Psychon Bull Rev (2012) 19:339–348
DOI 10.3758/s13423-012-0216-z

BRIEF REPORT

An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives

Guy E. Hawkins · Scott D. Brown · Mark Steyvers · Eric-Jan Wagenmakers

Published online: 3 February 2012

Abstract Decisions between multiple alternatives typically conform to Hick's Law: Mean response time increases log-linearly with the number of choice alternatives. We recently demonstrated context effects in Hick's Law, showing that patterns of response latency and choice accuracy were different for easy versus difficult blocks. The context effect explained previously observed discrepancies in error rate data and provided a new challenge for theoretical accounts of multialternative choice. In the present article, we propose a novel approach to modeling context effects that can be applied to any account that models the speed–accuracy trade-off. The core element of the approach is "optimality" in the way an experimental participant might define it: minimizing the total time spent in the experiment, without making too many errors. We show how this approach can be included in an existing Bayesian model of choice and highlight its ability to fit previous data as well as to predict novel empirical context effects. The model is shown to provide better quantitative fits than a more flexible heuristic account.

Keywords Decision making · Hick's Law · Context effect · Bayes · Model selection · Parameter space partitioning

G. E. Hawkins (✉) · S. D. Brown
University of Newcastle,
Callaghan, New South Wales, Australia
e-mail: guy.e.hawkins@gmail.com

M. Steyvers
University of California,
Irvine, CA, USA

E.-J. Wagenmakers
University of Amsterdam,
Amsterdam, the Netherlands

Most decision-making research has focused on binary decisions. For decisions between more than two alternatives, the typical result is Hick's Law (also known as the Hick–Hyman Law; Hick, 1952; Hyman, 1953), which states that mean response time (RT) increases linearly with the logarithm of the number of choice alternatives (K):

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

Hick's Law is a robust phenomenon that has received empirical support across many experimental paradigms, including absolute identification (e.g., Lacouture & Marley, 1995; Pachella & Fisher, 1972), eye saccades (e.g., see antisaccades in Kveraga, Boucher, & Hughes, 2002; Lee, Keller, & Heinen, 2005), rapid perceptual choice (e.g., Leite & Ratcliff, 2010), and expanded judgment (Brown, Steyvers, & Wagenmakers, 2009). However, there have also been counter examples to Hick's Law. For instance, in tasks with high stimulus-response compatibility RT does not increase with set size (e.g., Dassonville, Lewis, Foster, & Ashe, 1999; prosaccades in Kveraga et al. 2002; ten Hoopen, Akerboom, & Raaymakers, 1982; Wright, Marino, Belovsky, & Chubb, 2007), nor do RTs with extensively practiced stimulus sets (e.g., vocal responses to visually presented digits; Brainard, Irby, Fitts, & Alluisi, 1962; for general overview of Hick's Law, see Schweickert, 1993; Teichner & Krebs, 1974; Welford, 1980). Another limitation of the empirical support for Hick's Law is that it has rarely been tested for choices between more than eight alternatives, because of methodological limitations. Nevertheless, in the few existing examinations of Hick's Law with large set sizes, there is consistent evidence that the log-linearity in RTs remains (e.g., Beh, Roberts, & Pritchard-Levy, 1994; Brown et al., 2009; Hawkins, Brown, Steyvers, &

Wagenmakers, in press; Lee, Heo, & Chang, 2006; as well as the experiment described below).

Historically, Hick's Law was interpreted in terms of information theory (Hick, 1952; Shannon & Weaver, 1949), where mean RT is proportional to the amount of information in the stimulus. To control the amount of information processed, experimenters forced the observer to respond with perfect accuracy. However, more recent investigations that have allowed the observer to determine his or her own speed–accuracy trade-off, and hence commit errors, have still observed Hick's Law (e.g., Brown et al., 2009; Kveraga et al., 2002; Lacouture & Marley, 1995; Lee et al., 2005; Leite & Ratcliff, 2010). That is, mean RT increases linearly with the stimulus information actually processed by the observer, and also with the logarithm of choice set size (e.g., Hale, 1969; Pachella & Fisher, 1972).

In an experiment that allowed decision makers to determine their own speed–accuracy trade-offs, Hawkins et al. (in press) demonstrated context effects in Hick's Law that had powerful effects on qualitative patterns in response accuracy data. The context effect was due to the different conditions the participant experiences—their "decision context." For example, manipulating the number of choice alternatives on a within-subjects basis resulted in very different data than an otherwise-identical between-subjects manipulation. When manipulated within subjects, participants made choices across many different set sizes over trials, from two alternatives up to twenty (i.e., a variable context). The corresponding between-subjects manipulation required each participant to make decisions about only one set size (i.e., a nonvariable context). In the variable context, decision makers tended to "even out" their decision times by making faster but less accurate decisions for difficult conditions (many choice alternatives) and slower but more accurate decisions for the easy conditions (few choice alternatives), as compared with the nonvariable context.

These trade-offs between speed and accuracy suggest the accumulation of different amounts of evidence across set sizes: As compared with the nonvariable context, decision makers in the variable context waited for more evidence in easy-choice conditions and less evidence in hard-choice conditions. This account unified previously discrepant findings that sometimes accuracy declined as the number of choice alternatives increased (e.g., Brown et al., 2009; Lacouture & Marley, 1995; Leite & Ratcliff, 2010), whereas in other cases, accuracy remained constant (e.g., Hale, 1968; Rabbitt, 1968).

Modeling context effects in multialternative decisions

Different models of multialternative choice make different predictions about error rates, including: declining accuracy as choice set size increases (e.g., the max-minus-next model of Brown et al., 2009); constant, zero error rates (e.g., the ACT-R memory retrieval model of Schneider & Anderson, 2011, which, with modification, can provide a limited account of nonzero error rates); or constant, nonzero error rates (e.g., the Bayesian optimal observer of Brown et al., 2009, and the race model of Usher, Olami, & McClelland, 2002). Context effects can be included in any of these models in two different ways. First, models that naturally predict constant accuracy rates can assume a speed–accuracy trade-off in variable decision contexts. This allows those models to accommodate the declines in accuracy observed when participants experience multiple conditions, while still accommodating the constant accuracy observed in nonvariable contexts. Second, models that naturally predict decreasing accuracy rates can assume an opposite speed–accuracy trade-off across groups of participants in nonvariable decision contexts. This allows those models to accommodate the observed constant accuracy in those conditions.

We describe a new approach to explain these speed–accuracy trade-offs, extending the ideas of Hawkins et al. (in press). Our approach constrains and simplifies the models by replacing the free parameters associated with different speed–accuracy trade-offs in different conditions with a notion of optimality. We describe experiments in which participants made judgments in multiple conditions that each required a speed–accuracy trade-off setting. We show that the trade-offs can be explained—in an almost parameter-free manner—by assuming that participants attempt to finish the experiment as quickly as possible, conditional on maintaining a goal level of accuracy. The approach of minimizing RT can be applied to any model of multialternative choice with a speed–accuracy criterion parameter. We apply our approach to an existing Bayesian model to illustrate its ability to account for context effects. We use the Bayesian model only as a vehicle to demonstrate the utility of our RT minimizing approach, rather than to espouse the Bayesian model as a complete account of multialternative choice. Since declining accuracy rates are not the default prediction for the Bayesian model, we compare its quantitative fits with an alternative account, the max-minus-next heuristic, which naturally predicts declining accuracy rates. The RT minimizing approach allows the Bayesian model to provide a better account of data than a competing—and more flexible—model.

Minimizing total experiment time The Bayesian ideal observer is "optimal" in the sense that, for some predetermined accuracy rate (a "criterion," c), the expected decision time is minimized. At a sequence of discrete time steps during the decision process, the model calculates the posterior probability that each response alternative is the correct response and responds as soon as the largest of these probabilities exceeds the response criterion. Thus, by default, the Bayesian model

predicts constant accuracy at c for any number of choice alternatives (no more details of the Bayesian model are required to follow our approach, but for a full description see Brown et al., 2009).[1]

We illustrate our new approach to establishing speed–accuracy trade-off settings using the data from Experiment 1 reported by Hawkins et al. (in press). Participants in that task made judgments in many different set sizes that were randomized across trials, from $K = 2 \ldots 20$. Figure 1 shows that response accuracy (crosses) steadily declined as the number of choice alternatives increased.

We model these data by assuming that participants have a goal accuracy rate. For illustrative purposes, we assume a goal accuracy rate of 60%, the mean accuracy observed in Hawkins et al.'s (in press) data. If participants are free to set a different response criterion (c) for each set size, then there are many different ways one could achieve 60% accuracy. For example, the simplest approach to achieving 60% accuracy is to set the same response criterion, $c = .6$, for all set sizes, shown as a dotted gray line in Fig. 1. With this constant response criterion the Bayesian model predicts mean RT of 16.07 s across set sizes. An alternative approach is to set response criteria that steeply decline as set size increases, so that choices between few alternatives are far more accurate than decisions between many. The dashed gray line in Fig. 1 illustrates this approach, predicting an average RT of 16.32 s. These are just two examples of theoretically many possible combinations of response criteria across set sizes that result in 60% correct responses.

Out of all the many ways to set the response criteria, there is one setting that results in the fastest mean RT. This setting will satisfy the goal of meeting the required accuracy rate (60%) while minimizing the total time required to complete the experiment; the associated criterion values are optimal in this sense only. We refer to this approach as "Min-RT." To find these optimal criterion settings in our simulation studies, we conducted a brute force search over the range of response criterion settings for each set size to examine the predicted mean RT and accuracy of the model.

The criterion settings selected by Min-RT are completely determined by the goal accuracy (60%) because they are the unique set of criteria that minimize RT for that goal. The response criterion settings predicted by the Min-RT approach for a goal accuracy of 60% are shown by the black line in Fig. 1. Intriguingly, these criterion settings closely match the observed data (i.e., the black line is close to the crosses). The Min-RT approach is attractive because the goal of minimizing the total time subject to a goal accuracy rate is both simple and transparent. The approach is also highly constrained because for any goal accuracy rate, Min-RT selects just one response criterion setting for each set size. We return to the question of how participants might find the optimal settings in the Discussion section.

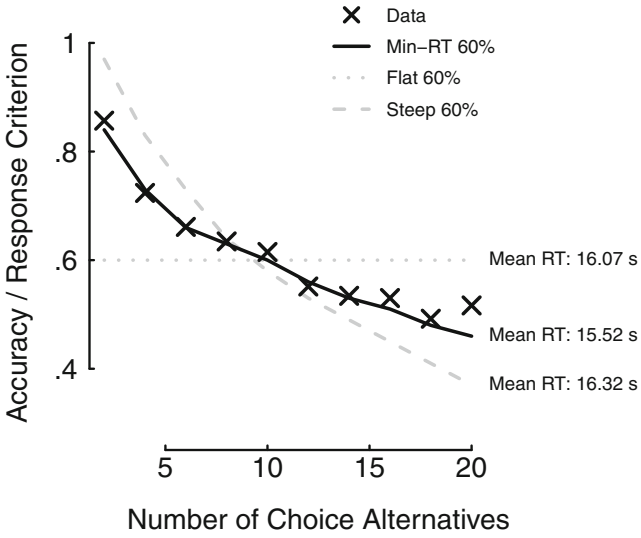

Fig. 1 Accuracy data (crosses) from Experiment 1 of Hawkins et al. (in press), as a function of the number of choice alternatives. The *three lines* demonstrate different approaches to setting response criteria conditional on a goal accuracy level of 60%: The *black line* shows the response criteria selected by the Min-RT procedure, the *dotted gray line* shows a constant response criterion of 60% across all set sizes, and the *dashed gray line* shows steeply declining response criteria that produce large differences in accuracy performance across set sizes. The right side of the figure shows the predicted mean response time (RT) for the three approaches

We conducted an experiment to test the predictions of the Min-RT approach. The context hypothesis predicts that the same stimulus presented in different contexts will elicit different responses. We tested this hypothesis by manipulating a specific set size—trials with $K = 10$ alternatives. We expected decisions for $K = 10$ alternatives to differ as a function of whether this set size was the smallest or largest number of choice alternatives experienced by a decision maker across trials. We therefore manipulated set size in two separate participant groups: a "low-K" group, who saw $K \in \{2, 4, 6, 8, 10\}$ alternatives across trials, and a "high-K" group, who made choices between $K \in \{10, 12, 14, 16, 18\}$. This means $K = 10$ choices were the slowest trials for the low-K context, but the fastest trials for the high-K context. At the ordinal level, our hypothesis suggests that the low-K group will respond faster and with lower accuracy to $K = 10$ trials than the high-K group, reflecting a participant-controlled speed–accuracy trade-off. At the quantitative level, Min-RT suggests that both groups will demonstrate the same mean accuracy, and as a consequence, mean RT will be faster in the low-K than in the high-K group. Min-RT also

[1] Since the Bayesian model approximates the continuous passage of time using discrete steps, predicted response accuracy will always be equal to or slightly greater than the response criterion, because the predicted RT is the first time step on which c is *exceeded*. This overshoot has implications for model fits to data described below.

makes different predictions about response accuracy across set sizes for the two participant groups. The low-*K* group will experience a greater range in RTs across set sizes than the high-*K* group (according to Hick's Law, Equation 1). Min-RT therefore predicts that the low-*K* group will demonstrate a greater decline in response accuracy across set sizes than the high-*K* group, since this will result in considerably shorter total experiment time.

Experiment

We based our experiment on the task employed by Hawkins et al. (in press). This paradigm involved a visual display of many squares representing the choice alternatives, randomly allocated into positions within a four-row × five-column grid. All choice alternatives began each trial as white squares with black borders. Over the course of a trial, each square randomly accumulated blue dots. On each 66-ms time step, there was a 40% chance of each square accumulating an extra dot, independently for each square. Just one square accumulated dots more quickly—the target square, which had a 50% chance. The participants' goal was to select this target square as quickly and accurately as possible. All the distractor elements had the same salience (i.e., accumulation rate), so we need not consider difficult questions about the statistical optimality of various settings of the perceptual template. In situations in which target-to-distractor similarity varies, such matters become important; see McMillen and Behseta (2010) for detailed discussion.

The slow accumulation of evidence ensures that a clear speed–accuracy trade-off emerges in this task: Early in the decision process, when only a few dots have accumulated, a distractor square is likely to be filled with more dots than the target, by random chance. A demonstration version of this experiment can be viewed online, at http://psych.newcastle.edu.au/~sdb231/buckets/vanillaR.html.

Method

Sixty-seven first-year psychology students from the University of Newcastle participated online for course credit. Each participant was randomly assigned to the low-*K* or high-*K* context. To equate total experiment time, participants in the low-*K* condition completed six blocks of 30 trials, and high-*K* participants completed seven blocks of 20 trials. The number of choice alternatives displayed on any trial was randomly selected from $K \in \{2, 4, 6, 8, 10\}$ for low-*K* participants and $K \in \{10, 12, 14, 16, 18\}$ for high-*K* participants, subject to the condition that each set size appeared equally often in each block.

Apart from the difference in set sizes and trial numbers, the experiment was identical for all participants. Each trial began with *K* squares with white backgrounds that were randomly allocated into positions within a 4 × 5 grid, each measuring 100 × 100 pixels. During each trial, time proceeded in discrete steps of 15 events per second. At each time step, a blue dot (2 × 2 pixels) had some chance of appearing at a random, unfilled location in each square. The probability of a dot appearing in each square was independent and equal for all squares at .4, except for one randomly selected target square, which had the probability of .5. The participants' task was to identify this target as quickly and accurately as possible. Participants were free to allow dots to accumulate until they felt confident with their decision. An example of different time points within a single trial with 10 alternatives is shown in Fig. 2. The maximum number of dots in each square was 2,500, meaning that no square could fill in less than approximately 3 min (which is much longer than any participant waited on any trial to make a response). After making a response, participants were provided with feedback in the form of many more time steps illustrated very rapidly, indicating which choice was the true target (which always ended up filling with the most dots). If the participant correctly identified the target, it turned the chosen square's border green. Incorrect identification of the target turned the chosen square's border red, whereas the true target square's border was turned green.

Results and discussion

We excluded data from 12 participants who made fewer than 33% correct responses, and two participants whose host computer displayed fewer than 13 time steps per second, on average. Of the remaining data, we removed 138 trials faster than 1 s, 25 trials slower than 100 s, and 122 trials in which the host computer displayed fewer than 13 time steps per second (3.45% of total trials).

Averaged RT and accuracy data are shown in Fig. 3 as functions of the number of choice alternatives. The left panel shows that both the low-*K* and high-*K* contexts demonstrated an approximately log-linear increase in mean response latency with increasing numbers of choice alternatives, in accordance with Hick's Law. The high-*K* group demonstrated slower RTs overall than did the low-*K* group, as predicted by Min-RT, and also slower responses to $K = 10$ trials, as predicted by the context hypothesis. The right panel of Fig. 3 shows that accuracy steadily declined as set size increased in both groups, although the decline was greater in the low-*K* group, as predicted by the Min-RT optimization hypothesis.

Using a two-way mixed ANOVA, we examined the effects of decision context (low- vs. high-*K*) and the number

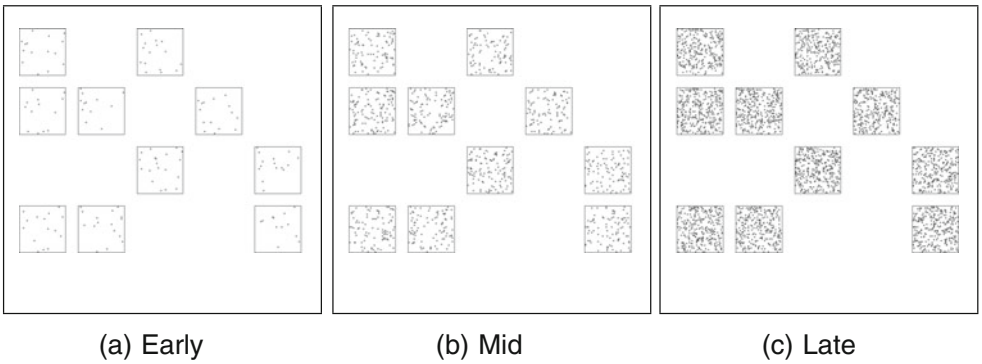

Fig. 2 Screenshots from the experiment at different time points during a choice between 10 alternatives. **a** depicts the early stages of the trial with relatively few dots in each square, **b** shows a mid point, and **c** shows the later stages of the trial when many dots appear in each square. The participants' goal is to identify the square that accumulates dots at the greatest rate, which in the current trial is in the third row of column three

of choice alternatives (where $K = 2$ was paired with $K = 10$, $K = 4$ paired with $K = 12$, and so on). There were significant interactions for both response latency and choice accuracy, reflecting greater changes in the low-K context than the high-K context. This is consistent with Hick's Law because the range of set sizes is greater for the low-K condition, on a log scale.

As expected, RTs were significantly slower in the high-K than in the low-K group ($F(1, 51) = 17.79$, $p < .001$). In contrast, the mean proportion of correct responses did not reliably differ as a function of context (low K: $M = .58$, SE = .03, IQR = .47–.64; high K: $M = .63$, $SE = .04$, IQR = .41–.76; for the main effect, $p = .26$). In both groups, RT increased and accuracy decreased with an increasing number of choice alternatives (for both main effects, $p < .001$). These results are consistent with the interpretation that the decision context promotes a speed–accuracy trade-off, resulting in declines in accuracy with more choice alternatives, while mean accuracy across groups remains relatively constant.

A critical prediction of the context effect hypothesis relates to $K = 10$ data common to both groups, which states that $K = 10$ choices will be treated differently across groups. Consistent with this expectation, RTs for $K = 10$ were slower and accuracy was higher in the high-K than in the low-K group, $t(51) = 2.42$, $p < .05$, and $t(51) = 3.60$, $p < .001$, respectively.

To investigate whether our results were due to taking averages across participants, we conducted individual-participant analyses on accuracy data. The Min-RT hypothesis predicts that low-K participants should make larger adjustments to their speed–accuracy trade-off settings across the range of choice set sizes than high-K participants. This is because the low-K participants experienced a larger range in RTs across set sizes, because of the logarithmic increase of Hick's Law. Consequently, the optimum settings for response criteria in the low-K group differ more across set sizes than in the high-K group. We calculated linear regressions of response accuracy against $\log_2(K)$ separately for each participant in each group (see Fig. 4). Nearly all low-K participants (27 of 28) had negative slope coefficients, suggesting that the speed–accuracy trade-off observed at group level was sufficiently robust to be observed at the individual-subject level. In the

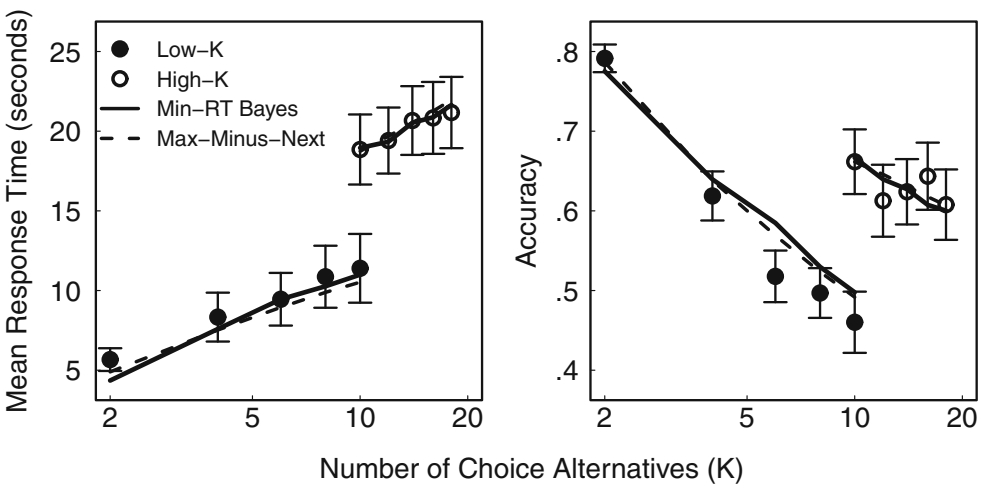

Fig. 3 Mean response time (left panel) and accuracy (right panel) as functions of the number of choice alternatives (K, on a log scale). *Filled and unfilled circles* represent the low-K and high-K contexts, respectively. The *error bars* represent ±1 between-subjects standard errors of the mean. Overlaid on data are mean RT and accuracy predictions of the Min-RT Bayesian model (*solid lines*) and max-minus-next heuristic (*dashed lines*)

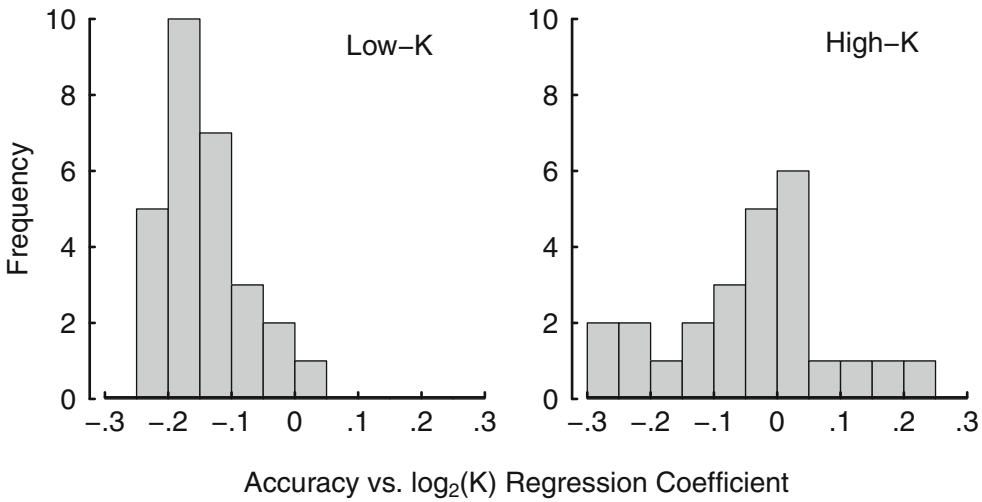

Fig. 4 Distributions of the slopes of lines of best fit for accuracy versus $\log_2(K)$ separately for the low-K and high-K conditions

high-K condition, there was a limited capacity to minimize experiment time by trading speed and accuracy, because mean RT did not differ very much across set sizes. In keeping with the Min-RT hypothesis, only a small majority of high-K participants (15 of 25) had negative slope coefficients. The regression slopes in the low-K group were steeper on average (M = -.14, SE = .01) than in the high-K group (M = -.04, SE = .03), and this difference was significant, one-tailed independent samples t test, $t(51) = 3.84$, $p < .001$. This difference is consistent with the Min-RT hypothesis and is based on accuracy versus set size functions estimated at an individual-subject level.

Model predictions

To test the quantitative fit of the Min-RT approach to data, we specified the goal accuracy parameter using the mean accuracy from data for the low-K group, and just below the mean accuracy for the high-K group (58% and 60%, respectively). We used a value slightly below observed accuracy for the high-K group to compensate for the accuracy overshoot problem described previously. Apart from these two values, there were no model parameters estimated from our data: Separate response criterion parameters (c) were determined for each set size by the model according to the Min-RT goal of minimizing overall experiment time. Mean RT and accuracy predictions of the model are shown as solid lines on data in Fig. 3. Following Hawkins et al. (in press), we scaled all model RTs by a factor of four (this is a fixed parameter in the model). Without this scaling, the optimal Bayesian model always responds much faster than humans. The scaling factor can be interpreted as capturing a perceptual limit: The very small dots in our display may have been perceptually grouped in fours. With these assumptions, the Bayesian Min-RT model provides a good account of all the data.

The operation of the Min-RT assumption can be made clearer by comparing its predictions to the standard Bayesian model predictions, with a fixed response criterion across all set sizes. In the low-K context, the Bayesian model using Min-RT criterion settings will complete the experiment in 10.3% less time than if a constant 58% accuracy criterion was established for each set size. In the high-K context, the benefits of adjusting response criteria are much smaller. Hence, this group had less to gain by trading accuracy across set sizes, and Min-RT predicts only approximately a 1% decrease in experiment time. This explains why the range in accuracy across set sizes in the high-K group was far smaller than that of the low-K group.

To compare the quantitative fit of the Min-RT approach as applied to the Bayesian model, we also compared its predictions with another single-parameter model of multi-alternative choice: the max-minus-next heuristic. By default, max-minus-next predicts accuracy rates that decline as set size increases, so it naturally suits the qualitative data patterns from our experiment. The purpose of comparing the max-minus-next heuristic with the Bayesian model is to demonstrate that the Min-RT idea is sufficiently powerful to rescue the Bayesian decision mechanism, which was previously shown (by Brown et al., 2009) to be inferior to the max-minus-next decision mechanism in data with context effects. The max-minus-next account proposes that a response is triggered as soon as the evidence for the most likely alternative exceeds the evidence for the second most likely alternative by some threshold amount, Δ. Dragalin, Tartakovsky, and Veeravalli (1999, 2000) demonstrated that this simple decision heuristic approximates the optimal multihypothesis sequential probability ratio test (Baum & Veeravalli, 1994) when error rates are low. Brown et al. found the max-minus-next heuristic to provide a good account of their data, demonstrating the Hick's Law regularity in RTs and also decreasing accuracy with increasing

numbers of choice alternatives, even though error rates were quite high.

We fit the max-minus-next model to our data by assuming that the decision evidence was just the number of filled dots in each stimulus. We applied the same 4× slow down imposed for the Bayesian model (on the assumption that a perceptual limitation ought to apply equally to the two models). The dashed lines in Fig. 3 show that the model fits the data quite well, assuming a response threshold of $\Delta = 3.2$ dots for the low-K context, and $\Delta = 4.5$ dots for the high-K context (obtained via optimization over a grid search in increments of $\Delta = .1$). Noninteger values for Δ might be interpreted as if participants employ a mixture of integer values.

Since both models provided excellent fits to mean RT data, we examined their predictions—under the same parameter settings—for the full distribution of RTs. For each participant, we calculated the 10%, 30%, 50% (i.e., median), 70%, and 90% percentiles of the RT distribution for each set size and then averaged these quantile estimates over participants in each group. Quantile averaging of this kind preserves distribution shape, under reasonable assumptions about between-subjects variability (Gilchrist, 2000). These distributions are shown separately for both groups as gray lines in Fig. 5. The lines depict, from bottom to top, the 10%, 30%, 50% (shown with larger symbols), 70%, and 90% percentiles. The upper lines are spread further apart than the lower lines, showing the positive skew of the RT distribution. Model predictions are shown in black lines. In the left panel, the max-minus-next heuristic predicts the central tendency (median) in data well; however, this model predicts too much variability in the distribution of RTs. In contrast, the Min-RT Bayesian model, shown in the right panel, quite closely predicts the percentiles of the RT distribution for both groups, except for some potential misfit in the smallest set sizes of the 70% and 90% percentiles of the low-K group.

Model Selection The aforementioned analyses suggest that both models account for the data reasonably well, but the Min-RT Bayesian model is better able to capture the full distribution of RTs. We then applied a model selection technique (parameter space partitioning: Pitt, Kim, Navarro, & Myung, 2006) to investigate whether the Bayesian model's improved fit was due to model flexibility. Parameter space partitioning involves simulating model predictions across the entire parameter space of a model and evaluating the qualitative data patterns the model is capable of producing. A model is preferred if it has little flexibility, if it predicts the same empirical trends across all parameter settings, and does not predict trends that do not occur in data. We examined whether the models predicted the basic finding—Hick's Law for RTs—across all parameter settings.

We were limited to simulating the predictions of the core Bayesian model (i.e., without the Min-RT approach to selecting response criteria). Since the predictions of the Bayesian model are probabilistic, and Min-RT depends entirely on these probabilistic predictions, the Min-RT predictions included too much simulation noise, given practical amounts of computer time. Simulation noise makes the model unsuitable for parameter space partitioning because the model's predictions then differ across the parameter space because of noise, rather than model flexibility. We also did not perform parameter space partitioning on accuracy predictions because there is no well-established equivalent to Hick's Law for accuracy data and thus no clear target for what the models should and should not predict. Furthermore, the core Bayesian model always predicts constant accuracy with increasing set size, and the max-minus-

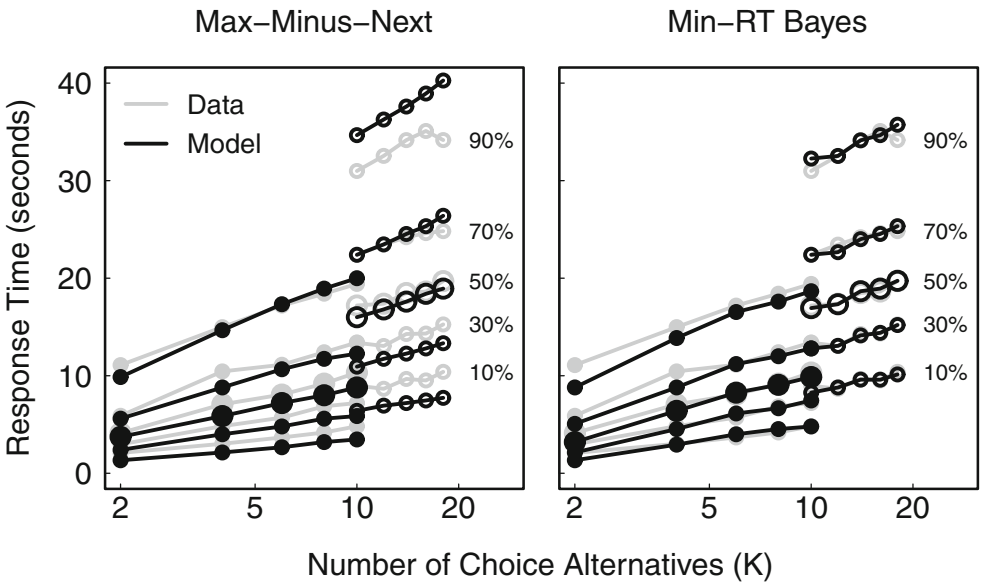

Fig. 5 Predicted response time distributions of the max-minus-next heuristic (left panel) and the Min-RT Bayesian account (right panel) overlaid on data. *Gray lines* show empirical data, and *black lines* show model fits. *Filled and unfilled circles* represent the low-K and high-K contexts, respectively. Within each group the lines represent, from bottom to top, the 10%, 30%, 50% (i.e., the median, shown with larger circles), 70%, and 90% percentiles of the distribution, shown by the text on the right side of both panels

next heuristic always predicts declining accuracy with increasing set size. Both patterns appear in data, depending on the decision contexts, so parameter space partitioning on accuracy predictions will be uninformative.

We simulated data from both models for $K \in \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$ to mimic the range of set sizes from our experiment. The Bayesian model was simulated across a response criterion grid from c = .26-.99 in increments of .01 (for c < .5 we removed the K = 2 data point, since this represents responses below chance accuracy), and the max-minus-next model from Δ = 2-20 in increments of 1. The range in parameter values for each model represents the maximum sensible range for the models. For each parameter setting, we evaluated whether Hick's Law was predicted by calculating the slope of the mean RT versus $\log_2(K)$ relationship for each successive pair of set sizes (i.e., K = 2 to K = 4, then K = 4 to K = 6, and so on). Hick's Law asserts these slopes should all be equal, so we classified the model as predicting Hick's Law whenever the range of the slopes was smaller than some tolerance value (to allow for small deviations from log-linearity). The tolerance values below are expressed as percentages of the mean slope estimates.

Parameter space partitioning on RTs indicated that the core Bayesian model is less flexible than the max-minus-next heuristic. Both models always predicted the generic ordering pattern expected in multialternative choice: RTs for $K = 2 < K = 4 < K = 6 \ldots$, which is a necessary, although not sufficient, requirement for Hick's Law. The max-minus-next heuristic never satisfied Hick's Law for tolerance values that were more strict than 36%. At that same tolerance value, the core Bayesian model predicted Hick's Law across most of its parameter space (69% of the range of c). These results are shown more completely, as receiver operating curves, in Fig. 6. The core Bayesian model predicted Hick's Law across almost all of its parameter space by a tolerance value of 66%. For the same tolerance, the max-minus-next model predicted Hick's Law across less than half of its parameter space and did not cover 95% until tolerance was at 82%.

The max-minus-next model performed rather poorly because it predicted smaller increases in RT at larger set sizes than Hick's Law suggests should occur. That is, max-minus-next predicted a flattening in RT across the larger set sizes. To illustrate, when examining only smaller set sizes, such as $K \leq 8$, the max-minus-next model predicts Hick's Law across 90% of the parameter space for 59% tolerance, with the same area covered with a tighter 29% tolerance for the core Bayesian model. This suggests the max-minus-next model performs acceptably for the set sizes typically studied in multialternative choice. However, Hick's Law has been demonstrated to hold for decisions between more than eight alternatives (e.g., Brown et al., 2009; Hawkins et al., in press).

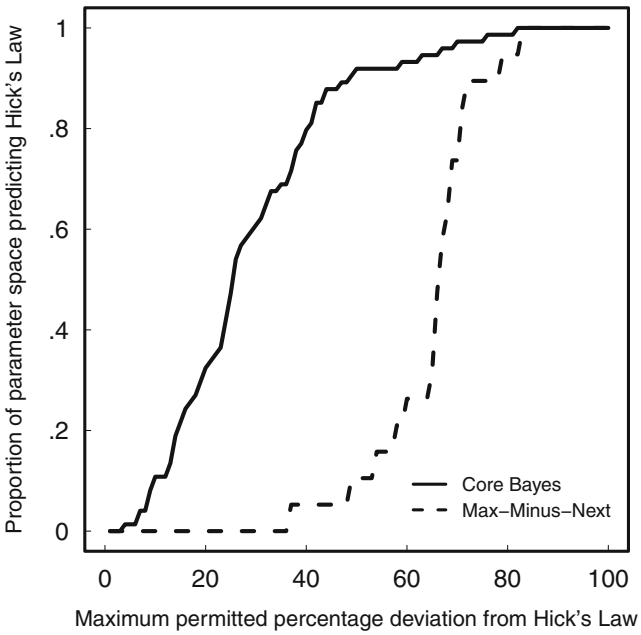

Fig. 6 Parameter space partitioning results for the proportion of the parameter space predicting Hick's Law for varying degrees of percentage deviation from Hick's Law (see the text for detail). *Solid and dashed lines* represent the core Bayesian and max-minus-next heuristic predictions, respectively

General discussion

In the present study, we developed a novel approach to modeling context effects in multialternative decisions and compared this approach to existing accounts. The RT minimizing approach (Min-RT) provides the first principled explanation for context effects. Min-RT constrains speed–accuracy trade-off settings across different set sizes in such a way as to minimize total experiment time, and can be applied to any model with a speed–accuracy criterion parameter. We demonstrated that Min-RT can account for a previous data set as well as novel empirical data. The experimental data reported presently also provide further support for Hick's Law, demonstrating that the log-linearity in RTs holds for judgments in large-choice set sizes, and is still observed despite high error rates (consistent with recent literature, e.g., Brown et al., 2009; Kveraga et al., 2002; Lacouture & Marley, 1995; Leite & Ratcliff, 2010). We also provided a novel empirical demonstration of context effects in multialternative choice.

The Min-RT Bayesian model and the max-minus-next heuristic both fit mean RT and accuracy data; however, the Min-RT Bayesian account provided a better fit to the RT distributions. Parameter space partitioning suggested the core Bayesian model is less flexible than the max-minus-next model: The Bayesian model predicted the constrained Hick's Law regularity considerably more often than did the max-minus-next model. The parameter space partitioning of the core Bayesian model compares favorably to the same model

flexibility analysis conducted by Schneider and Anderson (2011) on their ACT-R memory retrieval model. Schneider and Anderson examined $K \in \{2, 4, 8\}$ and found that their model predicted the generic $K = 2 < K = 4 < K = 8$ ordering of RTs across 88% of the parameter space, when using a tolerance for equality between set sizes of 20 ms, corresponding to about 5% of their shortest empirical response latencies. Using a similar 5% criterion in our parameter space partitioning over the same set sizes as in Schneider and Anderson, the core Bayesian model predicted the generic RT ordering in 100% of the parameter space, suggesting that the core Bayesian model is less flexible than the ACT-R model. This comparison, however, is not straightforward. Empirically, sometimes response latency does not differ across set sizes—for example, in tasks involving high stimulus–response compatibility or following extensive task practice (for a review, see Schweickert, 1993; Teichner & Krebs, 1974). This might suggest the ACT-R model is better suited to explaining a broader range of multialternative choice phenomena.

An interesting perspective on our notion that decision makers try to minimize total experiment time, conditional on maintaining a goal accuracy level, is to consider it as a generalization of the concept of "reward rate" optimization (e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). Reward rate is just the expected number of correct responses per unit time, so in any experimental design—and for given constraints on accuracy—there is a single speed–accuracy trade-off setting that maximizes reward rate. Bogacz et al. showed that most decision makers employed a too-careful speed–accuracy trade-off: They could have increased reward rate by making more frequent, but less accurate, decisions. With practice, however, most decision makers can eventually maximize reward rate (Balci et al., 2011; Simen et al., 2009; Starns & Ratcliff, 2010). The concept of reward rate is only directly applicable to experimental designs when the time available to participants is fixed, but the number of decisions is not: In the usual case, in which the number of decisions is fixed, being very careful will maximize the total number of correct decisions. One perspective on the Min-RT hypothesis is that it generalizes the notion of reward rate to the typical, fixed-trial, experimental designs. Without explicit instruction to do so, and without the structure imposed by a fixed-time experimental design, a plausible goal for participants is to leave the experiment as quickly as possible, without making socially unacceptable error rates.

The psychological plausibility of the Min-RT approach is unclear: Could decision makers feasibly arrive at those response criterion settings that minimize expected RT? It is possible that observers might approximate this minimum through a gradient descent search of some sort. For instance, a decision maker might begin (by default) with a constant response criterion across set sizes. In the Bayesian model, successive adjustments to criteria will maintain the goal accuracy as long as the sum of criterion increments across conditions is zero (e.g., if the criterion for one condition is raised by 1%, the criterion for another condition should be decreased by 1%). With this constraint, the task of the observer is simplified to just selecting which conditions should have their criteria raised, and which lowered. Some simple heuristics might get the decision maker close to the global optimum; for instance, perhaps one might increase the response criterion for the fastest condition and decrease it in the slowest condition. An alternative approach to the problem of psychological plausibility might be to relax the assumption that observers set a response criterion independently for every experimental condition. A lighter cognitive load would result instead from estimating one or two parameters of some simple function that approximates the optimal criterion settings (for a similar approach, see, e.g., Balci et al., 2011). For example, combining our present results with those from Hawkins et al. (in press) suggests that participants could get very close to the optimal solution by adjusting just a single parameter of a power function. Such process models of the search for optimal criteria are a topic for future research on this problem.

We have provided further empirical evidence for Hick's Law and context effects in multialternative choice, and developed the Min-RT approach that can be applied to most existing domain general models of these choices. We compared the goodness of fit of Min-RT as applied to a Bayesian ideal observer to the max-minus-next heuristic in explaining context effects and declining accuracy rates. The Bayesian model outperformed the heuristic account on two fronts: better fit to data and lower model flexibility. Min-RT provides a one-parameter approach to determining the speed–accuracy criterion settings across set sizes that minimize total experiment time. When applied to an existing domain general model of multialternative choice, Min-RT provided a strong account of context effects by describing mean RT, mean accuracy, and RT distributions.

Author note The present research was supported by the Keats Endowment Research Fund (Hawkins & Brown), ARC Discovery Project DP0878858 (Brown & Steyvers), and a Vidi grant from the Dutch Organization for Scientific Research (NWO, Wagenmakers). Correspondence concerning this article may be addressed to: G. Hawkins, School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia (e-mail: guy.e.hawkins@gmail.com).

References

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics, 73,* 640–657.

Baum, C. W., & Veeravalli, V. V. (1994). A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory, 40,* 1994–2007.

Beh, H. C., Roberts, R. D., & Pritchard-Levy, A. (1994). The relationship between intelligence and choice reaction time within the framework of an extended model of Hick's Law: A preliminary report. *Personality and Individual Differences, 16,* 891–897.

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.

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.

Dragalin, V. P., Tartakovsky, A. G., & Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests—part I: Asymptotic optimality. *IEEE Transactions on Information Theory, 45,* 2448–2461.

Dragalin, V. P., Tartakovsky, A. G., & Veeravalli, V. V. (2000). Multihypothesis sequential probability ratio tests—part II: Accurate asymptotic expansions for the expected sample size. *IEEE Transactions on Information Theory, 46,* 1366–1383.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. London, England: Chapman & Hall/CRC.

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.

Hawkins, G., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (in press). Context effects in multi–alternative decision making: Empirical data and a Bayesian model. *Cognitive Science*. doi:10.1111/j.1551-6709.2011.01221.x

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

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.

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

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

Lee, S., Heo, G., & Chang, S. H. (2006). Prediction of the human response time with the similarity and quantity of information. *Reliability Engineering and System Safety, 91,* 728–734.

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

McMillen, T., & Behseta, S. (2010). On the effects of signal acuity in a multi–alternative model of decision making. *Neural Computation, 22,* 539–580.

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.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113,* 57–83.

Rabbitt, P. M. A. (1968). Repetition effects and signal classification strategies in serial choice–response tasks. *Quarterly Journal of Experimental Psychology, 20,* 232–240.

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

Schweickert, R. (1993). Information, time, and the structure of mental events: A twenty–five–year review. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV* (pp. 535–566). Cambridge, MA: MIT Press.

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

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two–alternative decision making: Empirical tests of theoretical predictions. *Journal of Experimental Psychology. Human Perception and Performance, 35,* 1865–1897.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25,* 377–390.

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

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

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.

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