Decision Speed Induces Context Effects in Choice

Guy Hawkins[a], Scott D. Brown[a], Mark Steyvers[b], and 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

Abstract

The context in which a decision occurs can influence the decision-making process in many ways. In the lab, this is often evident in the effects of recent decisions. For instance, many experiments combine easy and difficult decisions, such as when word frequency is manipulated in lexical decision. The "blocking effect" describes how such decisions differ depending on whether the conditions are presented in pure blocks (comprised purely of easy or hard stimuli) or mixed blocks (also known as a "mixing cost"). We present a novel extension to these context effects, demonstrating in two experiments that they can be induced using conditions with identical difficulty, but different timing properties. This suggests that explanations of context effects based on task difficulty or error-monitoring alone might be insufficient, and suggest a role for decision time. In prior work, we suggested such a hypothesis under the assumption that observers minimize their decision time, subject to an accuracy constraint. Consistent with this explanation, we find that decisions in slower conditions were based on less evidence when they were experienced in mixed compared to pure blocks.

Keywords: Decision making, Evidence accumulation, Choice, Context, Mixing cost, Blocking effect, Reward rate

Introduction

Decisions about ostensibly identical stimuli are influenced by the difficulty of preceding decisions. Difficulty is often a factor of primary interest in decision tasks, for instance, with high frequency (easy) or low frequency (hard) words in lexical decision. If an experiment includes both easy and difficult conditions, these might either be presented separately

Correspondence concerning this article may be addressed to: Guy Hawkins, School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia; Email: guy.hawkins@newcastle.edu.au.

(in pure blocks or between-subjects) or randomly inter-mixed. Choices in a pure block of easy stimuli are faster but less accurate than when those same easy stimuli are presented in a mixed block, and choices in a pure block of hard stimuli are slower but more accurate than when those same hard stimuli are presented in a mixed block. This finding, referred to as a mixing cost, a blocking effect, or a context effect, has been observed in many speeded decision tasks, including sentence verification (Kiger & Glass, 1981), picture naming and arithmetic (Lupker, Kinoshita, Coltheart, & Taylor, 2003), and lexical decisions (Lupker, Brown, & Colombo, 1997; for review see Los, 1996).

In a similar vein, Hawkins, Brown, Steyvers, and Wagenmakers (in press) manipulated decision difficulty by changing the number of alternatives in a perceptual decision task – decisions between more alternatives are more difficult than those between fewer alternatives. In different experiments, participants made judgments either about many different choice set sizes randomized across trials (in mixed blocks) or about a single choice set size (in pure blocks). Decisions in a pure block of hard stimuli (larger set sizes) were slower but more accurate than when those same hard stimuli were presented in mixed blocks, similarly to the well-established blocking effects described above, even though these judgments were on a much longer timescale.

Hawkins et al. (in press) interpreted these results by supposing that participants engaged in a speed-accuracy tradeoff (Ratcliff, 1978; Wickelgren, 1977). That is, in the mixed condition participants elected to spend more time on easy decisions and less time on difficult decisions, perhaps in order to minimize the total amount of time spent in the experiment. The same speed-accuracy tradeoff patterns can emerge in mixed blocks even when participants are unable to elect their level of caution for easy and hard trials, because the task structure provides no warning about the class of upcoming stimuli. For instance, decision makers may commit to a goal accuracy rate which they will not go below, and want to perform as fast as possible while remaining above this accuracy criterion (Hawkins, Brown, Steyvers, & Wagenmakers, 2011). In mixed blocks, one could establish an intermediate response threshold that does not change across easy and hard trials. This approach can result in the same data patterns as "choosing" response thresholds separately for each experimental condition: more accurate choices for easy stimuli, and faster but less accurate choices for hard stimuli, compared to pure blocks. The idea of minimizing response times conditional on a goal accuracy rate is related to the idea that participants might adjust their speed-accuracy tradeoff to maximize "reward rate", which is just the rate of correct responses (for an overview see, e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006).

In contrast, previous theoretical accounts of blocking effects have been based on decision difficulty, rather than decision time. For instance, the self-regulating accumulator model (PAGAN, Vickers, 1979; Vickers & Lee, 1998, 2000), assumes that the decision maker adjusts performance by monitoring confidence, which is influenced by decision difficulty. Similarly, Jones, Mozer, and Kinoshita (2009) proposed that the decision maker estimates the average difficulty of decisions in recent trials, which differs between pure and mixed blocks. Jones et al. also questioned the feasibility of theoretical accounts based on speed-accuracy tradeoffs, a point we return to in the General Discussion.

These earlier theoretical approaches assume that differences in decision *difficulty* across conditions cause blocking effects, which contrasts with Hawkins et al.'s (in press)

assumption that blocking effects are caused by differences in decision *time* across conditions. In all of the studies reviewed above, the easier conditions produced faster choices than the harder conditions. For instance, the benchmark in multi-alternative choice is Hick's Law (Hick, 1952; Hyman, 1953), that response time increases linearly with the logarithm of set size (for general overview of Hick's Law research see Teichner & Krebs, 1974; Welford, 1980). Presumably, previous explanations of blocking effects have focused on decision difficulty rather than decision speed, simply because difficulty was the experimental parameter manipulated in those experiments. However, the confound between decision difficulty and decision time makes it difficult to distinguish the different theoretical accounts, and in particular, whether decision time may also play a role in eliciting blocking effects.

We report two experiments that examine the effects of decision time independent of decision difficulty. The different conditions in these experiments were created by manipulating how rapidly stimulus information is presented. We used the externalized evidence accumulation task of Brown, Steyvers, and Wagenmakers (2009), which involves comparing the height of a number of columns, and naturally permits such manipulations. We manipulated decision time randomly across trials (creating mixed blocks) in Experiment 1 and between-subjects (creating pure blocks) in Experiment 2. In the mixed blocks of Experiment 1 we observed that decisions from faster and slower conditions produced context effects similar to previously observed blocking effects from easier and harder decisions. In Experiment 2, we demonstrate that these context effects can be almost entirely eliminated by using a design with pure blocks. The changes observed across experiments are consistent with the time-minimization account based on speed-accuracy tradeoffs.

Experiment 1

We use the paradigm developed by Brown et al. (2009). Each decision involved a display of K columns that grew taller at different rates, by randomly accumulating increments of height (henceforth "bricks") at discrete time steps according to a simple statistical model. With time, one of the columns would grow taller than the others, on average, and the participant's goal was to identify this target column as quickly as possible. A difficult aspect of the task was to balance the tradeoff 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 (see Figure 1 for an example screenshot of the task).

In this paradigm, task speed can be adjusted by changing the delay between successive time steps, which we refer to as the "drop delay". In Experiment 1 we manipulated drop delay using mixed blocks, where the drop delay varied randomly from trial-to-trial. We limited decisions to just one set size ($K = 10$), which kept decision difficulty constant across trials. We make the assumption here that drop delay does not influence task difficulty. We highlight evidence indicating that this psychophysical assumption is supported in the discussion following Experiment 2. If context effects are induced by decision time (as proposed by Hawkins et al., in press), and not just decision difficulty, then accuracy and decision time (measured in discrete time steps) should both change across levels of the stimulus speed manipulation.

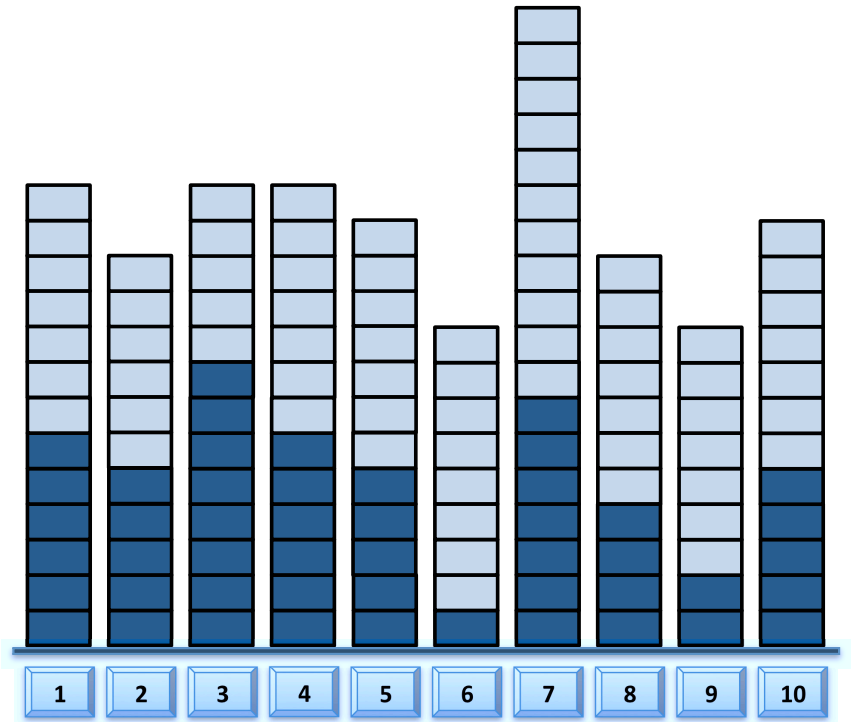

Figure 1. Example screenshot of a trial with $K = 10$ response alternatives. The dark and light blocks represent the number of bricks that have accumulated after 10 and 30 time steps, respectively. The participant's task is to select the target column, which is column 7 in this example.

Method

Fifty first-year psychology students from the University of Newcastle participated online for course credit. Each participant completed 6 blocks of 32 trials. Each trial consisted of $K = 10$ response buttons displayed along the base of the screen. The delay between each discrete time step was manipulated within-subjects, with the drop delay manipulation having eight levels. In the fastest condition a new row of evidence accumulation tokens (bricks) appeared every 266 milliseconds (msec), but this slowed to 276msec, 296msec, 326msec, 367msec, 436msec, 500msec and 625msec in the slower conditions. These settings meant that in the fastest condition participants saw evidence accumulate more than twice as quickly as in the slowest condition. The drop delay of any trial was randomly chosen from all drop delays, subject to the condition that each drop delay appeared equally often in each block.

On each trial, $K = 10$ response buttons of the same width were lined up abutting one another in the center of the base of the display, representing the different choice alternatives. A trial began with an empty display above the response buttons. At each time step, a new brick either fell from above onto the top of the response button, increasing the height of that column, or did not. The probability of a brick appearing on each column at any

time step was .35, independently for all columns, except for a randomly-selected target column which had an accumulation probability .5. The participant's task was to identify the target column as quickly and accurately as possible. Participants were informed that due to the random nature of the task if they respond too early they may incorrectly select a distractor column that had, by chance, accumulated the most bricks thus far in the trial. For example, in Figure 1 after 10 time steps have elapsed (indicated by dark bricks) column 3 was the tallest, even though column 7 was the target. However, after an additional 20 time steps (light bricks) column 7 had accumulated more bricks than the remaining response alternatives. Participants were free to sample information from the task environment until they felt confident with their decision. If a participant waited long enough, the tallest columns grew near to the top of the display. Whenever this occurred the column heights were smoothly re-scaled to remain within the display window.

Analysis Strategy

Because part of our argument requires support for a null effect of drop delay in pure blocks (in Experiment 2), we do not use null hypothesis significance testing. Instead, we fit multivariate normal distributions parameterized using general linear models of exactly the same form that would be applied under standard repeated-measures analysis of variance (ANOVA). Rather than use null hypothesis significance testing, from these models we calculate the Bayesian Information Criterion (BIC; Raftery, 1995; Schwarz, 1978) and use that to approximate posterior model probabilities based on a uniform prior across models, and on the assumption that the data-generating model was one of those under consideration. Similar approaches have been advocated as alternatives to regular null hypothesis significance testing under the ANOVA framework; approaches which circumvent some of the pervasive problems associated with null hypothesis testing (Glover & Dixon, 2004; Wagenmakers, 2007).

To reassure the reader that our results and conclusions are not specific to this analysis strategy, we repeated all primary analyses twice, once using the Akaike Information Criterion (AIC; Akaike, 1974) and once with null hypothesis significance tests accompanied with effect size estimates (see Appendix). Those analyses were consistent with the BIC analyses, although the AIC analyses are naturally less strict in their penalty of model complexity. For further detail and instruction on Bayesian data analysis we refer the reader to a few of the many excellent sources on the topic, including: Kruschke (2011), Rouder, Speckman, Sun, Morey, and Iverson (2009), and Wagenmakers, Lodewyckx, Kiruyal, and Grasman (2010).

In Experiment 1 we compared the fit of two nested, linear multivariate normal models. These models correspond to those considered in a standard one-way repeated-measures ANOVA: the null model with only a grand mean; and a model that also includes an effect of drop delay.[1] We estimated the model parameters (treatment effects and variance-covariance terms) using standard general linear model algorithms. In hierarchical models, such as the repeated-measures designs here, some subtlety is required in calculating the appropriate sample size to use in BIC calculations; in all cases, we used the methods of Pinheiro, Bates, DebRoy, Sarkar, and R Development Core Team (2011). We denote BIC approximations to posterior model probabilities with p^{BIC}.

[1] In all models in Experiments 1 and 2 we added a random effect for subjects, as in repeated-measures ANOVA.

Results

We excluded data from 5 participants who made fewer than 33% correct responses. The remaining data were screened for outlying trials resulting in the removal of 41 trials with responses faster than 4 time steps and 3 trials slower than 200 time steps (0.5% of total responses). Throughout the paper we refer to the number of discrete time steps that elapsed prior to response as the "step number". The step number is a quantifiable measure of the quantity of evidence used to inform a decision in our externalized accumulation task, and also does not confuse the outcome measure with elapsed time.

Mean step number and accuracy data are displayed in Figure 2 as functions of drop delay, using within-subjects standard error bars calculated according to Loftus and Masson (1994). The left panel suggests that mean step number decreased as the drop delay became longer. That is, in trials where evidence accumulated more slowly participants waited for fewer time steps before making a response. This finding was supported by BIC analysis, with the drop delay model strongly supported over the null model for both accuracy and step number (both $p^{BIC} = 1$).

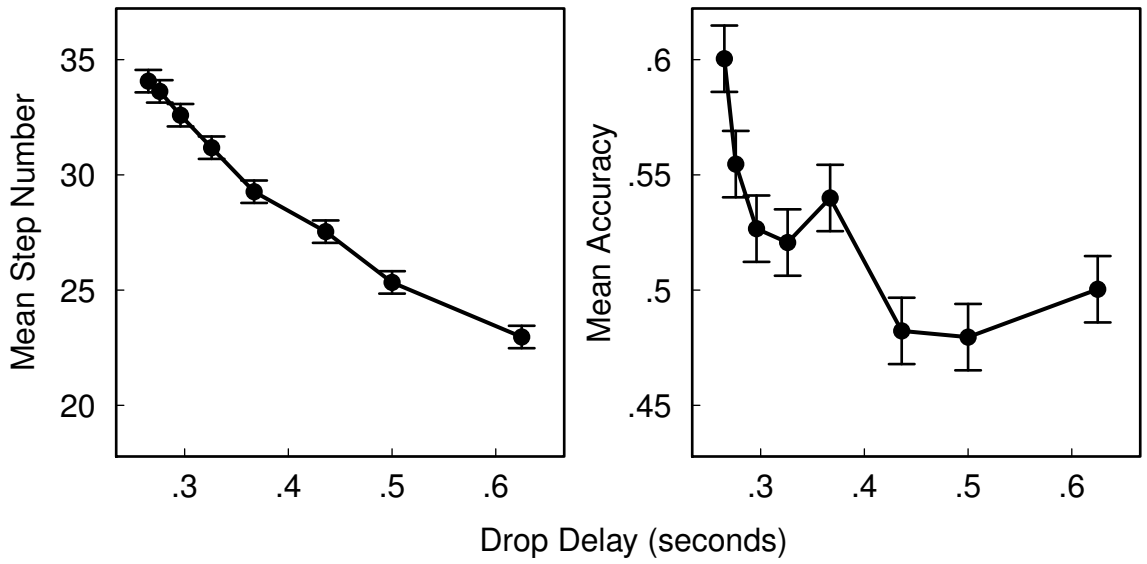

Figure 2. Mean step number (left panel) and accuracy (right panel) from Experiment 1, as functions of the drop delay. The error bars represent ±1 within-subjects standard errors of the mean.

Discussion

The results from Experiment 1 suggest that decisions were based on less evidence (fewer steps of the stimulus display) in conditions where the columns grew taller at slower rates compared to faster height accumulation. This represents a context effect similar to the well-known blocking effect, but with decision time taking the role of decision difficulty: the choice accuracy and information required for decisions (i.e., step number) both decreased for slow conditions relative to fast conditions. These results are consistent with decision

makers adjusting their speed-accuracy tradeoff settings between slower and faster choice conditions. Explanations of blocking effects based on changes in speed-accuracy tradeoff settings must assume that decision makers adjust their response criteria separately for each decision. These adjustments have previously been considered implausible (Jones et al., 2009; Ratcliff, 1978) because participants are not given advanced warning of the upcoming condition, and response thresholds presumably take longer to adjust than the amount of time available during a typical decision (a few hundred milliseconds). Such a criticism does not apply to our experiments, because decisions took many seconds, giving participants ample time for adjustment (e.g., Forstmann et al., 2008 found that adjustment took less than 1.5sec).

In Experiment 2 we manipulate drop delay between-subjects, analogous to the pure blocks described above. If our hypothesis holds, the effect of drop delay on choice accuracy and step number observed in Experiment 1 should be eliminated in Experiment 2.

Experiment 2

Hawkins et al. (in press) proposed that only differences in decision time drive context effects, which makes a strong prediction: if decision time is manipulated in pure blocks, it should have no effect on the amount of evidence that observers accumulate prior to response. This means that both the step number and response accuracy should be unaffected by changes in drop delay. Interestingly, this hypothesis is directly opposed to the predictions of theoretical accounts based on reward rate (i.e., the idea that people try to maximize their number of correct responses per unit time). Recent results have shown that people tend to produce more careful (slow and accurate) decisions than reward rate accounts would predict, but after some practice they become closer to optimal (e.g., Balci et al., 2011; Starns & Ratcliff, 2010). If decision makers maximized reward rate, accuracy should decrease as the drop delay becomes longer, since decisions in these conditions take longer (in real time), so encouraging faster guesses.

Method

A separate group of 172 first-year psychology students from the University of Newcastle participated online in Experiment 2 for course credit. The delay between each time step was manipulated between-subjects to create pure blocks, with each participant randomly assigned to one of seven drop delays: 276 msec, 296msec, 326msec, 367msec, 436msec, 500msec and 625msec. To moderate differences in the total duration of the experiment, participants in the three slowest conditions each completed 4 blocks of 40 trials, and participants in the four fastest conditions each completed 6 blocks of 40 trials (but only data from the first 4 blocks were analyzed below).

In Experiment 2, participants only experienced a single drop delay condition, which provided capacity for us to also manipulate the choice set size on a within-subjects basis. The number of choice alternatives shown on any trial was randomly selected 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.

Apart from the between-subjects manipulation of drop delay and within-subjects manipulation of set size, all other experimental details were the same as Experiment 1,

including the analysis approach. In Experiment 2 there were two experimental factors, hence we compared the fit of five nested, linear multivariate normal models, corresponding to those considered in a standard two-way ANOVA: (1) the null model with only a grand mean; (2) a single set size effect; (3) a single drop delay effect; (4) additive set size and drop delay effects; and (5) the saturated model with additive set size and drop delay effects as well as an interaction between the two.

Results

Data from 8 participants with fewer than 33% correct responses were excluded. The remaining data were screened for outlying trials, with the removal of 410 trials faster than 4 time steps and 41 trials slower than 200 time steps (1.36% of total responses).

Figure 3 illustrates the data, with mean response accuracy for each of the seven drop delay conditions graphed in the top panel. The middle panel shows the mean step number on which responses were made, and the bottom panel shows corresponding mean response times (in seconds). Focusing on the lower two panels, across all drop delay conditions we observed the pattern expected under Hick's Law; response time increased approximately log-linearly with set size.

Inspection of the upper two panels of Figure 3 suggests that drop delay had no systematic effects on choice accuracy or step number. Confirming this, the general linear model that included only an effect of set size was strongly supported in both dependent measures: $p^{BIC} \approx 1$ for response accuracy, and $p^{BIC} = .93$ for step number. Since there was no effect of drop delay on the step number on which participants responded, there was conversely a large and systematic effect of drop delay on the real time taken for each decision (i.e., response time in seconds, lower panel of Figure 3). This was confirmed by good support for the drop delay and set size interaction model, $p^{BIC} = .93$. The interaction model indicates that there was a greater increase in response latency (in seconds) with increasing set size for the slower drop delay groups than faster drop delay groups. Although response latency was bound to differ across drop delay groups given the step number and accuracy data, this result illustrates that there were strong and reliable differences in real time across groups, which apparently had no effect on decision accuracy. For instance, the slowest drop delay condition yielded response times between two and three times longer than the fastest condition.

Discussion

Manipulating the speed of information accumulation in the stimulus display systematically influenced the speed of participants' decisions, but apparently did not alter response accuracy, or the amount of information used to make a decision. This result supports our hypothesis that pure or mixed block presentation of decision time, and not just decision difficulty, can induce context effects in decision-making. The null effect of drop delay on mean step number and accuracy also provides support for our assumption that the speed of the stimulus display did not influence task difficulty.

There are theoretical implications from our finding that equivalent quantities of evidence were accumulated across the seven drop delay conditions of Experiment 2. Participants waited more than twice as long in the slowest condition than the fastest condition in order to make decisions with very similar accuracy. This is inconsistent with the idea

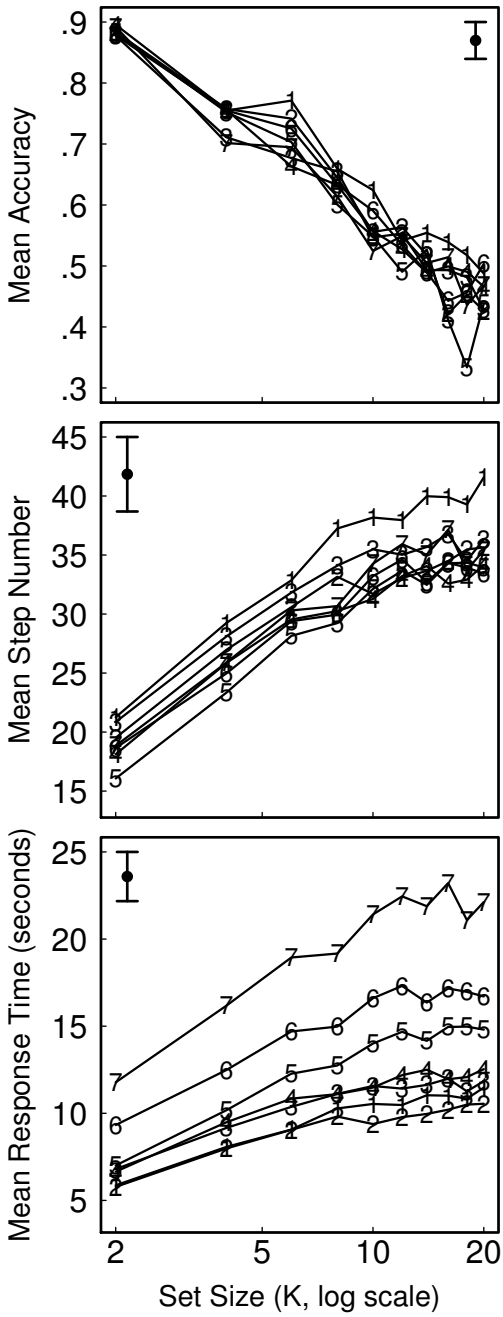

Figure 3. Mean accuracy, step number and response time (seconds; upper, middle and lower panels, respectively) from Experiment 2. The drop delay conditions are numbered from fastest to slowest (1 – 7). A pooled between groups error bar with ±1 standard errors of the mean is shown in each panel.

that decision makers try to optimize reward rate (Bogacz et al., 2006; Starns & Ratcliff, 2010). In slower conditions, reward rate – the rate of correct responses – can be improved by making faster, but less accurate, decisions compared with faster conditions, but this did not happen in Experiment 2. However, it is possible that either our design did not particularly encourage participants to maximize reward rate, as the total number of decision trials was fixed at the outset, or that many sessions of practice are required before participants can optimize reward rate (Balci et al., 2011; Starns & Ratcliff, 2010). Another possibility is that the range of drop delays we used was not large enough – perhaps reward rate maximization might have occurred if even slower drop delays were introduced. Finally, it could be that reward rate is more difficult to maximize in externalized evidence accumulation tasks compared to more traditional speeded choice tasks.

Our results are, however, consistent with decision makers optimizing a restricted notion of reward rate. As proposed by Hawkins et al. (2011), the data are consistent with the idea that participants set a goal accuracy rate (for the entire experiment), and then maximize reward rate subject to this accuracy goal, by minimizing their mean response time. For instance, suppose a participant aimed to achieve 60% correct responses throughout the experiment. The shortest total experiment time possible under this constraint is achieved by responding more accurately in the fastest conditions (small set sizes) and less accurately than the goal in the slowest conditions (large set sizes), as observed in the upper panel of Figure 3. The idea of a minimum acceptable accuracy rate neatly explains the consistent overall accuracy across drop delay conditions.

Quantitative Comparison of Experiments 1 and 2

We can provide a more direct comparison of the pure and mixed experiments by averaging the Experiment 2 data across set sizes, to produce a design closer to the design of Experiment 1. Similar, but noisier, results are obtained if only the $K = 10$ data are extracted from Experiment 2 rather than averaging across all set sizes. The averaged data from Experiment 2 are overlaid on data from Experiment 1 in Figure 4. The upper and middle panels show accuracy and step number data, respectively. As drop delay increased, participants in the mixed blocks waited for less information before making decisions that were subsequently less accurate. In contrast, in the pure blocks both accuracy and mean step number were almost unaffected by manipulation of drop delay.

We again used BIC, calculated from the lines of best fit, to approximate the posterior model probabilities, in order to examine the trends in data with increasing drop delay. For the mixed and pure blocks separately we compared a null model (i.e., the slope of the best fitting line is zero) against a drop delay model (i.e., non-zero slope). For the mixed blocks used in Experiment 1, this analysis strongly supported an effect of stimulus speed on accuracy and step number – BIC analyses clearly supported an effect of drop delay on step number ($p^{BIC} = 1$) and on choice accuracy ($p^{BIC} = .71$). For the pure blocks used in Experiment 2, in contrast, the evidence was in favor of the opposite hypothesis, constant slopes for accuracy ($p^{BIC} = .77$) and step number ($p^{BIC} = .76$).

Max-Minus-Next Heuristic

In this section, we use a simple cognitive model to test directly our hypothesis about the speed-accuracy tradeoff. Our external evidence accumulation paradigm affords direct

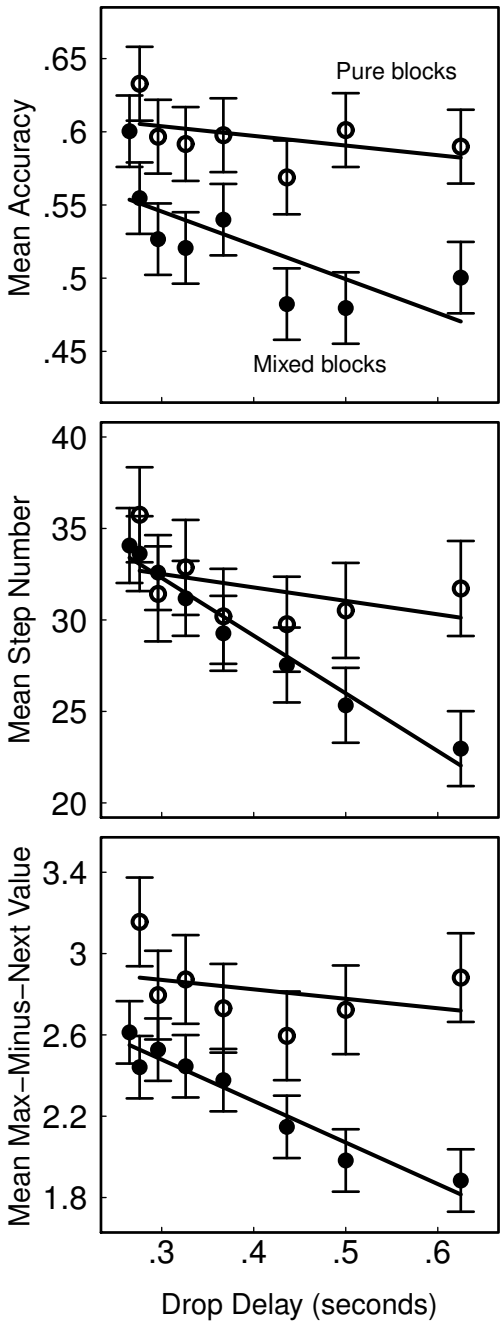

Figure 4. Mean accuracy, step number and max-minus-next value (upper, middle and lower panels, respectively) for the mixed blocks (black symbols) and pure blocks (white symbols). Pure block data are averaged across set sizes for comparison. All errors bars are ±1 between-subjects standard errors of the mean, pooled across the within- or between-subjects drop delay manipulation. The lines are regression lines of best fit, for illustrative purposes.

measurement of the quantity of evidence required to trigger a decision, by examining the accumulated bricks at the moment of response. These data can be naturally interpreted in terms of the "max-minus-next" heuristic, which holds 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 this simple decision rule approximated a statistically optimal multi-hypothesis sequential probability ratio test (Baum & Veeravalli, 1994), when error rates are low. Brown et al. (2009) interpreted the max-minus-next heuristic as a simple cognitive model, and showed that it provided a good account of data from an experiment similar to our Experiment 2.

An estimate of the max-minus-next model's parameter, Δ, is easily calculated from the data of each trial. At the point of decision in each trial, we find the two alternatives with the largest and second largest number of accumulated bricks, and calculate the difference in number of bricks between the two. The lower panel of Figure 4 shows the average Δ values from the pure and mixed blocks as functions of drop delay. As expected, in the mixed blocks there was very strong evidence that the max-minus-next threshold value (Δ) decreased as the drop delay became longer, with this model strongly supported over the null model, $p^{BIC} \approx 1$. This suggests that when the stimulus display evolved more slowly, choices were based on less careful decision criteria. This is consistent with a time-based context effect; decision makers wait for more evidence in faster choice conditions than slower conditions. In contrast, when drop delay was presented in pure blocks there was evidence in favor of the null model, $p^{BIC} = .82$. These findings reinforce our conclusions from the analyses of response time and accuracy data. When the rate of evidence accumulation (i.e., drop delay) is presented in a pure block we observe no effect of this parameter on response accuracy or the amount of information accumulated prior to choice.

Failing to Choose the Maximum Alternative

In our expanded judgment task, the column most likely to be the target, at any moment, is always the one tallest (the "maximum alternative"). Across participants and trials, 86.3% of responses in Experiment 1 responses and 87.6% of responses in Experiment 2 were to the maximum alternative. This implies that participants were operating in some clearly sub-optimal manner on about every seventh trial. To check that this behavior did not drive our results, we re-ran the above analyses, restricted to trials where the response was to the maximum alternative. For those analyses, there were no changes to the step number, accuracy or response time results reported for either experiment.

A more prosaic explanation for some of these sub-optimal responses is a delay between the point of decision and the time of response, a "response lag". Such a response lag is included in almost all decision models (called "non-decision time" in accumulator models such as: Brown & Heathcote, 2008; Ratcliff, 1978; Usher & McClelland, 2001) and the size of this delay is typically estimated indirectly, by estimating model parameters from response time data. The externalized nature of our evidence accumulation task allows us to estimate this lag more directly. As reported by Brown et al. (2009), we estimated the response lag by first assuming that participants' responses would more often align with the maximum alternative at the time the decision was made, than at the time the response was executed. We then identified the lag that maximized the agreement between response choice and the

maximum alternative, separately for each participant and each drop delay condition.

The estimated response lags were remarkably constant across participants and conditions, when measured in real time (rather than discrete steps), with an average of 1.27 seconds (SD = 0.91 seconds) in Experiment 1 and 1.06 seconds (SD = 0.73 seconds) in Experiment 2. We also confirmed that when the max-minus-next threshold parameter, Δ, was re-calculated taking into account the response lag, the patterns shown in Figure 4 were mostly unchanged (Δ decreased a little less across drop delay, but the decrease was still statistically reliable). Furthermore, when taking into account the estimated response lags, the proportion of responses to the maximum alternative increased markedly, to 97.6% in Experiment 1 and 96.2% in Experiment 2.

General Discussion

Decision context can exert a strong influence on choice behavior. Responses to the same stimulus differ as a function of preceding stimuli and responses. We provided a new example of context effects in choice elicited by a simple within- or between-subjects change in decision time, which we manipulated by altering the rate of evidence accumulation. This finding suggests that previous accounts of blocking effects founded on task difficulty alone might be incomplete, since a context effect emerged even when difficulty was held constant. We used a simple heuristic model to demonstrate that choices in the slower conditions of the mixed block (i.e., longer drop delays in Experiment 1) were based on less evidence than their faster condition counterparts. Finally, when the same drop delay parameter was presented in a pure block (as in Experiment 2), patterns of choice accuracy and information accumulation (step number) did not differ across any level of this between-subjects manipulation.

If difficulty-based explanations of context effects alone are not sufficient, how should we explain these findings? A possible alternative is through subject-controlled changes in the speed-accuracy tradeoff. Behavior in pure blocks was consistent with a single tradeoff setting established across the block, while in mixed blocks a different setting was apparently used depending on the expected decision time (e.g., an easy or hard trial, or a fast or slow drop delay). This explanation was supported by empirical observation of the speed-accuracy tradeoff parameter from the max-minus-next heuristic model, as applied to our data. Response threshold approaches have been rejected by some based on philosophical and empirical reasons (e.g., Jones et al., 2009). For instance, advance knowledge of the stimulus class of the present trial might be required in order to adjust the response threshold in time for the upcoming decision, and this information is not usually available. However, in our experiments the decision times were sufficiently slow that participants could easily have adjusted their decision thresholds during the trial, without advanced knowledge (e.g., Forstmann et al., 2008 found that participants were able to adjust their speed-accuracy tradeoff settings in less than 1.5sec., which is much faster than the decisions made by our participants).

In other work, we recently described a context effect caused by decision difficulty, and proposed an account for that effect based on speed-accuracy tradeoffs (Hawkins et al., 2011). This account assumed conditional optimality, generalizing the idea of optimizing reward rate. We assumed that participants maintain a goal accuracy rate (e.g., achieve 60% correct responses) and then adjust their response threshold settings wherever possible to minimize the time taken to complete the experiment while still achieving their goal accuracy

(we refer to this approach as "Min-RT"). The current data are qualitatively consistent with this explanation. For instance, the mixed blocks of Experiment 1 afforded participants the chance to alter their response thresholds between slower and faster conditions. For a fixed level of accuracy, shorter overall experiment times are achieved by responding more accurately than the goal rate for faster conditions (short drop delays) and less accurately than the goal rate for the slower conditions (long drop delays), as observed in the data. The same Min-RT explanation holds for Experiment 2, where the between-subjects manipulation of stimulus speed did not afford participants the chance to decrease total experiment time by changing response thresholds across levels of the drop delay manipulation. Assuming the goal accuracy rate of participants was not affected by the drop delay condition they were randomly allocated to, the fastest way to finish the experiment is to respond with the same (goal) accuracy for the whole experiment. We note, however, that the data from Experiment 2 were consistent with changing response thresholds across different choice set sizes. As in Hawkins et al., accuracy decreased as the number of choice alternatives increased, consistent with the Min-RT hypothesis.

Our data also provide proof-of-concept support for a standard assumption of many models of speeded decisions (Brown & Heathcote, 2005, 2008; Ratcliff, 1978; Usher & McClelland, 2001; Usher, Olami, & McClelland, 2002). These models assume that the time taken to make a decision depends only on the amount of evidence accumulated for and against the various alternatives. This assumption is difficult to test in traditional choice tasks where the process of evidence accumulation is assumed to be internal and implicit, for example, in perceptual decisions about the colour of a stimulus or a recognition memory judgment. In contrast, the external accumulation paradigm used here makes the accrual of evidence explicit and observable (see also: Busemeyer & Rapoport, 1988; Usher & McClelland, 2001; Vickers, 1979). It might be that the externalization of evidence accumulation imposed a different task strategy compared to internal and implicit choice tasks. At a minimum, however, the data from Experiment 2 support the notion that evidence, not decision time, was the determining factor for participants.

These data are therefore inconsistent with models in which decision time plays a role independent of decision difficulty. For example, Schneider and Anderson (2011) propose a model for multi-alternative choice based on the ACT-R cognitive architecture. Decisions in this model are composed of lower-order mental operations intrinsic to ACT-R (see Anderson et al., 2004), such as retrieving chunks and production rules from memory. Each subprocess is assumed to take a certain amount of time, and the predicted decision time is the sum of these parts. It could be that the data from expanded judgment tasks are produced by different underlying choice mechanisms than those proposed in the ACT-R memory model. Nevertheless, it is difficult to reconcile this model with the current data, which suggests that accumulated evidence, rather than elapsed time, is what determines response timing. It is possible that such a model could account for our Experiment 2 data by assuming that participants repeatedly run "micro-decisions" that are very fast, over and over again during the course of a trial, and make a response only when the result of one of these micro-decisions exceeds a criterion on some scale (such as confidence). However, it is hard to see how such an account can be separated from an information accumulation account.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

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.

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.

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.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117–128.

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*, 91–134.

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.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). The striatum facilitates decision–making under time pressure. *Proceedings of the National Academy of Science*, *105*, 17538–17542.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791–806.

Hawkins, G., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2011). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Manuscript submitted for publication.*

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.*

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.

Jones, M., Mozer, M., & Kinoshita, S. (2009). Optimal response initiation: Why recent experience matters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 785–792).

Kiger, J. I., & Glass, A. L. (1981). Context effects in sentence verification. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 688–700.

Kruschke, J. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS.* Oxford: Academic Press.

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

Los, S. A. (1996). On the origin of mixing costs: Exploring information processing in pure and mixed blocks of trials. *Acta Psychologica*, *94*, 145–188.

Lupker, S. J., Brown, P., & Colombo, P. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 570–590.

Lupker, S. J., Kinoshita, S., Coltheart, M., & Taylor, T. E. (2003). Mixing costs and mixing benefits in naming words, pictures and sums. *Journal of Memory and Language*, *49*, 556–575.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team. (2011). *nlme: Linear and nonlinear mixed effects models.*

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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.

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.

Vickers, D. (1979). *Decision processes in visual perception.* London: Academic Press.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self–regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*, 169–194.

Vickers, D., & Lee, M. D. (2000). Dynamic models of simple judgments: II. Properties of a self–organizing PAGAN (parallel, adaptive, generalized accumulator network) model for multi–choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences*, *4*, 1–31.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Lodewyckx, T., Kiruyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

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

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Appendix
Alternate Statistical Analyses

In this appendix we demonstrate that the results presented throughout the main text are not dependent on our use of BIC. We repeat our main analyses from Experiments 1 and 2 below, and report BIC (the values on which the approximations to posterior model probabilities, p^{BIC}, were based), the Akaike Information Criterion (AIC; Akaike, 1974) and null hypothesis signifcance tests, using analysis of variance (ANOVA: one-way repeated measures in Experiment 1, with drop delay manipulated within-subjects; and two-way mixed in Experiment 2, with drop delay manipulated between-subjects and set size manipulated within-subjects). To aid interpretation of AIC and BIC values, we also repeat below the BIC based approximations to posterior model probabilities (p^{BIC}) reported in the main text as well as the analogous measure for AIC, known as "Akaike weights" (Burnham & Anderson, 2004), denoted below with w^{AIC}.

The differences between the BIC analyses reported in the main text and the AIC and ANOVA analyses reported below are slight. Since ANOVA cannot directly test the null model in either Experiment, nor the additive model in Experiment 2, we appeal to a standard measure of effect size in general linear models (partial η^2) to illustrate that null models capture very little of the variance in the data. Furthermore, the nature of the complexity penalty in AIC means that AIC will be more lenient on complex models than BIC, for our data. AIC is given by $2k - 2\ln(L)$, and BIC by $k\ln(n) - 2\ln(L)$, where k is the number of parameters in the model, n is the number of data points in the observed data, and L is the maximized value of the likelihood function for the estimated model. For each additional parameter in the model, AIC adds a penalty term of 2, whereas BIC adds a penalty of $\ln(n)$, which is much larger than 2 in our experiments.

In Tables A1 and A2 we show that on the whole, the three analyses provide convergent conclusions. In Experiment 1, AIC, BIC and ANOVA all agree perfectly: they all provide strong evidence for the set size model, and do so for both accuracy and step number (Table A1). In Experiment 2, AIC and BIC agree perfectly for the accuracy data and the response time data. For the step number data, AIC shows some support for the model favored by the other approaches (an effect of set size only), but also shows support for a more complex model that also allows an additive effect of drop delay. In Experiment 2, BIC and ANOVA agree perfectly for step number and response time data, both favoring the set size only model and the saturated model, respectively. However, for the accuracy data, BIC (and AIC) prefer the simple set size only model, whereas the ANOVA approach suggests a small ($\eta^2 = .07$) interaction effect is also present. Inspection of Figure 3 suggests this may be due to noise in the $K = 18$ condition for drop delay 436ms (line 5 in Figure 3). This set size demonstrated particularly low mean accuracy, which was the result of a few participants responding close to chance in this set size who were not excluded from the overall analysis since their mean accuracy was above the exclusion criterion. To check this interpretation, for the 436ms drop delay group we replaced the mean accuracy for $K = 18$ trials with the median of the $K = 16$ and $K = 20$ trials, separately for each participant. This analysis did not change the AIC and BIC results, or the ANOVA main effects of set size and drop delay, but did reduce the strength of the drop delay by set size interaction: $F_{(54,1413)} = 1.46$, $p = .02$, partial $\eta^2 = .05$. Although this interaction would be declared sig-

nificant by $\alpha = .05$ convention, given the size of the tail probability and effect size estimate, relative to the other effects in reported in Table A2, we would be hesitant to consider this interaction as a reliable effect.

Table A1: Experiment 1 analyses of mean accuracy and step number. For AIC and BIC the preferred model is the one with the lowest value, denoted with an asterisk.

	Model (effect)	AIC (w^{AIC})	BIC (p^{BIC})	F-ratio (df)	p-value	Partial η^2
Accuracy	*Null*	-480.65 (0)	-469.00 (.001)	.	.	.
	Drop delay	-506.03^* (1)	-482.74^* (.999)	7.78 $(7, 308)$	$< .001$	.15
Step number	*Null*	2385.69 (0)	2397.34 (0)	.	.	.
	Drop delay	2070.12^* (1)	2093.41^* (1)	68.89 $(7, 308)$	$< .001$	.61

Table A2: Experiment 2 analyses of mean accuracy, step number, and response time in seconds. The additive model refers to the additive set size and drop delay effects. The saturated model refers to the additive set size and drop delay effects as well as the interaction between the two. For the ANOVA, the saturated model denotes the interaction effect. For AIC and BIC the preferred model is the one with the lowest value, denoted with an asterisk.

	Model (effect)	AIC (w^{AIC})	BIC (p^{BIC})	F-ratio (df)	p-value	Partial η^2
Accuracy	*Null*	−716.57 (0)	−700.37 (0)	.	.	.
	Set Size	−2099.55* (.994)	−2067.15* (1)	302.69 (9, 1413)	< .001	.66
	Drop delay	−706.95 (0)	−685.34 (0)	.60 (6, 157)	.73	.02
	Additive	−2089.49 (.006)	−2051.69 (0)	.	.	.
	Saturated	−2074.69 (0)	−2031.49 (0)	1.89 (54, 1413)	< .001	.07
Step number	*Null*	11693.22 (0)	11709.43 (0)	.	.	.
	Set size	10619.43 (.466)	10651.83* (.931)	202.51 (9, 1413)	< .001	.56
	Drop delay	11693.16 (0)	11714.77 (0)	.61 (6, 157)	.72	.02
	Additive	10619.23* (.513)	10657.04 (.069)	.	.	.
	Saturated	10625.62 (.021)	10668.82 (0)	1.03 (54, 1413)	.42	.04
Response time	*Null*	8944.38 (0)	8960.58 (0)	.	.	.
	Set size	7908.07 (0)	7940.48 (0)	178.95 (9, 1413)	< .001	.53
	Drop delay	8907.70 (0)	8929.30 (0)	9.30 (6, 157)	< .001	.26
	Additive	7890.23 (.005)	7928.04 (.069)	.	.	.
	Saturated	7879.64* (.995)	7922.84* (.931)	3.45 (54, 1413)	< .001	.12