How much data do we need to estimate computational models of decision-making? The COMPASS toolbox

Peer-reviewed author version

BEECKMANS, Maud; Huycke, Pieter; Verguts, Tom & Verbeke, Pieter (2023) How much data do we need to estimate computational models of decision-making? The COMPASS toolbox. In: Behavior Research Methods,.

DOI: 10.3758/s13428-023-02165-7
Handle: http://hdl.handle.net/1942/40645

How much data do we need to estimate computational models of decision making? The COMPASS toolbox

Maud Beeckmans[1], Pieter Huycke[1], Tom Verguts[1], Pieter Verbeke[1*]
[1] Department of experimental psychology; Ghent University; Belgium
[*]Corresponding author: pjverbek.verbeke@ugent.be

Abstract

How much data are needed to obtain useful parameter estimations from a computational model? The standard approach to address this question is to carry out a goodness-of-recovery study. Here, the correlation between individual-participant true and estimated parameter values determines when a sample size is large enough. However, depending on one's research question, this approach may be suboptimal, potentially leading to too small (underpowered) or too large (overcostly or unfeasible) sample sizes. In this paper, we formulate a generalized concept of statistical power, and use this to propose a novel approach toward determining how much data is needed to obtain useful parameter estimates from a computational model. We describe a Python-based toolbox (COMPASS) that allows determining how many participants are needed to fit one specific computational model, namely the Rescorla-Wagner model of learning and decision making. Simulations revealed that a high number of trials per person (more than number of persons) are a prerequisite for high-powered studies in this particular setting.

Keywords: Computational models, Statistical power, toolbox

Declarations

Funding: All authors were supported by the Research Foundation Flanders (FWO) / Fonds Nationale de la Reserche Scientifique EOS grant G0F3818N.

Conflicts of interest/Competing interests: The authors declare no competing interests.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publications: All authors consent for publication.

Availability of data and materials: There was no data collection for this paper.

Code availability: An installation guide for COMPASS can be found in the GitHub repository (https://github.com/CogComNeuroSci/COMPASS). On top of this installation guide, three types of files are provided. First, Anaconda environment files allow to create the PyPower coding environment. Second, python files provide the raw code to run COMPASS. These can be freely adapted by the researcher to implement different computational models or empirical designs. Third, csv templates are provided. The researcher should use these files to specify the parameters for power computations with COMPASS.

Introduction

How much data do I need to answer my research question? Nowadays, this question is high on the agenda of researchers, ethical committees, and funding agencies alike. And not without reason. Researchers collect data to arrive at valid conclusions about their theories. Too small samples lead to underpowered studies, as has been famously noted in the context of the replication crisis (Open Science Collaboration, 2012). In contrast, too large sample sizes also pose specific problems given that researchers do not have unlimited financial, time, and other resources. It is thus important to strike a balance between those two extremes. The current paper attempts to formulate an answer to the question of how much data is required in the context of fitting computational models to data. We describe a toolbox that applies it to one of the most popular models of learning and decision making, the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972).

The standard approach to answering the question of how much data is needed to obtain useful parameter estimations from a computational model, is by evaluating the goodness-of-recovery of parameter estimates (Wilson & Collins, 2019). Here, one computes the correlation between a set of true parameter values that were used to simulate the model and the set of estimated parameter values that resulted from fitting the model on the simulated data. Although this approach is widely used, a one-size-fits-all approach is not necessarily optimal. For example, suppose one is merely interested in whether a learning curve correlates with brain activity (measured using fMRI, in a neural area or voxel). Given that the resulting regressors are typically extremely highly correlated across several values of learning rate (Wilson & Niv, 2015), getting highly precise estimates of each individual participant's learning rate is probably not worth the cost and effort. Moreover, given these high financial and practical costs of fMRI, the standard approach will likely make answering the research question infeasible. In this case, a different approach might hence be preferred over a goodness-of-recovery study.

To address how much data is needed to obtain useful parameter estimations from a computational model, we start from the more conventional approach to determine one's sample size, which is the notion of statistical power (power for short). Usually, power refers to the probability of rejecting the null hypothesis, given a linear model, some effect size (for a linear contrast), some sample size, and several statistical assumptions (e.g., all errors are sampled independently and from identical distributions). But more generally, power can be taken to refer to the probability, given a specific model and sample size, that a well-chosen statistic exceeds a threshold. Formally, we could be interested in whether a statistic *T* reaches some cut-off value τ, conditional on a model (or hypothesis) *H*:

$$Pr(\ T \geq \ \tau \mid H) \tag{1}$$

Note that the standard notion of power is a straightforward application of (1). Indeed, one can choose as *H* the linear model together with a specific effect size (e.g., Cohen's d), *T* the standard *t*-statistic, and the threshold τ to be the 5% cutoff point in the (central) *t*-distribution (i.e., the threshold is chosen to maintain a fixed type-I error, typically 5%). We can thus define statistical power to be the probability (1) for any well-specified choice of *H*, *T*, and τ. For instance, *T* can be a descriptive statistic (e.g., a correlation) or a test statistic (t-value). Furthermore, *H* can be the linear model (with specific parameter settings), but also any computational model. Hence, one advantage of the general formulation of (1) is that it allows treating sample size determination in computational and in linear model fitting under the same conceptual umbrella. While previous work and toolboxes were limited to statistics such as the linear model and accompanying t-tests and F-tests (e.g., G-power; Faul et al., 2007), for which closed-forms solutions exist (Cohen, 1988), the general formulation in Equation (1) allows to compute power for a wide variety of situations for which closed-form solutions do not exist. Specifically, current work presents a novel Python-based toolbox which applies this approach to statistics derived from computational models, and the RW model in particular.

We next consider three different statistics *T* of interest. First, the standard approach to determining an appropriate sample size when fitting computational models is goodness-of-recovery (Lerche et al., 2017; Wilson & Collins, 2019). From here on, we refer to this approach as the Internal Correlation (IC) criterion. Thus, for the IC criterion, the statistic *T* is the correlation between true and estimated parameters. It is up to the researcher to determine the magnitude of threshold τ, regulating how high this correlation must be.

Second, statistic *T* can be the correlation between parameter values of the computational model and a measure external to the model. Again, the value of threshold τ indicates how high one desires this correlation to be. We refer to this criterion as the External Correlation (EC) criterion. Typical examples would be correlating the learning rate with the age of the participants (Xia et al., 2021), or with a questionnaire score (Goris et al., 2021).

Third, the criterion can be a difference in model parameter values (e.g., learning rate) such as for example between two groups (Rutledge et al., 2009) or two experimental conditions (Behrens et al., 2007). We refer to this criterion as the Group Difference (GD) criterion. This is most similar to the standard notion of power as typically employed in the context of the linear model.

Depending on how conservative researchers aim to be about rejecting a null hypothesis (absence of correlation or group difference), different cut-off values (τ) can be considered. As in standard power analysis, higher cut-offs decrease power and hence require more data to be collected. Also the underlying hypothesis (*H*) can be manipulated. As in standard power computations, the smaller the assumed effect (measured, for example, via variance of parameter values, group difference or correlation), the larger the sample must be to obtain a high power. Note that two levels of sample size can be distinguished. A first level considers the number of measurements for each participant, and the second level the number of participants. As we will demonstrate, both levels of sample size have distinctive influences on the obtained power.

Once the variables *H*, *T*, and τ are defined, any instance of power (under linear model, computational models, …) can be estimated using Monte-Carlo simulations, as summarized in Figure 1. Specifically, the algorithm entails to repeatedly (N_{reps} times) (1) sample a parameter from the assumed parameter distribution (as stipulated by hypothesis *H)*, (2) generate a dataset with this parameter set (also following *H*), (3) compute the statistic of interest *T* from the generated dataset. Finally, power is defined as the percentage for which *T* exceeds the cut-off τ.

In sum, current work presents a novel toolbox for COMputational Power Analyses using SimulationS (COMPASS) which implements three possible criteria (*T*) for computing power under the RW model. The cut-off values (τ) and Hypotheses (*H*) can be chosen by the user.

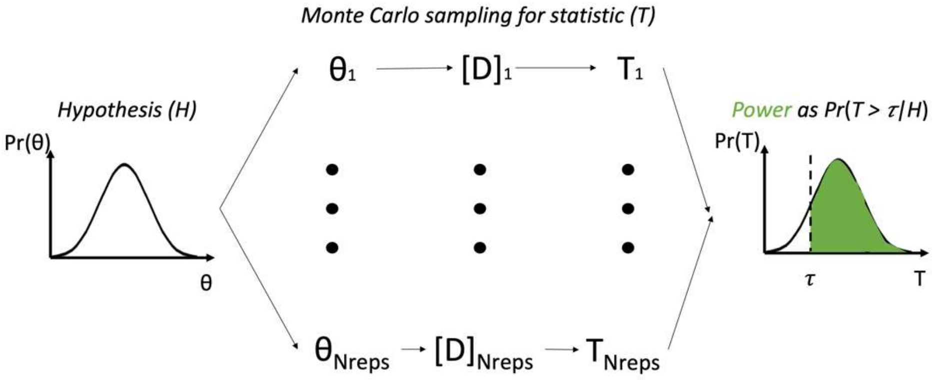

Figure 1. Power analysis for computational model evaluation. Visualization of a power analysis with a given hypothesis (*H*), statistic (*T*) and cut-off (τ). Here, one samples a parameter value (θ) under hypothesis *H*. From this parameter value, a dataset (*D*) is generated which allows to compute a

statistic (*T*). By doing this N_{reps} times, a distribution of *T* can be derived. Finally, power can be computed based on the distribution of *T* and a specified cut-off (τ).

Methods

The computational model

The COMPASS toolbox currently supports one computational model: the RW model (Rescorla & Wagner, 1972). In this model, stimulus-action associations are learned via

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\left(Rew_t - Q_t(s,a)\right) \tag{2}$$

in which the value of a given stimulus-action pair (*Q*(*s,a*)) on trial *t* is updated by the difference between the reward (*Rew*) on trial *t* and the current value of that stimulus-action pair. This update is scaled by the learning rate parameter (α). On each trial, the model selects an action via the Softmax decision (Sutton & Barto, 1998) rule described by

$$Pr(a) = \frac{exp(\gamma Q(s,a))}{\sum_{a'} exp(\gamma Q(s,a'))} \tag{3}$$

in which γ is an inverse temperature parameter that controls the degree of exploration. Lower values of γ imply a higher probability of selecting an action that does not have the highest value, and thus increased random responding.

The three power criteria

As described before, three criteria for power computations are implemented in COMPASS. First, for the IC criterion, the statistic of interest (*T*) is the Pearson correlation coefficient between a set of true parameter values and the estimated parameter values. This corresponds to the standard method to evaluate the goodness-of-recovery of model parameters. Note that even though this statistic contains true parameters, it is still a statistic because we condition on a model *H*; just like the standard t-statistic is compared to a (fictive but model-based) population value of 0. In COMPASS, the true parameter values are sampled from a normal distribution. The user has the freedom to specify a mean and standard deviation for each parameter (α and γ). Additionally, the user can define a cut-off value τ (between 0 and 1) for the correlation coefficient.

The second criterion is the EC criterion. Here, the statistic of interest (*T*) for the EC criterion is the Pearson correlation between an estimated parameter value from the computational model and an external measure such as for instance a questionnaire score or neurophysiological activity. Here, again a hypothetical parameter distribution must be specified. For this criterion, the values for the model parameter (learning rate) and external measure are sampled from a multivariate normal distribution. The user can determine the mean and standard deviation for the learning rate. The distribution for the external measure is fixed with a mean of 0 and standard deviation of 1. Note that these parameters simply scale the external measure and should not cause a loss of generality. As part of the model *H*, the user should also specify a hypothesized correlation between the learning rate and the external measure. From this correlation value, the covariance matrix of the multivariate normal distribution is computed, from which samples are taken to calculate *T* (N_{reps} times). For the EC, the cut-off τ is computed internally in such a way as to keep Type I error rate under control. Specifically, the user defines a Type I error rate. From this Type I error rate, and $N_{participants}$ a probability density (beta) function (see *Scipy.Stats.Pearsonr — SciPy v1.10.0 Manual*, n.d.) can be constructed which allows to derive the critical correlation coefficient to reach significance under a correlation of zero between model parameters and external criterion. This value functions as cut-off (τ) for the power computations.

Third, for the GD criterion, statistic *T* represents the difference between the mean estimated learning rates for two groups (t-statistic). To compute *T*, one first samples true parameter values from two group-specific normal distributions. The hypothesis (*H*) about the underlying true difference of the parameter values between the two groups can be implemented by specifying the means and standard deviation of each group. As for the EC criterion, the user needs to define a Type I error rate. From this error rate and $N_{participants}$ a critical t-value can be derived, which functions as cut-off value.

Power computations with COMPASS

Specifying parameters

COMPASS implements power computations for computational models. Here, all three criteria *T* for power computations are to different degrees influenced by the precision of parameter estimates in the computational model. Of course, precision of parameter estimates not only depends on the model, but also on the experimental design on which the model is tested, as well as the distribution of parameter values in the tested population (which collectively constitute the hypothesis *H*). One design factor that influences the precision of parameter estimates is the number of trials. However, we also consider the reward probability of selecting the optimal action and the number of reversals of stimulus-action associations. This allows the application of COMPASS power computations to a wide variety of probabilistic and/or reversal learning tasks. The distribution of parameter values can be defined by choosing a mean and standard deviation for both learning rate and inverse temperature which allows to construct a normal distribution for both model parameters. As will be demonstrated in the Simulations, choosing an appropriate population distribution of parameter values can significantly influence power computations.

Next, one should specify the remaining power parameters, namely the criterion (*T*) and the cut-off value (τ). Additionally, one must define the number of participants ($N_{participants}$) who will be empirically tested.

Monte Carlo simulations

Once the empirical design and power parameters are specified, power is estimated by using Monte Carlo simulations to compute (sample) multiple values for *T.* An algorithmic overview for power computations is given in Table 1.

A first step in the Monte Carlo simulations is to sample parameter values from the distributions that follow from hypothesis *H*. These parameters are then used to simulate the RW model on the experimental design. This results in a simulated dataset. Then, estimated parameter values are derived by maximizing the log likelihood on the simulated dataset. For this purpose, COMPASS uses the Nelder-Mead (Olsson & Nelson, 1975) method that is implemented in the SciPy package (version 1.9.3) in Python (version 3.10.6).

This sequential process of data simulation and parameter estimation is repeated $N_{participants}$ times. Then, the statistic *T* is computed. After repeating this process N_{reps} times, a distribution of statistics is computed. This process is illustrated in Figure 1. Power is then defined as the proportion of statistics that was equal to or larger than the cut-off value (see Equation (1)). COMPASS presents the user a value of power as well as a plot of the simulated distribution of statistics (see Applications).

Algorithm Power computations via Monte Carlo Simulations for computational models

```
Input:  csv data file with values for (1) creating simulated data (e.g., $N_{trials}$, $N_{reversals}$,
        $Pr(Rew)$, $N_{participants}$), (2) creating parameter distributions for $\alpha$ (learning rate)
        and $\gamma$ (inverse temperature) under hypothesis H (i.e., defining $\alpha_{mean}$, $\alpha_{sd}$, $\gamma_{mean}$,
        $\gamma_{sd}$, ...) and for (3) power computation ($\tau$, $N_{reps}$).
Process:
Create a design $K$ with $N_{trials}$, $N_{reversals}$ and $Pr(Rew)$
for $rep$ = 1: $N_{reps}$
    Sample true parameter values
    If criterion is EC they have a pre-specified correlation.
    If criterion is GD they have a pre-specified mean difference
    for $p$ = 1 : $N_{participants}$
        Simulate response data ($\boldsymbol{D}$) with true parameter values of participant $p$
        on design $K$
        $\boldsymbol{D}$ = Simulate_data($\alpha_p$, $\gamma_p$, $K$)
        Estimate parameters for participant $p$ that maximize log likelihood ($LL$)
        on simulated dataset ($\boldsymbol{D}$)
        $\widehat{\alpha_p}, \widehat{\gamma_p} = \underset{\alpha,\gamma}{argmax}\, LL(\alpha, \gamma | \boldsymbol{D})$
    end
    Compute statistic $T$ (correlation or t-value) for $\alpha$ across all simulated participants
    ($p$) for this repetition ($rep$)
end
Output: Approximate power as
        $\frac{1}{N_{reps}} \sum_{i=1}^{N_{reps}} I(T_i \geq \tau)$
where $I(.)$ is an indicator function equal to 1 if its argument is true and 0 otherwise.
```

Table 1. The power computation algorithm as implemented in COMPASS.

Simulations

Simulations were executed to validate COMPASS and increase our understanding of the parameters that are involved in power computations. Given the strong interest in RW models for probabilistic reversal learning tasks (Crawley et al., 2020; Goris et al., 2021; Manning et al., 2017; Verbeke et al., 2021), we defined a probabilistic reversal design. Here, each trial one of two possible stimuli is presented, and the participant/model must select one out of two response options. Each stimulus is mapped to one response option with a reward probability of 80% and this mapping reverses every 40 trials. As described before, we distinguish between two levels of sample size; number of trials and number of participants. As reasonable bounds to perform high-power empirical studies we varied the total number of trials from 80 to 880 and the number of participants from 40 to 200. The assumed true distribution of learning rates and temperatures were informed by previous work (Crawley et al., 2020; Goris et al., 2021; Verbeke et al., 2021).

The IC criterion

Power computations with different numbers of trials and participants were performed. In all simulations, the inverse temperature distribution has a mean of 1.5 and a standard deviation (SD) of 0.5. For the learning rate parameter, the mean of the distribution was 0.7. Simulations were executed once with different standard deviations for the learning rate distribution. This standard deviation could be small (SD = 0.05), medium (SD = 0.1) or large (SD = 0.2).

Figure 2 shows that, congruent with the well-known restriction of range phenomenon, power under the IC criterion dramatically depends on the *range* of learning rate parameters across participants (i.e., SD). As can be observed by comparing power in Figure 2A versus 2B and 2C, a distribution with a larger standard deviation will result in higher power.

Additionally, the power increases with an increasing number of trials. In contrast, it hardly (perhaps surprisingly; although see Discussion) depends on the number of participants included. Consider the results with a medium learning rate standard deviation of 0.1 (Figure 2B). One concrete conclusion that the applied researcher may draw from Figure 2B, is that if one wants an 80% probability of obtaining reliable parameter estimates (internal correlation >= 0.5), at least 560 trials per person would be needed. No matter how many participants are tested, if the number of trials is 400 or less, only a maximum power of 70% can be reached. In contrast, by increasing the number of trials from 400 to 560 power increases to at least 81%, even if only 40 participants are tested. Hence, under the IC criterion it is clearly optimal to prioritize large trial numbers over large participant numbers.

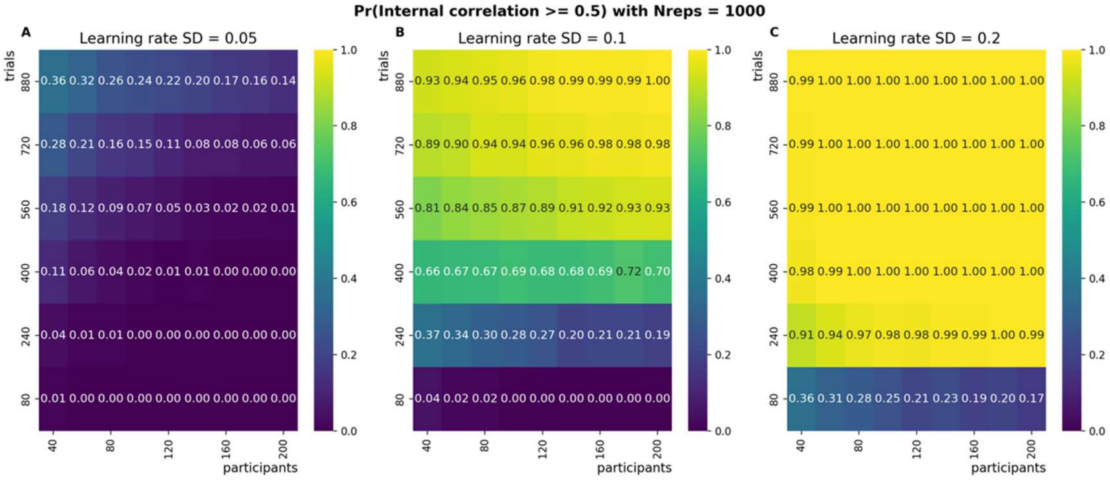

Figure 2. Power estimates under IC criterion. Heatmaps showing the power estimates with varying number of trials and participants when relying on the IC criterion. Number of participants is shown on the x-axis, number of trials on the y-axis. N_{reps} = 1000 repetitions were used to estimate the power for each cell in the grid. **A:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 and a small standard deviation of 0.05. **B:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 and a medium standard deviation of 0.1. **C:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 and a large standard deviation of 0.2.

The EC criterion

Under the EC criterion, we evaluate the power for obtaining a significant correlation between a parameter estimate (learning rate) with a measure that is external to the computational model (e.g., questionnaire score, neurophysiological measure, demographic variables, …). For simulations, the Type I error (which determines τ as mentioned above) was fixed at 0.05. Again, we evaluated power for a varying number of trials and participants. Additionally, three effect sizes were explored: We evaluated power under the assumption that the true correlation was small (ρ = 0.1), medium (ρ = 0.2) or large (ρ = 0.4) (Cohen, 1988).

In contrast to the IC criterion, power here increases both with increasing numbers of trials but also with increasing numbers of participants. Interestingly, under a low hypothesized correlation of 0.1 (Figure 3A), a lot of resources are required to achieve a decent power. In the most extreme case of our simulations with 200 participants and 880 trials, power reaches only a value of 31%. When the hypothesized correlation is 0.2 (Figure 3B) or 0.4 (Figure 3C), much less resources are required to achieve decent power.

Note however, that while for the IC criterion, the learning rate SD was manipulated, all EC simulations are performed under the assumption that the learning rate SD = 0.1. Also power under the EC criterion can be influenced by the hypothesized standard deviation.

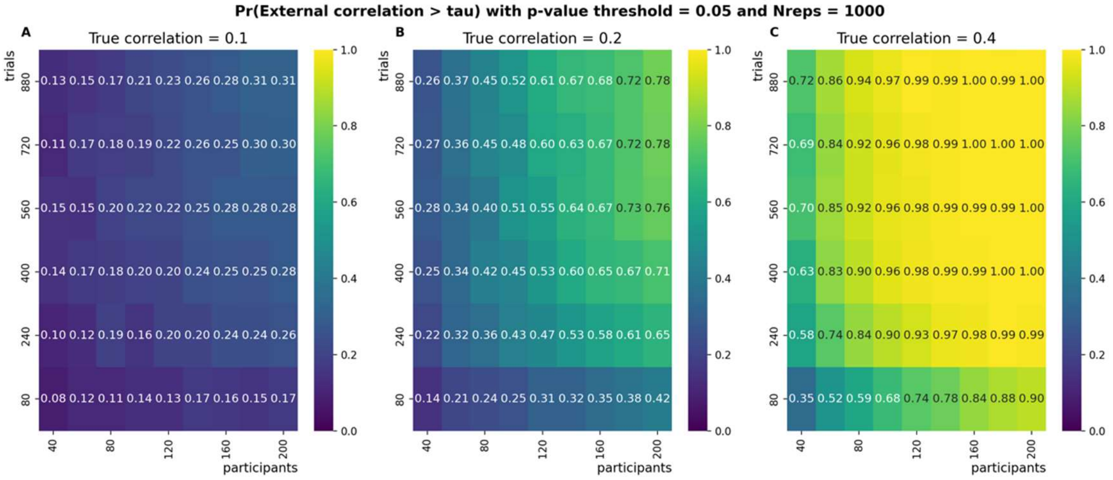

Figure 3. Power estimates under EC criterion. Heatmaps showing the power estimates with varying number of trials and participants when relying on the EC criterion. Number of participants is shown on the x-axis, number of trials on the y-axis. N_{reps} = 1000 repetitions were used to estimate the power for each cell in the grid. **A:** Power estimates when the assumed correlation between the learning rate and external measure is 0.1. **B:** Power estimates when the assumed correlation between the learning rate and external measure is 0.2. **C:** Power estimates when the assumed correlation between the learning rate and external measure is 0.4.

The GD criterion

Under the GD criterion, we compare parameter estimates between two groups of participants. The Type I error (which here also determines τ) was fixed at 0.05. Again, simulations were executed with a varying number of trials and participants. Here, we express effect size in terms of Cohen's *d* (Cohen, 1988), and we again explore small, medium and large effect sizes: *d* = 0.2, *d* = 0.4 and d = 0.6 respectively.

As can be expected, a much higher power can be obtained if the hypothesized effect size is stronger (Figure 3C versus Figure 3B versus Figure 3A). Furthermore, as for the EC criterion, results indicate that both an increase in the number of trials as well as an increase in the number of participants increases power under the GD criterion.

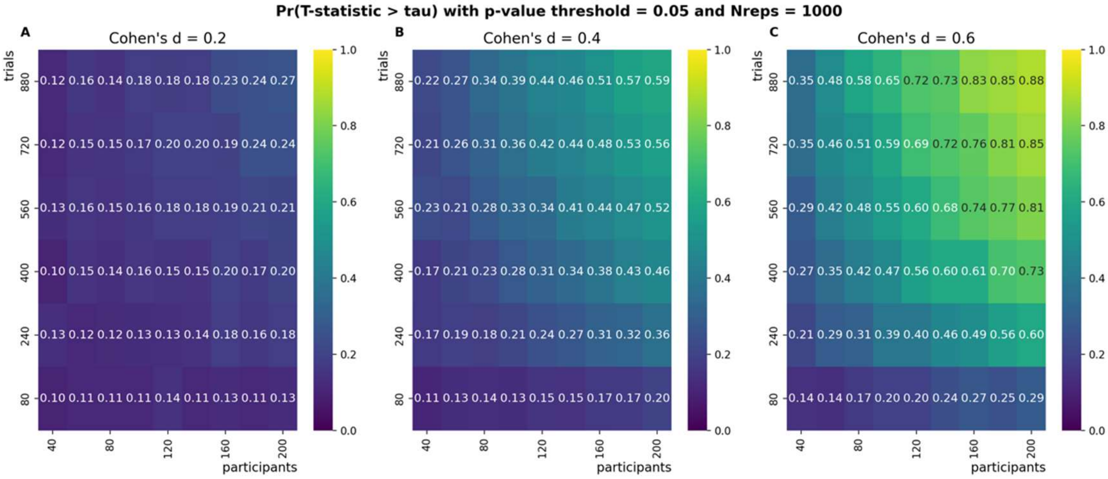

Figure 4. Power estimates under GD criterion. Heatmaps showing the power estimates with varying number of trials and participants when relying on the GD criterion. Number of participants is shown on the x-axis, number of trials on the y-axis. N_{reps} = 1000 repetitions were used to estimate the power for each cell in the grid. **A:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 +/- group difference (with Cohen's *d* = 0.2) and a pooled standard deviation of 0.1. **B:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 +/- group difference (with Cohen's d = 0.4) and a pooled standard deviation of 0.1. **C:** Power estimates when the assumed distribution of learning rates has a mean of 0.7 +/- group difference (with Cohen's d = 0.6) and a pooled standard deviation of 0.1.

Applications

This section explains how researchers can use COMPASS to estimate the number of participants and trials they need for their study. We consider a typical example as described in Goris et al., (2021). This study wanted to investigate whether learning rates in a probabilistic reversal learning task correlated with autistic traits. The RW model was used to estimate learning rate and temperature per participant. A questionnaire was used to measure autistic traits. As described in the Introduction, the standard approach to evaluate how many participants must be tested, would be to carry out a goodness-of-recovery study where true parameters are correlated with estimated ones (i.e., what we called the IC criterion). However, Goris et al. aimed to investigate the correlation between the learning rate and another (autistic trait) measure that was external to the computational model. Hence, evaluating power under the EC criterion could be more appropriate. Note that instead of correlating learning rate with the autistic trait score, researchers could choose to recruit participants with a (sub-)clinical autistic trait score and compare this clinical group to a control group. Then, a t-test can be used to evaluate the difference in learning rates between both groups which corresponds with our GD criterion.

Below, we use the Goris et al. (2021) study as an example to perform power computations with all three criteria implemented in the COMPASS toolbox. The Goris et al. (2021) study tested 150 participants on an empirical design with 90 trials. We performed power analyses with these settings. However, across all criteria, simulations revealed that at least 400 trials are needed to obtain decent power. Therefore, we also estimate power when the design would have contained 450 (90 x 5) trials. To illustrate the differential effect of participant and trial numbers we evaluate power with 150 participants as in the original study but also with five times less participants (30). This results in three situations, distinguishing between the two levels of sample size: (a) high number of participants (150) but low number of trials (90), as in the original study, (b) low number of participants (30) but high number of trials (450) and (c) high number of participants as well as trials.

After installation (see Code availability statement), COMPASS can be used. Using COMPASS for power analyses requires two steps. First, one needs to specify the parameters. This can be done by completing a csv file. In the GitHub repository (https://github.com/CogComNeuroSci/COMPASS), we provide a separate csv file for each of the three criteria. Each csv file has multiple columns which contain user-defined power computation. Seven parameters are present in each file: *ntrials, nreversals, npp, reward_prob, nreps, full_speed, output_folder.* In the Goris et al. study, 150 participants were tested on a probabilistic reversal learning task with 90 trials, 5 reversals and a reward probability of 90%. Hence, we fill in each csv file with *npp* = 150, *ntrials* = 90, *nreversals* = 5 and *reward_prob* = .9. Then, we define the number of repetitions (samples) in the Monte Carlo simulations. The more repetitions, the more precise the power estimate will be. Of course, more repetitions also significantly increase the computation time. As a decent standard, we propose nreps = 250 repetitions. Another variable that needs to be specified is the *output_folder*. This represents the folder to which output should be written. COMPASS will save two files to this folder. One is a csv file with all data from the Monte Carlo simulations. The second file is a .jpg image file showing the distribution of (nreps) statistics that were computed, and how this relates to the cut-off τ (see Figure 5).

To evaluate power under different conditions, we can use multiple rows in the csv files. As described above, we also evaluate power with more trials. Hence, we set *ntrials* = 450 and *nreversals* = 25 in the second (and third) row. We did this for 150 participants but also for the situation when only 30 participants were tested. Hence, in the third row we set *npp* = 30. All other variables were copy-pasted from the first row. To decrease computation time, a *full_speed* option is provided. If this is set to 1, Monte Carlo simulations will be performed in parallel, distributed over multiple logical cores of the computer. To avoid overhead, the *full_speed* option evaluates how many cores are available and leaves two cores out of the computation. We note that *full_speed* can be set to 0 at the expense of a significantly increased computation time.

Second, one must run COMPASS. For this purpose, one opens the terminal window, and runs the command

```
python PowerAnalyses.py $CRITERION$
```

Here, $CRITERION$ is the criterion that is evaluated (one of the following three options: IC, EC or GD). Since there are three rows in our csv file, COMPASS will perform three power analyses.

We next illustrate the use of the three criteria. First, we consider the IC criterion. Hence, we open the csv file: “InputFile_IC.csv”. On top of the standard parameters described above, one must define five additional parameters. The first four (*meanLR*, *sdLR,* *meanInverseTemperature* and *sdInverseTemperature*) represent the researcher’s hypothesis about the population distribution of learning rates and inverse temperatures. As in standard power analyses, it is recommended to use data from previous studies to inform the current hypothesis. Based on Goris et al., 2021 we use *meanLR* = 0.55, *sdLR* = 0.1, *meanInverseTemperature* = 1.5 and *sdInverseTemperature* = 0.5. The fifth parameter (*tau*) defines the cut-off (τ) that is used to evaluate whether parameter estimates are precise enough. Here, a default value is used of *tau* = 0.5. As described above, these values are entered in three rows of the csv file. The csv file “InputFile_IC.csv” is saved and in the terminal, we run the power analyses under the IC criterion. In the terminal window, output is printed for all three power analyses. Below, we show the output for one analysis.

```
(PyPower) MacBook-Pro-4:COMPASS USER$ python PowerAnalysis.py IC

Power estimation started at 2022-12-21 13:07:17.782379.

The power analysis will take ca. 28.0 minutes

Probability to obtain a correlation(true_param, param_estim) >= 0.5
with 90 trials and 150 participants: 0.0%

Power analysis ended at 2022-12-21 13:36:49.530127; run lasted
0:29:31.747748 hours.
```

As described in the documentation (https://github.com/CogComNeuroSci/COMPASS), running COMPASS can require significant computation time. This mostly depends on the number of Monte Carlo repetitions and the number of participants or trials. To help the user, COMPASS will soon after the start of its computations, provide an estimate of how long it will take (in this case, the estimate equals 28 minutes). Once power computations are finished it will provide a power estimate (in this case 0.0%) as well as a distribution plot. The distribution plots for all power computations described in this application are combined in Figure 5. Notably, simulations (see above) revealed that for the IC criterion, the most important variable is the number of trials. Consistently, power is also in this application estimated to be higher when high trial numbers are favored over high participant numbers. While power was 0% with 150 participants but only 90 trials (Figure 5A), it reached 71.2% with 30 participants but 450 trials (Figure 5B). When both participant numbers (150) and trial numbers (450) are high, power is 80.4% (Figure 5C). Hence, the increase in IC power is significant by administering more trials but limited by testing more participants.

Second, we consider the EC criterion. For this criterion, the csv file ("InputFile_EC.csv") holds six additional parameters. Again, the learning rate and inverse temperature distribution parameters should be specified. We define *meanLR* = 0.55, *sdLR* = .1, *meanInverseTemperature* = 1.5 and *sdInverseTemperature* = 0.5. Additionally, the hypothesized correlation should be specified. Here, we predict a medium effect size of *True_correlation* = 0.2. As a last parameter, the user should define a Type-I error rate, based on which the appropriate τ is internally calculated. Here, we define *TypeIerror* = 0.05. Again, we entered these values in three rows of the csv file. The csv file "InputFile_EC.csv" is saved and in the terminal, we run the power analyses under the EC criterion. In the terminal window, output is printed for all three power analyses. Again, we show one example below.

```
(PyPower) MacBook-Pro-4:COMPASS USER$ python PowerAnalysis.py EC

Power estimation started at 2022-12-21 10:34:17.198074.

The power analysis will take ca. 29.0 minutes

Probability to obtain a significant correlation under conventional
power implementation: 69.9%

Probability to obtain a significant correlation between model
parameter and an external measure that is 0.2 correlated with 90
trials and 150 participants: 22.4%

Power analysis ended at 2022-12-21 11:00:49.383942; run lasted
0:26:32.185868 hours.
```

For the EC criterion multiple outputs are presented in the terminal window. First, again a time estimation is given for the power computations. Second, COMPASS also provides an indication of what the conventional power would be to find a significant correlation under the true correlation that was specified in the csv file (e.g, as one would compute with G-power (Faul et al., 2007)). Third, the power of interest as computed by COMPASS is presented. This represents the power for finding a significant correlation between parameter estimates and an external measure if they correlate at *True_correlation* = 0.2. Notice that this is always lower than the power under the conventional implementation. This is because parameter estimates are not perfect and therefore there is additional noise introduced in the analyses. We observe for the EC criterion a power of 22.4% with 150 participants and 90 trials (Figure 5D). Interestingly, while for the IC criterion the number of participants hardly matters, there is a strong effect of the number of participants under the EC criterion. Here, a power of only 14% is obtained for 30 participants and 450 trials (Figure 5E), whereas a power of 51.6% is obtained for 150 participants and 450 trials (Figure 5F).

Third, the GD criterion is considered. For this criterion, the csv file (“InputFile_GD.csv”) holds five additional parameters. Again, the learning rate and inverse temperature distribution should be specified. However, here one needs two distributions, one for each group. For simplicity, we will assume that there is no difference between the groups in terms of inverse temperature. Thus, *meanInverseTemperature_g1* = *meanInverseTemperature_g2* = 1.5 and *sdInverseTemperature_g1* = *sdInverseTemperature_g2* = 0.5. Additionally, we assume that both groups have the same standard deviation for the learning rate. Hence, *sdLR_g1* = *sdLR_g2* = 0.1. Again, we hypothesize a medium effect size (*Cohen’s d* = 0.4). With a standard deviation of 0.1, this implies that the mean difference between the two groups should be 0.05. Therefore, we specify, *meanLR_g1* = 0.57 and *meanLR_g2* = 0.53. As for the EC criterion, also a statistical threshold should be specified. Here, we set *TypeIerror* = 0.05. Again, we entered these values in three rows of the csv file. Importantly, the group difference criterion requires inserting the number of participants in each group, rather than the total number of participants. Hence, instead of 150 or 30, we now insert 75 or 15 for *npp*. The csv file “InputFile_GD.csv'' is saved and in the terminal, we run the power analyses under the GD criterion . In the terminal window, output is printed for all three power analyses. Again, we show one example below.

```
(PyPower) MacBook-Pro-4:COMPASS USER$ python PowerAnalysis.py GD

Power estimation started at 2022-12-21 14:52:11.448771.

The power analysis will take ca. 27.0 minutes

Probability to obtain a significant group difference under conventional
power implementation: 68.2%
```

```
Probability to detect a significant group difference when the estimated
effect size d = 0.4 with 90 trials and 75 participants per group: 14.4%

Power analysis ended at 2022-12-21 15:21:34.069889; run lasted
0:29:22.621118 hours.
```

Also for the GD criterion, multiple outputs are presented in the terminal window. First, again a time estimation is given for the power computations. Second, COMPASS provides an indication of what the conventional power would be to find a significant group difference with the effect size that was specified in the csv file. Third, the power of interest as computed by COMPASS is presented. This represents the power for finding a significant group difference for learning rates under the hypothesized effect size. As for the EC criterion, this is always lower than power under the conventional implementation since the conventional implementation ignores the parameter estimation process of the computational model. Here, a power of 14.4% is obtained with 150 participants and 90 trials (Figure 5G). With only 30 participants but 450 trials, there is a slight increase in power to 22.4% (Figure 5H). Of course, power is highest when both the number of participants and the number of trials is high. In this case, a power can be obtained of 40.8%.

In sum, our application demonstrates that power critically relies on the criterion that is used. For the Goris et al. (2021) study, which aimed to correlate autistic trait scores with learning rate, the EC criterion could be considered most appropriate. Based on our analyses, we can conclude that their design with 150 participants and 90 trials resulted in a statistical power of 22.4 % (Figure 5D) to obtain a significant correlation between the learning rate and an autistic trait score. This is probably much lower than the researchers would have estimated themselves since, as COMPASS indicated, a conventional correlation-based power analysis would have told the researchers that the power to obtain a significant correlation was 69.9%. However, what was critically overlooked here, is the fact that many trials per participant are needed to obtain reliable learning rate estimates. We demonstrated that by increasing the length of their empirical design to 450 trials, a power could be achieved of 51.6% (Figure 5F).

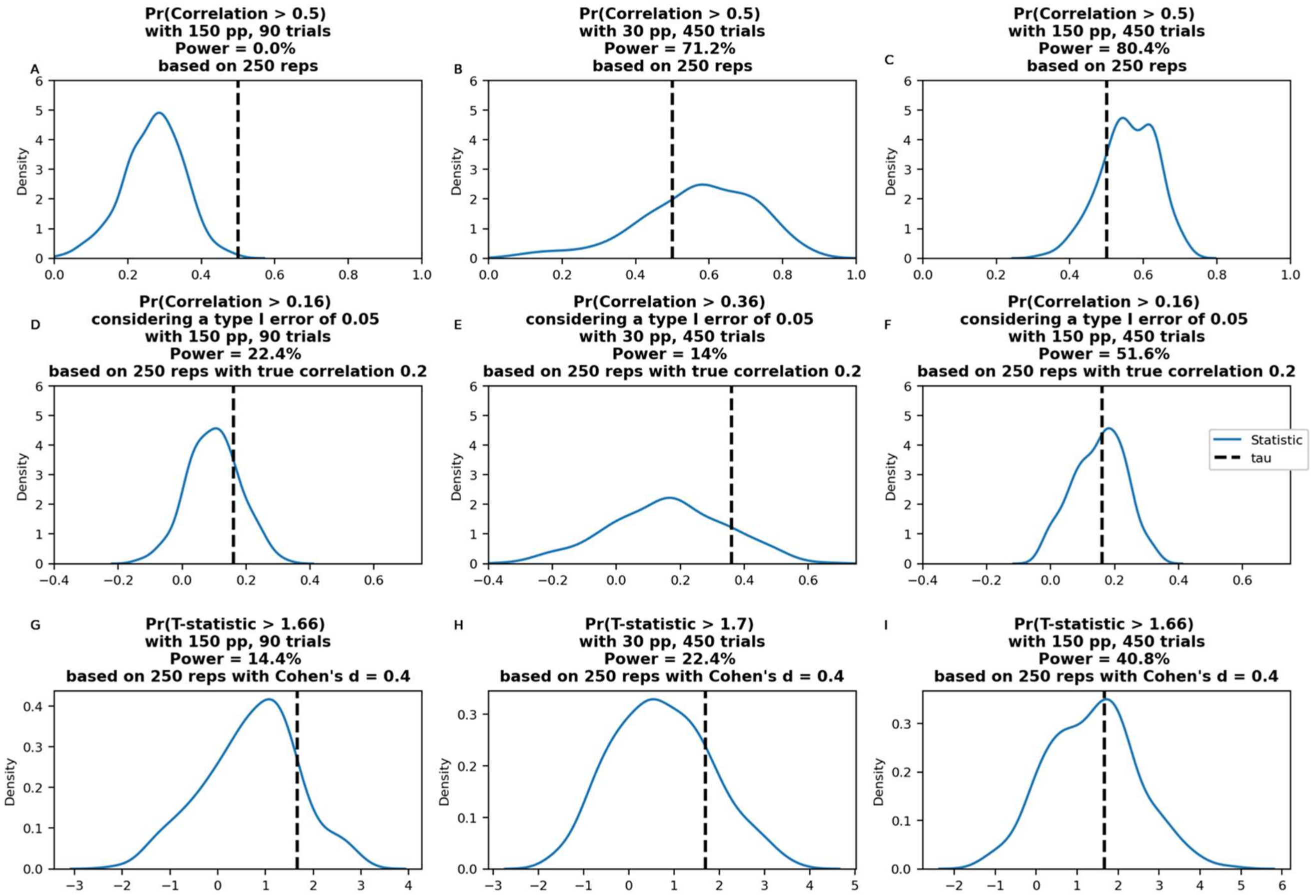

Figure 5. Output plots from COMPASS applications. Output plots show the distribution of computed statistics (*T*). Here, the upper row (panels A-C) shows distributions for the IC criterion, the middle row (panels D-F) illustrates distributions for the EC criterion and the lower row (panels G-I) shows distributions for the GD criterion. The vertical dashed line indicates the cut-off value (τ) that was specified by the user. Here, power is defined as the percentage of statistics right from (higher than) the cut-off.

Discussion

We presented a novel approach to determine how much data needs to be collected to obtain useful parameter estimations from computational models. We argued that this can be formulated in terms of the notion of statistical power, but that it should be tailored to the research question at hand. This approach encompasses the standard approach to sample size determination in computational models (goodness-of-recovery) with the traditional concept of statistical power (Cohen, 1988). We applied this approach to the Rescorla-Wagner model and described a toolbox that allows calculating the relevant power statistics in this case. This should allow applied modelers to make theory-driven design choices. Low statistical power has two complementary disadvantages: It makes it less likely to find true effects; but at the same time, found effects are more likely to be false positives (Button et al., 2013). On the other hand, if required power is overestimated, it may be unrealistic to even run a study (typically, in hard-to-sample contexts or populations). We thus consider it vital to provide a general systematic approach to determining how much data is required to reach sufficient power when using computational models.

Simulations revealed several practical notes for researchers when using the RW model. First, if one wants to obtain precise parameter estimates, as evaluated under the IC criterion, the number of trials in the experimental design is of much greater importance than the number of tested participants. Note that this is not entirely surprising since the IC criterion aims to find reliable parameter estimates for individual participants. Hence, while within-subject noise of the estimate should be minimized, between-subject noise is less relevant in this case. For the EC and GD criteria, where one aims to test correlations or differences across participants, reducing between-subject noise is (much) more relevant. As a result, the number of participants has a stronger influence for the EC and GD criteria than under the IC criterion. Nevertheless, also reducing within-subject noise remains relevant and much power can be gained by increasing the number of trials in the experimental design. This finding has important practical implications, because several previous studies using the RW model seemed to use the suboptimal approach to prioritize large participant numbers over large trial numbers (Goris et al., 2021; Mukherjee et al., 2020; Xia et al., 2021).

Second, as has been stressed in previous work (Wilson & Collins, 2019), and again demonstrated in our simulations under the IC criterion, the variance of learning rates in the tested population is an important factor that should not be ignored. Our simulations indicated that when learning rate distributions are narrow (small SD), a lot of data are needed to obtain precise parameter estimates. Given the fact that the distribution of learning rates in empirical studies typically has a small SD (Crawley et al., 2020; Goris et al., 2021; Verbeke et al., 2021), this might create a pessimistic image for reliably using computational models in empirical studies.

However, as a third important practical note, we demonstrated that precision of parameter estimates (as evaluated under the IC criterion) might be less important depending on the research question. For instance, while using 240 trials and 120 participants results in a power of 27% under the IC criterion (Fig. 2B), a power of 47% can be obtained with the same amount of data under the EC criterion (Fig. 3B).

An important factor that influences power but was not addressed in our simulations is the number of estimated parameters in the model. It is well known that the precision of parameter estimates decreases with an increasing number of model parameters (i.e., biasvariance trade-off; Pitt & Myung, 2002). As a result, also power will likely decrease with an increasing number of parameters. However, the exact amount of decrease strongly depends on the model, experimental design and task, and potentially even the tested population. Although a detailed investigation is beyond the current scope, for any specific model, task, and design, the toolbox allows evaluating the effect of adding extra parameters on statistical power.

The current toolbox is limited in terms of its underlying model (Rescorla-Wagner model); the number of statistical tests it carries out (currently three); the statistical framework it is formulated in (frequentist contrary to a Bayesian framework); and the type of power question it can address (*a priori* power analysis; Faul et al., 2007). However, one key take home message from the current work is that statistical power is not necessarily tied to the workhorse of statistical analysis in psychology, the linear model. Due to fast computing power, researchers with substantive hypotheses can now not only estimate parameters in their models, but also evaluate their (power) consequences. The general test (equation (1)) and algorithm (Table 1) should allow anyone with basic programming skills to adapt our code to their preferred model.

The replication crisis has received strong interest in recent years and underpowered studies have been diagnosed as one of its symptoms (Brysbaert, 2019; Button et al., 2013). However, several authors have proposed that in addition to the data replication crisis, a potentially equally severe theory crisis continues to plague psychology and neuroscience (Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Verguts, 2022). While the notion of power has typically been associated with the former, it is equally important for the latter. With this note, we intended to close this gap. Indeed, our approach allows answering how much data is needed to test any substantive hypothesis, not just for effect size in the linear model, but for any theoretical question formulated in a computational model.

Open practices statement: Our toolbox including all raw Python-code is available on the GitHub repository: https://github.com/CogComNeuroSci/COMPASS. All figures in the current paper can be reproduced using this code. As there was no experimental data collected, there was also no preregistration for the current study.

Acknowledgements: All authors were supported by the Research Foundation Flanders (FWO) / Fonds Nationale de la Reserche Scientifique EOS grant G0F3818N. We thank Marc Brysbaert for comments on an earlier draft of this paper. Additionally, we thank all two anonymous reviewers and Alessio Toraldo for their constructive feedback and helpful comments.

References

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 1. https://doi.org/10.5334/joc.72

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1988). *edition 2. Statistical power analysis for the behavioral sciences*. Hillsdale. Erlbaum.

Crawley, D., Zhang, L., Jones, E. J. H., Ahmad, J., Oakley, B., Cáceres, A. S. J., Charman, T., Buitelaar, J. K., Murphy, D. G. M., Chatham, C., Ouden, H. den, Loth, E., & Group, the E.-A. L. (2020). Modeling flexible behavior in childhood to adulthood shows age-dependent learning mechanisms and less optimal learning in autism in each age group. *PLOS Biology*, *18*(10), e3000908. https://doi.org/10.1371/journal.pbio.3000908

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

Goris, J., Silvetti, M., Verguts, T., Wiersema, J. R., Brass, M., & Braem, S. (2021). Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism*, *25*(2), 440–451. https://doi.org/10.1177/1362361320962237

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

Manning, C., Kilner, J., Neil, L., Karaminis, T., & Pellicano, E. (2017). Children on the autism spectrum update their behaviour in response to a volatile environment. *Developmental Science*, *20*(5), e12435. https://doi.org/10.1111/desc.12435

Mukherjee, D., Filipowicz, A. L. S., Vo, K., Satterthwaite, T. D., & Kable, J. W. (2020). Reward and Punishment Reversal-Learning in Major Depressive Disorder. *Journal of Abnormal Psychology*, *129*(8), 810–823. https://doi.org/10.1037/abn0000641

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

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Olsson, D. M., & Nelson, L. S. (1975). The Nelder-Mead Simplex Procedure for Function Minimization. *Technometrics*, *17*(1), 45–51. https://doi.org/10.1080/00401706.1975.10489269

Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/s1364-6613(02)01964-2

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99. https://doi.org/10.1101/gr.110528.110

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W. (2009). Dopaminergic Drugs Modulate Learning Rates and Perseveration in Parkinson's Patients in a Dynamic Foraging Task. *Journal of Neuroscience*, *29*(48), 15104–15114. https://doi.org/10.1523/JNEUROSCI.3524-09.2009

scipy.stats.pearsonr—SciPy v1.10.0 Manual. (n.d.). Retrieved January 17, 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

Sutton, R., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Verbeke, P., Ergo, K., De Loof, E., Verguts, T., Loof, E. D., & Verguts, T. (2021). Learning to synchronize: Midfrontal theta dynamics during rule switching. *Journal of Neuroscience*, *41*(7), 1–13. https://doi.org/10.1523/JNEUROSCI.1874-20.2020

Verguts, T. (2022). *Introduction to Modeling Cognitive Processes*. MIT Press.

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, 1–33. https://doi.org/10.7554/eLife.49547

Wilson, R. C., & Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLOS Computational Biology*, *11*(6), e1004237. https://doi.org/10.1371/journal.pcbi.1004237

Xia, L., Master, S. L., Eckstein, M. K., Baribault, B., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2021). Modeling changes in probabilistic reinforcement learning during adolescence. *PLoS Computational Biology*, *17*(7), 1–22. https://doi.org/10.1371/journal.pcbi.1008524