

Model discrimination through adaptive experimentation

Daniel R. Cavagnaro · Mark A. Pitt · Jay I. Myung

Published online: 16 November 2010
© Psychonomic Society, Inc. 2010

Abstract An ideal experiment is one in which data collection is efficient and the results are maximally informative. This standard can be difficult to achieve because of uncertainties about the consequences of design decisions. We demonstrate the success of a Bayesian adaptive method (adaptive design optimization, ADO) in optimizing design decisions when comparing models of the time course of forgetting. Across a series of testing stages, ADO intelligently adapts the retention interval in order to maximally discriminate power and exponential models. Compared with two different control (non-adaptive) methods, ADO distinguishes the models decisively, with the results unambiguously favoring the power model. Analyses suggest that ADO's success is due in part to its flexibility in adjusting to individual differences. This implementation of ADO serves as an important first step in assessing its applicability and usefulness to psychology.

Keywords Retention · Active learning · Model discrimination · Experimental design · Adaptive testing

Scientific progress depends on well-designed experiments. Researchers strive to make experiments maximally informative about the question of interest, but this goal can be difficult to attain because of the uncertain consequences of the many design decisions that must be made. Although some choices are likely to have minimal impact in an experiment (e.g., age of the experimenter), others can be

much more crucial (e.g., choice of treatment levels). In this paper, we demonstrate a new methodology that maximizes the efficiency of data collection in experiments by optimizing some of these design choices, taking the guesswork out of parts of the experimental design process.

Many of the practicalities of testing human participants make experiment efficiency, for better or worse, a necessary criterion in experimental design. For example, fatigue and the ethics of testing a participant more than is necessary are two factors that must be considered. Standard practices in experimental design tend to work against efficiency, however. Variables are factorially manipulated, creating all possible combinations of levels of the independent variables, even though it is rare for all of the cells in a design to be equally informative. Similarly, equal numbers of observations are collected in each cell of a design, even those that may be redundant. Large numbers of participants are often tested to combat large individual differences, which are almost always the source of most of the variability in performance.

Psychologists have been aware of these inefficiencies and difficulties for decades, and have developed heuristic methods for countering them. Most commonly, pilot studies are carried out to identify the most informative levels of the independent variables. In well-developed areas of research, previous experiments are used as an informal guide for setting design parameters, such as the number of observations to collect or the number of participants to test. Sometimes, the tacit knowledge that one has acquired over years in the laboratory can be the best guide. Psychological science would benefit from a principled means for making such design decisions (McClelland, 1997).

In the statistics literature, there is already a large body of research on formal methods for optimizing the design of an

D. R. Cavagnaro (✉) · M. A. Pitt · J. I. Myung
Department of Psychology, The Ohio State University,
1835 Neil Avenue,
Columbus, OH 43210, USA
e-mail: cavagnaro.2@osu.edu

experiment (Chaloner & Verdinelli, 1995). Among the most powerful of these optimization strategies are the adaptive methods, in which many design parameters of the experiment (e.g., stimulus levels, number of trials, number of observations per cell, number of conditions) are not fixed, but rather are allowed to evolve as the experiment progresses (Müller, Berry, Grieve, Smith, & Krams, 2007). The idea is to run an experiment as a sequence of stages, or mini-experiments, in which the design of the next stage is chosen based on the results of the previous stages. Thus, the information gained at each stage can be used to adapt the designs of subsequent stages to be maximally informative.

A general framework for implementing adaptive designs in experiments that compare models of cognition was proposed by Cavagnaro, Myung, Pitt, and Kujala (2010), extending previous work by Myung and Pitt (2009). In this simulation-based framework, called Adaptive Design Optimization (ADO), Bayesian decision theory is used to identify the most informative design at each stage of the experiment so that one can infer the characteristics of the underlying mental process in as few steps as possible. Essentially, each potential design is treated as a gamble whose payoff is determined by the outcome of a hypothetical experiment carried out with that design.

At the start of each stage of an ADO experiment, many hypothetical experiments with various candidate designs are simulated on computer and evaluated so that an “expected utility” (i.e., the expected payoff) of each design can be computed. The design with the highest expected utility is then chosen as the optimal design. Finally, empirical data are collected with that optimal design, and the results are used to inform design selection at the next stage of the experiment. This process repeats until a stopping criterion is reached.

The purpose of the present paper is to demonstrate the potential of ADO for improving experimentation in psychology by implementing it in an experiment with human participants. Specifically, we conduct an experiment in which ADO is used to extract the most informative data possible for discriminating power and exponential models of the time course of retention (i.e., forgetting). These models have proven notoriously difficult to discriminate in experiments because they make such similar predictions (Heathcote, Brown, & Mewhort, 2000; Brown & Heathcote, 2003), but they have been successfully discriminated in computer simulations using ADO (Cavagnaro et al., 2010), suggesting that ADO could be used to discriminate them in practice. Indeed the following experiment uses ADO and generates decisive evidence in favor of the power model. Analyses of the algorithm show how ADO adjusts to individual differences, constantly trying to optimize the informativeness of the next data point.

Discriminating models of retention

Over a century of research on forgetting has shown that a person’s ability to remember information just learned drops quickly for a short time after learning and then levels off as more and more time elapses (Ebbinghaus, 1964; Laming, 1992). The simplicity of this data pattern has led to the introduction of a number of models (equations) to describe the rate at which information is retained in memory (Rubin & Wenzel, 1996). A handful of these models have proven to be superior to the rest of the field (Wixted & Ebbesen, 1991).¹ Two strong competitors are the power model, $p = a(t + 1)^{-b}$, and the exponential model, $p = ae^{-bt}$, where in each equation the symbol p denotes the predicted probability of correct recall as a function of retention interval t with model parameters a and b . Despite the best efforts of researchers to design studies that were intended to discriminate them, results have been inconsistent. In some studies, well-established methods were employed to increase the statistical power of the experiments and thereby improve model discriminability. They included testing large numbers of participants and replicating the experiment using a range of different tasks or participant populations (e.g. Rubin, Hinton, & Wenzel, 1999). In other studies, advanced statistical tools for model selection were applied to existing datasets in order to squeeze as much information out of them as possible (e.g. Lee, 2004). However, the conclusions of these and other studies, as far as discriminating power and exponential models, have been only suggestive at best.

In the present study, we use ADO to improve the efficiency of data collection, with the goal of discriminating power and exponential models more conclusively. Specifically, we conduct a multi-stage retention experiment in which the retention interval (i.e., the time between the study phase and the test phase) is adapted at each stage. Our adaptive method breaks with most past studies in that we did not use a fixed set of retention intervals at which to probe memory. Rather, at each of the nine stages of this within-participant experiment, memory was probed at only a single (optimal) retention interval.

The ADO experiment, implemented within a Bayesian framework (Edwards, Lindman, & Savage, 1963; Lindley, 2000), begins with informative parameter priors and equal model priors (i.e., both models are treated as being equally likely at the outset). An optimal retention interval for discriminating the models with the given priors is found,

¹ Although more sophisticated models have been shown to give fuller accounts of forgetting (e.g., Oberauer and Lewandowski, 2008), the simplicity of the power and exponential functions along with the difficulty of discriminating between them provides an ideal setting in which to test ADO. When it has matured further, it can be applied to discriminating more complex models of serial and free recall.

and the first stage of testing is conducted using that retention interval. When the first stage is complete, Bayesian updating is used to improve the parameter estimates for each individual and re-assess the likelihood of each model based on the observed data pattern. Next, a new optimal retention interval for is found, independently for each individual, based on the updated parameter estimates and these new retention intervals are used in the second stage of testing. This process is repeated through nine stages, separately for each participant. Technical details can be found in the online supplement.

The choice of when to probe memory is critical for discriminating these models because their similarity varies significantly across retention intervals. ADO exploits to the fullest this variation in model discriminability by identifying the most differentiating retention interval at each testing stage. To illustrate, each graph in Fig. 1 depicts the power and exponential models for a narrow range of parameters. In the left graph, the models make nearly identical predictions at all retention intervals longer than 6 s. Testing in that range would not be informative because it would be impossible to obtain data that could be fit by one model and not the other. Testing at a retention interval between 0 and 5 s, on the other hand, where the power model predicts greater retention than the exponential, would better discriminate the models because observing higher recall would favor the power model while lower recall would favor the exponential model. In the right graph, which depicts the predictions of the two models for a different range of parameters, the predictions are nearly identical for all retention intervals shorter than about 15 s, after which they begin to diverge. The optimal retention interval for this range of parameters is between 15 and 40 s.

Using the preceding illustration as a backdrop, an experiment using ADO can be viewed as a game between ADO and the models. At each stage, ADO tries to pull the models apart by testing where they differ the most, and then the models adapt by updating their parameter estimates to fit the newly observed data point. ADO continues to press the models where they differ the most until the inferior model is exposed for being unable to fit the observed data pattern. More formally, ADO maximizes the odds (of the model posterior probabilities) that the observed data will be more likely under one model than the other.

Demonstration of ADO

We conducted a retention experiment to demonstrate the ability of ADO to discriminate the two models and the efficiency with which it can do so.² Aside from testing

² The technique scales easily up to three or more models by changing nothing more than the limits of the summation in Eq. (1).

across multiple stages, the methodology closely resembled that used by Wixted and Ebbesen (1991). There were three groups: ADO, random, and geometric. In the ADO group, the retention interval t at each stage was chosen by maximizing the function

$$U(t) = \sum_{m \in \{POW, EXP\}} p(m) \int \int p(y|\theta_m, t) p(\theta_m) \log \times \frac{p(m|y, t)}{p(m)} dy d\theta_m, \quad (1)$$

where y is a hypothetical observation at time t in the case that the true model is m with parameters θ_m . This utility function can be interpreted as an information-theoretic measure of the expected evidence that would be provided from collecting a data point at time t (see Cavagnaro et al., 2010, for technical details). Analysis of the data from the ADO group would address the goal of discriminating the models.

In the random group, the “optimization” part of ADO was turned off and the retention interval at each stage was chosen at random (uniformly between 1 and 40 s), making it possible to separate the effects of choosing an optimal retention interval from the effects of sequential testing with Bayesian updating. Comparison of these data with those obtained in the ADO group would provide an indication of the algorithm’s efficiency relative to a design strategy with no optimization built in.

Retention intervals in the geometric group were chosen at each stage to reflect the spacing of retention intervals that is common in retention experiments (e.g., Wixted & Ebbesen, 1991; Rubin & Wenzel, 1996). Specifically, retention intervals of 1, 2, 3, 5, 8, 12, 18, 27, and 40 s were used in a random order across the nine stages for each participant. The use of a roughly geometric or logarithmic spread of retention intervals is an informed choice that evolved through many years of first-hand experience and trial-and-error. We should therefore expect better results in it than in the random group. Comparison of the data in the ADO and geometric groups should demonstrate the additional benefit of an adaptive design, which in this instance has no knowledge of what designs worked well in the past.

Methodology

Twenty-four participants (from introductory psychology courses), eight per group, performed a variant of the Brown-Peterson task (Brown, 1958). In each trial, a target list of six words was randomly drawn from a pool of high frequency, monosyllabic nouns. These words were presented on a computer screen at a rate of two words per

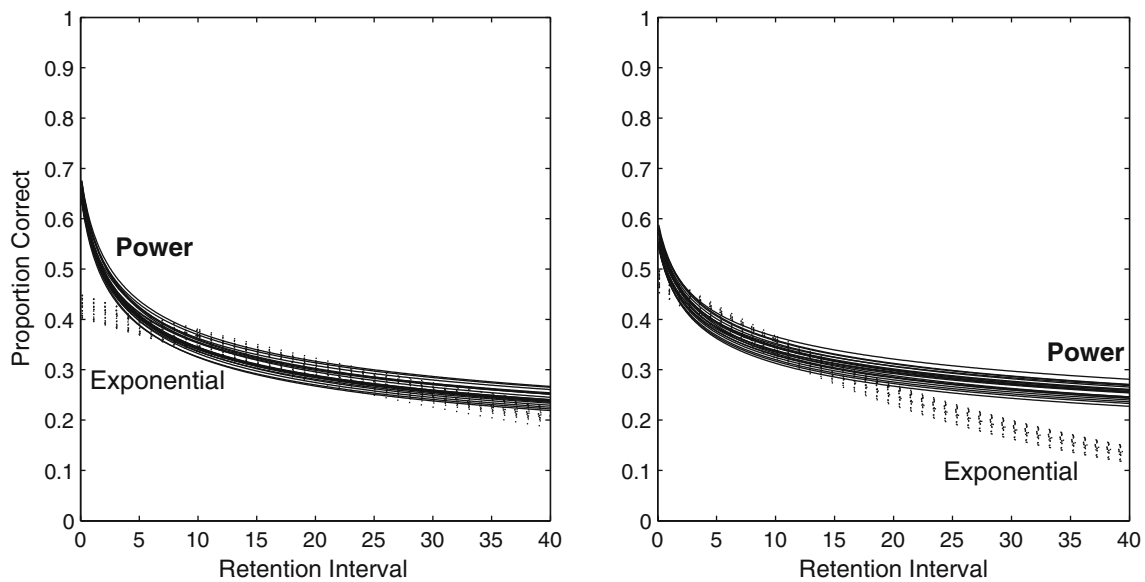


Fig. 1 Power and exponential curves for two different ranges of parameters show that the regions where the models are discriminable can change depending on the parameter estimates. On the left, parameter ranges of $0.65 < a < 0.70$ and $0.25 < b < 0.30$ for the power model and $0.40 < a < 0.45$ and $0.015 < b < 0.020$ for the exponential model create curves that overlap after about 5 s, making

them most discriminable for retention intervals shorter than 5 s. On the right, parameter ranges of $0.55 < a < 0.60$ and $0.20 < b < 0.25$ for the power model and $0.45 < a < 0.50$ and $0.030 < b < 0.035$ for the exponential model create curves that overlap for all retention intervals shorter than 15 s, making them most discriminable for retention intervals longer than 15 s

second, and served as the material that participants had to remember. Five seconds of rehearsal followed, after which the target list was hidden and distracter words were presented, one at a time at a rate of one word per second, for the duration of the retention interval. During the distracter phase, participants had to say each distracter word aloud into a microphone as it appeared on the computer screen. The purpose of the distracter task was to occupy the participant's verbal memory in order to prevent additional rehearsal of the target list during the retention interval. The distracter words were drawn from a separate pool of 2000 monosyllabic nouns, verbs, and adjectives. A different random word order was used for each participant. At the conclusion of the retention interval, participants were given up to 60 s for free recall of the words (typed responses on a keyboard) from the target list. A word was counted as being remembered only if it was typed correctly.

Each participant completed a single, 2-h session. The first 20 min were devoted to training and instructions. The next 90 min were broken up into 10-min stages, each consisting of a 5-min testing period followed by a 5-min break (necessary for algorithm computation). The same retention interval (range of 1–40 s) was used for all trials within a given stage. During each break, the data from the current stage were fed to another computer that estimated the optimal retention interval for the next testing stage. Participants read magazines during the break. The final 10 min was used for debriefing.

Results and analysis

We used the Bayes factor³ to measure the strength of evidence in favor of one model over the other. The Bayes factor, a standard method of model selection in Bayesian analysis, is defined as the ratio of the posterior marginal likelihoods of the two models, derived from Bayesian updating, and provides a direct and naturally interpretable metric for model selection (Kass & Raftery, 1995). A Bayes factor of ten, for example, means that the data are ten times more likely to have occurred under the one model than under the other. A low Bayes factor does not indicate that the models are performing poorly, however. The Bayes factor indicates relative model plausibility, not absolute model plausibility, so a value near one could also result from both models performing equally well.

Table 1 shows the Bayes factor for the power model over the exponential model for each participant at the conclusion of the experiment.⁴ In the ADO group, each of the eight participants produced data that unambiguously favor the

³ Readers unfamiliar with the Bayes factor should know that it is superior to measures that assess only goodness of fit, such as r^2 and percent variance accounted for (Myung, 2000).

⁴ The Bayes factors were computed from the same parameter-priors that were used to initiate the ADO experiments. Specifically, $a \sim \text{Beta}(1.03, 0.97)$, $b \sim \text{Beta}(0.97, 24.55)$ for the power model and $a \sim \text{Beta}(1.03, 0.97)$, $b \sim \text{Beta}(0.97, 24.55)$ for the exponential model. These distributions were derived via a method of moments from the results of a pilot experiment.

Table 1 Bayes factors, indicating the strength of evidence for the power model over the exponential model, for each participant at the conclusion of the experiment

ADO		Random		Geometric	
Participant	Bayes Factor	Participant	Bayes Factor	Participant	Bayes Factor
1	12.51	9	2.60	17	8.52
2	11.35	10	1.30	18	26.02
3	15.67	11	3.12	19	1.11
4	3.42	12	2.51	20	14.38
5	6.25	13	1.38	21	3.48
6	54.56	14	1.30	22	7.70
7	165.67	15	2.77	23	3.49
8	22.81	16	3.81	24	2.94
Average	36.53	Average	2.35	Average	8.46

power model. The average Bayes factor in the ADO group is 36.5, which easily exceeds the rule-of-thumb cut-off for “strong” evidence (Bayes factor of 10). In fact, six out of the eight ADO participants produced evidence that exceeds this cutoff. The data from the other two groups also support the power model, although far less conclusively. Six out of the eight participants in the random group yielded Bayes factors less than 3.0, which is “not worth more than a bare mention” according to the scale by Kass and Raftery (1995). The two participants that did exceed 3.0 did so by the thinnest of margins. As expected, participants in the geometric group yielded more informative data than those in the random group overall, but only two out of six participants exceeded the rule-of-thumb cut-off of 10, and the average Bayes factor of 8.5 is far below that obtained in the ADO group.

The retention intervals in the random group tended to be evenly spaced between 1 and 40 s, so the low Bayes factors could not be attributed to any noticeable pathology in the choice of retention intervals in those designs. For example, participant 13 was tested at 1, 3, 8, 10, 14, 15, 21, 30, and 35 s over the course of the experiment, and produced a Bayes factor of just 1.37. What is more, while seemingly very low, the strength of evidence that was obtained in the random group is comparable to that which has been obtained in previous experiments that employed fixed designs with typical choices for retention intervals (e.g., linearly or geometrically spaced). For example, Lee (2004) found a Bayes factor of 1.94 (in favor of the exponential model over the power model) for the data collected by Squire (1989). The same analysis also found Bayes factors between 1.0 and 5.0 (also in favor of exponential over power) for the data sets collated by Rubin and Wenzel (1996).

To give a sense of the rate at which evidence for the power model accumulated across stages, Fig. 2 shows the average Bayes factor at the end of each stage for each of the three groups. The plots are fairly flat for the first three

stages, indicating that participants in none of the groups generated evidence in favor of either model. This is not unexpected because both models should be able to fit the observed data pattern equally well when there are relatively few data points to constrain them. From stage four through six, the ADO and geometric curves gradually pull apart from the random curve, providing clear evidence in favor of the power model. The reason for the increasing Bayes Factor is different in the two designs. The geometric curve increases because the a-priori knowledge about ideal interval spacing begins to pay off. ADO, on the other hand, is learning on-the-fly, and its choice of discriminating retention intervals begins to pay dividends. Because this active learning continues through

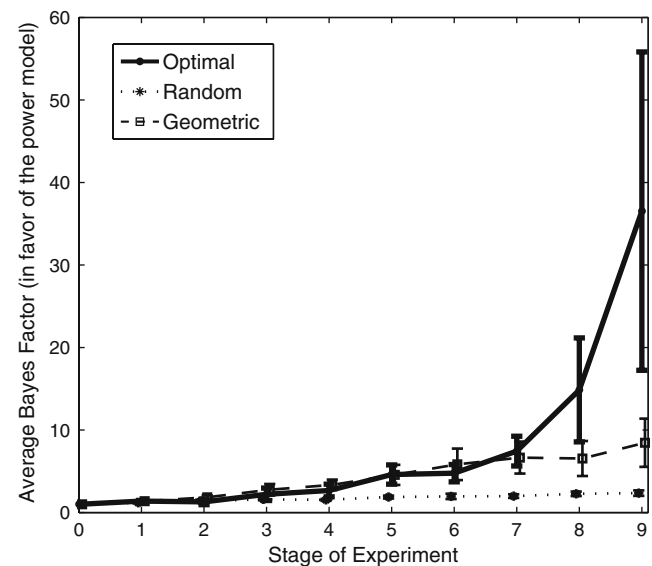


Fig. 2 Arithmetic mean of the Bayes factors obtained from all participants in each group at the end of each testing stage. Error bars show the standard error of the estimates; they are slightly offset from their respective points to make them more visible

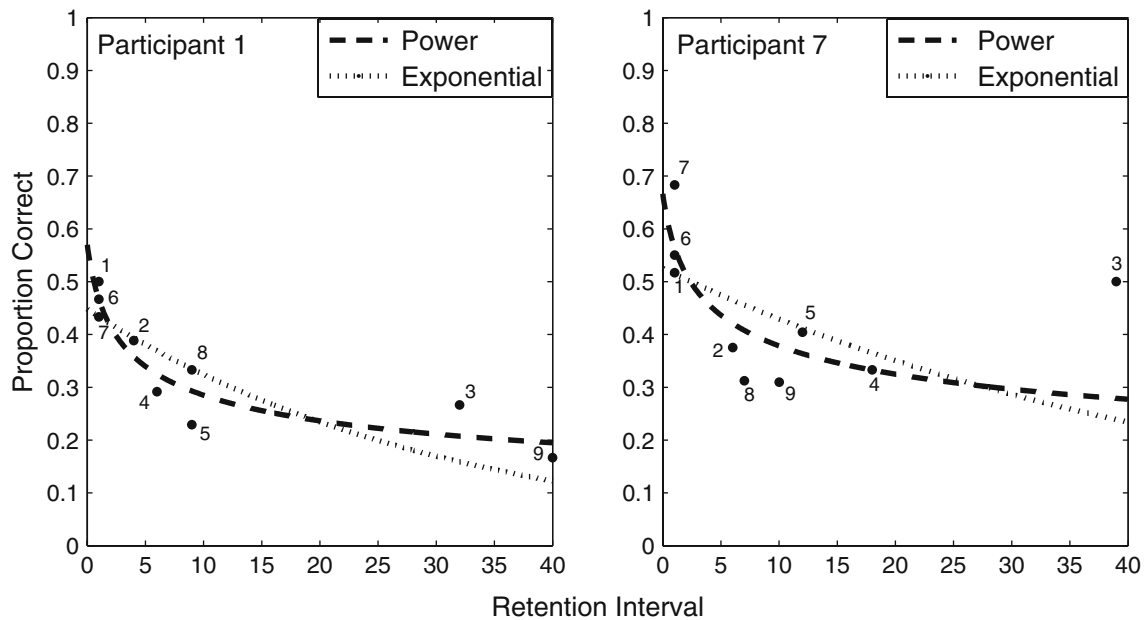


Fig. 3 Results from two participants in the ADO group. Each data point represents the proportion of correct responses observed in one testing stage. Numerical labels indicate the order of the testing stages

the remainder of the experiment, ADO continues to identify progressively more discriminating retention intervals. In stages seven through nine, the average Bayes factor rises sharply in the ADO group but remains fairly flat in the random and geometric groups. Moreover, the steepness of the ADO plot increases with each passing stage, indicating that the constraints placed on the models by the algorithm’s choice of each subsequent retention interval accelerated model discriminability significantly.

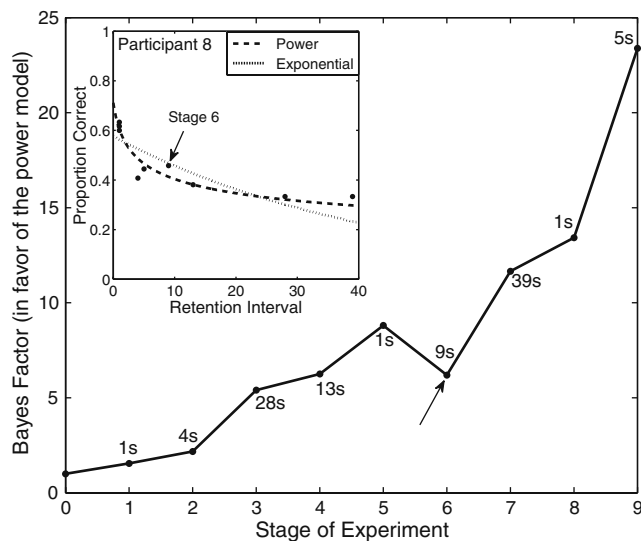


Fig. 4 Bayes factors at each stage for participant 8. Labels indicate the retention interval at the corresponding stage. The inset shows the best fitting power and exponential models for the data collected in all nine stages. The one drop in the Bayes factor (stage 6) corresponds to a point that was fitted best by the exponential model

The superiority of the power model across participants in the ADO group is due in part to ADO’s sensitivity to individual differences. This is illustrated in Fig. 3, which shows the best fitting power and exponential curves for two participants from the ADO group. In the left graph, the curves for participant 1 cross twice, once near 2 s and again near 18 s. Notice that the retention intervals tested by ADO all lie in regions where the models are visibly distinct (i.e., at one second, between 5 and 10 s, between 30 and 40 s). The curves in the right graph (participant 7) are similar except that the second point of intersection is at 28 s instead of 18. This can be attributed to unusually good recall at 39 s, which occurred in the third ADO stage, and which forced the exponential curve to flatten out. Accordingly, ADO selected a retention interval of 18 s in the fourth stage, a point that would not have been informative for participant 1, since that is where the curves for participant 1 cross, but is informative for participant 7 given the flatter estimate of the exponential curve. Further, ADO does not select a retention interval of 32 s for participant 7, as it does for participant 1, as this would be very close to the intersection point at 28 s.⁵

Figure 3 also shows that ADO finds it informative to return to the same retention interval across stages. For example, a retention interval of one second was used in three different stages for both participants. In fact, the one-

⁵ The curves shown in Fig. 3 are best fits to the data from all nine stages of the experiment, which reasonably approximates the information used by the algorithm to select optimal designs in the latter stages of the experiment.

second retention interval was revisited at least twice for each of the eight participants in the ADO group. This behavior of the algorithm nicely confirms what is believed by researchers in this field: a precise estimate of the initial retention rate is crucial for discriminating the models. ADO takes advantage of the differential informativeness of retention intervals by sampling more often at these intervals, which can be difficult in a fixed design.

It is important to note that ADO does not necessarily guarantee that an experiment will generate data that discriminate models; it merely sets up ideal conditions for that to occur. It is up to the participants to provide the discriminating data. Moreover, the same property of ADO that allows it to collect strong evidence very quickly (i.e., testing at retention intervals where the models differ the most) also makes it susceptible to being misled. This is illustrated in Fig. 4, which shows the data from participant 8 and the corresponding Bayes factors after each testing stage. Notice that there is a sharp drop in the Bayes factor at stage 6, which is indicated by an arrow. The reason for this drop can be seen in the graph inset, which shows the best fitting power and exponential models for this participant. The data point that was collected in stage six is also indicated with an arrow. Since this data point is more compatible with the exponential model than with the power model, it constitutes evidence against the power model, and hence the Bayes factor drops at the end of stage 6. This data point is an outlier, as the other eight points favor the power model. Note also that the drop in evidence for the power model is temporary. At each stage, the algorithm seeks retention intervals that maximize model discriminability. This is why the Bayes factor rises continuously from stages 6 through 9.

Conclusions

The complexity of computational models of cognition, even seemingly simple ones like retention functions, can be challenging to discriminate experimentally because it is difficult to identify conditions that lead to diverging predictions. ADO is an intelligent tool that exploits the mathematical precision inherent in models to identify these conditions. Across testing stages, the algorithm updated the experiment's design by identifying optimal retention intervals on the basis of participant responses. The data observed in each stage progressively constrained the models to a region of the response space where only one model could survive, in this case providing strong evidence in favor of the power model as a description of the time course of retention. ADO's ability to adapt the choice of retention intervals to each participant led to a consistent outcome, making this single experiment efficient and

informative. Although ADO is in its infancy and it is currently restricted to comparing mathematical models, there are few obstacles to expanding its scope to make it applicable to other types of models and a wider set of design questions.

Author Note Thanks to Yun Tang for her timely assistance with testing the participants.

References

- Brown, J. A. (1958). Some tests of the decay theory of immediate memory. *The Quarterly Journal of Experimental Psychology*, *10*, 12–21.
- Brown, S., & Heathcote, A. J. (2003). Bias in exponential and power function fits due to noise: Comments on Myung, Kim, and Pitt. *Memory & Cognition*, *31*, 656–661.
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Computation*, *22*, 887–905.
- Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*, 273–304.
- Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. New York: Dover. (Original work published in 1885.)
- Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.
- Heathcote, A. J., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
- Laming, K. (1992). Analysis of short-term retention: Models for Brown-Peterson experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1342–1365.
- Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, *48*, 310–321.
- Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–337.
- McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, *2*, 3–19.
- Müller, P., Berry, D., Grieve, A., Smith, M., & Krams, M. (2007). Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference*, *137*, 3140–3150.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.
- Myung, J. I., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *58*, 193–198.
- Rubin, D., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology*, *25*, 1161–1176.
- Rubin, D., & Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*, 734–760.
- Squire, L. R. (1989). On the course of forgetting in very long term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 241–245.
- Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.