0.06] | FALSE | RT | Single |
| 0.52 [0.39 ; 0.65] | 0.01 [-0.11 ; 0.10] | TRUE | RT | Single |
| 0.52 [0.41 ; 0.64] | 0.00 [-0.14 ; 0.17] | FALSE | RT+Conf | Single |
| 0.52 [0.40 ; 0.62] | 0.03 [-0.13 ; 0.22] | TRUE | RT+Conf | Single |
| 0.51 [0.41 ; 0.59] | 0.23 [0.01 ; 0.45] | TRUE | Binary | Hierarchical |
| 0.49 [0.39 ; 0.58] | 0.26 [0.09 ; 0.41] | TRUE | RT | Hierarchical |
| 0.48 [0.39 ; 0.57] | 0.21 [0.06 ; 0.36] | TRUE | RT+Conf | Hierarchical |
| 0.54 [0.49 ; 0.58] | 0.19 [-0.01 ; 0.42] | TRUE | Binary | Nested Hierarchical |
| 0.54 [0.51 ; 0.57] | 0.27 [0.07 ; 0.45] | TRUE | RT | Nested Hierarchical |
| 0.55 [0.53 ; 0.58] | 0.20 [0.06 ; 0.37] | TRUE | RT+Conf | Nested Hierarchical |

The latent or underlying effect might be our intervention or a difference between a healthy population and a patient population. The framing of power analyses is then to say that results are significant if the p-value is less than a particular value most often 5% and the probability that we detect the effect given that it is there is another arbitrary value with 80% being the standard. Moving to the more practical side of the power analysis our cognitive or statistical models will reject and fail to reject different rates of effects given magnitude of this effect. The table above is therefore quite misleading as in reality the dimension of reality is a continuous variable of size of the effect and our models have a particular probability of rejecting a hypothesis (given subjects and trials) at a particular effect size its tested on. An example of this might be that we want to detect whether there is an effect of gender on height in the human population. We assume that there is an underlying effect and observe X females and Y males and run a statistical analysis to determine whether we can reject the null hypothesis (there are no differences in height in the two genders). Compare this to the hypothesis that there is an effect of age (late adolescents and adults) on height. The former difference might in general be much larger than the former and therefore with all else being equal (trials, subjects, statistical model etc.) this difference will be easier to detect compared to the other difference. What is therefore done when conducting power analyses is that different observed effect sizes are simulated in differing number of trials and subjects and the ability of the statistical model to reject these simulated experiments are then accessed. Usually, this amounts to then counting the number of times, the model achieves “significant” results compared to non-significant results, which is 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 x percent be able call the results significant. The utility of this analysis is therefore to be able to examine how many subjects are needed to obtain a statistical power of usually 80% given that an effect size in the population is present, this effect size in the population might then be informed by previous studies and or meta-analyses. Extra assumptions are then needed to approximate the distribution of effect sizes as these statistical metrics also have uncertainty associated with them. This extra aspect many times disregarded or forgotten will be expanded upon later while showing how to incorporate it.

The power simulations here will be for a repeated measures design interested in a difference in threshold due to some intervention. I therefore simulate subjects, trials, and effect sizes in a variety of combinations (see full set of combinations in supplementary XXX or figure XYX). Here the effect size chosen was cohens’ d\_rm which formula can be seen in equation XX. This meant that the variance of the second condition (i.e. after the treatment) had to be specified, such that as simulated effect sizes increased the mean difference only increased. The choice here was that second session variances was 1.5 the variances of the first condition. See supplementary XYX for explanation for the choice and why this choice is arbitrary and does not meaningfully influence the interpretations further on. The simulation process followed the following procedure: first two set of agents were simulated from a multivariate normal distribution with group level parameters of the binary nested hierarchical model presented above (see supplementary material XXX for full table of these). The second set of agents had their threshold increased by a random variable that was drawn from the difference distribution defined by 1.5 times the first session variance and the effect size i.e. equation X and Y. To ensure a particular observed effect, size this process was repeated until an observed effect size of the desired value was obtained, this step of resampling for a particular effect size was mainly for visualization purposes later on. Next each agent was put through the pathfinder algorithm to get their trial-by-trial responses and stimulus values. The full trial-by-trial dataset was then fitted using a simple hierarchical model where the threshold was parameterized as a linear combination of intercept and session with dummy coding of session.

Mean and standard deviation of the difference distribution between the two sessions, where is the variance of session 1 is the variance of session 2. is the correlation between the two sessions.

## *Power analysis results*

Given that the space of trial and subject combination is in the extreme, infinite, and at even a practical level quite huge. What I will show here is that the variation in how well the model rejects the null hypothesis, given subject and trial combinations are quite stable over observed effect sizes of the given model, making it possible to give good predictions i.e. extrapolating from the simulations. I will use the decision threshold of saying that a result is significant if less than 5% of the posterior distribution of the difference in the threshold crosses 0 like setting an alpha value of 5% in a frequentist power analysis. This 5% is a reflection of what is currently used in the field as the standard decision threshold but can easily be modified (see github). To properly display the raw results of the power analysis where it’s possible to compare the effects of trials and subjects I will use the beta distribution to properly display a summary of the 100 simulations for each effect size. The beta distribution is a two-parameter distribution that can be parameterized in different ways (REF). The utility of the beta distribution to display the results of the power analysis is that one parameterization of this distribution involves how many times an event happened that we cared about and the other parameter being how many times this event did not happen. This therefore makes it possible to start with a uniform prior on the probability of rejecting the null hypothesis i.e. Beta(1,1) and then updating this probability density function with the amount of hits and misses here significant or non-significant results. This results in a PDF that contains all the information in each of the 100 binary points (i.e. significant or not). Figure XYX shows the resulting power curves for each trial subject combination. Three main things are of particular importance. The shape of the points very closely resembles a psychometric function where subjects and trials influence both the steepness and the location of the function. Secondly, Increasing the number of subjects has two important features, it shifts the points towards higher power with lower effect sizes, but it also seems to increase the sensitivity to the effect size, i.e. the slope of the curve is getting steeper with higher number of subjects. The amount that trials for each subject also matters for the shape of the curves, in figure XYX its quite clear that increasing trials if low i.e. 10, makes a big difference in the shape of the function, but the difference in going from high to very high i.e. 100 to 150, does not matter much. The tendency of the function to be less affected by ever increasing trials is also present for the number of subjects. This observation makes sense if one takes the function to its extremes in trials and subjects. Increasing subject and trials to infinitely many, we would expect, assuming the model has been shown to become increasingly better with increasing trials (like with the ICC metric presented above) that the model would be able to pick up on even the tiniest 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. In the other extreme were 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 reaching a step function in the limit when x goes to 0 and subjects and trials goes to infinity I.e.

These observations are what is going to be used in the next section in order to extrapolate the results from figure X. Making it possible to construct a model that will map trials, subjects and effect sizes to power.

## *Modeling of power analysis*

To use the information from above the latent psychometric function describing the relationship between subjects, trials and effect sizes needs to be investigated. Ideally a psychometric function that enforces the curve going through the origin, as no observed effect size should always entail no power. Next the parameters of these psychometric functions i.e. the threshold and slope need to be parameterized by the number of trials and subjects. Before fitting the general case and ensuring that a psychometric function is well fitting to the problem at hand, I start by fitting each set of trials and subject combinations independently to the parameters of the psychometric function. This amounts to fitting trials and subjects as factors in a linear regression framework (see supplementary for further explanation). This will help ensure that the fitted functions do pass through the points depicted in figure XYXX and increase the faith in the next type of modeling. For this it’s also possible to fit several kinds of psychometric functions and then compare them on their performance of predictability, because that is what we in the end care about here. One way to compare these models is using the Pareto smoothed importance sampling leave one out cross validation (PSIS-LOO-CV) (REF).

<https://projecteuclid.org/journals/bayesian-analysis/volume-13/issue-3/Using-Stacking-to-Average-Bayesian-Predictive-Distributions-with-Discussion/10.1214/17-BA1091.full> <https://link.springer.com/article/10.1007/s11222-016-9696-4> <https://arxiv.org/abs/1507.02646>

I therefore here fit three types of psychometric functions, 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 distributions is that the Weibull function is forced through the origin and its shape therefore quite different from the two other functions (see supplementary XXX for visualization of these differences). The choice of the cumulative normal or logistic function does not necessarily violate the assumptions laid out above because of the way that the parameters are going to be dependent on the trials and subjects. This is clear if one considers a non-zero asymptote for the threshold with an asymptote at y = 0 for the slope. This exactly matches the observation from above that the psychometric function moves closer and closer to a step-function and that the location of this step function approaches x = 0 but never reaches it if the asymptote for the threshold is not zero but close to. The results of this preliminary analysis can be seen in figure XYX where the independently fit logistic psychometric functions are overlaid on the observed datapoints from figure XXX. The figure clearly shows well fit for most of the trials and subject combinations.

## *Continuous mapping of the power analysis*

Moving to the continuous mapping of subjects and trials to the psychometric function that maps observed effect sizes to power, one needs to define the function that relates subjects and trials to these parameters. Given the observations above that the steepness of the function increases with increasing trials and subjects and that the threshold moves towards 0 a first choice of this mapping would 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.

Where is the value of the parameter when the number of trials and subjects approach infinity. is vector of parameters determining the steepness of the exponential decrease from the covariates in the matrix X, here trials subjects and their interaction. The parameter serve, together with , as the value of the parameter when trials and subjects are 0. See supplementary material for further explanation of this relationship and some of the interchange abilities between them.  
Another formulation of the dependency might be a power law equation XX.

Both approaches can produce the observed behavior and their difference depends on the underlying relationship between the parameters and X i.e. the trials and subjects. the exponential equation assumes that as trials and subjects increase by a fixed amount the parameters will decrease by a percentage, whereas the power law assumes that as trials and subjects increase by a percentage the parameters will decrease by a percentage. There are several ways of investigating which of these two approaches results in the better fit, firstly plotting the parameters of the independent fits vs trials and or subjects, which was conducted in the “modeling of power analysis” section, on two different coordinate systems either in (log(y),x) or (log(y),log(x)). Which of these produces the best-looking linear fit / line would be the best candidate (see supplementary XXYX for explanation for this). Figure XYX displays the three functions fitted independently on each of the two coordinate scales.

Another approach would be to fit both types of models and then compare them on LOO-CV. Doing this, displayed problems with 15, 25 and 3 % of observations for the normal, Weibull and logistic function respectively as the pareto k diagnostic value was above 1 for these percentages of datapoints. This essentially makes it meaningless to compare the functions. Moving forward only the logistic cumulative function was used as this was the only model that produced no problems with pareto k values when fitting trials and subjects as continuous variables, for a complete set of models including the normal and Weibull see supplementary XXX. The first logistic model was the exponentially decreasing function equation X. Four other models were fit with different ways of parameterizing equation Y. These four models displayed different ways of how the trials and subjects interact as there is no straightforward way of combining X and β in equation Y. The first was an additive model formulated by equation Z

The second was with a combination of additive and multiplicative operations:

The third was a multiplicative model without an interaction.

The last 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 (see supplementary ZZYA)

Comparing these four models using loo indicated that the best model was the last model but closely followed by the second model, which can be seen in table 5.

This 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 are both variations of the power law. To ensure that the tested models still capture the underlying data, figure XYX displays the winning model imposed on the data with 95 credibility intervals of the mean. As can be seen this closely resembles the individual independent fits, with the most drastic deviation in the 5 subjects 10 trials condition

The posterior distributions of the parameters of the winning model are displayed below in figure XXX.

Meaning that the resulting best guess of the underlying function transforming trials, subjects and observed effect sizes into a probability of rejecting a null hypothesis is as follows:

Where

Where each of these parameters are given by the distributions depicted above.

## *Utility of the power analysis*

As alluded to in the beginning section of the power analysis, the work presented here would be able to help independent researchers determine the probability of rejecting a particular observed effect size using this model, given trials and subjects. However, if this researcher wants to know the probability of rejecting a null hypothesis given that they assume a particular effect size in the population further assumptions needs to be made, as the effect size when conducting an experiment is not a fixed quantity. In practice this means that when we conduct an experiment, we assume that there is an underlying effect, but we cannot directly observe the size of this effect, but we do take a sample of it when conducting our experiment. Mathematically this means that the effect size that is observed in an experiment is given by a probability. The mean and standard deviation of this probability density function is given analytically by Cohen (REF) which could also be derived from bootstrapping as was done with the measurement uncertainty (see supplementary XXX for equivalence) .

Assuming that the effect size is normally distributed we get

The probability of rejecting this sampled effect size is given by the function that was obtained above.

What we ideally want to know here is the probability of observing a particular effect size and that we can reject the null hypothesis given this observed effect size. Probability theory and particularly conditional probabilities gives us the relationship between these quantities.

Here represents the probability that we are interested in, i.e. rejecting, and observing a particular effect size.

Integrating over all possible values of the effect size is now necessary to essentially integrating out the effect size, also known as marginalizing (REF).

Which becomes

Instead of trying to analytically solve this integral analytically, we again use the power of the computational resources to approximate the integral by taking draws of the normal distribution of the observed effect size and then putting them through which will give a draws from a probability distribution of rejecting the null hypothesis. As a last step it is then possible to calculate the proportion of rejected null hypotheses to the total number of draws giving us the power of the study assuming the mean difference and variance in the two sessions.

## *Practical implementation of the power analysis.*

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 here demonstrate how this can be done using what has been provided up until this point. This section will therefore hopefully provide a practical understanding of what different parts should go into a power analysis and how different factors will influence power. I will here assume 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, i.e. assuming that the intervention increases variation in the threshold, but that there is a clear effect (for reproducibility and ease of use the GitHub repository provides a function that does what is described below by inputting these assumptions).

Firstly, investigating 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:

Here and are given by large test-retest reliability analysis (here rounded) and are -8, 8 respectively. Given our assumptions and are therefore -3, 10. We can then vary the number of subjects i.e. draws from this multivariate normal and the correlation coefficient ρ to see the effect on the distribution of effect sizes i.e. p(d\_obs). Highlighting the sampling distribution of the effect size, and the factors influencing it. The results can be seen in figure XYX highlight the fact that both the sample size i.e. subjects, but also the correlation between sessions is vitally important for the variances of the observed effect size.

Now we can visualize how these observed effect size distributions fit into the probability of rejecting the null hypothesis i.e.  Note that the observed effect sizes above are not dependent on the number of trials in the experiment i.e. assuming that they are observed with perfect precision, what the function derived from the continuous power analysis function does is that it incorporates this information together with other factors that might change the precision of the parameters. As shown above the implications of the function can be visualized as psychometric functions in a ( , ) coordinate system with trials and subjects being fixed at values. Another more informative way to visualize these implications for the current purpose is to make a 3-dimensional grid of (Subjects OR trials , , ) with facets of the last variable of either subjects or trials. This visualization can also serve the purpose of projecting the above distributions of observing a particular effect-size unto the space of Figure XXX displays the projection of the histograms as ellipse where the vertical width of the ellipse (the major axis) is given by the 95% Highest density interval of the histograms above and the horizontal width (the minor axis) is for visualization purposes, to see the underlying probability of rejection. The correlation is of cause informed by the test-retest reliability study which was found to be 0.54 [0.49; 0.58] for the current model.

To further clarify the difference in approach, and results, between taking the prior probability of the observed effect size into account or not, as well as showing how an independent researcher might want to use the results provided. Two assumptions must be made, either a mean effect size or a mean difference of the intervention is assumed, and the variance introduced by the intervention. Here if the researcher expects a medium effect size of the intervention of and that the intervention does not increase variability meaning that the variance in both groups should be equal. To fully appreciate the power of this approach one could even imagine sampling these values i.e. 0.5 and the variances of the second session as random variables and not as point estimates, the most obvious case where this could be implemented is when effect sizes from meta-analyses are used for the best guess of an underlying effect size estimate for the study. These effect size estimates from meta-analyses namely come with uncertainties and neglecting this should not be advised!

Using equation XX, it is possible to derive the mean difference and therefore simulate observed effect sizes which are then put into equation XXX and the probability of rejecting that draw is calculated. Repeating this process over the 4000 draws of the posterior distribution of and calculating the ratio of rejected to failed rejected null hypotheses gives an estimate of power including all uncertainty. In the case of not including the prior probability of the effect size the effect size estimate is just repeatedly entered as 0.5. As can now be seen in figure XYX the observed effect size has been “integrated” out and a grid of subjects by trial span the space of power to reject the null hypothesis. Panel A and B of figure XYZ quite clearly display the difference between accounting for the sampling process of effect sizes. As a reference frame in figure XYX the red dashed line with subjects = 20 depicts the results from plugging the same assumptions i.e. Here and and , into the widely used and cited statistical software tool G\*power (REF)

<https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPower3-BRM-Paper.pdf>

. Further reiterating how and why uncertainty propagation is vital to designing studies of adequate power.

# Discussion

The current thesis has investigated how the handling of uncertainty in the field of cognitive science and especially in the developing field of cognitive modeling can be improved. The thesis has done this by demonstrating that using computational resources a deep mathematical understanding with rigorous closed end solutions is not necessary to get a deeper understanding of how uncertainties on each level can and will influence every statistical metric. The thesis outlined three critical types of uncertainty; measurement uncertainty being the lowest level of uncertainty that is often completely neglected in the field, even though it influences the resulting statistical metrics in very unpredictable ways if the data has influential datapoints which might be associated with particularly high measurement uncertainty. Researchers should firstly be aware that measurement uncertainty is always present and examining the extent to which it can be safely ignored in their statistical models must be determined. Even in measures like reaction times which is commonly used in cognitive science (REF), <https://www.jstor.org/stable/27828738> <file:///C:/Users/au645332/Zotero/storage/69LP2CDY/1986-00316-001.html> <file:///C:/Users/au645332/Zotero/storage/MF46PECL/1989-07414-001.html>

there are measurement uncertainties, which depending on the soft and hard-ware the experiment is run on might be a huge factor ([Crocetta & Andrade, 2015](#ref-crocetta_problem_2015); [Holden et al., 2019](#ref-holden_accuracy_2019); [Ohyanagi & Sengoku, 2010](#ref-ohyanagi_solution_2010)).

The thesis pointed to one aspect of cognitive science literature where measurement uncertainty is of difficulty, i.e. questionnaires, note however that the arguments laid out can be applied to any type of measure. Quite a literature exists on doing correlations or testing differences in populations groups on results from questionnaires, which begs a question of the certainty of these questionnaire scores for any one person. This is especially true when considering some quite important aspects in handling these statistical problems. Questionnaires are easy, fast, and cheap to conduct when performing a behavioral experiment, subjecting many questionnaires to be collected in order “to see if something is there”. This curiosity is sometimes what derives sciences forwards, however in cases like these it will inevitably lead to false positive findings that cannot be replicated as the pressure for publishing results might hide the multiple comparison correction that should have been made when finalizing a manuscript. Including some uncertainty into the questionnaires would serve to push the significance barrier higher and make it harder to find significant results in these types of analyses. Perhaps a reasonable comprise would be that the added uncertainty on questionnaire scores should be proportional or just in general related to the internal consistency, measured by ICC, Cronbach alpha, correlation coefficient etc. of the questionnaire i.e. their test re-test reliability uncertainty. Lastly estimation uncertainty was introduced as the uncertainty associated with doing computations is often displayed as standard error of statistical metrics. The main focus of the thesis was to use this understanding of uncertainties in the field of computational modeling and revise some of the methods and metrics used to validate these models. It was shown that using correlations, which has been used in many previous studies (REF), between simulated and recovered parameters values was not a sensible metric to determine the extent of internal model validity. Two reasons stood out, the decision of choosing what size and uncertainty of correlation coefficient to deem model parameters sensible is not straightforward, because the interpretation of the correlation coefficient in this regard is not straightforward. The other reason was that in instances where the simulated and recovered parameter values did show good dependency the correlation coefficient rapidly approached the asymptote at 1 even when more information could be gained by increasing the number of trials, due to its limited inclusion of estimation uncertainty. The study therefore suggests that using a variant of the intra class correlation coefficient (ICC) as the statistical metric for examining internal model validity which has recently been suggested in the literature (REF). The thesis found that this metric was much more sensitive to estimation uncertainty in the parameters, with a sensible interpretation of desirable to undesirable variance. With this new metric the thesis explored ways to decrease the undesirable variance and thereby increase the ICC metric, by incorporating smart experimental designs that are optimized for each individual on a trial-by-trial basis.

Furthermore, the study showed how thinking generatively about the origins of the responses given in an experiment can increase the ICC metric without the need for extra trials, the concrete example given in the thesis was incorporating reaction times into the cognitive model describing how stimulus intensities are transformed to binary forced choices. This highlights an idea of jointly modeling several dependent variables and their interactions, that has been around for quite some time but only now is slowly gaining traction in cognitive science literature (REF). The thesis highlights how all the above implementations and considerations do not necessarily have to rest on heavy mathematical understandings and proofs as computational resources has made it available for people with coding experience to gain these insights by the power of sampling; some of the implications of this will be discussed below. Lastly the thesis investigated and used data from a test-retest reliability study and showed that a reanalyzed could achieve better test-retest reliability by incorporating knowledge about the structure of how the data was gathered with incorporating information already represent in the data. This dataset was then used as an example of how power analyses of cognitive models could be conducted. This was done by first simulating and fitting the exact cognitive model that would be used in a study to many different simulated effect sizes in different trials and subject combinations. This approach allowed modelling the latent power curve that relates 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 exploratory power. Using this equation together with the many times overlooked aspect of sampling variability in the observed effect sizes when conducting power analyses, it was shown that incorporating sampling variability greatly increased the need for more subjects and trials to achieve the same amount of power. This section also highlighted why and where the test re-test reliability of these metric matters as increasing test re-test reliability shrinks the influence of sampling variability in the observed effect size. The thesis further showed that the sampling variability of the observed effect sizes is inversely proportional to the number of subjects in the study, making it vital to account for in studies with few subjects. Lastly the complete uncertainty propagated power analysis was compared to not accounting for sampling variability and the widely used statistical software tool G\*power (ref) in a concrete example. This comparison showed how G\*power estimation of sample size was equivalent to having around infinitely many trials and not accounting for sampling variability. With its 64417 citations G\*power which is aimed towards social, behavioral, and biomedical sciences one might wonder whether a contributing factor to why these fields of study have been some of the most attacked by the replication crisis (REF). One paradoxical aspect of this interaction between the replication crisis and the use of power analyses is that many times it is advised as one of the ways to increase the replicability of studies, as analyses of power of detecting small to medium effect size in social sciences have been found to be low to very low “<https://osf.io/yr8st/download>” (REF). Further, these investigations fail to account for the sampling variability in the effect sizes, further hiding the underlying issue. Interestingly, quite a large number of scientists have suggested that perhaps moving the arbitrary statistical significance threshold to 0.005 instead of the commonly used 0.05 could be one of the approaches used to combat this replication crisis (<https://www.nature.com/articles/s41562-017-0189-z>). Interestingly, lowing of the statistical threshold for significance would in practice lead to the conclusions drawn from this thesis, however from a very different reason as this lowering of the threshold would be a means to an end instead of addressing the underlying problem, which the authors do also acknowledge.

## *Scientific justifications or subjective nonsense*

Another interesting idea that coincides with the general theme of the thesis to combat the replication crisis is that of preregistration, registered reports, and blind analysis (REF). What all these types of interventions have in common is that they acknowledge the subjectivity in not only the data collection but also in the data analysis pipeline of scientific inquiry. This subjectivity is both what introduces biases into what is published, but also what derives novel ideas, and a tradeoff between exploration and exploitation might be necessary to fully guard against unwanted subjectivity. This subjectivity is ever present in all scientific fields and has been for many years (REF). What these interventions try to do is to have the analysis pipeline either fixed before data collection or have the data scrambled such that the results of the analyses are not known when producing the analysis pipeline. Hopefully it is clear that the checking and validation introduced in this thesis is not at stake with these interventions but facilitate them. Everything up until the power analysis was done on simulated data, meaning that all model checking, and validation can be done before collecting and therefore analyzing the experimental data. However, there are still considerations when analyzing the experimental data especially on the model convergence side, where in or excluding covariate or reparameterization of models might be necessary. This is where the blind analysis intervention might be a valuable insight from physics where experimental data is scrambled in various ways such that models, analysis pipelines can be done on data that resembles the collected data, but without being able to know the results before the data is unblinded. Decisions are therefore made on scientific justifications instead of completely subjective decisions to either make the experimental results fit a research paradigm or perhaps even worse produce significant results. The distinction between decision based on scientific justification and subjective nonsensical rationale is fuzzy and narrow, however keeping incentives, such as publishing pressure, fitting into a hypothesis or research paradigm, out of the equation can help with this distinction. This might even give rise to more rigorous methods and analysis pipelines because you do not necessarily stop when the results fit the preconceived notions of the scientific paradigm, but you stop the building and testing process when you are satisfied with the assumptions 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, which are taken as either ground truths or good approximations when they are at best noisy estimates. This could for instance be taking a sampled observed effect size from a previous study instead of relying on some further scientifically justified assumptions that the researchers hold in their domain that might be a much better approximation for the size of the underlying latent effect.

## *Why and how computational tools are vital in science.*

Perhaps cognitive or even computational modeling is the fresh start that is needed in sciences that have notoriously been relying on statistical models such as linear or generalized linear models. These more sophisticated models might be the steppingstone to engage in more theoretically driven modeling, however for this movement to not fall into the same traps and pitfalls as what statistical modeling has, it is essential that rigorous metrics are enforced from the beginning such that those models without any even provable, in principle, parameters or behaviors are discarded from the beginning. Example might arise where rigorous mathematical formulation of theories is developed but that in practice this formulation is not tractable from a computational perspective, it would be a shame to spent years investigating this model and its assumption in field of research just to discover that it in fact is intractable in practice. One might think that a necessity of these more complicated models is a need for deeper mathematical understanding in the people using them. This might be a valid concern however the field of cognitive science has rapidly picked up on sophisticated hierarchical / multi-level models without a need for a deeper understanding of the mathematical machinery. This is not to say that a better understanding of the machinery itself wouldn’t be helpful for researchers of various fields, but perhaps that instead of giving researchers and students a flowchart of when to use a particular statistical model, they should be getting the tools to understand and reflect on these statistical models and therefore also the tools to understand when they break. In the same way that a good scientific program does not teach students the right theories or hypotheses, it teaches them to think in a scientific way such that the individual can decide and test these themselves. Here the tools for understanding reflecting and experimenting with statistical models and concepts could be programming experience in statistics to conduct the types of data simulations presented in the current thesis. This would allow researchers to understand the assumptions that are being made when they go and try and stimulate the data generating process, which might even help spark new scientific ideas. This approach would have researchers more closely engaged in the statistical process of analyzing the data, instead of just picking an off the shelf model from a flowchart.

## *Nothing comes without a cost.*

All of the cognitive models used in the current paper were fitted using stan (REF) using full Bayesian statistical inference with Markov chain monte carlo (MCMC) sampling. As described in the introduction section about modeling definitions fitting and building models in this framework is extremely flexible however this does come with a computational cost of resources and time as these models are not fit at the same time as a linear or multilevel models in packages such as lme4, lmertest or GAMLSS to name a few quite flexible models fitting packages in R (REFS). This added time for doing the optimization to get estimate of parameters has drawbacks in a need for access to bigger machines to necessitate the need for parallelization of the computational burden, which is something that is growing in accessibility and already available to many universities or centers of research and has been correlated with research competitiveness (REFS)

# References

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>

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>

Cohen, S. (1994). *PERCEIVED STRESS SCALE*.

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.

Glover, G. H. (2011). Overview of Functional Magnetic Resonance Imaging. *Neurosurgery Clinics of North America*, *22*(2), 133–139. <https://doi.org/10.1016/j.nec.2010.11.001>

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>

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>

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>

Johnson, S. U., Ulvenes, P. G., Øktedalen, T., & Hoffart, A. (2019). Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. *Frontiers in Psychology*, *10*, 1713. <https://doi.org/10.3389/fpsyg.2019.01713>

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

Kuiper, N. A., Olinger, L. J., & Lyons, L. M. (1986). Global perceived stress level as a moderator of the relationship between negative life events and depression. *Journal of Human Stress*, *12*(4), 149–153. <https://doi.org/10.1080/0097840X.1986.9936781>

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>

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>

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>

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>

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>

Treadway, M. T., Buckholtz, J. W., & Zald, D. (2013). Perceived stress predicts altered reward and loss feedback processing in medial prefrontal cortex. *Frontiers in Human Neuroscience*, *7*. <https://doi.org/10.3389/fnhum.2013.00180>

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>

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, 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>

Xiao, T., Zhu, F., Wang, D., Liu, X., Xi, S.-J., & Yu, Y. (2023). Psychometric validation of the Perceived Stress Scale (PSS-10) among family caregivers of people with schizophrenia in China. *BMJ Open*, *13*(11), e076372. <https://doi.org/10.1136/bmjopen-2023-076372>

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>