Behavior Research Methods (2023) 55:58–76
https://doi.org/10.3758/s13428-022-01793-9

A practical guide for studying human behavior in the lab

Joao Barbosa[1,2] · Heike Stein[1,2] · Sam Zorowitz[3] · Yael Niv[3,4] · Christopher Summerfield[5] · Salvador Soto-Faraco[6] · Alexandre Hyafil[7]

Accepted: 4 January 2022 / Published online: 9 March 2022

Abstract

In the last few decades, the field of neuroscience has witnessed major technological advances that have allowed researchers to measure and control neural activity with great detail. Yet, behavioral experiments in humans remain an essential approach to investigate the mysteries of the mind. Their relatively modest technological and economic requisites make behavioral research an attractive and accessible experimental avenue for neuroscientists with very diverse backgrounds. However, like any experimental enterprise, it has its own inherent challenges that may pose practical hurdles, especially to less experienced behavioral researchers. Here, we aim at providing a practical guide for a steady walk through the workflow of a typical behavioral experiment with human subjects. This primer concerns the design of an experimental protocol, research ethics, and subject care, as well as best practices for data collection, analysis, and sharing. The goal is to provide clear instructions for both beginners and experienced researchers from diverse backgrounds in planning behavioral experiments.

Keywords Human behavioral experiments · Good practices · Open science · 10 rules · Study design

Introduction

We are witnessing a technological revolution in the field of neuroscience, with increasingly large-scale neurophysiological recordings in behaving animals (Gao & Ganguli, 2015) combined with the high-dimensional monitoring of behavior (Musall et al., 2019; Pereira et al., 2020) and causal interventions (Jazayeri & Afraz, 2017) at its forefront. Yet, behavioral experiments remain an essential tool to investigate the mysteries underlying the human mind (Niv, 2020; Read, 2015)—especially when combined with computational modeling (Ma & Peters, 2020; Wilson & Collins, 2019)—and constitute, compared to other approaches in neuroscience, an affordable and accessible approach. Ultimately, measuring behavior is the most effective way to gauge the ecological relevance of cognitive processes (Krakauer et al., 2017; Niv, 2020).

Here, rather than focusing on the theory of empirical measurement, we aim at providing a practical guide on how to overcome practical obstacles on the way to a successful experiment. While there are many excellent textbooks focused on the theory underlying behavioral experiments (Field & Hole, 2003; Forstmann & Wagenmakers, 2015; Gescheider, 2013; Kingdom & Prins, 2016; Lee &

Joao Barbosa and Heike Stein contributed equally to this work.

Significance Statement This primer introduces best practices for designing behavioral experiments, data collection, analysis, and sharing, as well as research ethics and subject care, to beginners and experienced researchers from diverse backgrounds. Organized by topics and frequent questions, this piece provides an accessible overview for those who are embarking on their first journey in the field of human behavioral neuroscience.

✉ Joao Barbosa
palerma@gmail.com

[1] Brain Circuits & Behavior lab, IDIBAPS, Barcelona, Spain

[2] Laboratoire de Neurosciences Cognitives et Computationnelles, INSERM U960, Ecole Normale Supérieure - PSL Research University, 75005 Paris, France

[3] Princeton Neuroscience Institute, Princeton University, Princeton, USA

[4] Department of Psychology, Princeton University, Princeton, USA

[5] Department of Experimental Psychology, University of Oxford, Oxford, UK

[6] Multisensory Research Group, Center for Brain and Cognition, Universitat Pompeu Fabra Barcelona, Spain, and Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[7] Centre de Recerca Matemàtica, Bellaterra, Barcelona, Spain

Wagenmakers, 2013), the practical know-how, which is key to successfully implementing these empirical techniques, is mostly informally passed down from one researcher to another. This primer attempts to capture these practicalities in a compact document that can easily be referred to. This document is based on a collaborative effort to compare our individual practices for studying perception, attention, decision-making, reinforcement learning, and working memory. While our research experience will inevitably shape and bias our thinking, we believe that the advice provided here is applicable to a broad range of experiments. This includes any experiment where human subjects respond through stereotyped behavior to the controlled presentation of stimuli in order to study perception, high-level cognitive functions, such as memory, reasoning, and language, motor control, and beyond. Most recommendations are addressed to beginners and neuroscientists who are new to behavioral experiments, but can also help experienced researchers reflect on their daily practices. We hope that this primer nudges researchers from a wide range of backgrounds to run human behavioral experiments.

The first and critical step is to devise a working hypothesis about a valid research question. Developing an interesting hypothesis is the most creative part of any experimental enterprise. How do you know you have a valid research question? Try to explain your question and why it is important to a colleague. If you have trouble verbalizing it, go back to the drawing board—that nobody did it before is not a valid reason in itself. Once you have identified a scientific question and operationalized your hypothesis, the steps proposed below are intended to lead you towards the behavioral dataset needed to test your hypothesis. We present these steps as a sequence, though some steps can be taken in parallel, whilst others are better taken iteratively in a loop, as shown in Fig. 1. To have maximal control of the experimental process, we encourage the reader to get the full picture and consider all the steps before starting to implement it.

Step 1. Do it

There are many reasons to choose human behavioral experiments over other experimental techniques in neuroscience. Most importantly, analysis of human behavior is a powerful and arguably essential means to studying the mind (Krakauer et al., 2017; Niv, 2020). In practice, studying behavior is also one of the most affordable experimental approaches. This, however, has not always been the case. Avant-garde psychophysical experiments dating back to the late 1940s (Koenderink, 1999), or even to the nineteenth century (Wontorra & Wontorra, 2011), involved expensive custom-built technology, sometimes difficult to fit in an office room (Koenderink, 1999). Nowadays, a typical human behavioral experiment requires relatively inexpensive equipment, a few hundred euros to compensate voluntary subjects, and a hypothesis about how the brain processes information. Indeed, behavioral experiments on healthy human adults are usually substantially faster and cheaper than other neuroscience experiments, such as human neuroimaging or experiments with other animals. In addition, ethical approval is easier to obtain (see Step 4), since behavioral experiments are the least invasive approach to study the computations performed by the brain, and human subjects participate voluntarily.

Time and effort needed for a behavioral project

With some experience and a bit of luck, you could implement your experiment and collect and analyze the data

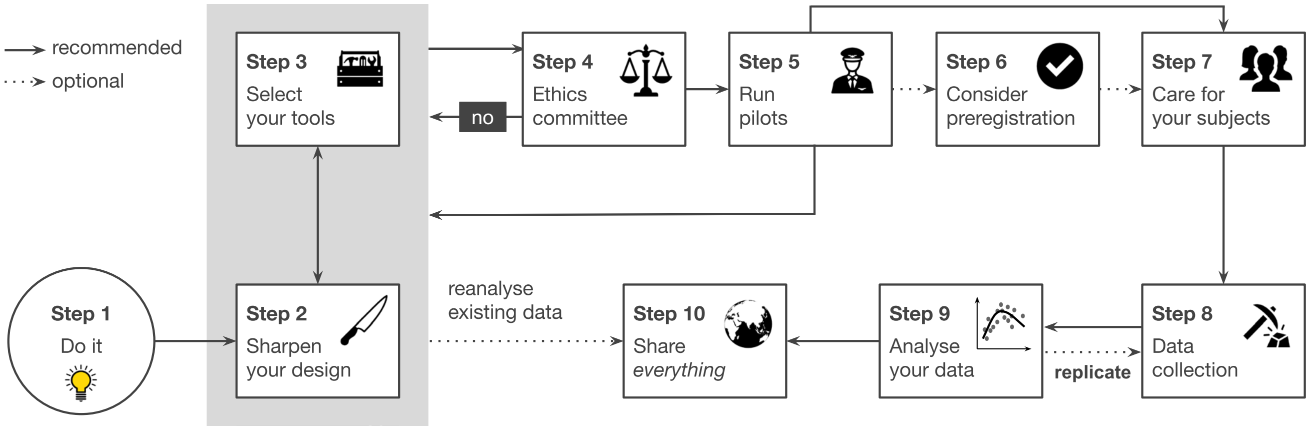

Fig. 1 Proposed workflow for a behavioral experiment. See main text for details of each step.

in a few months. However, you should not rush into data collection. An erroneous operationalization of the hypothesis, a lack of statistical power, or a carelessly developed set of predictions may result in findings that are irrelevant to your original question, unconvincing, or uninformative. To achieve the necessary level of confidence, you will probably need to spend a couple of months polishing your experimental paradigm, especially if it is innovative. Rather than spending a long time exploring a potentially infinite set of task parameters, we encourage you to loop through Steps 2–5 to converge on a solid design.

Reanalysis of existing data as an alternative to new experiments

Finally, before running new experiments, check for existing data that you could use (Table 1), even if only to get a feeling for what real data will look like, or to test simpler versions of your hypothesis. Researchers are increasingly open to sharing their data (Step 10), either publicly or upon request. If the data from a published article is not publicly available (check the data sharing statement in the article), do not hesitate to write an email to the corresponding author politely requesting that data to be shared. In the best-case scenario, you could find the perfect dataset to address your hypothesis, without having to collect it. Beware however of data decay: the more hypotheses are tested in a single dataset, the more spurious effects will be found in this data (Thompson et al., 2019). Regardless, playing with data from similar experiments will help you get a feeling for the kind of data you can obtain, potentially suggesting ways to improve your own experiment. In Table 1 you can find several repositories with all sorts of behavioral data, both from human subjects and other species, often accompanied by neural recordings.

Table 1 Open repositories of behavioral data. In parentheses, specification of how to contribute to the database or repository. Legend: o, open for anyone to contribute; c, contributions restricted to specific community; p, peer-review process necessary

Database	Type of data	URL
Generic data		
DataverseNL (c)	All types of data, including behavior	dataverse.nl
Dryad (o)	Data from different fields of biology, including behavior	datadryad.org
Figshare (o)	All types of data, including behavior	figshare.com
GIN (o)	All types of data, including behavior	gin.g-node.org
Google Dataset Search (o)	All types of data, including behavior	datasetsearch.research.google.com
Harvard Dataverse (o)	All types of data, including behavior	dataverse.harvard.edu
Mendeley Data	All types of data, including behavior	data.mendeley.com
Nature Scientific Data (o,p)	All types of data, including behavior	nature.com/sdata
OpenLists (o)	All types of electrophysiology, including behavior	github.com/openlists/ElectrophysiologyData
OSF (o)	All types of data, including behavior and neuroimaging. Preregistration service	osf.io
Zenodo (o)	All types of data, including behavior	zenodo.org
Human data		
OpenData (o)	A collection of publicly available behavioral datasets curated by the Niv lab.	nivlab.github.io/opendata
CamCan (c)	Cognitive and neuroimaging data of subjects across adult lifespan	cam-can.org
Confidence database (c)	Behavioral data with confidence measures	osf.io/s46pr
Human Brain Project (p)	Mostly human and mouse recordings, including behavior	kg.ebrains.eu/search
Oasis (c)	Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer's Disease	oasis-brains.org
Open Neuro (o)	Human neuroimaging data, including behavior	openneuro.org
PsychArchives (o)	All fields of psychology	psycharchives.org
The Healthy Brain Network (c)	Psychiatric, behavioral, cognitive, and lifestyle phenotypes, as well as multimodal brain imaging of children and adolescents (5-21)	fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network
Animal data		
CRCNS (o)	Animal behavior and electrophysiology	crcns.org
International Brain Lab (c)	Mouse electrophysiology and behavior	data.internationalbrainlab.org
Mouse Bytes (c)	Mouse cognition, imaging and genomics	mousebytes.ca

Step 2. Aim at the optimal design to test your hypothesis

Everything should be made as simple as possible, but not simpler

After you have developed a good sense of your hypothesis and a rough idea of what you need to measure (reaction times, recall accuracy on a memory task, etc.) to test it, start thinking about how you will frame your arguments in the prospective paper. Assuming you get the results you hope for, how will you interpret them and which alternative explanations might account for these expected outcomes? Having the paper in mind early on will help you define the concrete outline of your design, which will filter out tangential questions and analyses. As Albert Einstein famously did *not* say, "everything should be made as simple as possible, but not simpler." That is a good mantra to keep in mind throughout the whole process, and especially during this stage. Think hard on what is the minimal set of conditions that are absolutely necessary to address your hypothesis. Ideally, you should only manipulate a small number of variables of interest, which influence behavior in a way that is specific to the hypothesis under scrutiny. If your hypothesis unfolds into a series of sub-questions, focus on the core questions. A typical beginner's mistake is to design complex paradigms aimed at addressing too many questions. This can have dramatic repercussions on statistical power, lead to overly complicated analyses with noisy variables, or open the door to "fishing expeditions" (see Step 6). Most importantly, unnecessary complexity will affect the clarity and impact of the results, as the mapping between the scientific question, the experimental design, and the outcome becomes less straightforward. On the other hand, a rich set of experimental conditions may provide richer insights into cognitive processes, but only if you master the appropriate statistical tools to capture the complex structure of the data (Step 9).

At this stage, you should make decisions about the type of task, the trial structure, the nature of stimuli, and the variables to be manipulated. Aim at experimental designs where the variables of interest are manipulated orthogonally as they allow for the unambiguous attribution of the observed effects. This will avoid confounds that will be difficult to control for a posteriori (Dykstra, 1966; Waskom et al., 2019). Do not be afraid to innovate if you think this will provide better answers to your scientific questions. Often, we cannot address a new scientific question by shoehorning it into a popular design that was not intended for the question. However, innovative paradigms can take much longer to adjust than using off-the-shelf solutions, so make sure the potential originality gain justifies the development costs. It is easy to get over-excited, so ask your close colleagues for honest feedback about your design—even better, ask explicitly for advice (Yoon et al., 2019). You can do that through lab meetings or contacting your most critical collaborator that is good at generating alternative explanations for your hypothesis. In sum, you should not cling to one idea; instead, be your own critic and think of all the ways the experiment can fail. Odds are it will.

Choosing the right stimulus set

For perceptual or memory studies, a good set of stimuli should have the following two properties. First, stimuli must be easily parametrized, such that a change in a stimulus parameter will lead to a controlled and specific change in the perceptual dimension under study (e.g. motion coherence in the random-dot kinematogram will directly impact the precision of motion perception; number of items in a working memory task will directly impact working memory performance). Ideally, parameters of interest are varied continuously or over at least a handful of levels, which allows for a richer investigation of behavioral effects (see Step 9). Second, any other sources of stimulus variability that could impact behavior should be removed. For example, if you are interested in how subjects can discriminate facial expressions of emotion, generate stimuli that vary along the "happy–sad" dimension, while keeping other facial characteristics (gender, age, size, viewing angle, etc.) constant. In those cases where unwanted variability cannot be removed (e.g. stimulus or block order effects), counterbalance the potential nuisance factor across sessions, participants, or conditions. Choose stimulus sets where nuisance factors can be minimized. For example, use synthetic rather than real face images or a set of fully parameterized motion pulses (Yates et al., 2017) rather than random-dot stimuli where the instantaneous variability in the dot sequence cannot be fully controlled. Bear in mind, however, that what you gain in experimental control may be lost in ecological validity (Nastase et al., 2020; Yarkoni, 2020). Depending on your question (e.g. studying population differences of serial dependence in visual processing [Stein et al., 2020] vs. studying the impact of visual serial dependence on emotion perception [Chen & Whitney, 2020]) it might be wise to use naturalistic stimuli rather than synthetic ones.

For cognitive studies of higher-level processes, you may have more freedom over the choice of stimuli. For example, in a reinforcement-learning task where subjects track the value associated with certain stimuli, the choice of stimuli might seem arbitrary, but you should preferably use stimuli that intuitively represent relevant concepts (Feher da Silva & Hare, 2020; Steiner & Frey, 2021) (e.g. a deck of cards to illustrate shuffling). Importantly, make sure that stimuli are effectively neutral if you do not want them to elicit distinct learning processes (e.g. in a reinforcement-learning framework, a green stimulus may be a priori associated with a higher value than a red stimulus) and matched at the perceptual level (e.g. same luminosity and contrast level), unless these perceptual differences are central to your question.

Varying experimental conditions over trials, blocks, or subjects

Unless you have a very good reason to do otherwise, avoid using different experimental conditions between subjects. This will severely affect your statistical power as inter-subject behavioral variability may take over condition-induced differences (for the same reason that an unpaired *t*-test is much less powerful than a paired *t*-test). Moreover, it is generally better to vary conditions over trials than over blocks, as different behavioral patterns across blocks could be driven by a general improvement of performance or fluctuations in attention (for a detailed discussion see Green et al. (2016)). However, specific experimental conditions might constrain you to use a block design, for example if you are interested in testing different types of stimulus sequences (e.g. blocks of low vs. high volatility of stimulus categories), or preclude a within-subject design (e.g., testing the effect of a positive or negative mood manipulation). In practice, if you opt for a trial-based randomization, you need to figure out how to cue the instructions on each trial without interfering too much with the attentional flow of the participants. It can be beneficial to present the cue in another modality (for example, an auditory cue for a visual task). Beware also that task switching incurs some significant behavioral and cognitive costs and will require longer training to let participants associate cues with particular task instructions.

Pseudo-randomize the sequence of task conditions

Task conditions (e.g. stimulus location) can be varied in a completely random sequence (i.e. using sampling with replacement) or in a sequence that ensures a fixed proportion of different stimuli within or across conditions (using sampling without replacement). Generally, fixing the empirical distribution of task conditions is the best option, since unbalanced stimulus sequences can introduce confounds difficult to control a posteriori (Dykstra, 1966). However, make sure the randomization is made over sequences long enough that subjects cannot detect the regularities and use them to predict the next stimulus (Szollosi et al., 2019). Tasks that assess probabilistic learning, such as reinforcement learning, are exceptions to this rule. Because these tasks are centered around learning a probabilistic distribution, you should sample your stimuli randomly from the distribution of interest (Szollosi et al., 2019).

Carefully select the sample size

Start early laying out specific testable predictions and the analytical pipeline necessary for testing your predictions (Steps 6 and 9). This will help you find out which and how much data you need to gather for the comparisons of interest. It might be a good idea to ask a colleague with statistics expertise to validate it. If you plan on testing your different hypotheses using a common statistical test (a *t*-test, regression, etc.), then you can formally derive what is the minimum number of subjects you should test to be able to detect an effect of a given size, should it be present with a given probability (the *power* of a test; Fig. 2a) (Bausell & Li, 2002). As can be seen in Fig. 2, more power (i.e. confidence that if an effect exists, it will not be missed) requires more participants. The sample size can also be determined based on inferential goals other than the power of a test, such as estimating an effect size with a certain precision (Gelman & Carlin, 2014; Maxwell et al., 2008). For tables of sample size for a wide variety of statistical tests and effect size, see Brysbaert (2019); some researchers in the field prefer to use the G*Power software (Faul et al., 2009). Either way, be aware that you will often find recommended sample sizes to be much larger than those used in previous studies, which are likely to have been underpowered. In practice, this means that despite typical medium to small effect sizes in psychology (Cohen, 1992), the authors did not use a sample size sufficiently large to address their scientific question (Brysbaert, 2019; Kühberger et al., 2014). When trying to determine the recommended sample size, your statistical test might be too complex for an analytical approach (Fig. 2), e.g. when you model your data (which we strongly recommend, see Rule 9). In such cases the sample size can be derived using simulation-based power analysis (Fig. 2b). That implies (a) simulating your computational model where your effect is present for many individual "subjects," (b) fitting your model to the synthetic data for each subject, and (c) computing the fraction of times that your effect is significant, given the sample size.

Whether based on analytical derivations or simulations, a power analysis depends on the estimation of the effect size. Usually that estimation is based on effect sizes from related studies, but reported effect sizes are often inflated due to publication biases (Kühberger et al., 2014). Alternatively, a reverse power analysis allows you to declare what is the minimum effect size you can detect with a certain power given your available resources (Brysbaert, 2019; Lakens, 2021). In simulations, estimating the effect size means estimating a priori the value of the model parameters, which can be challenging (Gelman & Carlin, 2014; Lakens, 2021). To avoid such difficulties, you can decide the sample size a posteriori using Bayesian statistics. This

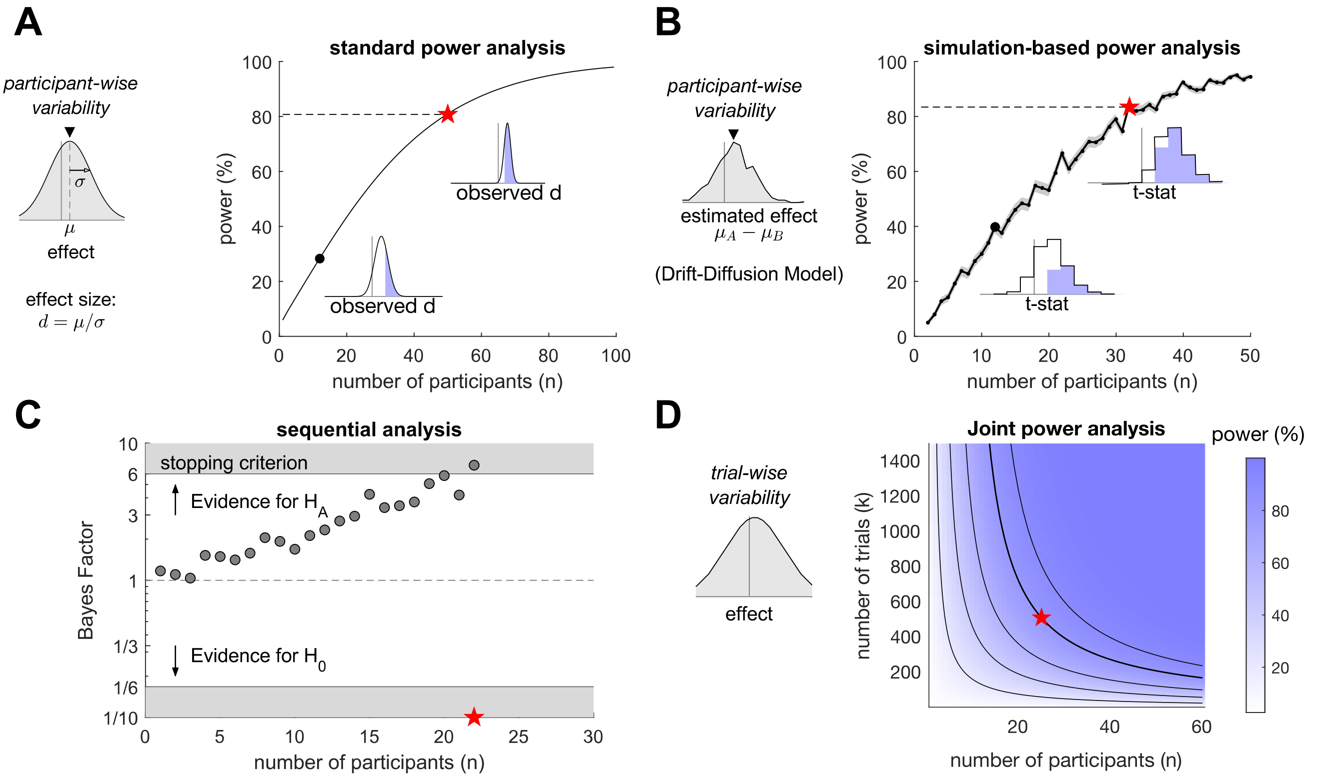

Fig. 2 Methods for selecting the sample size. **a** Standard power analysis, here applied to a *t*-test. An effect size (represented by Cohen's *d*, i.e. the differences in the population means over the standard deviation) is estimated based on the minimum desired effect size or previously reported effect sizes. Here, we compute the power of the *t*-test as a function of sample size assuming an effect size of $d=0.4$. Power corresponds to the probability of correctly detecting the effect (i.e. rejecting the null hypothesis with a certain α, here set to 0.05). The sample size is then determined as the minimal value (red star) that ensures a certain power level (here, we use the typical value of 80%). Insets correspond to the distribution of estimated effect sizes (here, *z* statistic) for example values of sample size (solid vertical bar: $\hat{d}=0$). Blue area represents the significant effects. **b** Simulations of the drift-diffusion model (DDM, Ratcliff & McKoon, 2008) with a condition-dependent drift parameter μ ($\mu_A=0.8$, $\mu_B=0.7$); other parameters: diffusion noise $\sigma=1$, boundary $B=1$, non-decision time $t=300$ ms). We simulated 60 trials in each condition and estimated the model parameters from the simulated data using the PyDDM toolbox (Shinn et al., 2020). We repeated the estimation for 500 simulations, with the corresponding distribution of estimated effect sizes ($\hat{\mu}_A - \hat{\mu}_B$) shown in the left inset (black triangle marks the true value μ_A-$\mu_B=0.1$). The power analysis is then performed by computing a paired *t*-test for subsamples of *n* simulations between the estimated drift terms (100 subsamples have been used for each value of *n*). **c** Bayes factor (BF) as a function of sample size in a sequential analysis where the sample size is determined a posteriori (see main text). In these simulations, BF steadily increases as more subjects are included in the analyses, favoring the alternative hypothesis. Data collection is stopped once either of the target BF value of 6 or 1/6 is reached (here 6, i.e. very strong evidence in favor of the alternative hypothesis). Adapted from Keysers et al. (2020). The sample size is thus determined a posteriori. **d** A power analysis can also determine jointly the number of participants *n* and number of trials per participant *k*. The power analysis is based on assuming an effect size ($d=0.6$) with a certain trial-wise (or within-subject) variability σ_w and across-subject variability σ_b. Here we used $\sigma_w=20$ *and* $\sigma_b=0.6$. The same logic as for the standard power analysis applies to compute the power analytically for each value of *n* and *k*, depicted here in the two-dimensional map. Contour plots denote power of 20%, 40%, 60%, 80%, and 90% (the thick line denotes the selected 80% contour). Points on a contour indicate distinct values of the pair (*n*,*k*) that yield the same power. The red star indicates a combination that provides 80% power and constitutes the preferred trade-off between the number of trials and number of subjects. Adapted from Baker et al. (2021). See https://shiny.york.ac.uk/powercontours/ for an online tool. Code for running the analysis is available at https://github.com/ahyafil/SampleSize.

allows you to stop data collection whenever you reach a predetermined level of confidence in favor of your hypotheses, or the null hypothesis (Fig. 2c) (Keysers et al., 2020). This evidence is expressed in a Bayesian setting and computed as the Bayes factor, i.e. the ratio of the marginal likelihood for *both* the null and alternative hypotheses, which integrates the evidence provided by each participant. While a sequential analysis can also be performed with a frequentist approach, here you will have to correct for the sequential application of multiple correlated tests (Lakens, 2014; Stallard et al., 2020). Finally, another possibility is to fix the sample size based on heuristics such as rules or thumbs or replicating sample sizes from previous studies, but this is not recommended as publication biases lead to undersized samples (Button et al., 2013; Kvarven et al., 2020; Lakens, 2021).

More trials from fewer subjects vs. fewer trials from more subjects

The number of trials per subject and number of subjects impacts the length and cost of the experiment, as well as its statistical power. Striking a balance between the two is a challenge, and there is no silver bullet for it, as it depends largely on the origin of the effect you are after, as well as its within-subject and across-subject variance (Baker et al., 2021). As a rule of thumb, if you are interested in studying different strategies or other individual characteristics (e.g. Tversky & Kahneman, 1974), then you should sample the population extensively and collect data from as many subjects as possible (Waskom et al., 2019). On the other hand, if the process of interest occurs consistently across individuals, as it is often assumed for basic processes within the sensory or motor systems, then capturing population heterogeneity might be less relevant (Read, 2015). In these cases, it can be beneficial to use a small sample of subjects whose behavior is thoroughly assessed with many trials (Smith & Little, 2018; Waskom et al., 2019). Note that with joint power analysis, you can determine the number of participants and number of trials per participant together (Fig. 2d) (Baker et al., 2021).

Step 3. Choose the right equipment and environment

Often, running experiments needs you to carefully control and/or measure variables such as luminance, sound pressure levels, eye movements, timing of events, or the exact placement of hardware. Typical psychophysical setups consist of a room in which you can ideally control, or at least measure, these factors. Think whether any of these could provide a variable of interest to your study or help to account for a potential confound. If so, you can always extend your psychophysics setup with more specialized equipment. For instance, if you are worried about unconstrained eye movements or if you want to measure pupil size as a proxy of arousal, you will need an eye tracker.

Eye trackers and other sensors

You can control the impact of eye movements in your experiment either by design or by eliminating the incentive of moving the eyes, for example by using a fixation cross that minimizes their occurrence (Thaler et al., 2013). If you need to control eye gaze, for example to interrupt a trial if the subject does not fixate at the right spot, use an eye tracker. There are several affordable options, including those that you can build from scratch (Hosp et al., 2020; Mantiuk et al., 2012) that work reasonably well if ensuring fixation is all you need (Funke et al., 2016). If your lab has an EEG setup, electrooculogram (EOG) signals can provide a rough measure of eye movements (e.g. Quax et al., 2019). Recent powerful deep learning tools (Yiu et al., 2019) can also be used to track eye movements with a camera, but some only work offline (Bellet et al., 2019; Mathis et al., 2018).

There are many other “brain and peripheral sensors” that can provide informative measures to complement behavioral outputs (e.g. heart rate, skin conductivity). Check OpenBCI for open-source, low-cost products (Frey, 2016). If you need precise control over the timing of different auditory and visual events, consider using validation measures with external devices (toolkits, such as the Black Box Toolkit [Plant et al., 2004] can be helpful). Before buying expensive equipment, check whether someone in your community already has the tool you need, and importantly, if whether is compatible with the rest of your toolkit, such as response devices, available ports, eye trackers, but also software and your operating system.

Scaling up data collection

If you conclude that luminosity, sounds, eye movements, and other factors will not affect the behavioral variables of interest, you can try to scale things up by testing batches of subjects in parallel, for example in a classroom with multiple terminals. This way you can welcome and guide the subjects through the written instructions collectively, and data collection will be much faster. For parallel testing, make sure that the increased level of distractions does not negatively affect subjects’ performance, and that your code runs 100% smoothly (Table 2). Consider running your experiment with more flexible and maybe cheaper setups, such as tablets (Linares et al., 2018). Alternatively, you can take your experiment online (Gagné & Franzen, 2021; Lange et al., 2015; Sauter et al., 2020). An online experiment speeds up data collections by orders of magnitude (Difallah et al., 2018; Stewart et al., 2017). However, it comes at the cost of losing experimental control, leading to possibly noisier data (Bauer et al., 2020; Crump et al., 2013; Gagné & Franzen, 2021; Thomas & Clifford, 2017). In addition, you will need approximately 30% more participants in an online experiment for the same statistical power if done in the lab, although this estimation depends on the task (Gillan & Rutledge, 2021). Making sure your subjects understand the instructions (see tips in Step 7) and filtering out those subjects that demonstrably do not understand the task (e.g. by asking comprehension questions after the instructions) also has an impact on data quality. Make sure you control for more technical aspects, such as enforcing full-screen mode and guarding your experiments against careless responding (Dennis et al., 2018; Zorowitz et al. 2021). Keep in mind that online crowdsourcing experiments come with their own

Table 2 Top 10 most common coding and data handling errors committed by the authors when doing psychophysics, and how to avoid them. These are loosely sorted by type of error (crashes, incorrect runs, saving data issues), not by frequency

Common mistake	How to avoid it
1) The code breaks in the middle of a session, and all data are lost.	Save your data at the end of each block or, if possible, at the end of each trial.
2) Your code breaks when a certain key is hit, or when secondary external hardware (e.g. eye tracker) unexpectedly stops sending signals.	Check which keys are assigned in your paradigm, and which lead to the interruption of the program. Check in advance what happens if external hardware problems emerge. Perform a *crash test* of your code to make sure it is resilient to wrong keys being hit, or keys being hit at the wrong time.
3) You made an "improvement" just before the experimental session. Your code now breaks unexpectedly or doesn't run at all during data collection.	Never use untested code.
4) Some software sends a notification, such as software updates, in the middle of a session. The experiment is interrupted, and the subject might not even notify you.	Switch off all software you don't need, disable automatic updates. Disable the internet connection.
5) The randomization of stimuli or conditions is wrong, or identical for all subjects.	Make sure to use a different random seed (whenever you want your data to be independent) and save it along with the other variables. Inspect the distribution of conditions in your data generated by your randomization.
6) Your subject is not doing what they should and you don't notice.	Have a control screen or a remote connection to mirror the subject's display (e.g. with Chrome Remote Desktop), but make sure it will not introduce delays. There, also print ongoing performance measures.
7) You overwrite data/code from earlier sessions or subjects. These data are now lost.	Add a line of code that checks whether the filename where you want to store the data already exists. Backup output directory regularly through git. Alternatively or additionally, use timestamps in file names (with second resolution) and and back up your data automatically (e.g., on the cloud).
8) You save participant data with the wrong identifier and later cannot assign it correctly.	Use multiple identifiers to name a file: subject and session ID + date and time + computer ID, for example.
9) You decided at some point to adjust "constant" experimental parameters during data collection. Now, which participants saw what?	Define all experimental parameters at the beginning of your code, preferably in a flexible format such as a python *dictionary,* and save them in a separate log file for each session or include them in your table repeatedly for each trial.
10) After data collection, you start having doubts about the timing of events, and the temporal alignment with continuous data, possibly stored on another device (fMRI, eye tracking).	Save "time stamps" in your data table for each event of a trial (fixation onset, stimulus onset, etc.). Make sure your first event is temporally aligned to the onset of continuous data.

set of ethical concerns regarding payment and exploitation and an extra set of challenges regarding the experimental protocol (Gagné & Franzen, 2021) that need to be taken into consideration (Step 4).

Use open-source software

Write your code in the most appropriate programming language, especially if you do not have strong preferences yet. Python, for example, is open-source, free, versatile, and currently the go-to language in data science (Kaggle, 2019) with plenty of tutorials for all levels of proficiency. PsychoPy is a great option to implement your experiment, should you choose to do it in Python. If you have strong reasons to use Matlab, Psychtoolbox (Borgo et al., 2012) is a great tool, too. If you are considering running your experiment on a tablet or even a smartphone, you could use StimuliApp (Marin-Campos et al., 2020). Otherwise check Hans Strasburger's page (Strasburger, 1994) that has provided a comprehensive and up-to-date overview of different tools, among other technical tips, for the last 25 years.

Step 4. Submit early to the ethics committee

This is a mandatory yet often slow and draining step. Do not take this step as a mere bureaucratic one, and instead think actively and critically about your own ethics. Do it early to avoid surprises that could halt your progress. Depending on the institution, the whole process can take several months. In your application, describe your experiment in terms general enough as to accommodate for later changes in the design that will inevitably occur. This is of course without neglecting potentially relevant ethical issues, especially if your target population can be considered vulnerable (such as patients or minors). You will have to describe factors concerning the sample, such as the details of participant recruitment, planned and justified sample sizes (see Step 3), and details of data anonymization and protection, making sure you comply with existing regulations (e.g. General Data Protection Regulation in the European Union). For some experiments, ethical concerns might be inherent to the task design, for example when you use instructions that leave the subject in the dark about the concepts that are studied, or purposefully distract their attention from them (*deceiving participants*; Field & Hole, 2003). You should also provide the consent form that participants will sign. Each committee has specific requirements (e.g. regarding participant remuneration, see below), so ask more seasoned colleagues for their documents and experiences, and start from there. Often, the basic elements of an ethics application are widely recyclable, and this is the one of the few cases in research where copy-pasting is highly recommendable. Depending on your design, some ethical aspects will be more relevant than others. For a more complete review of potential points of ethical concern, especially in psychological experiments, we refer the reader to the textbook by Field and Hole (Field & Hole, 2003). Keep in mind that as you want to go through the least rounds of review as possible, you should make sure you are abiding by all the rules.

Pay your subjects generously

Incentivize your subjects to perform well, for example by offering a bonus if they reach a certain performance level, but let them know that it is normal to make errors. Find the right trade-off between the baseline show-up payment and the bonus: if most of the payment is show-up fee, participants may not be motivated to do well. If most of it is a performance bonus, poorly performing participants might lose their motivation and drop out, which might introduce a selection bias, or your study can be considered exploitative. Beware that the ethics committee will ultimately decide whether a specific payment scheme is considered fair or not. In our experience, a bonus that adds up to 50–100% of the show-up fee to the remuneration is a good compromise. Alternatively, social incentives, such as the challenge to beat a previous score, can be effective in increasing the motivation of your subjects (Crawford et al., 2020). Regardless of your payment strategy, your subjects should always have the ability to leave the experiment at any point without losing their accumulated benefits (except for some eventual bonus for finishing the experiment). If you plan to run an online experiment, be aware that your subjects might be more vulnerable to exploitation than subjects in your local participant pool (Semuels, 2018). The average low payment in crowdsourcing platforms biases us to pay less than what is ethically acceptable. Do not pay below the minimum wage in the country of your subjects, and significantly above the average wage if it is a low-income country. It will likely be above the average payment on the platform, but still cheaper than running the experiment in the lab, where you would have to pay both the subject and the experimenter. Paying well is not only ethically correct, it will also allow you to filter for best performers and ensure faster and higher data quality (Stewart et al., 2017). Keep in mind that rejecting a participant's experiment (e.g., because their responses suggest they were not attentive to the experiment) can have ramifications to their earning ability in other tasks on the hosting platform, so you should avoid rejecting experiments unless you have strong reason to believe the participant is a bot not a human.

Step 5. Polish your experimental design through piloting

Take time to run pilots to fine-tune your task parameters, especially for the most innovative elements in your task. Pilot yourself first (Diaz, 2020). Piloting lab mates is common practice, and chances are that after debriefing, they will provide good suggestions on improving your paradigm, but it is preferable to hire paid volunteers also for piloting instead of coercing (even if involuntarily) your lab mates into participation (see Step 7). Use piloting to adjust design parameters, including the size and duration of stimuli, masking, the duration of the inter-trial interval, and the modality of the response and feedback. In some cases, it is worth considering using online platforms to run pilot studies, especially when you want to sweep through many parameters (but see Step 2).

Trial difficulty and duration

Find the right pace and difficulty for the experiment to minimize boredom, tiredness, overload, or impulsive responses (Kingdom & Prins, 2016). If choices represent your main behavioral variable of interest, you probably want subjects' overall performance to fall at an intermediate level far from perfect and far from chance, so that changes in conditions lead to large choice variance. If the task is too hard, subjects will perform close to chance level, and you might not find any signature of the process under study. If the task is too easy, you might observe *ceiling effects* (Garin, 2014), and the choice patterns will not be informative either (although reaction times may). In the general case, you might simulate your computational model on your task with different difficulty levels to see which provides the maximum power, in a similar way that can be done to adjust the sample size (see Step 2 and Fig. 2b). As a rule of thumb, some of the authors typically find that an average performance roughly between 70 and 90% provides the best power for two-alternative forced-choice tasks. If you want to reduce the variability of subject performance, or if you are interested in studying individual psychophysical thresholds, consider using an adaptive procedure (Cornsweet, 1962; Kingdom & Prins, 2016; Prins, 2013) to adjust the difficulty for each subject individually. Make conscious decisions on what aspects of the task should be fixed-paced or self-paced. To ease subjects into the task, you might want to include a practice block with very easy trials that become progressively more difficult, for example by decreasing the event duration or the stimulus contrast (i.e. *fading*; Pashler & Mozer, 2013). Make sure you provide appropriate instructions as these new elements are added. If you avoid overwhelming their attentional capacities, the subjects will more rapidly automatize parts of the process (e.g. which cues are associated with a particular rule, key-response mappings, etc.).

Perform sanity checks on pilot data

Use pilot data to ensure that the subjects' performance remains reasonably stable across blocks or experimental sessions, unless you are studying learning processes. Stable estimates require a certain number of trials, and the right balance for this trade-off needs to be determined through experience and piloting. A large *lapse rate* could signal poor task engagement (Fetsch, 2016) (Step 8); in some experiments, however, it may also signal failure of memory retrieval, exploration, or another factor of interest (Pisupati et al., 2019).

Make sure that typical findings are confirmed (e.g. higher accuracy and faster reaction times for easier trials, preference for higher rewards, etc.) and that most responses occur within the allowed time window. Sanity checks can reveal potential bugs in your code, such as incorrectly saved data or incorrect assignment of stimuli to task conditions (Table 2), or unexpected strategies employed by your subjects. The subjects might be using superstitious behavior or alternative strategies that defeat the purpose of the experiment altogether (e.g. people may close their eyes in an auditory task while you try to measure the impact of visual distractors). In general, subjects will tend to find the path of least resistance towards the promised reward (money, course credit, etc.).

Finally, debrief your pilot subjects to find out what they did or did not understand (see also Step 7), and ask open-ended questions to understand which strategies they used. Use their comments to converge on an effective set of instructions and to simplify complicated corners of your design.

Exclusion criteria

Sanity checks can form the basis of your exclusion criteria, e.g. applying cutoff thresholds regarding the proportion of correct trials, response latencies, lapse rates, etc. Make sure your exclusion criteria are orthogonal to your main question, i.e. that they do not produce any systematic bias on your variable of interest. You can decide the exclusion criteria after you collect a cohort of subjects, but always make decisions about which participants (or trials) to exclude *before* testing the main hypotheses in that cohort. Proceed with special caution when defining exclusion criteria for online experiments, where performance is likely more heterogeneous and potentially worse. Do not apply the same criteria to online and in-lab experiments. Instead, run a dedicated set of pilots to define the appropriate criteria. All excluded subjects should be reported in the manuscript and, and their data shared together with the other subjects (see Step 10).

Including pilot data in the final analyses

Be careful about including pilot data in your final cohort. If you decide to include pilot data, you must not have tested your main hypothesis on that data; otherwise it would be considered scientific malpractice (see *p-hacking,* Step 6). The analyses you run on pilot data prior to deciding whether to include it in your main cohort must be completely orthogonal to your main hypothesis (e.g. if your untested main hypothesis is about the difference in accuracy between two conditions, you can perform sanity checks to assess whether the overall accuracy of the participants is in a certain range, if your untested main hypothesis is about the difference in accuracy between two conditions). If you do include pilot data in your manuscript, be explicit about what data were pilot data and what was the main cohort in the paper.

Step 6. Preregister or replicate your experiment

An alarming proportion of researchers in psychology reports to have been involved in some form of questionable research practice (Fiedler & Schwarz, 2016; John et al., 2012). Two common forms of questionable practices, p-hacking and HARKing (Stroebe et al., 2012), increase the likelihood of obtaining false positive results. In p-hacking (Simmons et al., 2011), significance tests are not corrected for testing multiple alternative hypotheses (Benjamini & Hochberg, 2000). For instance, it might be tempting to use the *median* as the dependent variable, after having seen that the *mean* gave an unsatisfactory outcome, without correcting for having performed two tests. HARKing refers to the formulation of a hypothesis *after* the results are known, pretending that the hypothesis was a priori (Kerr, 1998). Additionally, high-impact journals have a bias for positive findings with sexy explanations, while negative results often remain unpublished (Ioannidis, 2005; Rosenthal, 1979). These practices posit a substantial threat to the efficiency of research, and they are believed to underlie the replication crisis in psychology and other disciplines (Open Science Collaboration, 2015). Ironically, the failure to replicate is highly replicable (Klein et al., 2014, 2018).

Preregistration

This crisis has motivated the practice of preregistering experiments before the actual data collection (Kupferschmidt, 2018; Lakens, 2019). In practice, this consists of a short document that answers standardized questions about the experimental design and planned statistical analyses. The optimal time for preregistration is once you finish tweaking your experiment through piloting and power analysis (Steps 2–5). Preregistration may look like an extra hassle before data collection, but it will actually often save you time: writing down explicitly all your hypotheses, predictions, and analyses is itself a good sanity check, and might reveal some inconsistencies that lead you back to amending your paradigm. More importantly, it helps to protect from the conscious or unconscious temptation to change the analyses or hypothesis as you go. The text you generate at this point can be reused for the introduction and methods sections of your manuscript. Alternatively, you can opt for registered reports, where you submit a prototypic version of your final manuscript without the results to peer review (Lindsay et al., 2016). If your report survives peer review, it is accepted in principle, which means that whatever the outcome, the manuscript will be published (given that the study was rigorously conducted). Many journals, including high-impact journals such as *eLife*, Nature Human Behavior, and Nature Communications already accept this format. Importantly, preregistration should be seen as "a plan, not a prison" (DeHaven, 2017). Its real value lies in clearly distinguishing between planned and unplanned analyses (Nosek et al., 2019), so it should not be seen as an impediment to testing exploratory ideas. If (part of) your analyses are exploratory, acknowledge it in the manuscript. It does not decrease their value: most scientific progress is not achieved by confirming a priori hypotheses (Navarro, 2020).

Several databases manage and store preregistrations, such as the popular Open Science Framework (OSF.io) or AsPredicted.org, which offers more concrete guidelines. Importantly, these platforms keep your registration private, so there is no added risk of being scooped. Preregistering your analyses does not mean you cannot do exploratory analyses, just that these analyses will be explicitly marked as such. This transparency strengthens your arguments when reviewers read your manuscript, and protects you from involuntarily committing scientific misconduct. If you do not want to register your experiment publicly, consider at least writing a private document in which you detail your decisions before embarking on data collection or analyses. This might be sufficient for most people to counteract the temptation of questionable research practices.

Replication

You can also opt to replicate the important results of your study in a new cohort of subjects (ideally two). In essence, this means that analyses run on the first cohort are exploratory, while the same analyses run on subsequent cohorts are considered confirmatory. If you plan to run new experiments to test predictions that emerged from your findings, include the replication of these findings in the new experiments. For most behavioral experiments, the cost of running new cohorts with the same paradigm is small in comparison to the great benefit of consolidating your results. In general, unless we have a very focused hypothesis or have limited resources, we prefer replication over pre-registration. First, it allows for less constrained analyses on the original cohort data, because you don't tie your hands until replication. Then, by definition, replication is the ultimate remedy to the replication crisis. Finally, you can use both approaches together and preregister before replicating your results.

In summary, preregistration (Lakens, 2019) and replication (Klein et al., 2014, 2018) will help to improve the standards of science, partially by protecting against involuntary malpractices, and will greatly strengthen your results in the eyes of your reviewers and readers. Beware, however—preregistration and replication cannot replace a solid theoretical embedding of your hypothesis (Guest & Martin, 2021; Szollosi et al., 2020).

Step 7. Take care of your subjects

Remember that your subjects are volunteers, not your employees (see Step 4). They are fellow *homo sapiens* helping science progress, so acknowledge that and treat them with kindness and respect (Wichmann & Jäkel, 2018). Send emails with important (but not too much) information well in advance. Set up an online scheduling system where subjects can select their preferred schedule from available slots. This will avoid back and forth emails. If you cannot rely on an existing database for participant recruitment, create one and ask your participants for permission to include them in it (make sure to comply with regulations on data protection). Try to maintain a long and diverse list, but eliminate unreliable participants. For online experiments, where you can typically access more diverse populations, consider which inclusion criteria are relevant to your study, such as native language or high performance in the platform.

Systematize a routine that starts on the participants' arrival

Ideally, your subjects should come fully awake, healthy, and without the influence of any non-prescription drugs or medication that alter perceptual or cognitive processes under study. Have them switch their mobile phones to airplane mode. During the first session, participants might confuse the meaning of events within a trial (e.g. what is fixation, cue, stimulus, response prompt, feedback), especially if they occur in rapid succession. To avoid this, write clear and concise instructions and have your participants read them before the experiment. Ensuring that your subjects all receive the same written set of instructions will minimize variability in task behavior due to framing effects (Tversky & Kahneman, 1989). Showing a demo of the task or including screenshots in the instructions also helps a lot. Conveying the rules so that every participant understands them fully can be a challenge for the more sophisticated paradigms. A good cover story can make all the difference, or example, instead of a deep dive into explaining a probabilistic task, you could use an intuitive "casino" analogy. Allow some time for clarifying questions and comprehension checks, and repeat instructions on the screen during the corresponding stages of the experiment (introduction, practice block, break, etc.). Measure performance on practice trials and do not move to the subsequent stage before some desired level of performance is reached (unless you are precisely interested in the learning process or the performance of naive subjects). If your experimental logic assumes a certain degree of naïveté about the underlying hypothesis (e.g. in experiments measuring the use of different strategies), make sure that your subject does not know about the logic of the experiment, especially if they are a colleague or a friend of previous participants: asking explicitly is often the easiest way. If a collaborator is collecting the data for you, spend some time training them and designing a clear protocol (e.g. checklist including how to calibrate the eye tracker), including troubleshooting. Be their first mock subject, and be there for their first real subject.

Optimize the subjects' experience

Use short blocks that allow for frequent quick breaks (humans get bored quickly), for example every ~5 minutes (Wichmann & Jäkel, 2018). Include the possibility for one or two longer breaks after each 30–40 minutes, and make sure you encourage your subjects between blocks. One strategy to keep motivation high is to gamify your paradigm with elements that do not interfere with the cognitive function under study. For instance, at the end of every block of trials, display the remaining number of blocks. You can also provide feedback after each trial by displaying coin images on the screen, or by playing a sequence of tones (upward for correct response, downward for errors).

Providing feedback

In general, we recommend giving performance feedback after each trial or block, but this aspect can depend on the specific design. At minimum, provide a simple feedback to acknowledge that the subject has responded and what they responded (i.e. an arrow pointing in the direction of the subject's choice). This type of feedback helps maintain subject engagement through the task, especially in the absence of outcome feedback (e. g. (Karsh et al. 2020)). However, regarding outcome feedback, whereas this is mandatory in reinforcement-learning paradigms, it can be counterproductive in other cases. First, outcome feedback influences the next few trials due to win-stay-lose-switch strategies (Abrahamyan et al., 2016; Urai et al., 2017) or other types of superstitious behavior (Ono, 1987). Make sure this nuisance has a very limited impact on your variables of interest, unless you are precisely interested in these effects. Second, participants can use this feedback as a learning signal (Massaro, 1969) which will lead to an increase in performance throughout the session, especially for strategy-based paradigms or paradigms that include confidence reports (Schustek et al., 2019).

Step 8. Record everything

Once you have optimized your design and chosen the right equipment to record the necessary data to test your hypothesis, record everything you can record with your equipment. The dataset you are generating might be useful for others or your future self in ways you cannot predict now. For example, if you are using an eye tracker to ensure fixation, you may as well record pupil size (rapid pupil dilation is a proxy for information processing [Cheadle et al., 2014] and decision confidence [Urai et al., 2017]). If subjects respond with a mouse, it may be a good idea to record all the mouse movements. Note, however, that logging everything does not mean you can analyze everything without having to correct for multiple tests (see also Step 6).

Save your data in a tidy table (Wickham, 2014) and store it in a software-independent format (e.g. a .csv instead of a .mat file) which makes it easy to analyze and share (Step 10). Don't be afraid of redundant variables (e.g. response identity and response accuracy); redundancy enables robustness to correct for possible mistakes. If some modality produces continuous output, such as pupil size or cursor position, save it in a separate file rather than creating Kafkaesque data structures. If you use an eye-tracker or neuroimaging device, make sure you save synchronized timestamps in both data streams for later data alignment (see Table 2). If you end up changing your design after starting data collection—even for small changes—save those version names in a lab notebook. If the lab does not use a lab notebook, start using one, preferably a digital notebook with version control (Schnell, 2015). Mark *all* incidents there, even those that may seem unimportant at the moment. Use version control software such as GitHub to be able to track changes in your code. Back up your data regularly, making sure you comply with the ethics of data handling (see also Steps 4 and 10). Finally, don't stop data collection after the experiment is done. At the end of the experiment, debrief your participants. Ask questions such as "Did you see so-and-so?" or "Tell us about the strategy you used to solve part II" to make sure the subjects understood the task (see Step 5). It is also useful to include an informal questionnaire about the participant at the end of the experiment, e.g. demographics (should you have approval from the ethics committee).

Step 9. Model your data

Most statistical tests rely on some underlying linear statistical model of your data (Lindeløv, 2019). Therefore, data analysis can be seen as modeling. Proposing a statistical model of the data means turning your hypothesis into a set of statistical rules that your experimental data should comply with. Using a model that is tailored to your experimental design can offer you deeper insight into cognitive mechanisms than standard analyses (see below). You can model your data with different levels of complexity, but recall the "keep it simple" mantra: your original questions are often best answered with a simple model. Adding a fancy model to your paper might be a good idea, but only if it adds to the interpretation of your results. See Box 1 for general tips on data analysis.

Box: General tips for data analysis.

- Each analysis should *answer a question*: keep the thread of your story in mind and ask one question at a time.
- Think of several analyses that could falsify your current interpretation, and only rest assured after finding a coherent picture in the cumulative evidence.
- Start by visualizing the results in different conditions using the simplest methods (e.g. means with standard errors).
- Getting a feeling for a method means understanding its assumptions and how your data might violate them. Data violates assumptions in many situations, but not always in a way that is relevant to your findings, so know your assumptions, and don't be a slave to the stats.
- Nonparametric methods (e.g. bootstrap, permutation tests, and cross-validation; see Model fitting day at 't Hart et al., 2021), the Swiss knife of statistics, are often a useful approach because they do not make assumptions about the distribution of the data—but see Wilcox & Rousselet (2018).
- Make sure that you test for interactions when appropriate (Nieuwenhuis et al., 2011).
- If your evidence coherently points to a null finding, use Bayesian statistics to see whether you can formally accept it (Keysers et al., 2020).
- Correct your statistics for multiple comparisons, including those you end up not reporting in your manuscript (e.g. Benjamini & Hochberg, 2000).

Learn how to model your data

Computational modeling might put off the less experienced in statistics or programming. However, modeling is more accessible than most would think. Use regression models (e.g. linear regression for reaction times or logistic regression for choices; Wichmann & Hill, 2001a, b) as a descriptive tool to disentangle different effects in your data. If you are looking for rather formal introductions to model-based analyses (Forstmann & Wagenmakers, 2015; Kingdom & Prins, 2016; Knoblauch & Maloney, 2012), the classical papers by Wichmann & Hill (2001a, b) and more recent guidelines (Palminteri et al., 2017; Wilson & Collins, 2019) are a good start. If you prefer a more practical introduction, we recommend *Statistics for Psychology: A Guide for Beginners (and everyone else)* (Watt & Collins, 2019) or going through some hands-on courses, for example Neuromatch Academy ('t Hart et al., 2021), BAMB! (bambschool.org), or Model-Based Neuroscience Summer School (modelbasedneurosci.com).

The benefits of data modeling

Broadly, statistical and computational modeling can buy you four things: (i) *model fitting*, to quantitatively estimate relevant effects and compare them between different conditions or populations, (ii) *model validation*, to test whether your conceptual model captures how behavior depends on the experimental variables, (iii) *model comparison,* to determine quantitatively which of your hypothesized models is best supported by your data, and (iv) *model predictions* that are derived from your data and can be tested in new experiments. On a more general level, computational modeling can constrain the space of possible interpretations of your data, and therefore contributes to reproducible science and more solid theories of the mind (Guest & Martin, 2021). See also Step 2.

Model fitting

There are packages or toolboxes that implement model fitting for most regression analyses (Bürkner, 2017; Seabold & Perktold, 2010) and standard models of behavior, such as the DDM (Shinn et al., 2020; Wiecki et al., 2013) or reinforcement learning models (e.g. Ahn et al., 2017; Daunizeau et al., 2014). For models that are not contained in statistical packages, you can implement custom model fitting in three steps: (1) Formalize your model as a series of computational, parameterized operations that transform your stimuli and other factors into behavioral reports (e.g. choice and/or response times). Remember that you are describing a probabilistic model, so at least one operation must be noisy. (2) Write down the likelihood function, i.e. the probability of observing a sequence of responses under your model, as a function of model parameters. Lastly, (3) use a maximization procedure (e.g. function *fmincon* in matlab or *optimize* in python, or learn how to use Bayesian methods as implemented in the cross-platform package *Stan* [mc-stan.org]) to find the parameters that maximize the likelihood of your model for each participant individually—the so-called maximum-likelihood (ML) parameters. This can also be viewed as finding the parameters that minimize the *loss function*, or the model error on predicting subject behavior. Make sure your fitting procedure captures what you expect by validating it on synthetic data, where you know the true parameter values (Heathcote et al., 2015; Palminteri et al., 2017; Wilson & Collins, 2019). Compute uncertainty (e.g. confidence intervals) about the model parameters using bootstrap methods (parametric bootstrapping if you are fitting a sequential model of behavior, classical bootstrapping otherwise). Finally, you may want to know whether your effect is consistent across subjects, or whether the effect differs between different populations, in which case you should compute confidence intervals across subjects. Sometimes, subjects' behavior differs qualitatively and cannot be captured by a single model. In these cases, Bayesian model selection allows you to accommodate the possible heterogeneity of your cohort (Rigoux et al., 2014).

Model validation

After fitting your model to each participant, you should validate it by using the fitted parameter values to *simulate* responses, and compare them to behavioral patterns of the participant (Heathcote et al., 2015; Wilson & Collins, 2019). This control makes sure that the model not only fits the data but can also perform the task itself while capturing the qualitative effects in your data (Palminteri et al., 2017).

Model comparison

In addition to your main hypothesis, always define one or several "null models" that implement alternative hypotheses and compare them using model-comparison techniques (Heathcote et al., 2015; Wilson & Collins, 2019). In general, use cross validation for model selection, but be aware that both cross validation and information criteria (Akaike/Bayesian information criterion [AIC/BIC]) are imprecise metrics when your dataset is small (<100 trials; Varoquaux, 2018); in this case, use fully Bayesian methods if available (Daunizeau et al., 2014). In the case of sequential tasks (e.g. in learning studies), where the different trials are not statistically independent, use block cross-validation instead of cross-validation (Bergmeir & Benítez, 2012). For nested models—when the complex model includes the simpler one—you can use the likelihood-ratio test to perform significance testing.

Model prediction

Successfully predicting behavior in novel experimental data is the Holy Grail of the epistemological process. Here, one should make predictions about the cognitive process in a wider set of behavioral measures or conditions. For example, you might fit your model on reaction times and use those fits to make predictions about a secondary variable (Step 8), such as choices or eye movements, or generate predictions from the model in another set of experimental conditions.

Step 10. Be transparent and share

Upon publication, share everything needed to replicate your findings in a repository or shared database (see Table 1). That includes your data and code. Save your data in a tidy table (Wickham, 2014) with one trial per line and all the relevant experimental and behavioral variables as columns. Try to use a common data storage format, adopted within or outside your lab. Aim at properly documented data and code, but don't let that be the reason not to share. After all, bad code is better than no code (Barnes, 2010; Gleeson et al., 2017). If possible, avoid using proprietary software for your code, analyses, and data (e.g. share a .csv instead of a .mat file). We recommend the use of python or R notebooks (Rule et al., 2019) to develop your analyses and git for version control (Perez-Riverol et al., 2016). Notebooks make it easier to share code with the community, but also with advisors or colleagues, when asking for help.

Discussion

Our goal here was to provide practical advice, rather than illuminating the theoretical foundations for designing and running behavioral experiments with humans. Our recommendations, or steps, span the whole process involved in designing and setting up an experiment, recruiting and caring for the subjects, and recording, analyzing, and sharing data. Through the collaborative effort of collecting our personal experiences and writing them down in this manuscript, we have learned a lot. In fact, many of these steps were learned after painfully realizing that doing the exact opposite was a mistake. We thus wrote the "practical guide" we wished we had read when we embarked on the adventure of our first behavioral experiment. Some steps are therefore rather subjective, and might not resonate with every reader, but we remain hopeful that most of them are helpful to overcome the practical hurdles inherent to performing behavioral experiments with humans.

Acknowledgements The authors thank Daniel Linares for useful comments on the manuscript. The authors are supported by the Spanish Ministry of Economy and Competitiveness (RYC-2017-23231 to A.H.), the "la Caixa" Banking Foundation (Ref: LCF/BQ/IN17/11620008, H.S.), the European Union's Horizon 2020 Marie Skłodowska-Curie grant (Ref: 713673, H.S.), and the European Molecular Biology Organization (Ref: EMBO ALTF 471-2021, H.S.). JB was supported by the Fyssen Foundation and by the Bial Foundation (Ref: 356/18). S.S-F. is funded by Ministerio de Ciencia e Innovación (Ref: PID2019-108531GB-I00 AEI/FEDER), AGAUR Generalitat de Catalunya (Ref: 2017 SGR 1545), and the FEDER/ERFD Operative Programme for Catalunya 2014-2020.

References

't Hart, B. M., Achakulvisut, T., Blohm, G., Kording, K., Peters, M. A. K., Akrami, A., Alicea, B., Beierholm, U., Bonnen, K., Butler, J. S., Caie, B., Cheng, Y., Chow, H. M., David, I., DeWitt, E., Drugowitsch, J., Dwivedi, K., Fiquet, P.-É., Gu, Q., & Hyafil, A. (2021). *Neuromatch Academy: a 3-week, online summer school in computational neuroscience*. https://doi.org/10.31219/osf.io/9fp4v

Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M., & Gardner, J. L. (2016). Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(25), E3548-57. https://doi.org/10.1073/pnas.1518786113

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24–57. https://doi.org/10.1162/CPSY_a_00002

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/10.1037/met0000337

Barnes, N. (2010). Publish your computer code: it is good enough. *Nature*, *467*(7317), 753. https://doi.org/10.1038/467753a

Bauer, B., Larsen, K. L., Caulfield, N., Elder, D., Jordan, S., & Capron, D. (2020). *Review of Best Practice Recommendations for Ensuring High Quality Data with Amazon's Mechanical Turk*. https://doi.org/10.31234/osf.io/m78sf

Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511541933

Bellet, M. E., Bellet, J., Nienborg, H., Hafed, Z. M., & Berens, P. (2019). Human-level saccade detection performance using deep neural networks. *Journal of Neurophysiology*, *121*(2), 646–661. https://doi.org/10.1152/jn.00601.2018

Benjamini, Y., & Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60–83. https://doi.org/10.3102/10769986025001060

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Borgo, M., Soranzo, A., & Grassi, M. (2012). Psychtoolbox: sound, keyboard and mouse. In *MATLAB for Psychologists* (pp. 249–273). Springer New York. https://doi.org/10.1007/978-1-4614-2197-9_10

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), 16. https://doi.org/10.5334/joc.72

Bürkner, P.-C. (2017). brms: an *R* package for bayesian multilevel models using *stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

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

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–1441. https://doi.org/10.1016/j.neuron.2014.01.020

Chen, Z., & Whitney, D. (2020). Perceptual serial dependence matches the statistics in the visual world. *Journal of Vision*, *20*(11), 619. https://doi.org/10.1167/jov.20.11.619

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Cornsweet, T. N. (1962). The Staircase-Method in Psychophysics. *The American Journal of Psychology*, *75*(3), 485. https://doi.org/10.2307/1419876

Crawford, J. L., Yee, D. M., Hallenbeck, H. W., Naumann, A., Shapiro, K., Thompson, R. J., & Braver, T. S. (2020). Dissociable effects of monetary, liquid, and social incentives on motivation and cognitive control. *Frontiers in Psychology*, *11*, 2212. https://doi.org/10.3389/fpsyg.2020.02212

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *Plos One*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, *10*(1), e1003441. https://doi.org/10.1371/journal.pcbi.1003441

DeHaven, A. (2017, May 23). *Preregistration: A Plan, Not a Prison*. Center for Open Science. https://www.cos.io/blog/preregistration-plan-not-prison

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). Mturk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3233954

Diaz, G. (2020, April 27). *Highly cited publications on vision in which authors were also subjects*. Visionlist. http://visionscience.com/pipermail/visionlist_visionscience.com/2020/004205.html

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*, 135–143. https://doi.org/10.1145/3159652.3159661

Dykstra, O. (1966). The orthogonalization of undesigned experiments. *Technometrics : A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *8*(2), 279. https://doi.org/10.2307/1266361

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y

Fetsch, C. R. (2016). The importance of task design and behavioral control for understanding the neural basis of cognitive functions. *Current Opinion in Neurobiology*, *37*, 16–22. https://doi.org/10.1016/j.conb.2015.12.002

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Field, A., & Hole, G. J. (2003). *How to Design and Report Experiments* (1st ed., p. 384). SAGE Publications Ltd.

Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). (2015). *An Introduction to Model-Based Cognitive Neuroscience*. Springer New York. https://doi.org/10.1007/978-1-4939-2236-9

Frey, J. (2016). Comparison of an Open-hardware Electroencephalography Amplifier with Medical Grade Device in Brain-computer Interface Applications. *Proceedings of the 3rd International Conference on Physiological Computing Systems*, 105–114. https://doi.org/10.5220/0005954501050114

Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R., & Menke, L. (2016). Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1240–1244. https://doi.org/10.1177/1541931213601289

Gagné, N., & Franzen, L. (2021). *How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience*. https://doi.org/10.31234/osf.io/nt67j

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, *32*, 148–155. https://doi.org/10.1016/j.conb.2015.04.003

Garin, O. (2014). Ceiling Effect. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 631–633). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_296

Gelman, A., & Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gescheider. (2013). *Psychophysics: The Fundamentals*. Psychology Press. https://doi.org/10.4324/9780203774458

Gillan, C. M., & Rutledge, R. B. (2021). Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience*. https://doi.org/10.1146/annurev-neuro-101220-014053

Gleeson, P., Davison, A. P., Silver, R. A., & Ascoli, G. A. (2017). A commitment to open source in neuroscience. *Neuron*, *96*(5), 964–965. https://doi.org/10.1016/j.neuron.2017.10.013

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*(3), 750–763. https://doi.org/10.3758/s13423-015-0968-3

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience* (pp. 25–48). Springer New York. https://doi.org/10.1007/978-1-4939-2236-9_2

Hosp, B., Eivazi, S., Maurer, M., Fuhl, W., Geisler, D., & Kasneci, E. (2020). RemoteEye: An open-source high-speed remote eye tracker : Implementation insights of a pupil- and glint-detection algorithm for high-speed remote eye tracking. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01305-2

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jazayeri, M., & Afraz, A. (2017). Navigating the neural space in search of the neural code. *Neuron*, *93*(5), 1003–1014. https://doi.org/10.1016/j.neuron.2017.02.019

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kaggle. (2019). *State of Data Science and Machine Learning 2019*. https://www.kaggle.com/kaggle-survey-2019

Karsh, N., Hemed, E., Nafcha, O., Elkayam, S. B., Custers, R., & Eitam, B. (2020). The Differential Impact of a Response's Effectiveness and its Monetary Value on Response Selection. *Scientific Reports*, *10*(1), 3405. https://doi.org/10.1038/s41598-020-60385-9

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. https://doi.org/10.1038/s41593-020-0660-4

Kingdom, F., & Prins, N. (2016). *Psychophysics* (p. 346). Elsevier. https://doi.org/10.1016/C2012-0-01278-1

Klein, Richard A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R A, Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. Springer New York. https://doi.org/10.1007/978-1-4614-4475-6

Koenderink, J. J. (1999). Virtual Psychophysics. *Perception*, *28*(6), 669–674. https://doi.org/10.1068/p2806ed

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *Plos One*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Kupferschmidt, K. (2018). More and more scientists are preregistering their studies. Should you? *Science*. https://doi.org/10.1126/science.aav4786

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, Daniël. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, Daniel. (2019). *The value of preregistration for psychological science: A conceptual analysis*. https://doi.org/10.31234/osf.io/jbh4w

Lakens, Daniel. (2021). *Sample Size Justification*. https://doi.org/10.31234/osf.io/9d3yf

Lange, K., Kühn, S., & Filevich, E. (2015). "just another tool for online studies" (JATOS): an easy solution for setup and management of web servers supporting online studies. *Plos One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Linares, D., Marin-Campos, R., Dalmau, J., & Compte, A. (2018). Validation of motion perception of briefly displayed images using a tablet. *Scientific Reports*, *8*(1), 16056. https://doi.org/10.1038/s41598-018-34466-9

Lindeløv, J. K. (2019, June 28). *Common statistical tests are linear models*. Lindeloev.Github.Io. https://lindeloev.github.io/tests-as-linear/

D. S. Lindsay, D. J. Simons, Scott O. Lilienfeld. (2016). Research Preregistration 101 – Association for Psychological Science – APS. *APS Observer*.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. *ArXiv*.

Mantiuk, R., Kowalik, M., Nowosielski, A., & Bazyluk, B. (2012). Do-It-Yourself Eye Tracker: Low-Cost Pupil-Based Eye Tracker for Computer Graphics Applications. *Lecture Notes in Computer Science (Proc. of MMM 2012)*, *7131*, 115–125.

Marin-Campos, R., Dalmau, J., Compte, A., & Linares, D. (2020). *StimuliApp: psychophysical tests on mobile devices*. https://doi.org/10.31234/osf.io/yqd4c

Massaro, D. W. (1969). The effects of feedback in psychophysical tasks. *Perception & Psychophysics*, *6*(2), 89–91. https://doi.org/10.3758/BF03210686

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Musall, S., Urai, A. E., Sussillo, D., & Churchland, A. K. (2019). Harnessing behavioral diversity to understand neural computations for cognition. *Current Opinion in Neurobiology*, *58*, 229–238. https://doi.org/10.1016/j.conb.2019.09.011

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Navarro, D. (2020). *Paths in strange spaces: A comment on preregistration*. 10.31234/osf.io/wxn58

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Niv, Y. (2020). *The primacy of behavioral research for understanding the brain*. https://doi.org/10.31234/osf.io/y8mxe

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Ono, K. (1987). Superstitious behavior in humans. *Journal of the Experimental Analysis of Behavior*, *47*(3), 261–271. https://doi.org/10.1901/jeab.1987.47-261

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(4), 1162–1173. https://doi.org/10.1037/a0031679

Pereira, T. D., Shaevitz, J. W., & Murthy, M. (2020). Quantifying behavior to understand the brain. *Nature Neuroscience*, *23*(12), 1537–1549. https://doi.org/10.1038/s41593-020-00734-z

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., & Vizcaíno, J. A. (2016). Ten simple rules for taking advantage of git and github. *PLoS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2019). Lapses in perceptual judgments reflect exploration. *BioRxiv*. https://doi.org/10.1101/613828

Plant, R. R., Hammond, N., & Turner, G. (2004). Self-validating presentation and response timing in cognitive paradigms: how and why? *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *36*(2), 291–303. https://doi.org/10.3758/bf03195575

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

Quax, S. C., Dijkstra, N., van Staveren, M. J., Bosch, S. E., & van Gerven, M. A. J. (2019). Eye movements explain decodability during perception and cued attention in MEG. *Neuroimage*, *195*, 444–453. https://doi.org/10.1016/j.neuroimage.2019.03.069

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Read, J. C. A. (2015). The place of human psychophysics in modern neuroscience. *Neuroscience*, *296*, 116–129. https://doi.org/10.1016/j.neuroscience.2014.05.036

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *Neuroimage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Computational Biology*, *15*(7), e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4). https://doi.org/10.3390/brainsci10040251

Schnell, S. (2015). Ten simple rules for a computational biologist's laboratory notebook. *PLoS Computational Biology*, *11*(9), e1004385. https://doi.org/10.1371/journal.pcbi.1004385

Schustek, P., Hyafil, A., & Moreno-Bote, R. (2019). Human confidence judgments reflect reliability-based hierarchical integration of contextual information. *Nature Communications*, *10*(1), 5430. https://doi.org/10.1038/s41467-019-13472-z

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Semuels, A. (2018, January 23). *The Online Hell of Amazon's Mechanical Turk* . The Atlantic. https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Shinn, M., Lam, N. H., & Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife*, *9*. https://doi.org/10.7554/eLife.56938

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Stallard, N., Todd, S., Ryan, E. G., & Gates, S. (2020). Comparison of Bayesian and frequentist group-sequential clinical trial designs. *BMC Medical Research Methodology*, *20*(1), 4. https://doi.org/10.1186/s12874-019-0892-8

Stein, H., Barbosa, J., Rosa-Justicia, M., Prades, L., Morató, A., Galan-Gadea, A., Ariño, H., Martinez-Hernandez, E., Castro-Fornieles, J., Dalmau, J., & Compte, A. (2020). Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nature Communications*, *11*(1), 4250. https://doi.org/10.1038/s41467-020-18033-3

Steiner, M. D., & Frey, R. (2021). Representative design in psychological assessment: A case study using the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001036

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Strasburger, H. (1994, July). *Strasburger's psychophysics software overview* . Strasburger's Psychophysics Software Overview. http://www.visionscience.com/documents/strasburger/strasburger.html

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Szollosi, A., Liang, G., Konstantinidis, E., Donkin, C., & Newell, B. R. (2019). Simultaneous underweighting and overestimation of rare events: Unpacking a paradox. *Journal of Experimental Psychology: General*, *148*(12), 2207–2217. https://doi.org/10.1037/xge0000603

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Thompson, W. H., Wright, J., Bissett, P. G., & Poldrack, R. A. (2019). Dataset Decay: the problem of sequential analyses on open datasets. *BioRxiv*. https://doi.org/10.1101/801696

Tversky, A, & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, Amos, & Kahneman, D. (1989). Rational choice and the framing of decisions. In B. Karpak & S. Zionts (Eds.), *Multiple criteria decision making and risk analysis using microcomputers* (pp. 81–126). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-74919-3_4

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*, 14637. https://doi.org/10.1038/ncomms14637

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*(Pt A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Waskom, M. L., Okazawa, G., & Kiani, R. (2019). Designing and interpreting psychophysical investigations of cognition. *Neuron*, *104*(1), 100–112. https://doi.org/10.1016/j.neuron.2019.09.016

Watt, R., & Collins, E. (2019). *Statistics for Psychology: A Guide for Beginners (and everyone else)* (1st ed., p. 352). SAGE Publications Ltd.

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

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329. https://doi.org/10.3758/BF03194545

Wichmann, F. A., & Jäkel, F. (2018). Methods in Psychophysics. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–42). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn507

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). https://doi.org/10.18637/jss.v059.i10

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Wilcox, R. R., & Rousselet, G. A. (2018). A guide to robust statistical methods in neuroscience. *Current Protocols in Neuroscience*, *82*, 8.42.1-8.42.30. https://doi.org/10.1002/cpns.41

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

Wontorra, H. M., & Wontorra, M. (2011). Early apparatus-based experimental psychology, primarily at Wilhelm Wundt's Leipzig Institute

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., & Huk, A. C. (2017). Functional dissection of signal and noise in MT and LIP during decision-making. *Nature Neuroscience*, *20*(9), 1285–1292. https://doi.org/10.1038/nn.4611

Yiu, Y.-H., Aboulatta, M., Raiser, T., Ophey, L., Flanagin, V. L., Zu Eulenburg, P., & Ahmadi, S.-A. (2019). DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of Neuroscience Methods*, *324*, 108307. https://doi.org/10.1016/j.jneumeth.2019.05.016

Yoon, J., Blunden, H., Kristal, A. S., & Whillans, A. V. (2019). Framing Feedback Giving as Advice Giving Yields More Critical and Actionable Input. *Harvard Business School*

Zorowitz, S., Niv, Y., & Bennett, D. (2021). *Inattentive responding can induce spurious associations between task behavior and symptom measures*. https://doi.org/10.31234/osf.io/rynhk

Open Practices Statement

No experiments were conducted or data generated during the writing of this manuscript. The code used to perform simulations for power analysis is available at https://github.com/ahyafil/SampleSize.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.