Computational Brain & Behavior
https://doi.org/10.1007/s42113-019-00035-0

ORIGINAL PAPER

Maximizing the Expected Information Gain of Cognitive Modeling via Design Optimization

Daniel W. Heck[1] · Edgar Erdfelder[1]

Abstract
To ensure robust scientific conclusions, cognitive modelers should optimize planned experimental designs a priori in order to maximize the expected information gain for answering the substantive question of interest. Both from the perspective of philosophy of science, but also within classical and Bayesian statistics, it is crucial to tailor empirical studies to the specific cognitive models under investigation before collecting any new data. In practice, methods such as design optimization, classical power analysis, and Bayesian design analysis provide indispensable tools for planning and designing informative experiments. Given that cognitive models provide precise predictions for future observations, we especially highlight the benefits of model-based Monte Carlo simulations to judge the expected information gain provided by different possible designs for cognitive modeling.

Keywords Optimal design · Power analysis · Recovery simulation · Experimental design · Bayes factor design analysis

Recently, Lee et al. (in press) provided recommendations of how to increase the robustness of cognitive modeling. We especially appreciated Lee et al.'s recommendation to rely on preregistration and *Registered Modeling Reports* in cognitive modeling projects. Such tools are especially valuable for testing confirmatory hypotheses that are ubiquitous in cognitive modeling, for instance, when testing the goodness of fit of a model, when validating model parameters by selective influence, when selecting between competing models, or when testing which parameters are affected by specific experimental manipulations.

However, before collecting any new data in a confirmatory (possibly preregistered) study, cognitive modelers should regularly consider the question of how to design the study optimally such that it maximizes the expected information gain for answering the substantive question of interest given a model-based analysis of the data (Myung and Pitt 2009). By doing so, researchers can exploit limited resources such as money, number of participants, or study duration most efficiently. Even though optimizing the experimental design a priori is crucial for ensuring informative and robust scientific conclusions through cognitive modeling, Lee et al. (in press) mentioned this aspect only very briefly, merely stating that model recovery studies "can help diagnose issues like (weak) identifiability with respect to the type and amount of information likely to be available" (p. 9). In a similar vein, Lee et al. (in press) argued against the usefulness of power analysis in general (p. 5). However, when viewed more broadly, considerations of the expected information gain of a planned study (e.g., in terms of statistical power) *are* important. In the present commentary, we highlight the importance of optimizing experimental designs before data collection and elaborate on the practical benefits of Monte Carlo simulations for improving the informativeness of cognitive-modeling studies a priori.

Informative Experimental Designs

Irrespective of whether a cognitive modeling project is confirmatory or exploratory, researchers should always consider whether a planned study is informative for answering a scientific question of interest. Here, we use the broad term of an "informative design" since it encompasses several more specific concepts both in theories of philosophy of science and in various statistical frameworks (e.g., classical or Bayesian statistics).

✉ Daniel W. Heck
heck@uni-mannheim.de

[1] University of Mannheim, Schloss (Room EO 255), 68131 Mannheim, Germany

Published online: 01 August 2019

According to Popper's (2005) logic of scientific discovery, the principle of falsificationism states that theories or models should make nontrivial testable predictions that, when tested rigorously, nevertheless hold empirically. In turn, to test a cognitive model rigorously, it is necessary to optimize the experimental design in a way to increase chances of falsification in case the model is actually false. In a cognitive modeling project, this means that the researcher should select a paradigm, experimental conditions, stimuli, and number of participants that generate data structures for which the cognitive model makes precise, in principle falsifiable predictions. Only then, a statistical test of the model's fit constitutes an informative test of the theory. Within psychology, this important aspect of scientific inference and statistical modeling has been emphasized in a now classical paper by Roberts and Pashler (2000): "only when both theory and data provide substantial constraints does this [good model fit] provide significant evidence for the theory" (p. 359). Importantly, the degree to which theory and data provide constraints completely depends on the specific experimental design. As an extreme example, consider the case of testing a theory of memory decay without actually manipulating the retention interval. Obviously, such an experimental design is not suited to test the theory at all. However, the general principle underlying this (admittedly rather absurd) example directly generalizes to realistic scenarios, since even basic choices of the experimental design (e.g., the type and the number of stimuli or conditions) determine how strongly the range of possible outcomes is restricted by a specific theory.

The importance of informative experimental designs becomes even more obvious when considering Platt's (1964) principle of strong inference. According to Platt, researchers should aim to design an *experimentum crucis* for which two competing theories make exactly opposite predictions. Thereby, the researcher maximizes the expected information gain of the planned experimental design before collecting any data. Within cognitive modeling, this goal can be achieved by deriving critical tests based only on core assumptions of a set of competing cognitive models (Kellen and Klauer 2015). By testing informative data patterns that discriminate between the competitors, such critical tests circumvent the reliance on auxiliary (parametric) assumptions that are usually required by goodness-of-fit tests or model-selection methods. For instance, Kellen and Klauer (2015) derived distinct predictions for confidence-rating receiver operating characteristic (ROC) curves in recognition memory for two general classes of models (i.e., signal detection theory and high-threshold models). Thereby, they obtained a critical test that was tailored to the research question at hand (i.e., discriminating between the two model classes) and thus was more informative than the more common approach of comparing how well specific parametric versions of the models fit to ROC curves.

Statistical Criteria and Approaches for Assessing Information Gain

In practice, the question emerges how to assess the expected information gain of an experimental design before any data are collected. The answer to this question depends on the choice of the statistical framework (e.g., classical or Bayesian statistics) and the statistical criterion of interest (e.g., hypothesis testing, model selection, or precision of estimation).

Power Analysis and Optimal Design

Within classical statistics, the information provided by experimental designs is intimately linked to statistical power analysis within the Neyman-Pearson framework (Neyman and Pearson 1933; for applications to standard tests, see Faul et al. 2007) and to accuracy in parameter estimation (i.e., controlling the width of the confidence interval, Maxwell et al. 2008; or other aspects of optimal experimental design, Berger and Wong 2009). In a typical a priori power analysis, the required overall sample size is determined by computing the minimum sample size for which the Type I and Type II error probabilities of statistical decisions are below the prespecified thresholds α and β, respectively, given a minimal to-be-detected "effect size" (i.e., deviation from the null hypothesis model) in a specific design. However, next to such calculations of the required overall sample size, power analysis also provides an appropriate framework for planning various additional aspects of an informative experimental design to test a cognitive model: the proportions of the total sample that should be assigned to different experimental conditions, the optimum number of items per participant in within-subject designs, the gain in power by including covariates into a model, the selection of the most powerful statistical test, and the calibration of context conditions that maximize the power of the hypothesis test of main interest.

What is more, classical a priori power analysis (Cohen 1988) can be refined and improved in a cognitive-modeling framework by formalizing to-be-detected effects in terms of psychologically meaningful parameters of the relevant model. In contrast, power analysis for standard hypothesis tests usually rests on conventions for standardized effect sizes that are notoriously difficult to agree upon (and that actually have different meanings in different designs). These difficulties largely diminish if to-be-detected effects can be defined in terms of the model's parameters as shown in the "Example: Optimal Design for the Pair-Clustering Model" section (see also Erdfelder et al. 2005; Moshagen 2010). Given the many benefits of power analysis for planning details of an experimental design (and not just for deciding about the total sample size), we disagree with Lee et al. (in press) statement that "unless a modeling project depends critically on a null

hypothesis test and there is only one opportunity for data to be collected, a-priori power analysis does not serve a necessary role" (p. 5).

Model Selection and Model-Recovery Simulations

If more than two hypotheses or models are compared, statistical inference often relies on model-selection methods such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or the minimum description length principle (Myung et al. 2000). In such a scenario, the information provided by an experimental design may be assessed by classification accuracy, the probability of selecting the true, data-generating model out of a set of competing models when using a specific model-selection criterion. Focusing on classification accuracy as one of many possible criteria, Myung and Pitt (2009) developed a formal framework to maximize the utility of an experimental design for discriminating between multiple cognitive models. Essentially, the proposed method searches for the global optimum of all possible levels of a design factor (e.g., the spacing of multiple retention intervals) by maximizing a high-dimensional integral of the utility function over both the parameter space and the data space (weighted by the prior distribution and the likelihood function, respectively). However, even though the authors adapted a simulation-based approach to address the issue of computational complexity (Müller et al. 2004), the implementation of such a highly formalized framework may be intractable or too time-intensive in practice for many researchers.

As a remedy, cognitive modelers can often rely on the less sophisticated, but much simpler approach of Monte Carlo simulations to estimate the expected information gain for a small number of possible experimental designs. As an example of a particularly simple method, Algorithm 1 shows pseudo-code for a standard model-recovery simulation: A set of cognitive models M_1, …, M_J is used to generate and fit simulated data, thus enabling the estimation of the classification accuracy for a specific experimental design d when using a specific model-selection criterion C (e.g., the AIC or BIC). By performing multiple such simulations while varying details of the design d, researchers can easily improve the expected information gain of a planned study in terms of the model-recovery rate (for similar frameworks, see Navarro et al. 2004; Wagenmakers et al. 2004).

Even though heuristic simulations do in no way guarantee that the optimal design is found, they may serve as a simple and effective means of improving the informativeness of a planned study without much effort. For instance, Heck et al. (2017) classified individuals as users of different decision-making strategies using the Bayes factor and a specific version of the minimum description length (i.e., the normalized maximum likelihood). By varying the number of trials in a model-recovery simulation similar to Algorithm 1, the minimum number of responses per participant was determined to ensure high classification accuracies for both methods.

Going beyond the specific criterion of model recovery, simulations are also useful to assess the expected information gain of experimental designs with respect to any other quantity of interest. For instance, Gelman and Carlin (2014) proposed to regularly consider Type S errors (i.e., sign errors that parameter estimates are in the wrong direction) and Type M errors (i.e., that the magnitude of an effect is overestimated). The probabilities for both types of errors can easily be estimated by means of simulations, in turn facilitating the search for an experimental design that minimizes Type S and Type M error rates. As a beneficial side effect of simulating possible outcomes of a study before data collection, the results of a simulation-based search for an informative design may be reported as a justification for choosing a specific design and sample size—which is especially useful for (but not limited to) *Registered Modeling Reports*.

Algorithm 1. Pseudo-code for a standard model-recovery simulation.

1. **for** all replications $r = 1, \ldots, R$ and all models $M_1, \ldots, M_J$ **do**
   1.1. Select a data-generating model M_j.
   1.2. Define plausible model parameters θ_r (e.g., fixed values or sampled from prior distribution).
   1.3. Simulate data y_r for a specific experimental design d conditional on θ_r.
   1.4. Fit all models $M_1, \ldots, M_J$ under consideration.
   1.5. Compute statistical criterion C_r for all models (e.g., AIC, BIC, Bayes factor).
2. Summarize the distribution of C_r (e.g., probability of recovering the data-generating model).

Note. By changing details of the simulated experimental design d across simulations (e.g., sample size, relative number of trials per condition, or data-generating parameters), one can estimate the expected information gain of different experimental designs.

Expected Information Gain in the Bayesian Framework

A priori measures of design optimization do not become less important within the Bayesian framework. Essentially, the planned experimental design should maximize the expected information gain obtained by updating the prior to the posterior distribution (for a formal definition based on the Shannon entropy, see Lindley 1956). For example, in case of a Bayesian t-test, researchers may assess the expected sample size required for a high a priori probability of observing a sufficiently large Bayes factor for or against the null hypothesis before collecting any data (Schönbrodt and Wagenmakers 2018).

The general idea of searching for an experimental design that maximizes the expected evidence of a study (Lindley 1956) easily transfers to the application of Bayes factors in

cognitive modeling. By using the Bayes factor as the statistical criterion C in the simulation sketched in Algorithm 1, researchers can easily search for the minimum sample size or number of trials that are necessary to ensure an expected value of the Bayes factor that represents convincing evidence for or against the substantive question of interest. Simulated distributions of Bayes factors can even be useful if the data have already been collected. For instance, in a large-scale reanalysis, Heck et al. (2018b) obtained ambiguous evidence (i.e., Bayes factors close to one) when regressing dishonest behavior on several basic personality traits (e.g., Conscientiousness, Extraversion) even though the merged dataset included a large number of participants ($N = 5002$). To judge the possible benefits of future studies linking personality to dishonesty, the authors plotted the distribution of simulated Bayes factors as a function of sample size, thereby showing that unfeasibly large sample sizes would be required to draw any firm conclusions.[1]

If the focus is on Bayesian parameter estimation instead of hypothesis testing, simulation studies allow researchers to judge the expected precision of the parameter estimates of a cognitive model for a specific experimental design. As an example, the approach by Arnold et al. (2019) provides a case study of what Lee et al. referred to as a *Registered Modeling Report*. Arnold et al. used a well-established multinomial model of multidimensional source memory (Batchelder and Riefer 1990; Meiser 2014) to investigate whether the binding of context features is a robust phenomenon (as opposed to an aggregation artifact due to fitting traditional, complete-pooling models) and whether different types of encoding during the study phase affect the binding parameter. Before collecting any data, the authors submitted a detailed preregistration plan that specified the substantive hypotheses of interest and details of the experimental design (e.g., sample size and number of trials), but also specifics of the cognitive model (i.e., a Bayesian hierarchical multinomial model with predetermined parameter constraints and prior distributions). By using a Monte Carlo simulation similar to that in Algorithm 1 with the difference of focusing on parameter estimates instead of model selection, the authors assessed the expected precision for estimating the difference in the binding parameters across the two between-subjects conditions (Heck et al. 2018a). Thereby, the sample size and the number of trials were chosen in a way to ensure sufficiently precise parameter estimates in terms of the Bayesian credibility interval.

[1] Notably, even though simulated distributions of Bayes factors may be relevant for planning specific design considerations for future studies, the results do not affect the interpretation of an observed Bayes factor after data collection. This is due to the fact that the Bayes factor obeys the likelihood principle, which means that only the actually observed data (as opposed to hypothetical data) are relevant for statistical inference (Berger and Wolpert, 1988).

Design Choices in Cognitive Modeling

Irrespective of which specific statistical framework and criteria are used, the information gain provided by an experimental design for cognitive modeling hinges on many design choices. Most prominently, the total sample size is a critical variable in cognitive modeling, for instance, when testing selective influence of experimental manipulations on model parameters (e.g., by showing that a standard memory-strength manipulation affects only a targeted memory parameter but not any remaining response-bias parameters). A priori considerations of the required sample size are even important in cognitive-modeling projects that rely on sequential tests and optional stopping to increase efficiency of a statistical test compared to fixed-N designs (e.g., using the Bayes factor or the Wald test, cf. Schnuerch and Erdfelder in press; Schönbrodt et al. 2017; Wald 1947). Given that the resources for empirical studies are always limited, it is crucial to check whether the planned experiment has a realistic chance of answering the substantive question of interest. If the expected information gain provided by the planned design for cognitive modeling is not assessed a priori, a sequential sampling plan might merely provide ambiguous conclusions even after collection of hundreds of participants.

Importantly, the amount of information provided by a study does not only depend on the total sample size but also on other design factors such as the relative sample size per condition in between-subject designs or the number of trials in within-subject designs (for an example, see the "Example: Optimal Design for the Pair-Clustering Model" section). Moreover, the informativeness of cognitive modeling often depends on the presentation format and specific features of the presented stimuli. As an example of modeling long-term memory, the information gain in discriminating between power and exponential models of forgetting curves depends crucially on the spacing of the retention intervals (Myung and Pitt 2009; Navarro et al. 2004). As a second example in judgment and decision-making, consider the classification of participants as users of different decision-making strategies such as take-the-best or weighted-additive (Bröder and Schiffer 2003). In typical choice tasks for investigating strategy use across different environments, the presence or absence of item features (e.g., which of two choice options is recommended by different experts) must be carefully chosen to ensure that each strategy predicts a unique response pattern across items. Next, the decision strategies under consideration are usually formalized as statistical models to classify participants as users of different strategies. However, the discriminatory power of empirical strategy classification depends crucially on the specific combinations of item features used in a study (Heck et al. 2017). As a remedy, instead of searching for informative item configurations heuristically, Jekel et al. (2011) employed Euclidian diagnostic task selection to search for the most diagnostic cue patterns for a given set of decision strategies or models.

Whereas sample-size planning and diagnostic stimulus selection is relevant for statistical inference in general (Maxwell et al. 2008), cognitive modeling in particular offers some unique benefits for design optimization in terms of calibrating the substantively meaningful model parameters. For instance, the parameters in multinomial processing tree models are defined as conditional probabilities of entering different cognitive states or processes (Batchelder and Riefer 1990). Hence, the information that can be gained about a single test-relevant parameter may heavily depend on the specific values of the remaining (nuisance) parameters. For example, in standard source-memory paradigms (Arnold et al. 2019), participants have to learn items from two sources (e.g., words presented in blue and red). In the test phase, the studied items are then presented intermixed with lures while omitting the source information (e.g., by showing all words in black). In the corresponding source-monitoring model, the probability of source recognition d (i.e., detecting whether a word was presented in blue or red) is defined conditionally on target detection D (i.e., detecting whether a word was studied at all). If a researcher is mainly interested in source memory, this implies that the length and the content of the study list should be chosen in a way that ensures a high probability of target detection (i.e., large D) which in turn results in a high accuracy of the estimate for source recognition d.

In a cognitive model, a calibration of the psychologically meaningful parameters can be achieved by relying on parameter estimates for specific experimental designs, stimulus materials, or populations of participants in previous studies. Moreover, the specification of design factors is facilitated by the fact that the parameters of a cognitive model usually represent meaningful psychological concepts about which researchers can usually make informed guesses or even define precise prior distributions (Lee and Vanpaemel 2018).

Example: Optimal Design for the Pair-Clustering Model

In this subsection, we provide an example of how to optimize the design of an empirical study to maximize the expected information gain (here, statistical power) in practice. For this purpose, we rely on the pair-clustering model, a classical multinomial processing tree model for disentangling storage and retrieval in memory (Batchelder and Riefer 1980, 1986). In the standard pair-clustering paradigm, participants first study a list of words; some of which are semantically related to a second word (*pairs*; e.g., dog and cat), whereas others are unrelated to all remaining words in the study list (*singletons*). In a free-recall test, individuals then have to remember as many of the studied words as possible.

To model the data, the following categories are defined: For pairs, we distinguish the recall of both items in direct succession (C_{11}), the recall of both items separated by at least one other item (C_{12}), the recall of only one item of a pair (C_{13}), and the recall of none of the items (C_{14}). For singletons, we only distinguish between items that are recalled and those that are not (C_{21} and C_{22}, respectively). The pair-clustering model assumes that the semantic association of two paired items may result in a memory benefit due to clustering, meaning that the two words are jointly stored and retrieved. To specify this assumption more precisely, the model (shown in Fig. 1) assumes that a pair is stored as a cluster with probability c, in which case it is retrieved during free recall with conditional probability r, resulting in a C_{11} response, or not retrieved with probability $1–r$, resulting in recall failure (i.e., C_{14}). If clustering fails with probability $1–c$, in contrast, participants store and retrieve each item of a pair independently with probability u. Similarly, each singleton is stored and retrieved independently with probability a.

In most applications of the pair-clustering model, it is assumed that storage and retrieval of single words are identical for pairs and singletons. To test this hypothesis H_0: $u = a$, Batchelder and Riefer (1986) developed a likelihood ratio test. However, the statistical power (and thus, the expected information gain) of this test depends on two crucial design factors: first, on the proportion of pairs vs. singletons in the study list, and second, on the population value of the probability of clustering c, which in turn depends on the difficulty of detecting semantic associations. To illustrate this, we computed the power of the likelihood ratio test for testing the null hypothesis H_0: $u = a$ at the significance level $\alpha = 5\%$. For multinomial processing tree models, the power can easily be approximated

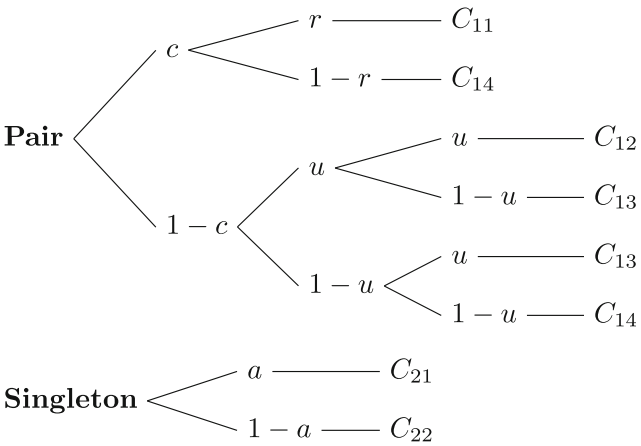

Fig. 1 The pair-clustering model (Batchelder and Riefer 1980). The parameters refer to c = probability of storing a semantically related pair of items as a cluster, r = probability of retrieving a clustered pair, u = probability of recalling a single item of a non-clustered pair, and a = probability of recalling a singleton

by (1) specifying the population model under H_1 with all parameter values fixed at plausible values, (2) defining the number of pairs and singletons (N_1 and N_2, respectively), (3) computing the expected frequencies under H_1, (4) fitting the model under H_0 to these expected frequencies by minimizing the likelihood ratio statistic G^2, and (5) computing the statistical power $1–\beta$ using the noncentral $\chi_1^2(\lambda)$ distribution with the minimized G^2 as noncentrality parameter λ (Erdfelder et al. 2005; Moshagen 2010).

Figure 2 shows the statistical power of testing H_0: $u = a$ as a function of the proportion of pairs in the study list (left panel) and the probability of clustering c (right panel). First, with respect to the proportion of pairs, it is evident that the power is highest if the study list contains more pairs than singletons. Depending on the parameter values in the population, the optimal proportion of pairs is 65% if $c = r = 0.50$ and 75% if $c = r = 0.80$, resulting in a statistical power of 84.8% and 58.2%, respectively. Hence, depending on the expectation about the most plausible value for c and r (which may be based on prior studies with similar participants), one can adapt the design accordingly. Second, for a fixed proportion of 50% pairs, the second panel in Fig. 2 shows the statistical power as a function of the probability of clustering c. Irrespective of the population value of the test-relevant parameter a, the power for testing the hypothesis H_0: $u = a$ is maximized for small c. If researchers are interested in testing this hypothesis rigorously, they should thus ensure that the probability of clustering is relatively small (e.g., by using pairs with weak semantic associations).

Quite often researchers will be interested in several research questions simultaneously. For example, cognitive aging researchers might be interested in (1) testing the pair clustering model for both young and older adults (i.e., H_0: $u_y = a_y$ and $u_o = a_o$) and also in (2) assessing whether retrieval parameters differ between age groups or not (i.e., H_0: $r_y = r_o$; cf. Riefer and Batchelder 1991). We have already seen that the first research question can be addressed most efficiently when c is small. By the same logic, it is possible to show that the optimal design for the second research question requires c to be large. How to resolve this conflict? There are basically two ways to go. First, one could design two independent experiments based on the same underlying populations, with encoding conditions aiming at a small versus a large c, respectively, to answer each of the two research questions most efficiently. Second, one could design a single experiment aiming at a medium parameter of, say, $c = 0.50$ and compensate for the loss in statistical power due to non-optimal c by increasing the sample sizes N_1 and N_2 for pairs and singletons, respectively. Again, a priori design planning will enable researchers to decide which of these two possible ways minimizes resources (i.e., time and money) required for answering both research questions of interest with a predetermined level of statistical power.

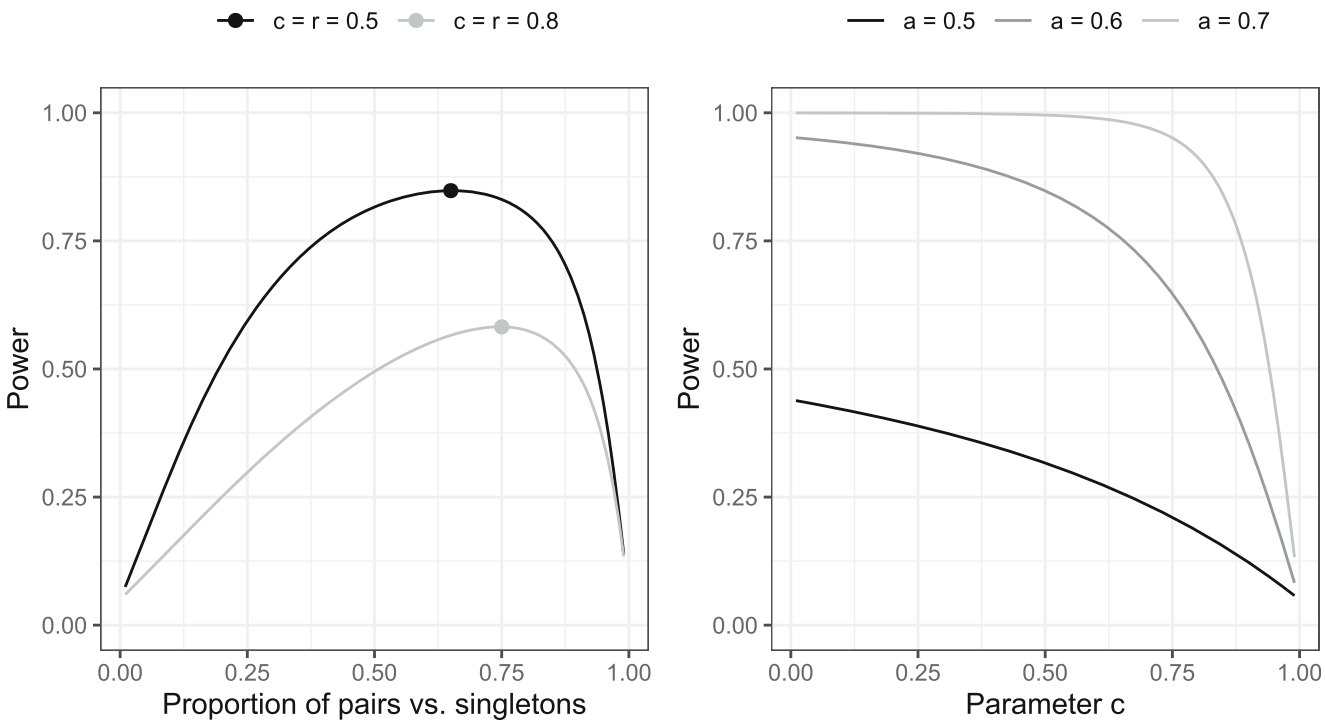

Fig. 2 Statistical power of testing the hypothesis H_0: $u = a$ at the significance level $\alpha = 5\%$. Panel A is based on the alternative hypothesis $u = 0.40$ and $a = 0.60$ and a total sample size of $N = 480$. In panel B, the sample sizes for pairs and singletons are $N_1 = 320$ and $N_2 = 160$, respectively, with a probability of retrieval of $u = 0.40$ and $r = 0.80$

Overall, our example shows how the expected information gain of cognitive modeling can be improved by optimizing details of the experimental design. This is achieved by assessing the statistical criterion of interest (e.g., the statistical power) as a function of one or more design factors (i.e., the relative proportion of pairs and the population values of the meaningful model parameters). More generally, the example also shows that the utility of power analysis extends beyond planning the total sample size for a fixed experimental design.

Conclusion

Before collecting any new data for a confirmatory (possibly preregistered) test in cognitive modeling, it is important that researchers consider whether the planned study and experimental design is informative with respect to the cognitive models of interest. In order to optimize the design, a formal analysis allows researchers to maximize, for instance, the expected utility in terms of model discrimination (Myung and Pitt 2009), the statistical power of a hypothesis test (Cohen 1988), or the discrepancy between prior and posterior distribution (Lindley 1956). As a much simpler, less time-intensive alternative approach, Monte Carlo simulations allow researchers to perform any planned analysis a priori across many replications to assess the impact of various design choices on the expected gain in information (see Algorithm 1). Going beyond the scope of a priori design considerations, the benefits of design optimization can be exploited to an optimal degree by selecting the most informative stimuli while the experiment is running by means of adaptive designs (Myung et al. 2013).

However, whereas optimizations of the design should regularly be considered for confirmatory tests in cognitive modeling, this is more difficult or may even be impossible in exploratory contexts. First, when reanalyzing existing data (e.g., to develop a new model), design optimization is obviously not possible. Nevertheless, researchers can judge post hoc whether the implemented design was informative for model development and use this knowledge as a basis for designing more informative follow-up studies (Navarro et al. 2004). Second, if the aim is to collect new data for exploratory modeling, design optimizations are more difficult but still possible. In such a case, researchers usually cannot formalize expected outcomes in terms of model equations or parameter values—a prerequisite required for methods such as power analysis or model-recovery simulations. However, cognitive modelers usually collect new data only if they have at least some idea of what to model (e.g., whether certain design factors affect retention curves in memory). Instead of assessing information gain in light of established cognitive models, researchers can still rely on classical power analysis for standard statistical models (e.g., ANOVA, correlation, or regression analysis) to estimate the number of observations required to detect relevant systematic patterns in the data empirically.

In experimental psychology more generally, the importance of optimizing experimental designs was recently emphasized within the statistical guidelines of the Psychonomic Society (2019), advising researchers to "do what you reasonably can to design an experiment that allows a sensitive test." Within cognitive modeling, methods such as design optimization, power analysis, Bayesian design analysis, and simulations of model recovery and parameter estimation are crucial tools to assess the expected information gain provided by an experimental design before collecting any data. At least under ideal conditions as those generated in a simulation or assumed by an analytical analysis, the chosen experimental design should ensure a high a priori probability that the models under consideration provide insights about the psychological theories and cognitive processes of interest.

Funding information This research was supported by the German Research Foundation (DFG; Research Training Group "Statistical Modeling in Psychology," grant GRK 2277).

Data availability All R scripts are available on the Open Science Framework at https://osf.io/xehk5/.

References

Arnold, N. R., Heck, D. W., Bröder, A., Meiser, T., & Boywitt, C. D. (2019). Testing hypotheses about binding in context memory with a hierarchical multinomial modeling approach: a preregistered study. *Experimental Psychology, 66*(3), 239–251. https://doi.org/10.1027/1618-3169/a000442.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review, 87*(4), 375–397. https://doi.org/10.1037/0033-295X.87.4.375.

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology, 39*, 129–149. https://doi.org/10.1111/j.2044-8317.1986.tb00852.x.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564. https://doi.org/10.1037/0033-295X.97.4.548.

Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. Haywood, CA: The Institute of Mathematical Statistics.

Berger, M. P. F., & Wong, W.-K. (2009). *An introduction to optimal designs for social and biomedical research*. Hoboken: John Wiley & Sons.

Bröder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making, 16*(3), 193–213. https://doi.org/10.1002/bdm.442.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1565–1570). https://doi.org/10.1002/0470013192.bsa491.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: a flexible statistical power analysis program for the social, behavioral,

and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/bf03193146.

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.

Heck, D. W., Hilbig, B. E., & Moshagen, M. (2017). From information processing to decisions: formalizing and comparing probabilistic choice models. *Cognitive Psychology, 96*, 26–40. https://doi.org/10.1016/j.cogpsych.2017.05.003.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018a). TreeBUGS: an R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7.

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018b). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making, 13*(4), 356–371.

Jekel, M., Fiedler, S., & Glöckner, A. (2011). Diagnostic task selection for strategy classification in judgment and decision making: theory, validation, and implementation in R. *Judgment and Decision Making, 6*(8), 782–799.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: a critical test with minimal assumptions. *Psychological Review, 122*(3), 542–557.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*(1), 114–127. https://doi.org/10.3758/s13423-017-1238-3.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., et al. (in press). Robust modeling in cognitive science. *Computational Brain & Behavior.* https://doi.org/10.1007/s42113-019-00029-y.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics, 27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069.

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*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735.

Meiser, T. (2014). Analyzing stochastic dependence of cognitive processes in multidimensional source recognition. *Experimental Psychology, 61*(5), 402–415. https://doi.org/10.1027/1618-3169/a000261.

Moshagen, M. (2010). multiTree: a computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42–54. https://doi.org/10.3758/BRM.42.1.42.

Müller, P., Sansó, B., & Iorio, M. D. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association, 99*(467), 788–798. https://doi.org/10.1198/016214504000001123.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116*(3), 499–518. https://doi.org/10.1037/a0016104.

Myung, J. I., Forster, M. R., & Browne, M. W. (2000). Guest editors' introduction: special issue on model selection. *Journal of Mathematical Psychology, 44*, 1–2. https://doi.org/10.1006/jmps.1999.1273.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology, 57*, 53–67. https://doi.org/10.1016/j.jmp.2013.05.005.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49*(1), 47–84. https://doi.org/10.1016/j.cogpsych.2003.11.001.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289–337.

Platt, J. R. (1964). Strong inference. *Science, 146*(3642), 347–353.

Popper, K. (2005). *The logic of scientific discovery.* https://doi.org/10.4324/9780203994627.

Psychonomic Society. (2019). Statistical guidelines. Retrieved from https://www.psychonomic.org/page/statisticalguidelines

Riefer, D. M., & Batchelder, W. H. (1991). Age differences in storage and retrieval: a multinomial modeling analysis. *Bulletin of the Psychonomic Society, 29*(5), 415–418. https://doi.org/10.3758/BF03333957.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358.

Schnuerch, M., & Erdfelder, E. (in press). Controlling decision errors with minimal costs: the sequential probability ratio t test. *Psychological Methods*.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339. https://doi.org/10.1037/met0000061.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*(1), 28–50. https://doi.org/10.1016/j.jmp.2003.11.004.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.