COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 498–516

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01221.x

Context Effects in Multi-Alternative Decision Making: Empirical Data and a Bayesian Model

Guy Hawkins,[a] Scott D. Brown,[a] Mark Steyvers,[b] Eric-Jan Wagenmakers[c]

[a]*School of Psychology, University of Newcastle*
[b]*Department of Cognitive Sciences, University of California, Irvine*
[c]*Department of Psychology, University of Amsterdam*

Received 20 December 2010; received in revised form 10 May 2011; accepted 16 May 2011

Abstract

For decisions between many alternatives, the benchmark result is Hick's Law: that response time increases log-linearly with the number of choice alternatives. Even when Hick's Law is observed for response times, divergent results have been observed for error rates—sometimes error rates increase with the number of choice alternatives, and sometimes they are constant. We provide evidence from two experiments that error rates are mostly independent of the number of choice alternatives, unless context effects induce participants to trade speed for accuracy across conditions. Error rate data have previously been used to discriminate between competing theoretical accounts of Hick's Law, and our results question the validity of those conclusions. We show that a previously dismissed optimal observer model might provide a parsimonious account of both response time and error rate data. The model suggests that people approximate Bayesian inference in multi-alternative choice, except for some perceptual limitations.

Keywords: Multi-alternative choice; Hick's Law; Context effect; Speed–accuracy tradeoff; Bayesian; Optimal observer

1. Introduction

Many decisions require selection of a single response from many possible alternatives. These decisions can range from simple perceptual choices to complex evaluations of multi-attribute information. A fundamental result for such choices is that the average decision time depends on the number of choice alternatives, according to Hick's Law (Hick, 1952;

Correspondence should be sent to Guy Hawkins, School of Psychology, University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: guy.hawkins@newcastle.edu.au

Hyman, 1953). Hick's Law can be expressed in a number of ways, but the most simple is that the mean time taken to select a response (i.e., $\overline{\mathrm{RT}}$) and the logarithm of the number of choice alternatives (K) are linearly related:

$$\overline{\mathrm{RT}} = a + b\log_2(K). \tag{1}$$

Hick's Law describes data from a wide range of paradigms, including speeded perceptual judgments (e.g., Leite & Ratcliff, 2010), eye saccades (e.g., anti-saccades in Kveraga, Boucher, & Hughes, 2002; Lee, Keller, & Heinen, 2005), absolute identification (e.g., Lacouture & Marley, 1995; Pachella & Fisher, 1972), manipulations of stimulus–response compatibility (e.g., Brainard, Irby, Fitts, & Alluisi, 1962; Dassonville, Lewis, Foster, & Ashe, 1999), and it has even been observed in monkeys (Laursen, 1977) and pigeons (Vickrey & Neuringer, 2000; for additional examples in other paradigms see Brown, Steyvers, & Wagenmakers, 2009; Teichner & Krebs, 1974; ten Hoopen, Akerboom, & Raaymakers, 1982).

Hick's Law was initially interpreted with recourse to information theory (Hick, 1952; Shannon & Weaver, 1949), where the time taken to judge between alternatives is proportional to the amount of stimulus information to be processed. For this reason, traditional investigations of Hick's Law emphasized error-free responding: Experimenters forced observers to process all of the information provided by the stimulus. Later investigations allowed observers to set their own tradeoffs between response speed and accuracy, and still observed Hick's Law: Mean response time was linearly related to the amount of stimulus information that observers actually processed (Hale, 1969; Pachella & Fisher, 1972).

1.1. Speed–accuracy tradeoff in decisions between multiple alternatives

When observers are allowed to set their own speed–accuracy tradeoff, it is often observed that response accuracy declines steadily as the number of choice alternatives increases. This result has been observed in paradigms using eye saccades (anti-saccades in Kveraga et al., 2002; Lee et al., 2005; Thiem, Hill, Lee, & Keller, 2008), absolute identification (Lacouture & Marley, 1995), and externalized evidence accumulation displays (Brown et al., 2009; for more examples in other paradigms, see Churchland, Kiani, & Shadlen, 2008; Leite & Ratcliff, 2010; ten Hoopen et al., 1982; Wright, Marino, Belovsky, & Chubb, 2007). The variability observed in patterns of accuracy has had important implications for theoretical accounts of Hick's Law, with some models explicitly targeting constant accuracy rates (Usher, Olami, & McClelland, 2002) and others explicitly targeting decreasing accuracy rates (Brown et al., 2009). There has been no attempt to unify these different theoretical accounts.

We explore the possibility that a free response protocol can induce participants to alter their speed–accuracy tradeoff across different numbers of choice alternatives. We suggest that the observer's goal accuracy for each set size is influenced by the experimental context: the other set sizes they have experienced. As described by Hick's Law, choices between a small number of alternatives are relatively fast, and choices between a larger number of alternatives are slower. We propose that the observer tries to mitigate this variability by speeding up on the slowest decisions and slowing down on the fastest ones. This has the effect of making decisions between small numbers of alternatives more accurate and choices between many alternatives less accurate, as is sometimes observed in data. A side effect of this tradeoff, and a possible explanation for it, is that the tradeoff also results in the observer taking less time (overall) to achieve the same average response accuracy, as if they did not adjust their speed–accuracy tradeoff.

To test our hypothesis, in two experiments we employ the same decision task but manipulate the number of choice alternatives differently. In Experiment 1, we manipulate the number of choice alternatives within-subjects: Each participant experiences some choices with only two alternatives, and some with 20 alternatives, and many values in between. According to our hypothesis, this should induce participants to be more accurate (but slower than they otherwise might be) for choices between few alternatives, and less accurate (but faster than they might otherwise be) for choices between many alternatives. In Experiment 2, each participant experiences just one set size—for example, one participant might only ever be given choices between $K = 8$ alternatives. If our hypothesis holds, this design should not induce a speed–accuracy tradeoff between different set sizes; we expect accuracy to remain constant across all choice alternatives. The results from these experiments have important theoretical implications. As we will show, they provide new constraints on models for Hick's Law and, more generally, models for speeded, multi-alternative decision tasks. These constraints change the way previous models have been evaluated.

2. Experiment 1

Our experimental paradigm is based on Brown et al.'s (2009). In their study, each decision involved a display of columns that grew taller at different rates by randomly accumulating increments of height (bricks) according to a simple statistical model. One of the columns tended to accumulate bricks more quickly than the others, and the participant's goal was to select this target column as quickly as possible. Observers were free to choose their own balance between speed and accuracy: Early in the process, when only a few bricks have accumulated, a distractor column is likely to be taller than the target by random chance.

Brown et al.'s (2009) paradigm was unusual because it made the process of gradual evidence accrual both explicit and external. This process is assumed to occur, even if it cannot be directly observed, in most theoretical accounts of choice response time. Although the task was a little unusual, the data from Brown et al.'s paradigm were quite standard: Hick's Law was observed in response time, and accuracy declined with increasing numbers of choice alternatives. Our first experiment expands on Brown et al.'s study to examine choices between even more alternatives (up to 20).

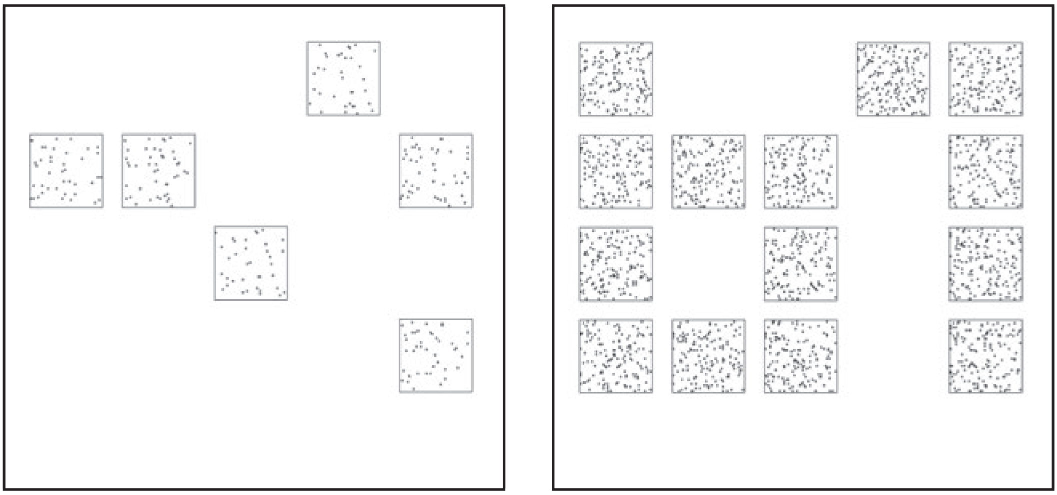
Another novel aspect of Experiment 1 is a new instantiation of Brown et al.'s (2009) experimental paradigm. Their paradigm used column heights as the primary display feature, which may have biased the data in favor of certain theoretical accounts (e.g., Brown et al. found that the best description of their data was provided by a theory that operated on the difference in height between the tallest and second-tallest columns). For Experiment 1, we re-cast the falling bricks from a novel perspective, as when watching drops of paint fall on a canvas. A demonstration version of this experiment can be viewed online, at http://psych.newcastle.edu.au/~sdb231/buckets/vanillaR.html.

2.1. Method

Fifty-seven first-year psychology students from the University of Newcastle participated online in Experiment 1 for course credit. Each participant completed six blocks of 30 trials. Each decision trial began with K empty squares, randomly placed into 20 locations on a 4×5 grid, with each square measuring 100×100 pixels (plus a 2-pixel border, see Fig. 1). The number of squares shown on any trial was randomly chosen from $K \in \{2,4,6,8,10,12,14,16,18,20\}$, subject to the condition that each K appeared equally often in each block.

During each trial, time proceeded in discrete steps at the rate of 15/s. On each time step, each square either accumulated a new dot or not. The chance of each square accumulating a new dot was independent and equal for all squares, at $\theta_{(d)} = 0.4$, except that one ''target'' square had a higher probability of $\theta_{(t)} = 0.5$. The participant's task was to identify the target as quickly as possible.

Squares began with a completely white background (unfilled), and each time a new dot was accumulated, a 2×2 pixel area within the square changed to a dark blue color. The position of the new dot was chosen randomly from the remaining unfilled area of the square. Participants were free to allow dots to accumulate until they felt confident with their



(A) $K = 6$ Trial

(B) $K = 14$ Trial

Fig. 1. Screenshots from Experiment 1 for (A) a choice between six alternatives and (B) a choice between fourteen alternatives. Panel (A) depicts the early stages of a trial with relatively few dots in each square, and panel (B) shows the later stages of a trial when many dots appear in each square. The task is to identify the square with the fastest rate of accumulating dots.

decision. The maximum number of fill events that could be accommodated by any square was 2,500, meaning that no square could completely fill in less than about 3 min (this was much longer than any participant ever waited to make a response).

After the participant chose a target square, feedback on his or her response was provided by a fast animation that illustrated many more time steps very quickly. If the participant correctly identified the target square, its border was turned green. If an incorrect selection was made, the incorrect square's border was turned red and the true target square's border was turned green.

2.2. Results and discussion

We imposed strong exclusion criteria, resulting in the removal of data from 19 participants who made fewer than 45% correct responses and two participants whose computers displayed an average of fewer than 13 time steps per second (for a discussion of these criteria, see Appendix). The remaining data were screened for outlying trials resulting in the removal of 9 trials with response times faster than 1 s, 6 trials slower than 100 s, and 25 trials during which the host computer displayed fewer than 13 time steps per second (0.62% of trials in total).

Mean response time and accuracy data are displayed in Fig. 2 as functions of the number of choice alternatives, using within-subjects standard error bars calculated according to Loftus and Masson (1994). Mean response time increased approximately linearly with log(*K*), in accordance with Hick's Law. A one-way within-subjects analysis of variance indicated a highly significant effect of the number of choice alternatives, $F(9, 315) = 49.8$, $p < .001$, with a significant linear trend, $F(1, 315) = 365$, $p < .001$; higher order trends were non-significant. There was some evidence that the $K = 20$ trials were treated differently than others; for example, the mean accuracy for the $K = 20$ condition was slightly higher than for $K = 18$, which did not occur for any other pair of conditions.

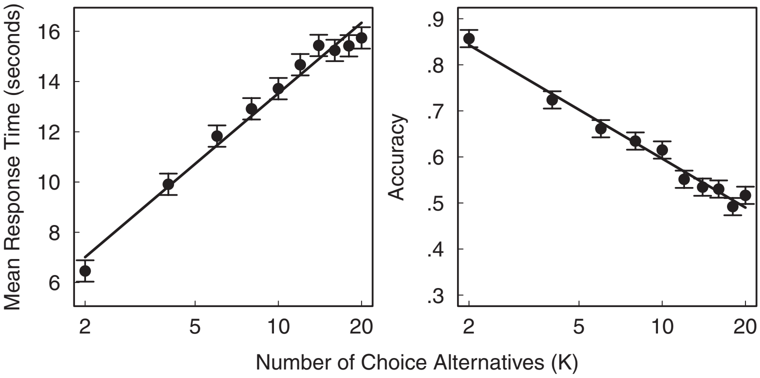


Fig. 2. Mean response time (left panel) and accuracy (right panel) from Experiment 1, as functions of the number of choice alternatives, *K*. The error bars represent ±1 within-subjects standard errors of the mean. The lines represent regression lines of best fit.

The right panel of Fig. 2 shows that response accuracy steadily declined as the number of choice alternatives increased, $F(9, 315) = 36.4$, $p < .001$. This gradual decline in accuracy differs from traditional investigations of Hick's Law that demanded errorless performance (Teichner & Krebs, 1974), but it replicates trends in data from more recent experiments where participants were allowed to make errors (e.g., Brown et al., 2009; Kveraga et al., 2002; Lacouture & Marley, 1995; Lee et al., 2005; Leite & Ratcliff, 2010).

3. Experiment 2

Our hypothesis is that lower accuracy is observed for higher-K choices because participants alter their speed–accuracy tradeoff settings in an attempt to even out the time taken for their different decisions. In Experiment 2, we examined the same range of K as in Experiment 1, except that we no longer used any trials with $K = 20$ (as we worried that the decision task was treated differently when there were no gaps in the 4×5 stimulus grid). The key change for Experiment 2 was that we manipulated set size on a between-subjects basis. For example, one participant was only asked for judgments about $K = 4$ alternatives, while another was only asked to make choices between $K = 16$ alternatives. If declining accuracy rates with increasing numbers of choice alternatives were due to a context effect, we should observe constant error rates across all set sizes for Experiment 2.

3.1. Method

A separate group of 159 first-year psychology students from the University of Newcastle participated online in Experiment 2 for course credit. Each participant was randomly allocated to one of nine conditions defined by the number of choice alternatives, making judgments about only one level of $K \in \{2,4,6,8,10,12,14,16,18\}$. Participants in each condition completed a different number of trials to balance the total duration of the experiment. All blocks were 15 trials in length, with $K = 2$ participants each completing 11 blocks, $K = 4$ completing 10 blocks, $K = 6$ and $K = 8$ each completing 9 blocks, $K = 10$ completing 8 blocks, $K = 12$ and $K = 14$ each completing 7 blocks, and $K = 16$ and $K = 18$ each completing 6 blocks. All other experimental details were the same as Experiment 1.

3.2. Results and discussion

Data from 57 participants who made fewer than 45% correct responses and 4 participants whose computer displayed an average of fewer than 13 time steps per second were excluded from analysis. The exclusion criteria did not exclude differential proportions of participants across set sizes, $\chi^2(8) = 13.8$, $p > .05$ (see Appendix for more discussion of excluded data). Remaining data were screened for outlying trials, resulting in the removal of 113 trials with response times faster than 1 s, 25 trials slower than 150 s, and 149 trials displaying fewer than 13 time steps per second (2.19% of responses).

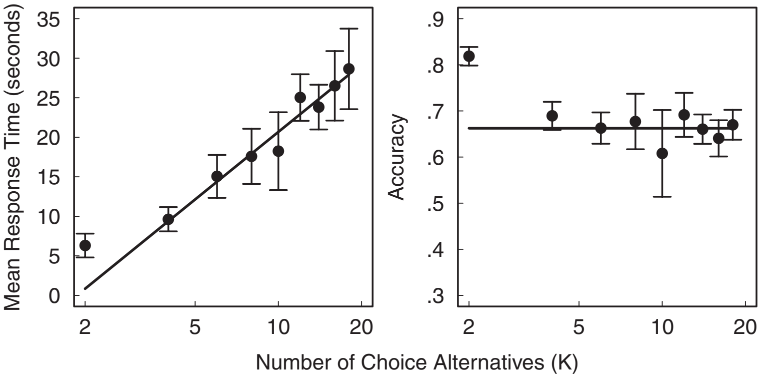


Fig. 3. Mean response time (left panel) and accuracy (right panel) from Experiment 2, as functions of the number of choice alternatives, K. The error bars represent ±1 between-subjects standard errors of the mean. The lines represent the regression line of best fit (left panel) and mean accuracy across set sizes (right panel), both excluding $K = 2$ data.

The left panel of Fig. 3 demonstrates that mean response time increased approximately linearly with log(K), consistent with the results of Experiment 1, supporting Hick's Law. A one-way between-subjects analysis of variance indicated a highly significant effect of the number of choice alternatives on response latency, $F(8, 89) = 6.29, p < .001$, with a significant linear trend, $F(1, 89) = 39.3$, $p < .001$; higher order polynomial trends were nonsignificant. Consistent with our hypothesis about a speed–accuracy tradeoff, mean response times for decisions in the larger set sizes were much slower in Experiment 2 than in Experiment 1, up to twice as long for $K = 18$.

The right panel of Fig. 3 illustrates response accuracy as a function of the number of choice alternatives. This time, accuracy was relatively constant at approximately 66% across all set sizes, apart from $K = 2$, which does not fit with the constant accuracy trend (we return to this point in Section 4 below). Supporting this claim, a one-way between-subjects analysis of variance indicated a significant effect of the number of choice alternatives on response accuracy, $F(8, 89) = 2.88$, $p < .01$; however, when the $K = 2$ group were excluded from the analysis, there was no longer a significant difference in accuracy across set sizes, $F < 1$.

4. Theoretical implications

The critical difference between Experiment 2—where accuracy was approximately constant—and Experiment 1—where accuracy decreased with increasing number of choice alternatives—was the number of different conditions experienced by the participants. These results are consistent with the hypothesis of a context-induced speed–accuracy tradeoff, and they set a new challenge for models of Hick's Law: to accommodate Hick's Law for response times when accuracy is independent of the number of choice alternatives, and also when accuracy declines with increasing number of choice alternatives. Further, since our data suggest that these patterns are functions of a speed–accuracy tradeoff, models should account for the two different patterns in terms of changes in tradeoff settings. We now briefly review domain general accounts for Hick's Law in light of our findings, and then demonstrate that an existing Bayesian model can be extended to account for our data by adjusting its speed–accuracy tradeoff mechanism.

The earliest domain-general accounts for Hick's Law were couched in communication theory (Hick, 1952; Hyman, 1953; for review, see Teichner & Krebs, 1974), but this approach has been heavily criticized, beginning with Laming (1966). Since then, most accounts of Hick's Law have cast each multi-alternative decision as a race between accumulators that collect evidence in favor of different responses. Only a few such models have been proposed as domain-general explanations (Brown et al., 2009; Schneider & Anderson, 2011; Usher et al., 2002).

Schneider and Anderson (2011) modeled multi-alternative decisions using the ACT-R cognitive architecture (Anderson et al. 2004), with a focus on memory retrieval of stimulus–response mappings. Our experiments have no obvious memory component, so it is not clear that Schneider and Anderson's model could be applied to our task. On the other hand, Usher et al. (2002) modeled multi-alternative decisions in terms of evidence accumulation, with a race model (Usher & McClelland, 2001). In its simplest form, the race model predicts that mean response times will be faster as more choice alternatives are added, but Usher et al. showed that the model can predict Hick's Law and constant accuracy rates, if the decision threshold parameter increases with the number of choice alternatives. Brown et al. (2009) examined two different accumulator models (an optimal Bayesian algorithm and the max-minus-next heuristic model) and showed that both produced Hick's Law, even when no parameters were changed across set sizes. While both models predicted Hick's Law, they made different predictions for response accuracy: The heuristic model predicted decreasing accuracy with increasing number of choice alternatives, and the Bayesian model predicted constant accuracy.

In principle, any of these models could predict both flat and decreasing accuracy rates by allowing changes in speed–accuracy tradeoff settings. For example, both the heuristic model from Brown et al. (2009) and Usher et al.'s (2002) accumulator model predict decreasing accuracy with more choice alternatives. These models could be extended to predict flat accuracy rates (as observed in Experiment 2) by assuming that participants in the different between-subjects groups of that experiment used different speed–accuracy tradeoff settings. An alternative version of this assumption would be to assume subjects in different conditions use a constant goal accuracy level and adjust their speed–accuracy tradeoff settings to achieve that goal. Conversely, Brown et al.'s (2009) Bayesian model predicts constant accuracy by default, and so might accommodate declining accuracy rates by allowing the response threshold parameter to decrease with increased number of choice alternatives.

Using Brown et al.'s (2009) Bayesian model, we demonstrate here that it is possible to account for the empirical context effect within an existing domain-general model of multi-alternative choice. We present these model fits as an initial exploration into modeling context effects. This is not to say, however, that the Bayesian model is to be preferred, or that it provides the best theoretical account of context effects in Hick's Law. To do this, future work must further explore context effects across various multi-alternative choice paradigms to provide a comprehensive test bed for comparison of domain-general models.

4.1. An ideal observer account of Hick's Law

Brown et al.'s (2009) Bayesian model is ''optimal'' in the sense that, for some predetermined accuracy rate (say, c), the expected decision time is minimized. At each time step the optimal decision maker must choose between terminating evidence accumulation and selecting the current-best alternative or observing more evidence. To this end, the model calculates—at each time step—the posterior probability that each response alternative is the target, and it makes a response as soon as the largest of these posterior probabilities exceeds c. Thus, the predicted response accuracy is c, regardless of the number of choice alternatives. With no more assumptions, this model predicts Hick's Law for response times. Choices between many alternatives are slowed because (a) the prior probabilities start lower, at $\frac{1}{K}$, and (b) the posterior probabilities rise more slowly because the current-best alternative is more likely to be similar to one of the others due to sampling noise. We briefly describe the calculations for the model below, but for full details see Brown et al. (2009).

Eq. 2 denotes the hypothesis that the ith choice alternative is the target with H_i and observed data with D. According to Bayes' theorem, the posterior model probability that choice alternative h is the target is

$$p(H_h|D) = \frac{p(D|H_h)p(H_h)}{\sum_j p(D|H_j)p(H_j)}. \tag{2}$$

We assume the a priori probabilities for each alternative are equal, simplifying Eq. 2 to

$$p(H_h|D) = \frac{p(D|H_h)}{\sum_j p(D|H_j)}. \tag{3}$$

The model calculates posterior probabilities assuming full knowledge of the statistical data generating process. For alternative i, let s_i denote the number of ''successes'' (i.e., the number of dots alternative i has accumulated) and $f_i = n - s_i$ denote the number of ''failures'' (i.e., the number of time steps where alternative i did not accumulate a dot) at time step n. Considering the hypothesis that choice alternative h is the target, H_h, the data s_h from n time steps originated from a binomial process with rate parameter $\theta_{(t)} \in [0,1]$: $p(s_h, n|H_h) = \binom{n}{s_h}\theta_{(t)}^{s_h}(1-\theta_{(t)})^{f_h}$. Correspondingly, the data for each of the remaining choice alternatives s_j, where $j \neq h$, originated from binomial processes with rate parameter $\theta_{(d)}$. This means the likelihood of the data under H_h can be given as $p(D|H_h) = \binom{n}{s_h}\theta_{(t)}^{s_h}(1-\theta_{(t)})^{f_h}\prod_{j\neq h}\binom{n}{s_j}\theta_{(d)}^{s_j}(1-\theta_{(d)})^{f_j}$. The same calculations apply to the hypothesis that any other alternative is the target (i.e., has the largest rate parameter), producing, after some simplification:

$$p(H_h|D) = \frac{\left(\theta_{(t)}/\theta_{(d)}\right)^{s_h}\left[\left(1-\theta_{(t)}\right)/\left(1-\theta_{(d)}\right)\right]^{f_h}}{\sum_k \left\{\left(\theta_{(t)}/\theta_{(d)}\right)^{s_k}\left[\left(1-\theta_{(t)}\right)/\left(1-\theta_{(d)}\right)\right]^{f_k}\right\}}, \tag{4}$$

where $k = 1, \ldots,$ and K indexes the hypotheses entertained by the decision maker. For full derivation of Eq. 4, see Brown et al. (2009).

Fig. 4 illustrates predictions of the Bayesian optimal observer overlaid on accuracy and response time data from both experiments. Experiment 2 is shown as the left column, and the Bayesian model's predictions for response accuracy (with $c = 0.66$) fit the data quite well, except for the unusually high accuracy for binary decisions ($K = 2$).

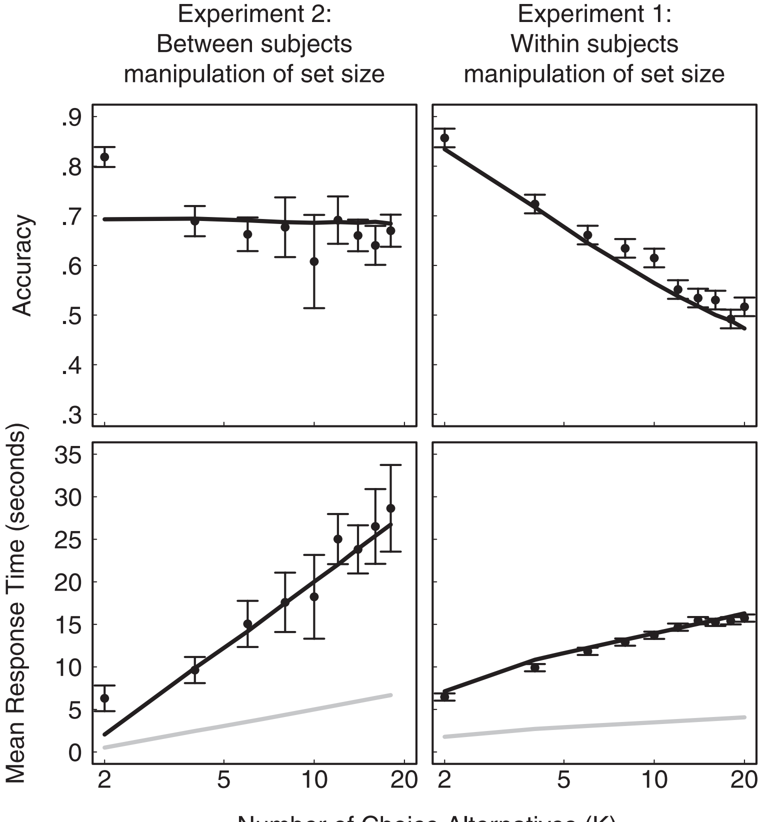


Fig. 4. Mean accuracy and response time data (upper and lower panels, respectively) for the between- and within-subjects manipulations of set size (left and right panels, respectively). The error bars represent ±1 between- or within-subjects standard errors of the mean, depending on the experimental design. Accuracy predictions of the Bayesian optimal observer are shown by black lines in the upper panels. Unscaled and scaled response time predictions are shown in the lower panels by gray and black lines, respectively.

The lower-left panel shows the Bayesian model's predictions for response times as a gray line. The model predicts much faster responses than the participants gave, which is to be expected of an optimal model. The black line on this panel shows the model's predicted response times slowed down by a factor of 4×, and these predictions match data. We propose that this four-fold slowing may arise because observers suffer from a perceptual limitation in this task. The stimulus display was created from very small dots, each just 2×2 pixels in size. We suggest that participants were unable to resolve such small differences and instead grouped neighboring dots into 4×4 pixel regions. Suppose these larger regions (comprising four of the smaller dots) could only be perceived as completely filled or unfilled, and the probability of perceiving the region as filled was given by the proportion of filled dots within it (e.g., if three of the four dots inside the region were filled, the participant would perceive the entire region as filled with probability .75, and as unfilled with probability .25). With these assumptions, the distribution of small dots at time t—binomial, with rate θ and size t—is four-fold scaled to make a binomial distribution for the larger dots—still with rate θ, but with size $\frac{t}{4}$. With this perceptual limitation, Eq. 4 predicts the same posterior probabilities at time $\frac{t}{4}$ as would otherwise be predicted for time t. A natural prediction of the assumed perceptual limitation is that response times should be unaffected by using even smaller dots than our current task, assuming those dots are then grouped to the same perceptual limit (4×4 pixels).

4.2. Modeling decreasing accuracy with increasing number of choice alternatives

The very simplest way to allow the Bayesian model to predict different accuracy across different numbers of choice alternatives is to estimate a free parameter for the response threshold (c) for each set size. That approach perfectly accommodates the response accuracy data by setting c equal to the observed accuracy at each set size. Even so, the approach is constrained because it need not fit the response time data. We have explored this approach and found that it accounts for both accuracy and response time data from Experiment 1 quite well, with no other changes to the model.

Estimating a free parameter for the response threshold at each set size addresses the question of *what* participants were doing (altering their speed–accuracy tradeoff across set sizes) but not *why* they were doing it. We investigated this question using an alternative definition of optimality: that the observer's goal was to finish the experiment in minimum time, subject to reaching some pre-set accuracy goal. For example, suppose participants had in mind a pre-set accuracy goal of 56%. In Experiment 2, where each participant experienced just one set size, this goal simply implies that every participant should set $c = 0.56$, which would lead to the observed flat accuracy profile. However, the within-subjects design of Experiment 1 provides participants with more freedom. One could attain an overall accuracy of 56% in many ways, by balancing higher and lower accuracy across set sizes. One possible approach is to vary accuracy across set sizes in a way that attains the desired overall accuracy but minimizes the time taken. For a fixed accuracy goal, and with the fixed number of trials in our experiment, the goal of minimizing experiment time is equivalent to the well-studied goal of maximizing the ''reward rate'' for that fixed level of accuracy[1] (see, e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). We investigated the model through simulation and found that total experiment time is approximately minimized, subject to an overall accuracy of 56%, by setting response thresholds according to a simple power function: $c = K^{-0.27}$. This model provides a good account for the response accuracy data from Experiment 1 and also predicts that participants will finish the experiment about 6% quicker than if they set the equivalent fixed accuracy across all conditions ($c = 0.56$). Using this power function to fit the response accuracy data, and making no other changes, the model still predicts the response time data from Experiment 1 quite closely, as shown in the right column of Fig. 4.

The notion of a ''goal accuracy'' rate may also help explain why the model failed to fit the $K = 2$ data from the between-subjects manipulation of set size. The model misfit may be due to asymmetry between shifting up and down speed–accuracy settings. For instance, by default, the $K = 18$ condition results in very low accuracy (these trials are very difficult), so those participants would need to increase their response threshold to move accuracy performance up to the goal rate. In contrast, the $K = 2$ participants easily exceed the goal accuracy rate (since these trials are easy), and it appears that these participants stick with their higher-than-goal accuracy performance. Hence, it may be that decision makers do not mind performing better than their goal rate, and do not trade their accuracy performance down, but are not happy to perform worse than the goal. In this sense, the ''goal'' accuracy may instead be better thought of as a ''minimum acceptable'' accuracy. Although this explanation is speculative, it is open to empirical testing. For example, one could design an experiment with a between-subjects manipulation of an experimental parameter that has some conditions that are far easier than others, which would produce performance above the goal rate for those groups, and test whether those participants will choose to trade their accuracy down.

The Bayesian ideal observer not only describes mean response times but also their distribution. The lines in Fig. 5 depict, from bottom to top, the 10%, 30%, 50% (i.e., median), 70%, and 90% quantiles of the response time distributions for the between- and within-subjects manipulations of set size (Experiments 2 and 1; left and right panels, respectively). Data are represented by gray lines and model predictions by black lines. Apart from $K = 2$ data for the between-subjects manipulation, the model does quite well at predicting the entire response time distribution.

4.3. Limitations of the model

Although the Bayesian model provides a good fit to the data, some of its assumptions are almost certainly too simplistic. For example, we do not consider non-decision components of choice, such as the time taken to encode stimuli and perform a motor response. These components are almost always included as a fixed offset parameter in models of response time (e.g., Usher et al., 2002). Such an approach could be incorporated in the Bayesian model easily enough. We omitted this assumption for simplicity because the model fit the data well enough without it.

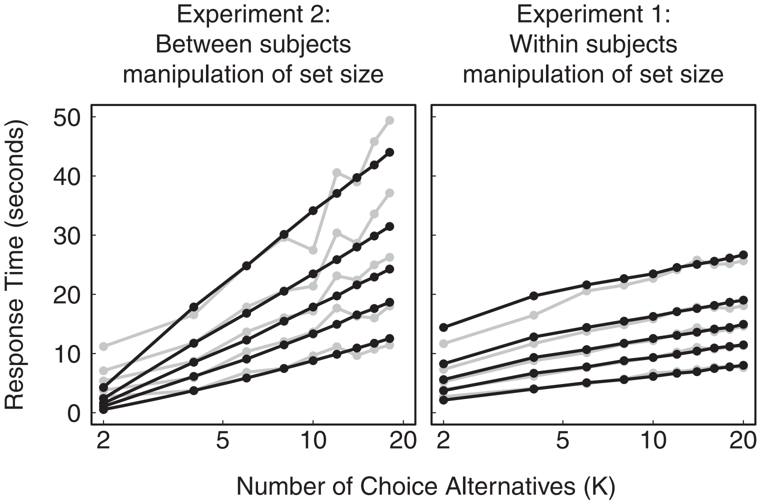


Fig. 5. Response time quantiles for between- and within-subjects manipulations of set size (left and right panels, respectively). Data and predictions of the Bayesian optimal observer are shown in gray and black lines, respectively. The panels show five percentiles of the response time distribution (from bottom to top, 10th, 30th, 50th, 70th, 90th) as functions of the number of choice alternatives, *K*.

Our model also does not include a mechanism for visual scanning of the display, or switching of attention between the response alternatives. While these elements almost certainly influence the data, we again opted for the simplicity obtained by not including corresponding model assumptions. It is likely that the model still fits the data even without modeling attention-switching processes because of the particularly long response times in our experiments: The average response time was about 12 s. In the task switching and visual search paradigms, the speed of visual scanning and attention switching are usually estimated to be orders of magnitude faster than this. Thus, visual scanning and attention switching could occur repeatedly during each trial in our experiment without appreciably increasing the observed response time.

Another possible limitation of the model is that the reasoning behind our four-fold scaling of model response time predictions has yet to be confirmed in data. The externalized evidence accumulation paradigm we use here has the benefit of allowing such detailed hypotheses to be empirically tested, at least in principle (e.g., using experiments that manipulate the speed of the display and the size of the dots). However, we note that most mathematical models of decision time include an arbitrary mechanism to convert model time into real time, and it is a strength of our model that this scaling factor is testable and explicit.

5. General discussion

Recent explorations of multi-alternative choice have reported declining response accuracy with increasing numbers of response alternatives. This important empirical result has been incorporated into the architecture of some theoretical accounts of choice (e.g., Brown et al., 2009; Lacouture & Marley, 1995) or can be modeled under various parameter settings in others (e.g., Usher et al., 2002). We propose the increase in error rates may be due to a speed–accuracy tradeoff induced by within-subjects manipulations of experimental parameters. The two experiments presented here support this assertion: A within-subjects manipulation of the number of choice alternatives produced response accuracy that steadily declined across set sizes, whereas the same parameter manipulated between subjects demonstrated approximately constant accuracy. As a consequence, response times were much slower in difficult conditions of the between-subjects experiment than in the corresponding conditions of the within-subjects experiment.

These results are consistent with previous research. We identified nine studies that reported declining accuracy rates with increasing numbers of choice alternatives (Brown et al., 2009; Churchland et al., 2008; Kveraga et al., 2002; Lacouture & Marley, 1995; Lee et al., 2005; Leite & Ratcliff, 2010; ten Hoopen et al., 1982; Thiem et al., 2008; Wright et al., 2007) and found that all of them employed within-subjects manipulations of set size. This is in stark contrast to the approximately constant error rates for between-subjects manipulations of set size observed here and elsewhere (Hale, 1968).

Studies in which within-subject manipulations of set size produced close-to-constant accuracy provide a challenge to our account. However, all such studies that we reviewed have used experimental paradigms known not to conform to Hick's Law. For instance, response latency is almost independent of the number of choice alternatives when stimulus–response compatibility is very high (such as saccades to one target among many), and so in that case it is not surprising that accuracy also does not change across set sizes (e.g., see pro-saccades in Kveraga et al., 2002; Kveraga & Hughes, 2005; Lawrence, 2010; Lawrence & Gardella, 2009; Lawrence, John, Abrams, & Snyder, 2008). Similar data are observed for movements toward a target, using a joystick or stylus (Pellizzer & Hedges, 2003; Wright et al., 2007). Such data can be accounted for within the Bayesian model presented here. For instance, high stimulus–response compatibility is akin to a target that is highly distinct from distractors, which can be represented in the model with a large value for $\theta_{(t)}$ and a low value for $\theta_{(d)}$. Such settings lead to model predictions that qualitatively match the data—fast response times that are only marginally affected by increasing set size and close-to-constant accuracy.

Some more traditional approaches to Hick's Law also observed constant accuracy even when the number of response alternatives was manipulated within subjects (e.g., Teichner & Krebs, 1974; Welford, 1980). However, these studies employed explicit measures to enforce close-to-perfect accuracy in all conditions, which preclude the speed–accuracy tradeoff we propose. This suggests some natural extensions of our experiments to further test our hypothesis. For example, our hypothesis implies that participants should produce constant error rates across set sizes, even when this is manipulated within subjects, if task timing were carefully adjusted such that different set sizes resulted in similar decision times. For instance, in our paradigm we could vary the target and distractor fill rates across set sizes in such a manner to produce approximately constant response times across set sizes. If the declining accuracy with increasing number of choice alternatives is really due to observers' attempts to minimize total experiment time, the effect should be attenuated or even removed by careful adjustment of the stimulus speed. Interestingly, this experimental test applies not just to the Bayesian model we outline, but to the more general hypothesis of a context-induced speed–accuracy tradeoff—no matter which model framework it is implemented within.

This proposed experiment would also provide an interesting challenge for some theories of Hick's Law. The Bayesian model we developed predicts constant accuracy by default and predicts decreasing accuracy for within-subjects manipulations through an assumption about observer's minimizing their total experiment time. The above experiment, which would manipulate total experiment time, should influence (negate) the assumed speed–accuracy tradeoff. On the other hand, models such as the max-minus-next heuristic naturally predict decreasing accuracy rates. Such a model could be naturally extended to predict constant accuracy in the between-subjects experiment, as described earlier. However, this model is bound to *always* predict decreasing accuracy for the within-subjects manipulation of set size, and so it predicts that response accuracy should be unaffected by the experiment proposed above, in which stimulus speed is manipulated. This is because, in the max-minus-next model, decreasing accuracy occurs simply due to the statistical properties of the stimulus, with no consideration of the time taken to make responses. Experiments to investigate these differential predictions are currently underway in our lab.

Future work should explore whether the context effects observed here can be generalized to more traditional Hick's Law tasks that do not involve possible perceptual limitations (and instead may impose memory requirements) on the decision maker. It is possible that the context effect we report may be specific to tasks similar to our paradigm, where the accrual of evidence is both explicit and external. One could explore this hypothesis by manipulating the experimental context in more traditional choice tasks. For instance, Schneider and Anderson (2011) designed a task where a series of letters were each paired with a number, and the decision maker was later presented with one of these letters and was asked to recall the number previously paired with that letter. In this paradigm, one could incorporate a within- and between-subjects manipulation of set size, where some participants would only experience a couple of letter–number pairs (i.e., a $K = 2$ group), while others would make judgments about multiple letter–number pairs (e.g., a $K = 6$ condition). This kind of additional experimentation will be instructive in determining the generality of the context effects we propose here, as well as guide processes of model comparison and selection. Such experiments are also currently under way in our laboratory.

In conclusion, our results provide a possible mechanism to integrate previously divergent empirical findings from Hick's Law through the observation of a simple design consideration and a plausible hypothesis about observers' speed–accuracy tradeoff behavior. In addition to reconciling discrepant findings, our work has important theoretical implications for models of Hick's Law. We demonstrated that a previously dismissed Bayesian model described the data well and provided a parsimonious account of the data. Future research must explore the extent to which contexts effects emerge in other multi-alternative choice paradigms, and it must determine whether the Bayesian model provides the best quantitative account of context effects relative to competing models of decisions between multiple alternatives.

Note

[1] With fixed accuracy and number of trials, there is also a fixed number of correct responses to be made.

Acknowledgments

This research was supported by the Keats Endowment Research Fund. We thank Keith Nesbitt for his help in programming the tasks, and Marius Usher and an anonymous referee for their helpful comments on earlier drafts of this manuscript.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two–alternative forced choice tasks. *Psychological Review*, *113*, 700–765.

Brainard, R. W., Irby, T. S., Fitts, P. M., & Alluisi, E. A. (1962). Some variables influencing the rate of gain of information. *Journal of Experimental Psychology*, *63*, 105–110.

Brown, S., Steyvers, M., & Wagenmakers, E-J. (2009). Observing evidence accumulation during multi-alternative decisions. *Journal of Mathematical Psychology*, *53*, 453–462.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*, 693–702.

Dassonville, P., Lewis, S. M., Foster, H., & Ashe, J. (1999). Choice and stimulus-response compatibility affect duration of response selection. *Cognitive Brain Research*, *7*, 235–240.

Hale, D. J. (1968). The relation of correct and error responses in a serial choice reaction task. *Psychonomic Science*, *13*, 299–300.

Hale, D. J. (1969). Speed-error tradeoff in a three-choice serial reaction task. *Journal of Experimental Psychology*, *81*, 428–435.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11–26.

ten Hoopen, G., Akerboom, S., & Raaymakers, E. (1982). Vibrotactual choice reaction time, tactile receptor systems and ideomotor compatibility. *Acta Psychologica*, *50*, 143–157.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188–196.

Kveraga, K., Boucher, L., & Hughes, H. C. (2002). Saccades operate in violation of Hick's Law. *Experimental Brain Research*, *146*, 307–314.

Kveraga, K., & Hughes, H. C. (2005). Effects of stimulus-response uncertainty on saccades to near-threshold targets. *Experimental Brain Research*, *162*, 401–405.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology*, *39*, 383–395.

Laming, D. R. J. (1966). A new interpretation of the relation between choice–reaction time and the number of equiprobable alternatives. *British Journal of Mathematical and Statistical Psychology*, *19*, 139–149.

Laursen, A. M. (1977). Task dependence of slowing after pyramidal lesions in monkeys. *Journal of Comparative and Physiological Psychology*, *91*, 897–906.

Lawrence, B. M. (2010). An anti-Hick's effect for exogenous, but not endogenous, saccadic eye movements. *Experimental Brain Research*, *204*, 115–118.

Lawrence, B. M., & Gardella, A. L. (2009). Saccades and reaches behaving differently. *Experimental Brain Research*, *195*, 413–418.

Lawrence, B. M., John, A. S., Abrams, R. A., & Snyder, L. H. (2008). An anti-Hick's effect in monkey and human saccade reaction times. *Journal of Vision*, *8*, 1–7.

Lee, K-M., Keller, E. L., & Heinen, S. J. (2005). Properties of saccades generated as a choice response. *Experimental Brain Research*, *162*, 278–286.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy data of multiple-alternative decisions. *Attention, Perception & Psychophysics*, *72*, 246–273.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Pachella, R. G., & Fisher, D. (1972). Hick's Law and the speed-accuracy trade-off in absolute judgment. *Journal of Experimental Psychology*, *92*, 378–384.

Pellizzer, G., & Hedges, J. H. (2003). Motor planning: Effect of directional uncertainty with discrete spatial cues. *Experimental Brain Research*, *150*, 276–289.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Author: Vienna, Austria (ISBN 3-900051-00-3).

Schneider, D. W., & Anderson, J. R. (2011). A memory-based model of Hick's Law. *Cognitive Psychology*, *62*, 193–222.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Teichner, W. H., & Krebs, M. J. (1974). Laws of visual choice reaction time. *Psychological Review*, *81*, 75–98.

Thiem, P. D., Hill, J. A., Lee, K.-M., & Keller, E. L. (2008). Behavioral properties of saccades generated as a choice response. *Experimental Brain Research*, *186*, 355–364.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's Law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704–715.

Vickrey, C., & Neuringer, A. (2000). Pigeon reaction time, Hick's Law, and intelligence. *Psychonomic Bulletin & Review*, *7*, 284–291.

Welford, A. T. (1980). *Reaction times*. London: Academic Press.

Wright, C. E., Marino, V. F., Belovsky, S. A., & Chubb, C. (2007). Visually guided, aimed movements can be unaffected by stimulus-response uncertainty. *Experimental Brain Research*, *179*, 475–496.

Appendix: Exclusion criteria

Given the long response latencies possible in our experimental paradigm, we observed very large between-participant variability, particularly in Experiment 2, even within a single experimental condition. For example, in Experiment 2 an unusually high proportion of poorly performing participants were randomly allocated to the $K = 10$ condition. To reduce the high within-condition variance, we imposed a strong exclusion criterion: removing data from participants with mean accuracy below 45%. The motivation behind this exclusion criterion was to produce cleaner data for analysis, as long as the qualitative trends in the data were not altered. To check this, we reproduced the figures of mean response time and accuracy data from the main text using a more lenient exclusion criterion: Participants were only removed if they had mean accuracy lower than 25% (gray

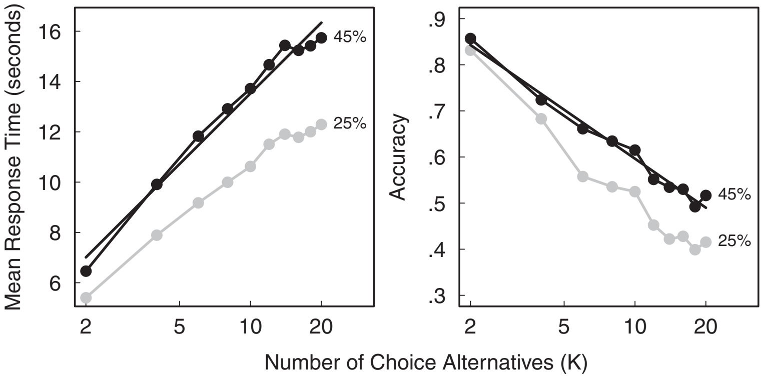


Fig. A1. Mean response time (left panel) and accuracy (right panel) from Experiment 1, as functions of the number of choice alternatives, *K*. The dotted lines represent exclusion criteria of mean accuracy below 45% (black lines) and 25% (gray lines). The straight lines represent regression lines of best fit for exclusion criterion 45%.

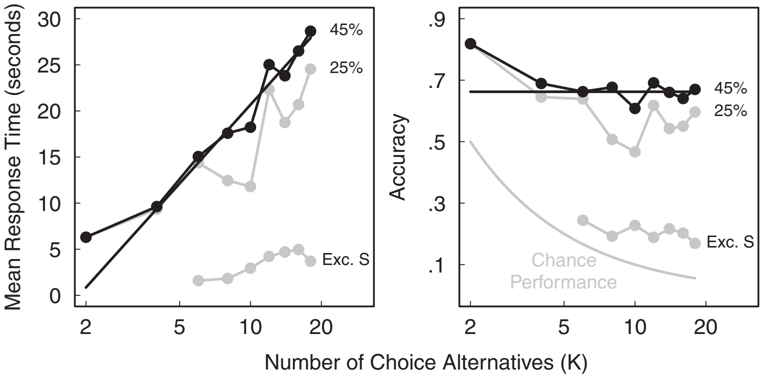


Fig. A2. Mean response time (left panel) and accuracy (right panel) from Experiment 2, as functions of the number of choice alternatives, *K*. The dotted lines represent exclusion criteria of mean accuracy below 45% (black lines) and 25% (gray lines). The participants excluded from analysis when using mean accuracy 25% exclusion are plotted as the lower gray line (labeled Exc. S). The straight lines represent the regression line of best fit to data (left panel) and mean accuracy across set sizes for exclusion criterion 45% (right panel), both excluding $K = 2$.

dotted lines). We compare the data using this exclusion criterion to the analysis from the main text, using the strict exclusion criterion (black dotted lines), for both Experiment 1 (Fig. A1) and Experiment 2 (Fig. A2). Error bars have been omitted to keep the figures readable (error bars for the more lenient analysis are wider than the standard ones shown in the main text).

Importantly, altering the exclusion criterion did not alter the general patterns in data in either experiment. In Experiment 1, there was a global decline in response time and accuracy when using the more lenient exclusion criterion. When employing a 25% exclusion criterion, the only data excluded from analyses were from the two participants whose host computer displayed fewer than 13 time steps per second (these data are not displayed).

In Experiment 2, the more lenient exclusion criterion only altered the number of participants excluded for choices between six or more alternatives ($K \geq 6$). For these conditions, the more lenient exclusion criterion resulted in generally faster response times and lower accuracy. The data from the excluded participants (lowest gray lines marked ‘‘Exc. S’’ in Fig. A2) showed response times that were almost independent of the number of choice alternatives, and response accuracy not much above chance—but, importantly, their mean accuracy was still relatively constant as the number of choice alternatives increased.

Readers who wish to explore our data and exclusion criteria in more detail can find the raw data files from Experiments 1 and 2 in the ‘‘publications’’ section of the first and second authors’ website at http://www.newcl.org/, as well as code to read the raw data files into an interpretable format in the freely available R language (R Development Core Team, 2011).