# Introduction

The so-called reproducibility crisis has shaken Psychology to the core. Many effect that the scientific community had deemed established, could not be reproduced in highly powered replication studies. Among these are, prominently, Ego Depletion (Hagger et al., 2016) and Terror Management Theory (R. Klein, 2013). (Klein et al., 2018) tested a whole series of effects, some of which were not reproduced at all, and some of which were reproduced in terms of statistical significance, but with smaller effect sizes. Different reasons for the lack of reproducibility have been suggested: wide-spread p-hacking, Hypothesizing After Results are Known ("HARKing"), publication bias, lacking power and lacking theory.

Interestingly, Cognitive Psychology, and more specifically Cognitive Psychology that relies on psychophysical methods, has generated less attention in terms of the reproducibility of its results. (Hesse, 1986) compared different psychophysical methods along criteria such as threshold estimates, efficiency (that is, how many trials it takes to achieve reliable results) and their intra-subject reproducibility in the auditory domain. The reproducibility in the psychophysical measurement of pain sensations has received some attention (Nilsson et al., 2014; Rosier, Iadarola, & Coghill, 2002).

Prima facie, the study of perception is a methodological outlier in several aspects: the number of subjects tested in a typical experiment is quite low, starting from two or three in older papers. And even the typical sample size in modern studies is rarely much higher than 10. On the other hand, each subject typically performs large numbers of trials, starting from around 50 per condition up to several hundred. Furthermore, psychophysical studies are much more likely to be between-subject designs, which lowers the random variability in responses, thus raising power.

## Different Approaches to Null Hypothesis Testing in Psychophysics

There are different approaches to Null Hypothesis Testing in Psychophysics. Classically, psychometric functions (Cumulative Gaussian or Weibull functions) are fitted for each condition and participant to obtain the Points of Subjective Equality (PSEs) and Just Noticeable Differences (JNDs). This yields one data point per subject and condition, over which a t test or an ANOVA is performed to test for statistical significance. This approach neglects that each PSE and JND is based on a large number of trials and thus fails to account for the added reliability of the measures provided. Depending on the experimental design, this approach sacrifices vast amounts of statistical power. As a solution, @Moscatelli2012 have suggested the use of General Linear Mixed Modelling (GLMM). GLMM allows to fit population parameters across all data, while still taking into account that responses within each condition and participant are correlated more strongly than across conditions and participants. In the following, we will estimate power for both types of analyses, to quantify how much power is lost with the two-level approach.

Classically, psychometric functions (Cumulative Gaussian or Weibull functions) are fitted for each condition and participant to obtain the Points of Subjective Equality (PSEs) and Just Noticeable Differences (JNDs). The mean of the Cumulative Gaussian corresponds to the PSE and its standard deviation corresponds to the JND. This yields one data point per subject and condition, over which a t-test or an ANOVA are performed to test for statistically significant differences between conditions.

[FIGURE WITH PSYCHOMETRIC FUNCTION]

This approach neglects that each PSE and JND is based on a large number of trials, which leads to a loss in statistical power. @Moscatelli2012 have suggested the use of General Linear Mixed Modelling (GLMM). GLMM allows to obtain population-wide parameters for PSEs and JNDs, while still accounting for inter-subject variability in responses. Mixed Modelling is a more flexible form of linear regression. It allows to fit regression coefficients across the whole population for some parameters, while allowing the coefficients for other parameters to vary within subgroups of the dataset. A classic example is the modelling of the efficacy of a learning intervention on population of students from different classes in one school. Mixed models can account for inherent performance differences between classes, but fit a population-wide coefficient for the efficacy of the intervention. General Linear Mixed Modelling extends this principle by allowing to fit not only linear regression lines, but also other functions – such as cumulative Gaussians, which are commonly used as approximations for psychometric functions. You can find a more thorough explanation and examples in @Moscatelli2012.

One of improvements that have been demanded in the wake of the reproducibility crisis is a more thorough and meticulous study planning. Researchers need to be more aware of underlying theoretical considerations, specify hypotheses before analyzing the data, make precise predictions of how their hypotheses should manifest in their data and formulate statistical models to test these hypotheses. One important step in this process is to make sure that the experiment has sufficient statistical power to detect the postulated effects. We will first show that many experiments using the two-level approach of hypothesis testing described above often lacks power and indicate how much power could be gained for the same experimental designs by using @Moscatelli2002's GLMM approach. In the second part of this paper, we will demonstrate how to plan the sample size in a psychophysical experiment, levering the advantages of the GLMM approach.

# Power analyses for the GLMM approach using simulations in R

In the following we will provide an example of how to compute the power for common psychophysical designs, using the GLMM approach for analysis. Further below, we will also compare the power we obtain for the same designs when using the Two-Level approach. Please note that we advise strongly to read this document together with the R script available in the GitHub Repository. Words in bold and between quotation marks refer to variables in the script. For some of the variables, we demonstrate how to derive them from existing datasets. To this end, we will use published data on velocity judgements about horizontal motion and motion in depth (Aguado & López-Moliner, 2019).

## Assumptions

This method requires all relevant parameters. Some pertain to the stimuli, some can be taken from the literature, and some must be guessed (educatedly).

**“ID”** is a vector containing one ID for each subject we want to simulate.

**“ConditionOfInterest”** is a vector containing IDs for a binary categorical variable related to the main hypothesis of the experiment. For example: Is there a pictorial background scene?

**“StandardValues”** is a vector containing values for a categorial variable that serves as comparison stimuli. It can contain one value if you want to determine PSEs/JNDs for only one stimulus intensity, but typically you will have several, e. g. when you want to diversify your stimuli to show that a certain effect is not tied to one specific stimulus strength.

**“reps”** is a vector containing an ID for each trial, the maximum number being the average number of trials we expect for any given staircase.

**“PSE\_Difference”** is a value that indicates the percentage to which the PSEs differ between test and standard condition. It can be zero if the condition of interest is not expected to influence PSEs.

**“JND\_Difference”** is a value that indicates the percentage to which the JNDs differ between test and standard condition. It can be zero if the condition of interest is not expected to influence JNDs.

**“Mean\_Standard”** is the Mean of the psychometric function expected for the standard condition. In many cases, this is the stimulus strength of the comparison stimulus.

**“Multiplicator\_SD\_Standard”** is the Standard Deviation of the psychometric function expected for the standard condition, normalized to a mean of 1. We later multiply this normalized standard deviation by the Mean of the psychometric function we aim to simulate. That is, we assume that Weber fractions are constant across the tested stimulus range, which is generally assumed to hold for many cases. While this has been put into doubt (Krueger, 1989) and we recommend to verify to what extent Weber’s law holds for the stimulus in question, we believe this to be a reasonable simplification.

The standard deviation is thus proportional to the relevant Weber fraction and JNDs, which are available in the literature for many different stimulation types. Weber fractions and JNDs can be converted into standard deviations of psychometric functions and vice-versa. The JND is that difference in stimulus intensity that leads the participant to choose the correct stimulus in 75 % of the cases. Weber fractions are normalized versions of this value. Normalization is achieved by dividing it by the intensity of the standard stimulus. To obtain the standard deviation, convert JNDs first into Weber fractions. The Weber fraction is that distance to the mean where the psychometric function yields 25% or 75% correct responses. With the Weber Fraction given, we thus need to determine the appropriate standard deviation given these constraints.

**“SD\_Standard”** is then the standard deviation of the psychometric function for each stimulus intensity (**Multiplicator\_SD\_Standard** \* **Mean\_Standard**).

**“Type\_ResponseFunction”** describes the function the stimulus strengths are chosen from by the method. It can take the values "normal", "Cauchy" and "uniform". "Normal" and "Cauchy" are recommended when you are using a staircase procedure, while "uniform" corresponds to methods of constant stimuli. For a comparison between the three options, see further below. Figure 1 visualizes different response distributions. A Gaussian distribution with an adequate standard deviation should be accurate enough for most intents and purposes when staircase procedures are used. The Cauchy distribution has more heavy tails and could be used if the starting values are relatively far away from the expected PSEs, and the initial step sizes are small. For the Method of Constant Stimuli, no randomness is involved in how the presented stimulus strengths are chosen. In this case, we use the values chosen for your stimulus.

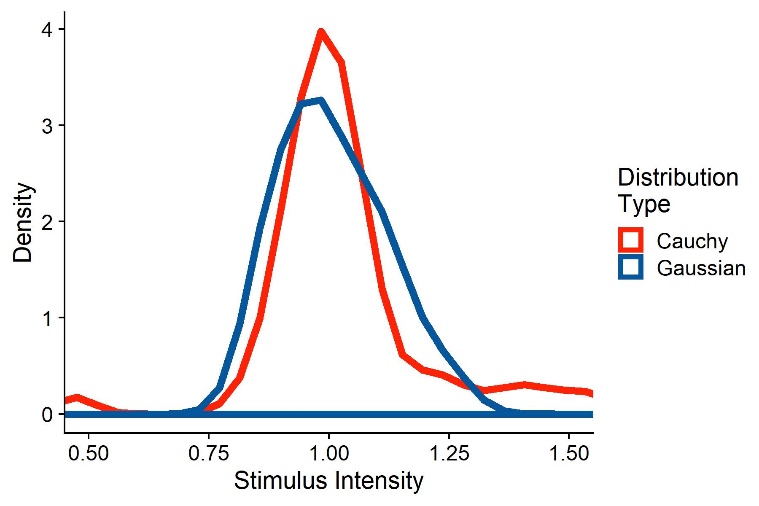


Figure : Two sample distributions of stimulus strengths, representative of stimulus intensities presented when using a staircase procedure. The red distribution corresponds to stimulus strengths drawn from a Cauchy function with a mode of 1 and a scale of 0.05. The blue distribution are responses drawn from a Gaussian distribution with a mean of 1 and a standard deviation of 0.1.

**“SD\_ResponseFunction”** further describes the describes the function the stimulus strengths are chosen from. For normal distributions, this value corresponds to its standard deviation; for Cauchy distributions, this corresponds to its scale; and for uniform distributions, this corresponds to a vector with the values tested.

We assume that there is between-participant variability in the means of the psychometric functions. “**Mean\_Variability\_Between”** sets the standard deviation of the normal distribution these PSEs are drawn from. This normal distribution has a mean of 1, that is, the standard deviation needs to be set accordingly.

We assume that there is between-participant variability in the standard deviations of the psychometric functions. “**SD\_Variability\_Between”** sets the standard deviation of the normal distribution these standard deviations are drawn from. This normal distribution has a mean of 1, that is, the standard deviation has to be set accordingly.

### Establishing the Psychometric Functions

Next, we simulate one whole data set based on the above values. We first create a data frame with one row for each trial. Then, we draw multiplicators for PSEs and JNDs per subject, accounting for between-subject differences in biases and precision. Omitting this step amounts to the assumption that the effect of interest is equally strong in each participant. This can be a valid assumption, but it should not be the default. Rather, the value chosen here should be justified, independently of whether it is zero or above zero. Next, we simulate means and standard deviations of the psychometric functions for each condition and we add between-subject variability, and factor between-subject variability in.

Next, we draw the stimulus strengths likely to be presented in our experiment. As mentioned above, this varies depending on the way the experiment is controlled. For staircase procedures, the responses are more akin to normal distributions with relatively low standard deviations or Cauchy distributions with low scales. A good way to determine the most appropriate function would be to plot the distribution of presented stimulus strengths for pilot data and compare them to different distributions. For the method of constant stimuli, the responses are typically uniformly distributed across 5 to 9 values around the standard stimulus strength. We then use these multipliers ("**staircase\_facto**r") to compute the test stimulus strengths presented in the experiment ("**Presented\_TestStimulusStrength**"). Lastly, we compute the difference between test stimulus and standard stimulus for each trial ("**Difference**").

Then, we compute the probability on each trial to judge the test stimulus intensity as higher (e. g. the test stimulus was faster, brighter, longer, ...) by feeding the simulated test stimulus strengths in a cummulative Gaussian with the mean and the standard deviations calculated above. We then use this value ("**AnswerProbability**") to simulate binary answers ("**Answer**") by drawing responses from a Bernoulli distribution.

As a next step, we bring the data into the format necessary for the glmer() function: We first remove extreme outliers (e. g. by a simple criterion such as excluding trials in which the difference between test and standard stimulus was higher than half the standard stimulus strength). Then, we compute the number of "Test stimulus intensity was higher" responses for each Condition and difference between test and comparison stimulus strength and the number of total observerations for each condition and difference in intensities.

Now, we can inspect these psychometric functions visually to verify whether the values chosen above give rise to the expected psychometric functions in terms of PSE and slopes. Figure 1 illustrates the simulated psychometric functions for the above values. The vertical lines indicate the PSE for each participant and stimulus strength. We can see that the PSEs for Condition of Interest: 1 are shifted towards the right. Furthermore, the curves for Condition of Interest: 1 are more shallow, indicating higher JNDs.

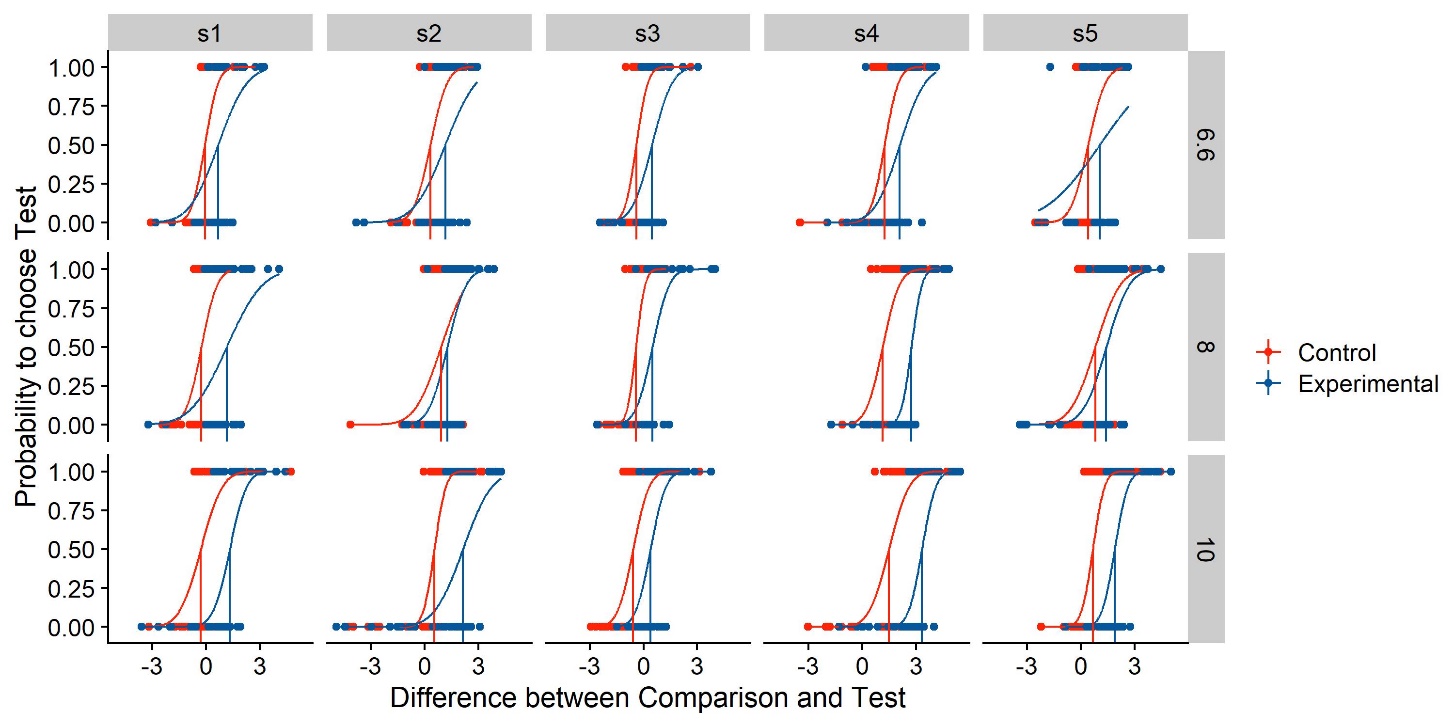


Figure : Simulated psychometric functions based on the example values chosen above. We plot the difference in stimulus intensity between test and standard stimulus (x axis) against the participants’ probability to choose the test stimulus as more intense (y axis). Different panels are the psychometric functions per participant (columns) and per standard stimulus intensity (rows). The psychometric functions are color-coded blue for the experimental Condition of Interest, and red for the control condition without manipulation. The red and blue vertical lines indicate the Points of Subjective Equality, while the vertical and horizontal grey dashed lines denote a difference between test and comparison of 0, and a probability of 0.5 to choose either stimulus. Their intersection thus indicates perfect accuracy, with a PSE of 0. The curves are cumulative Gaussians fitted to the data, while the dots indicate the answer (0 or 1) for each trial.

## Estimating population parameters of the psychometric functions with the GLMM approach

### Accuracy

Next, we establish the statistical models we use to test our hypotheses. Following (Moscatelli, Mezzetti, & Lacquaniti, 2012), we use Generalized Linear Mixed Models for this purpose. For differences in PSEs in our simulated data set, the GLMM could have “ConditionOfInterest” (a binary categorical variable with the values “1” for “Test Condition” and “0” for “Standard Condition”) as fixed effect, and random intercepts and slopes for the Difference between test stimulus and standard stimulus (“Diff”) per Participant (“ID”) and value of the standard stimulus (“StandardValue”).

|  |  |
| --- | --- |
|  | (1) |

Note that there are different ways of specifying the model depending on assumptions about the data and the hypotheses. This sample specification:

* Assumes
  + That we are interested in a population-wide estimate of the the impact of the condition of interest on PSEs (fixed effect of Condition of Interest)
  + That we are not interested in its population-wide impact on JNDs (no interaction between "Condition of Interest" and "Difference")
* Allows intercepts and slopes to vary per participant. Intercepts correspond to PSEs, while slopes correspond to JNDs. That is, we acommodate individual differences in sensitivity and accuracy.
* Allows intercepts and slopes to vary per standard value. It thus acommodates that different standard values might lead to differences in PSEs and JNDs. For example, higher standard values should lead to more shallow slopes (in absolute terms) if Weber Fractions hold for the stimulus type under investigation.

Applying the summary() function to the statistical model yields estimates for the coefficients, along with standard errors. Furthermore, the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) provides the possibility to compute p values using the Satherthwaite degrees of freedom method. The authors of the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) recommend to rely on coefficients and their standard errors alone to estimate the impact of the Condition of Interest. p values from the lmerTest package should thus be regarded as complimentary rather than essential tool. Nonetheless, we believe that p values for the variables of interest are an appropriate, simple proxy for detection of differences between conditions, and allow a quick-and-dirty judgement of whether the ConditionOfInterest has a significant impact in a simulated dataset. After loading lmerTest, we can inspect coefficient, standard errors and p values with the summary() function from R core (R Core Team, 2017).

### Precision

Keep in mind that, in the above model, the coefficient of "Difference" corresponds to the slope of the psychometric function. To capture to what extent a manipulation impacts the slope (as a proxy for precision), one needs to determine how much the slope differs between two conditions. This corresponds to the interaction between "Condition of Interest" and "Difference" in our model. We thus add an interaction between “ConditionOfInterest” and “Difference” to the GLMM we used above to test for PSEs differences.

|  |  |
| --- | --- |
|  | (1) |

As above, we can recover the p value for the interaction between “ConditionOfInterest” and “Difference” with the summary() function after loading the lmerTest package. Importantly, the same disclaimers about the interpretation of p values apply as above.

Computing the Power from Simulations

To simulate the power with a given set of parameters, we need to execute the above procedure sufficient times (we recommend at least 1000 times, although this might be too time consuming for studies with a high count of subjects and/or trials; further below we give some recommendations on how to speed up the relatively slow glmer() model fitting implementation in the R package lme4), and calculate the ratio of simulations in which the test model is significantly better than the test model, given a certain false positive rate (typically 0.05). To this end, we establish functions containing the above procedure. Then, we determine a range of subject counts for which we want to compute the power. Typically, participant counts between 10 and 20 should allow to detect most relevant effects. Fewer are generally not recommended unless there is strong evidence that between-participant variability is really low. Then, we execute thiss method "nIterations" times for each number of subjects specified above, both for precision and accuracy. Once the simulation process is finished, we can plot the data for each number of subjects, which might ressemble our example illustrated in Figure XX.

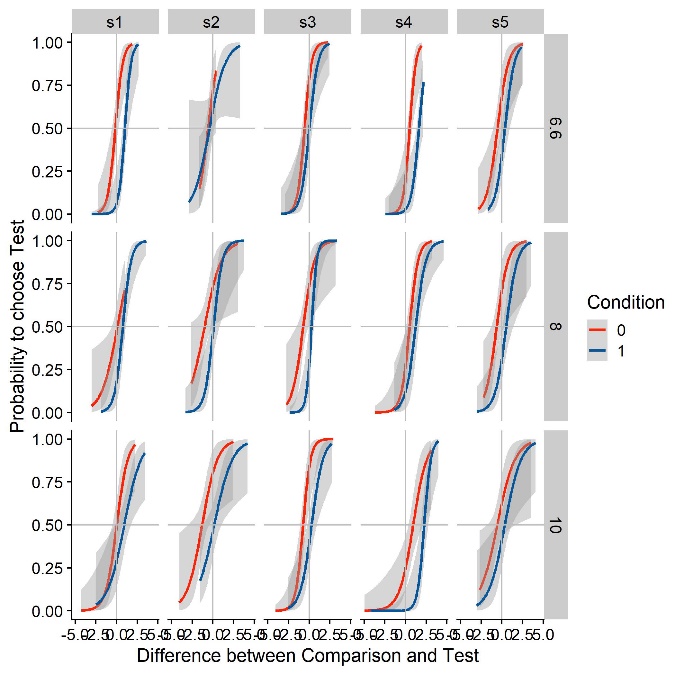


Figure 3:

# Comparing the power for the GLMM and the Two-Level approach

Moscatelli & Lacquaniti (2012) argue that power is lost when using the Two-Level approach. While this is an intuitive notion, it has, to our knowledge, not been confirmed analytically nor synethetically. We thus use the above power simulations to quantify just how much power is lost when using the Two-Level approach in comparison to the GLMM approach. We will use the above procedure to simulate power for the GLMM approach.

## Estimating population parameters of the psychometric functions with the Two-Level approach

For the Two-Level approach, one would first fit psychometric functions for each condition and participant. Then, one would conduct a t.test or an anova to test whether they are different. While there are different methods to fit psychometric functions that each have their own benefits, we use a direction likelihood maximization method (Prins & Kingdom 2010; Knoblauch & Maloney 2012), implemented in the R package quicksy (Linares 2017). The bootstrap option is used to compute confidence intervals, which allow for statistical comparisons. However, the quickpsy package currently does not include an option to estimate population-wide parameters. We thus deactive the bootstrap option, which speeds up the fitting process significantly. Then, we extract the parameters and bring the output of quickpsy into the adequate format for ANOVA analysis. Finally, we perform ANOVAs over means and standard deviations of the fitted psychometric functions, with Condition of Interest and Standard Values as main effects. We are mainly interest in the main effect of Condition of Interest, so we extract the p value for this main effect for both means and standard deviations of the psychometric function.

## Comparing the statistical power of each approach

We pack these computations into functions and follow the same approach as above to compute the power for PSE and JND differences with the GLMM approach and the Two-Level approach. We use the same values we chose above. Figure XX visualizes the simulated power for each approach (GLMM/Two-Level) and measure (PSE/JND). We can see that the power for PSEs is generally so high that even the Two-Level approach achieves a power of nearly 1 for the whole range of participant counts. For the JNDs, the panorama is quite different. While the GLMM approach obtains a satisfactory power almost from 10 participants on out, the Two-Level approach fails to detect population differences for many cases. To quantify by how much power differs across a more diverse range of power levels, we repeat this process another three times with different parameters. We then obtain the false negative rate (1-power) for each approach, measure and participant count, and obtain the ratio as a measure of how much power is lost when using the Two-Level approach with regards to the GLMM approach. Figure XX illustrates that this relationship is, …