

## **Anticipating Cognitive Effort: Roles of Perceived Error-likelihood and Time Demands**

Timothy L. Dunn<sup>1</sup>

Michael Inzlicht<sup>2,3</sup>

Evan F. Risko<sup>1</sup>

1. Department of Psychology, University of Waterloo

2. Department of Psychology, University of Toronto

3. Rotman School of Management, Toronto

All data and corresponding code are freely available via the Open Science Framework at [osf.io/2szy3](https://osf.io/2szy3)

Address correspondence to: Timothy L. Dunn ([t2dunn@uwaterloo.ca](mailto:t2dunn@uwaterloo.ca)), Department of Psychology, University of Waterloo, 200 University Ave W, Waterloo, ON, N2L 3G1

Word count: 13824

### **Abstract**

Why are some actions evaluated as effortful? In the present set of experiments we address this question by examining individuals' perception of effort when faced with a trade-off between two putative cognitive costs: how much time a task takes versus how error-prone it is. Specifically, we were interested in whether individuals anticipate engaging in a small amount of hard work (i.e., low time requirement, but high error-likelihood) versus a large amount of easy work (i.e., high time requirement, but low error-likelihood) as being more effortful. In between-subject designs, Experiments 1 through 3 demonstrated that individuals anticipate options that are high in perceived error-likelihood (yet less time consuming) as more effortful than options that are perceived to be more time consuming (yet low in error-likelihood). Further, when asked to evaluate which of the two tasks was (a) more effortful, (b) more error-prone, and (c) more time consuming, effort-based and error-based choices closely tracked one another, but this was not the case for time-based choices. Utilizing a within-subject design, Experiment 4 demonstrated overall similar pattern of judgments as Experiments 1 through 3. However, both judgments of error-likelihood and time demand similarly predicted effort judgments. Results are discussed within the context of extant accounts of cognitive control, with considerations of how error-likelihood and time demands may independently and conjunctively factor into judgments of cognitive effort.

*Keywords:* anticipated effort, cognitive costs, control, heuristics, decision-making

### **Anticipating Cognitive Effort: Roles of Perceived Error-likelihood and Time Demands**

Why are effortful actions evaluated as being effortful? The question of what constitutes the cost of effort has recently come to the fore in research focused on the control of human behavior (e.g., Inzlicht, Bartholow, & Hirsh, 2015; Kurzban, 2016). Effortful actions are often attributed to those instances when behavioral control is deployed, are most often thought to be inherently costly in decision-making, be aversive (Dreisbach & Fischer, 2012; Kool, McGuire, Rosen, & Botvinick, 2010; Shenhav et al., 2017), and invoke an urge to disengage even when such actions may be considered adaptive (Kurban, 2016). For example, individuals willingly avoid lines of action associated with more effort when given a choice (Kool et al., 2010) and require higher levels of reward to engage in such actions (Westbrook, Kester, & Braver, 2013).

Though the discussion of effort as a key determinant in decision-making and control is ubiquitous, what specifically constitutes effort, and the cost therein, as a construct to be empirically tested is still an open question. Though controversial, (Kurzban, Duckworth, Kable, & Myers, 2013; Navon, 1984), perhaps the most pervasive conception of effort treats the construct as being *exerted* or *invested* when situated within a demanding task (e.g., Ackerman & Thompson, 2017; Navon & Gopher, 1979; Kahneman, 1973; Verguts, Vassena, & Silvetti, 2015; Wickens 2002). That is, cognitive effort is a type of capacity or resource in-and-of itself that can be deployed in response to task demand requirements, much like conceptions of physical effort. This process of exerting cognitive effort is often indexed by the modulation of performance in response to control-demanding conditions (Akçay & Hazeltine 2007, Dreisbach & Fischer, 2011; Kerns et al., 2004), the engagement of the executive control system (for a review see

Botvinick & Braver, 2015), and modulated peripheral physiological responses (e.g., pupil dilation, Bijleveld, Custers, & Aarts, 2009; Diede & Bugg, 2017; Kahneman & Beatty, 1966; Van Steenbergen & Band, 2013; skin conductance responses, Botvinick & Rosen, 2009; Kahneman, Tursky, Shapiro & Crider, 1969; Naccache et al., 2005). Recently, Shenhav and colleagues (2017) proposed a formalization of cognitive effort as a type of capacity that can be exerted to mediate between a task and the potential capacity/fidelity of controlled processing.

A theoretical alternative to the above conceptualization of cognitive effort is that this effort is a subjective phenomenon used to control behavior (Dunn, Lutes, & Risko, 2016; Kool et al., 2010; Kurzban et al., 2013; Westbrook & Braver, 2015). Importantly, in at least some of these accounts, cognitive effort is not directly tied to the cognitive demands of a task. This argument comes from accumulating findings that demonstrate dissociations between various measures of increased control (i.e., performance, neural and physiological measures) and measures of subjective effort (Chong et al., 2017; Desender, Van Opstal, & Van den Bussche, 2017; Desender, Buc Calderon, Van Opstal, & Van den Bussche, 2017; Dunn et al., 2016; Dunn & Risko, 2016; Dunn & Risko, submitted; Kool et al., 2010; Gold et al., 2014; McGuire & Botvinick, 2010; Naccache et al., 2005; Westbrook, Kester, & Braver, 2013). Dunn and colleagues (2016; see also Dunn & Risko, submitted) explicitly liken cognitive effort to a subjective and largely inferential metacognitive evaluation of demand used to control action selection whether situated in a task (e.g., disengaging from an action through a retrospective evaluation of effort) or not (e.g., avoiding an action outright through a prospective evaluation of effort).

Though specification of what leads to some tasks being perceived as effortful within this metacognitive account is lacking.

Given the competing formalizations of cognitive effort, testing factors that relate to the *effortfulness* (i.e., the inherent cost of engaging in a demanding task) of tasks remains an important theoretical problem to address. In the following, we highlight two separate (but sometimes related) basic perceived cognitive costs often associated with effortfulness: time-requirements and error-likelihood.

### **Costs Related to Effortfulness**

*Time requirements.* The claim that processes or actions that incur a time cost relative to alternatives are more effortful has enjoyed a successful history within psychology. For example, the Soft Constraints Hypothesis (Gray, Sims, Fu, & Schoelles, 2006) posits that, at relatively fast time scales, the cognitive system selects routines of actions that minimize time costs while achieving expected benefits. Griffiths and colleagues (2015) suggest that the system selects between operations by predicting and attempting to maximize the Value of Computation (VOC) of a process, which consists of the reward of the computation discounted by the cost of the computation in terms of time.

The general idea of more time being associated with effortfulness is prevalent at varying levels in several opportunity cost frameworks of behavioral control as well. An opportunity cost can generally be conceived of as the process of engaging in some choice at the cost of some foregone alternative choice (e.g., “To go without fish to get game or the raising of wheat upon terms foregoing the raising of corn...”, Davenport, 1911, p. 725). Opportunity costs express the basic relation between scarcity and choice, and as such, provides a useful construct in understanding cost-benefit analyses in behavioral

control. For example, Niv and colleagues (2007) propose that an average rate of reward serves as a cue for opportunity cost in evaluations of physical effort. If the average rate of reward is high, then every second that a reward is not delivered is costly. Thus, there is a benefit of performing at a quicker rate even if the energetic costs of doing so are greater. Within this context, the average reward rate approximates the opportunity cost of time and the system may apply this rate across many types of decision contexts (Boureau, Sokol-Hessner, & Daw, 2015).

Unsurprisingly, the opportunity cost approach to understanding behavioral control has been applied to specific accounts of cognitive effort. Kurzban and colleagues (2013; Kurzban, 2016; see also Kool & Botvinick, 2014) propose that opportunity costs arise as a function of parallel-processing capacity being finite, scarce (Baddeley & Hitch, 1974; Kahneman, 1973; Kurzban et al., 2013; Navon & Gopher, 1979; Shiffrin & Schneider, 1977; Wickens, 2002), and dynamically allocated across processes. The feeling of effort arises as an output of mechanisms that compute the costs and benefit of engaging in a task relative to alternative tasks that the same processes may be applied. Costly, and therefore effortful tasks, are those that engage work from multiple cognitive processes (Botvinick & Braver, 2015) in the absence of an offsetting reward, and are often conceptualized as inherently placing high levels of demand on the executive control system (e.g., task-switching, Monsell, 2003; the flanker task, Eriksen, 1995; the Simon task, Lu & Proctor, 1995; the Stroop task, MacLeod, 1991). As an example, the opportunity cost of task-switching can be indexed by the need to shift resources across attention, classification, and memory retrieval between different attributes of the tasks. Correspondingly, these shifts across processes generally require more time by the system

to execute relative to not switching (i.e., switch costs; Monsell, 2003). This increase in processing time may then prevent additional similar processes from being carried out, given some amount of capacity is being held up by the multiple processes associated with task-switching. Put simply, factors that make current goals take longer to obtain are expected to influence opportunity costs and effort (Westbrook & Braver, 2015). Thus, switching should be felt as more effortful than not switching in a situation where no reward offsets engaging in switching.

Comparable to explicit opportunity cost accounts, some motivational accounts of self-control similarly assign a cost to increased time requirements. For example, Inzlicht, Schmeichel, and Macrae's (2014) shifting-priorities process model of self-control hypothesizes that, given time is a scarce resource, the system attempts to optimally balance a trade-off between cognitive work and cognitive rest, with the former often requiring some external reward to engage in and the latter often being more intrinsically rewarding. Cognitive work continuing beyond some expected reward over time becomes aversive. This time cost accumulates and is tracked leading to increased subjective experiences of signals such as mental fatigue and effort. These signals are then used by the system to amplify the urge to disengage in favor of more rewarding behaviors such as exploration, leisure, or a "want-to" (rather than "have-to") goal. It is important to note that in both this model and the opportunity cost model proposed by Kurzban and colleagues (2013), the effortfulness of a task is always relative to the motivation of the agent. That is, inherently demanding tasks would be expected to be felt as less effortful if offset by some form of reward.

*Error-likelihood.* Beyond time requirements, error likelihood can also be considered as a potential determinant of effortfulness. At the neural level, error commission leads to a fast-negative deflection in a fronto-centrally located event-related potential (ERP) component known as the Error Related Negativity (ERN or Ne; Falkenstein, Hoormann, Christ, & Hohnsbein, 2000; Gehring, Goss, Coles, Meyer, & Donchin, 1993; Holroyd & Coles, 2002; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001). The ERN is thought to serve as a reinforcement-learning signal used to optimize performance (Holyrod & Coles, 2002; c.f., Brown & Braver, 2005; Yeung, Botvinick, Cohen, 2004) and is expected to play a key role in driving behaviors. Upon error commission the ERN is generated by activity from the mesencephalic dopamine system located within the anterior cingulate cortex (ACC) signaling that the consequences of an action are worse (or better) than expected by the system (i.e., a temporal difference error; Schultz, Dayan, & Montague, 1997). This difference between the expected- and experienced-reward functions as a signal in action and outcome learning that increases a behaviors' reinforcement likelihood (Gläscher, Hampton, & O'Doherty, 2009; Montague, Dayan, Sejnowski, 1996). Thus, when a person commits an error (i.e., a deviation from intended behavior), an error signals that effortful control processes may be required for behavioral adjustment (e.g., post error slowing; Rabbitt, 1966) or entail the reassessment of an entire behavioral plan (e.g., avoid a line of action or disengage from a current action; Taylor, Stern, & Gehring, 2007). For example, Frank and colleagues (2005) demonstrated that the magnitude of the ERN predicts learning from errors, and that more negative ERNs are associated with a higher avoidance of negative stimuli. Westbrook and Braver (2016) recently offered a formalization of the



relation between effort and errors by hypothesizing that a specific form of error-related signal (i.e., reward prediction errors) carries effort-discounted signals for use in decision-making.

An alternative, but not necessarily mutually exclusive possibility, is that error likelihood can serve as a signal of whether the system is approaching capacity limitations when situated in a task. Such a view can be grounded in the distinction between automatic and controlled processing, with the former being argued to be relatively effortless and the latter being more effortful (for recent reviews see Botvinick & Cohen, 2014; Shenhav et al., 2017). Recent work has suggested that the source of these capacity limitations lie in cross-talk produced by the use of shared representation by different processes (Feng, Schwemmer, Gershman, & Cohen, 2014). Such processing bottlenecks may then require the intercession of control mechanisms to manage and minimize cross-talk (Shenhav et al., in 2017). From a limited-capacity perspective (Baddeley & Hitch, 1974; Kahneman, 1973; Kurzban et al., 2013; Navon & Gopher, 1979; Shiffrin & Schneider, 1977; Wickens, 2002), error-likelihood may then signal the need to reconfigure processes through control to avoid situations of cross-talk.

Beyond response monitoring accounts of the error monitoring system (e.g., Dehaene, Posner, & Tucker, 1994; Holyrod & Coles, 2002), several alternative accounts suggest that the ERN also reflects a negative affective response to errors that is sensitive to motivational states and traits (Hajcak, Moser, Yeung, & Simons, 2005; Luu, Collins, & Tucker, 2000; Luu, Tucker, Derryberry, Reed, & Poulsen, 2003; Maier, Scarpazza, Starita, Filogamo, & Ladavas, 2016). Within this framework errors can be considered broadly as maladaptive responses that, upon commission, may place an organism in

danger and threaten its safety (Hajcak & Foti, 2008). To illustrate this notion, Hajcak and Foti (2008) demonstrated that defensive startle responses (i.e., the reflexive contracting of the body into a defensive posture) were larger upon error commission. The authors thus argue that errors prompt defensive responses, serving a basic motivational function. Situated within accounts of control, errors can then be considered to be particularly aversive, generate strong emotional responses upon commission, and require greater adjustments of effortful control to resolve (Inzlicht et al., 2015). Tasks associated with a higher likelihood of errors relative to an alternative course of action then can be considered an additional factor leading to effortfulness.

### **Present Investigation**

To better understand how time requirements and error likelihood together are related to effortfulness, here, we pitted both costs directly against one another in a decision-making task. Specifically, we contrasted individuals' perception of anticipated effort when faced with a trade-off between engaging in a task associated with relative high error-likelihood but low time requirements versus a task associated with relative low error-likelihood but high time requirements. Individuals made choices between two explicitly presented alternative tasks with respect to which was more effortful, had a higher likelihood of an error, or was more time demanding. Though many accounts of effort largely focus on why actions become subjectively effortful over time while situated within a task, less attention has been paid to why some actions can be perceived as effortful at the point of initial evaluation prior to engaging.

We focused on individuals' evaluations of *anticipated* effort, time, and errors. Anticipated effort, as opposed to *experienced* effort, has been argued to be crucial in

decision-making processes (Dunn, Lutes, & Risko, 2016; Payne, Bettman, & Johnson, 1993). Investigating judgments of anticipated effort, errors, and time affords the opportunity to gain valuable insight into the extra-experimental biases that individuals bring to bear when making effort-based judgments. This is not to devalue the utility of investigating experienced effort. Rather, prospective evaluations of anticipated effort and “online” evaluations of experienced effort both deserve researchers’ attention. For example, the former likely plays a major role in decisions about whether to take on a task at all (i.e., one cannot experience the effortfulness of a task if it is avoided because of a prospective evaluation of effortfulness). Furthermore, we chose to elicit judgments in a between-subject design where individuals made only one judgment based on effort, errors, or time in isolation of explicit information about the other dimensions. Utilizing a between-subject design allows for a purer test of heuristic reasoning in judgment and neutralizes the potential for individuals simply aligning choices to be consistent (i.e., across the effort, error, and time dimensions) which is an issue with within-subject designs (Kahneman & Tversky, 1996).

To vary time and errors we used basic manipulations of stimulus rotation and set size. Individuals were presented with a single word rotated 110° and two upright words. Critically, based on past research (Jordan & Huntsman, 1990; Koriat & Norman, 1984) reading a single word rotated 110° aloud generates more errors relative to two upright words, whereas reading two upright words takes longer to read relative to a single rotated word. We confirmed this general pattern of performance data below in Experiment 1a with an in-lab behavioral task where participants read the above stimuli aloud (i.e., Experiment 1a served as a type of manipulation check to confirm the trade-off pattern). In

Experiment 1b then, individuals were faced with a trade-off between a faster option associated with a higher likelihood of an error and a slower option associated with a lower likelihood of an error when making anticipated effort judgments.

Experiment 1c utilized the same choice context and stimuli but manipulated the basis of individuals' choices between-subject. Here, individuals were asked to make *more time demanding* or *higher error likelihood* choices. If effort judgments are associated with time, then the 2-words/0° display should elicit greater *more effortful* and *more time demanding* choices but lower *higher error likelihood* choices relative to the 1-word/110° display. Alternatively, if effort judgments are more associated with error-likelihood, then the 1-word/110° display should elicit greater *more effortful* and *higher error likelihood* choices, but lower *more time demanding* choices relative to the 2-words/0° display.

### **Experiments 1a, 1b, and 1c**

#### **Method**

In the following we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012).

#### **Participants**

*Initial Sample Size Determination for Experiment 1a.* Twenty University of Waterloo undergraduates participated for course credit. This sample size allows for the detection of at the least a medium effect size across conditions for a within-subjects design. No individuals were removed from the below analyses.

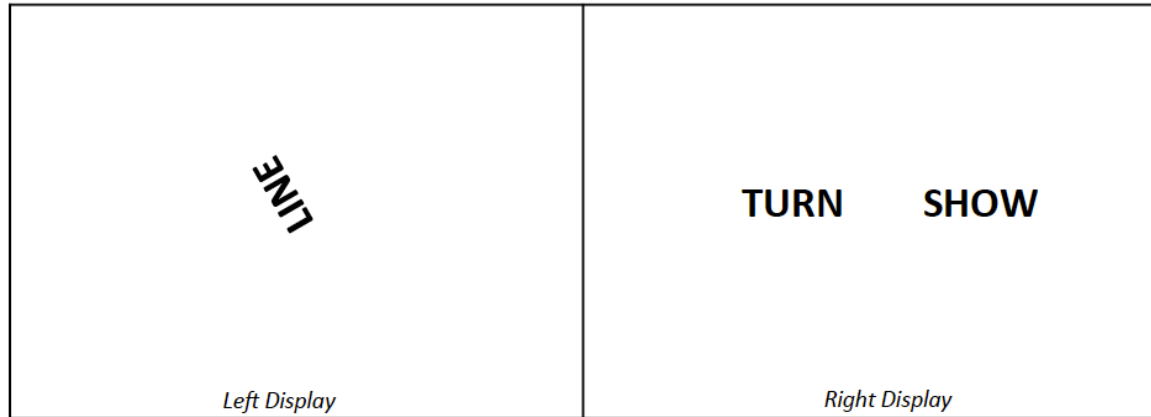
*Initial Sample Size Determination for Experiments 1b and 1c.* A pilot study was conducted based on Experiment 4 from Dunn, Koehler, and Risko (2017) where

individuals similarly made effort-based choices between displays including stimulus rotation and set size. Results from the pilot study demonstrated an effect of 63% for effort choices favoring the stimulus rotation condition,  $g = .13$ ,  $BF_{Alt} = 1.40$ , suggesting a sample size of 93 was needed per condition (based on null hypothesis significance testing; NHST). The initial sample size was set at  $n = 96$  per rating condition to ensure complete counterbalancing of the stimulus lists (see stimuli below). Given the size of the effects for accuracy and time choices are unknown for Experiment 1c, the sample size from Experiment 1b was carried over for each of the dimensions.

*Current Sample for Experiments 1b and 1c.* Ninety-six Amazon Mechanical Turk (MTurk) workers were recruited in Experiment 1b for the online study (see Buhrmester, Kwang, & Gosling, 2011) and compensated \$1 USD for participating. Twenty-five percent of individuals failed an attention check embedded in the survey (see procedure below) resulting in a final  $N$  of 72 ( $Median_{Age} = 29$  years,  $Min_{Age} = 20$  years,  $Max_{Age} = 61$  years, 54% male participants, and 54% reported completing a Bachelor's degree or higher).

One hundred and ninety-six MTurk workers were recruited ( $n = 96$  per dimension) in Experiment 1c for the online study and compensated \$1 USD for participating. Nine percent of individuals failed the attention check embedded in the survey resulting in a final  $N$  of 174 ( $Median_{Age} = 31$  years,  $Min_{Age} = 20$  years,  $Max_{Age} = 68$  years, 57% male participants, and 48% reported completing a Bachelor's degree or higher).

Figure 1. Example of Choice Screen with each Option



*Note:* Both options, the 1-word/110° display and 2-words/0° display, were presented side-by-side to individuals. Individuals were instructed to choose which option they felt would be more effortful (Experiments 1a and 2a), more time demanding (Experiments 1b and 2a), or less accurate (Experiments 1b and 2a) to read aloud.

### **Design**

For Experiment 1a, a one-factor (Display Condition: 1-word/110°; 2-words/0°; 3-words/0°; 4-words/0°; 5-words/0°) within-subjects design was employed. For Experiment 1b, a one-factor (Choice Option: 1-word/110°; 2-words/0°) design was employed, where individuals only made effort-based choices (see Figure 1 for an example). For Experiment 1c, a one factor (Rating Dimension: Error-likelihood, Time) between-subjects design was employed where individuals either made an error- or time-based choice when contrasting the 1-word/110° and 2-words/0° display conditions.

### **Apparatus**

Experiment 1a was deployed using DMDX software (Forster & Forster, 2003). Stimuli were presented on a 24" LCD monitor with individuals sitting approximately 70cm away. A standard QWERTY keyboard was used for manual responses.

**Stimuli**

Stimuli consisted of a single word presented at  $\pm 110^\circ$  and two words presented at upright ( $0^\circ$ ). Words consisted of three high frequency nouns: “LINE”, “TURN”, and “SHOW”, *Mean Written Word Frequency* = 273 per thousand. In addition, an arrow was placed between the words in the 2-word display to draw attention to reading direction. Twelve unique lists were constructed and counterbalanced such that each word appeared in every position across the left and right displays (see Figure 1). All stimuli were similar in Experiment 1b, though the arrow was removed from the 2-word stimuli given several participants in Experiment 1a reported that it was unclear whether they were to imagine naming the word “ARROW” in the display. This removal resulted in a better rate of individuals passing the attention check.

**Procedure**

*Experiment 1a.* Individuals entered the testing room and were seated approximately 70 cm away from the monitor. Instructions stated that individuals were to read each presented display aloud as quickly and as accurately as possible and to press the “B” button when they were finished. Extra emphasis was added to be sure that they had fully finished reading aloud prior to pressing the “B” button to avoid spoiled trials. In addition, individuals were asked to maintain an upright head position while loosely remaining in a headrest. Individuals were not required to fully set their chin into the headrest to ensure that they could comfortably respond aloud. Individuals completed 16 trials of each of the choice option conditions for a total of 80 trials. The entire experiment took approximately 15 minutes to complete.

*Experiments 1b and 1c.* MTurk workers selected the task and provided informed consent electronically. Instructions stated that the task to-be-completed would be to choose which out of two different tasks presented would be “*More Effortful (i.e., difficult or demanding)*” to complete. Individuals were further instructed that they were to imagine that the specific task they would be asked to do would be to name the word or all of the words presented to them aloud. In addition, individuals were presented with a sample display of three upright words and instructed that if they were presented the display, then we would want them to imagine that they would be expected to read all three words in the display in a natural left-to-right manner. Once confirming that they understood the instructions as stated, participants were randomly presented one display from the list of 12. To make their “*More Effortful*” choice, individuals selected one of two radio buttons labeled “*The Left Display*” and “*The Right Display*”. For Experiment 1b, instructions for the time dimension stated that the task to-be-completed would be to choose which out of two different tasks presented would be “*More Time Demanding (i.e., take more time)*” to complete. Instructions for the accuracy dimension stated that the task to-be-completed would be to choose which out of two different tasks presented you would be “*Less Accurate (i.e., make more errors)*” at completing.

Once participants made their choice, an attention check was presented displaying the same choice screen the participant received and asked, “*If we asked you to name the words on the left/right, then how many words total would you have named?*”. The specific “*left/right*” designation was always to the 2-word display, thus the correct answer was “2” for every participant. Individuals then completed demographic



information and were given a unique code to enter back into Mechanical Turk to receive payment. Completion of the study took approximately five minutes.

## Results

Results are reported first for display option performance (i.e., response times and accuracy), followed by individuals' effort, time, and error choices. Bayes Factors (BF) were computed using the BayesFactor package (Morey & Rouder, 2015) in R (R Core Team, 2014). Evidential strength categories for Bayes Factors (i.e., in favor of the alternative hypothesis) follows the criteria outlined by Lee and Wagenmakers (2013; see similarly Jeffreys, 1961): 1-3 "*Anecdotal*", 3-10 "*Moderate*", 10-30 "*Strong*", 30-100 "*Very Strong*", > 100 "*Extreme*". A default fixed  $r$  scale =  $\frac{1}{2}$  was used in the calculation of BFs for the ANOVAs computed for performance. These BFs are presented as referenced to the random effect error model as the null model.

Bayesian analyses of the maximum a posteriori estimate (i.e., mode value; MAP) of the  $\theta$  parameter (i.e., successes and failures for the choice data) and 95% Highest Density Intervals (HDI) around  $\theta$  (Kruschke, 2013) were generated in R. The determination of priors for the Beta distribution shape parameters  $\beta(a,b)$  (where  $a$  = successes and  $b$  = failures within the sample) used for HDIs and BFs are outlined preceding the reporting of results. Furthermore, Binomial and chi-square test results are presented alongside Bayesian analyses where applicable. For Experiments 1b and 1c default priors were used for BFs (i.e.,  $r$  scale = .707), and priors for estimation were set for  $\beta(a,b)$  as  $a = 4$  and  $b = 4$ . The latter prior closely approximates the former for BFs, thus the two priors are fairly commensurate across the BF and estimation analyses.

### Experiment 1a

The ezANOVA (Lawrence, 2015) package was utilized for ANOVA analyses. Performance coding was completed using CheckVocal software (Protopapas, 2007). Approximately 2% of trials were removed as spoiled (i.e., hitting the “B” button prior to finishing the response). Results are reported first for response times (RT) followed by accuracy (see Table 1).

All error trials were removed for RT analyses. One trial was removed as an extreme outlier based on *Z*-scoring ( $Z = 6.40$ ). Upon removal, the RT distribution showed little signs of extreme skewness (.58) or kurtosis (2.99). A one-way BF ANOVA demonstrated positive evidence for the alternative (i.e., an effect of display condition on RTs relative to an error only model),  $BF_{Alt} = 5.34$ ,  $F(4, 76) = 296.33$ ,  $MSE = 28595.49$ ,  $p < .001$ ,  $\eta^2 = .94$ . Further, BFs were computed for the 1-word/110° condition relative to all other conditions. Extreme evidence for the alternative (i.e., a difference between the display conditions) was demonstrated for each of the four comparisons, minimum  $BF_{Alt} = 102.28$ , minimum  $d = .57$  for the 1-word/110° x 2-word/0° comparison. Thus, the 1-word/110° condition was faster to read aloud relative to all other conditions with the smallest effect being medium in size.

For accuracy, A one-way Bayes Factor (BF) ANOVA demonstrated strong evidence for the null (i.e., effect of display condition on RTs relative to an error only model),  $BF_{Null} = 333.33$ ,  $F(4, 76) = .91$ ,  $MSE = .004$ ,  $p > .1$ ,  $\eta^2 = .05$ , demonstrating that accuracy did not vary across the five display conditions. Qualitatively the 1-word/110° condition produced the lowest accuracy (i.e., more errors) relative to all condition with the exception of the 5-word/0° condition.

Table 1. Mean Performance Results

	1-words/110°	2-words/0°	3-words/0°	4-words/0°	5-words/0°
Response Times (ms)	1321 (392)	1543 (298)	1999 (378)	2461 (425)	2914 (492)
Accuracy	96% (19%)	98% (15%)	98% (15%)	98% (14%)	96% (19%)

*Note:* standard deviations are presented in parentheses.

### Experiments 1b and 1c

All results in the following are expressed as the overall proportion of choices for the 1-word/110° display relative to the 2-words/0° display alternative. First for Experiment 1b, individuals chose the 1-word/110° display as the more effortful option relative to the 2-words/0° display 75% of the time,  $BF_{Alt} > 1,000$ , MAP = 73%, 95% HDI [63%, 82%],  $p < .001$  Binomial test. For the error-likelihood dimension in Experiment 1c individuals chose the 1-word/110° display as the less accurate option relative to the 2-words/0° display 79% of the time,  $BF_{Alt} > 100,000$ , MAP = 77%, 95% HDI [68%, 85%],  $p < .001$  Binomial test. For the time dimension, individuals chose the 1-words/110° display as the more time demanding option relative to the 2-words/0° display 50% of the time,  $BF_{Null} = 3.85$ , MAP = 50%, 95% HDI [40%, 60%],  $p > .1$  Binomial test.

Furthermore, BF chi-square tests were conducted to test 1-word/110° display choices across the effort, error-likelihood, and time dimensions. Results demonstrated moderate evidence for the null hypothesis (i.e., the frequencies are similar) for 1-word/110° choices across the effort (75%) and accuracy (79%) dimensions,  $BF_{Null} = 5.00$ ,  $\chi^2(1) = .17$ ,  $p > .1$ . This offers moderate support for the notion that effort ratings closely track ratings of error-likelihood in the between-subject design. Comparisons of choices

for the effort and error-likelihood dimensions against the time dimension (50%) demonstrated very strong evidence for the alternative (i.e., the frequencies are different) in both cases,  $BF_{Alt} = 36.21$ ,  $\chi^2(1) = 9.40$ ,  $p < .001$ ,  $BF_{Alt} = 583.54$ ,  $\chi^2(1) = 14.78$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time, respectively. Overall, this offers very strong support for the notion that ratings of effort and error-likelihood did not track ratings of time-requirements.

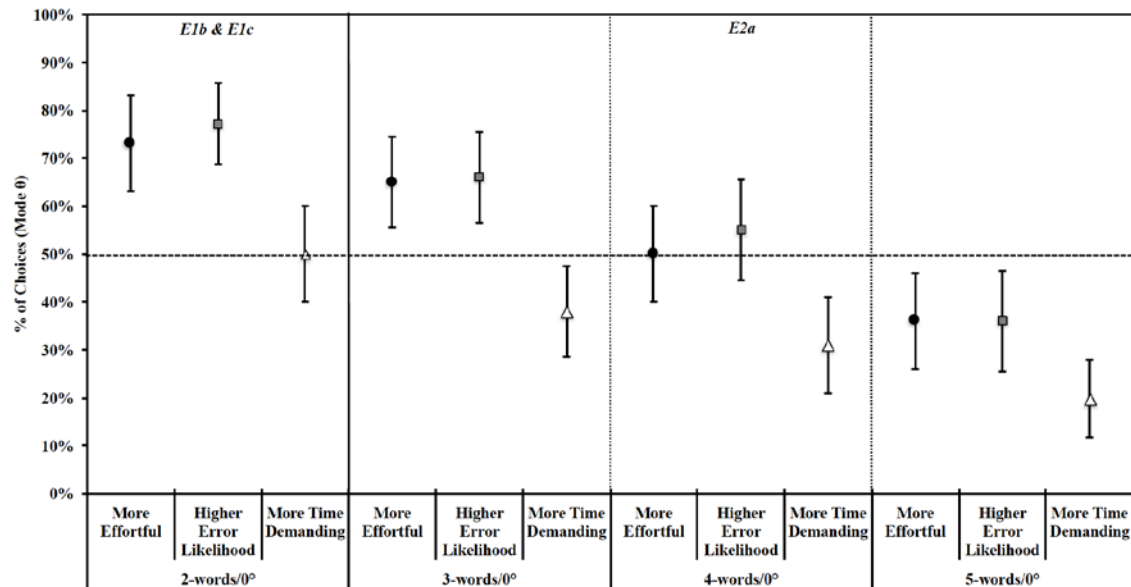
In sum, in a between-subject design, individuals similarly chose the 1-word/110° option as the more effortful and more error-prone option relative the 2-words/0° display. In contrast, by individuals showing no difference on the dimension of time requirement, they contrasted sharply with effort and error-likelihood choices (see Figure 2).

## **Discussion**

Experiment 1 demonstrated varying patterns of choices across anticipated effort, error-likelihood, and time judgments. Individuals anticipated that the 1-word/110° display would be more effortful and more error-prone to read aloud relative to the 2-words/0° display, though accuracy was relatively equivalent across the options based on actual performance estimated from the separate Experiment 1a sample. Moreover, individuals showed no difference in choices when evaluating the displays based on anticipated time requirements. This was the case even though actual performance in Experiment 1a demonstrated a moderate time cost (222 ms;  $d = .57$ ) for the 2-words/0° display relative to the 1-word/110° display. These findings coincide with much previous work highlighting dissociations between subjective reports of performance and actual performance (e.g., Bryce & Bratzke, 2014; Dunn & Risko, 2016; Dunn et al., 2016; Marti, Sackur, Sigman, & Dehaene, 2010; Miller, Vieweg, Kruize & McLea, 2010).

Therefore, in the between-subject design used across Experiments 1b and 1c, individuals' anticipated effort choices showed a close relation to anticipated error-likelihood choices, but not with time requirement choices.

Figure 2. Individuals' More Effortful, Higher Error-Likelihood, and More Time Demanding Choices for the 1-word/110° display in Experiments 1b, 1c, and 2a.



*Note:* All presented data points are for choices of the 1-word/110° display; the alternative choice relative to the 1-word/110° display is plotted on the X-axis. That is, choices below chance (50%) would reflect a tendency to more often choose the alternative choice denoted on the X-axis. Data for the 1-word/110° x 2-words/0° comparison was collected in Experiments 1b and 1c. All other data were collected in Experiment 2a. The mode  $\theta$  values are based on the posterior distribution. Error bars represent 95% Highest Density Intervals (HDI).

### Experiments 2a and 2b

Experiment 2a aimed to further explore the observed pattern of judgments between the effort, error-likelihood, and time dimensions (i.e., effort = errors > time). Again, using a between-subject design, individuals completed the judgment task as described in Experiments 1b and 1c. However, in Experiment 2 the set size manipulation

varied from 3-words to 5-words while keeping the 1-word/110° display constant across all contrasts. By increasing set size, time judgments would be expected to fully dissociate from effort and error judgments (i.e., more items should increase the likelihood that individuals judge more words as having greater time requirements). A specific hypothesis is not forwarded as to which set size comparison these differences will occur, but rather by increasing set size, the likelihood that this difference occurs should increase accordingly. To foreshadow, Experiment 2a demonstrated a clear dissociation between effort and error-likelihood judgments relative to time judgments when specifically contrasting the 1-word/110° and 3-words/0° displays. Given the importance of this finding to our main goal, a registered replication was completed of the specific 1-word/110° versus 3-words/0° choice condition for all choice dimensions in Experiment 2b (please see <https://osf.io/2szy3/registrations> for the replication protocol).

## **Method**

### **Participants**

*Initial Sample Size Determination for Experiment 2a.* The sample sizes of  $n = 96$  per condition from Experiments 1b and 1c were used in Experiment 2 (i.e., 96 participants in each of the nine cells of the between-subject design.).

*Current Sample for Experiment 2a.* Eight hundred and sixty-four MTurk workers were recruited for the online study and compensated \$1 USD for participating. Eleven percent of individuals failed the attention check embedded in the survey resulting in a final  $N$  of 769 ( $Median_{Age} = 33$  years,  $Min_{Age} = 18$  years,  $Max_{Age} = 82$  years, 48% male participants, and 49% reported completing a Bachelor's degree or higher).

*Initial Sample Size Determination for Experiment 2b.* The replication used optional stopping methods (Rouder, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017) to determine the final sample size. A Bayes Factor of 5 favoring either the null or the alternative was used as the cut-off for data collection. Sub-samples of 32 individuals were run until this cut-off was met (see below for final BFs). Three-hundred and twenty MTurk workers were recruited for the online study and compensated \$1 USD for participating.

*Current Sample for Experiment 2b.* Ninety-six individuals were run in the effort and accuracy dimensions and 128 individuals were run for the time dimension. Seven percent of individuals failed the attention check embedded in the survey resulting in a final  $N$  of 279 ( $Median_{Age} = 32$  years,  $Min_{Age} = 19$  years,  $Max_{Age} = 70$  years, 48% male participants, and 54% reported completing a Bachelor's degree or higher).

## **Design**

A 3 (Comparison Condition: 1-word/110° vs. 3-words/0°, 1-word/110° vs. 4-words/0°, 1-word/110° vs. 5-words/0°) x 3 (Rating Dimension: Effort, Error-likelihood, Time) between-subjects design was used in Experiment 2a. Experiment 2b utilized a one-factor (Rating Dimension: Effort, Time, Accuracy) between-subject design, where individuals either made an effort-, error-, or time-based choice when only contrasting the 1-word/110° and 3-words/0° display conditions

## **Stimuli**

The stimuli closely followed Experiment 1. However, a larger word list was needed to complete full counterbalancing. For both experiments, words consisted of six

high frequency nouns: “LINE”, “TURN”, “SHOW”, “FEET”, “PAST”, and “HALF”,  
*Mean Written Word Frequency* = 276 per thousand words.

### **Procedure**

All portions of the procedure followed Experiments 1b and 1c.

### **Results**

Reporting of results follow above. Results are first presented for Experiment 2a (see Figure 2) followed by Experiment 2b. All of the following analyses utilized priors as described in Experiments 1b and 1c. All results in the following are expressed as the overall proportion of choices for the 1-word/110° display relative to the described alternative.

### **Experiment 2a**

*1-word/110° vs. 3-words/0°*. First, a BF demonstrated extreme evidence for the alternative that each column of the data had different frequencies,  $BF_{Alt} = 558.31$ ,  $\chi^2(2) = 19.94$ ,  $p < .001$ . For the effort dimension, individuals chose the 1-word/110° display as the more effortful option relative to the 3-words/0° display 66% of the time,  $BF_{Alt} = 23.51$ , MAP = 65%, 95% HDI [55%, 74%],  $p < .01$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as being associated with a higher error-likelihood relative to the 3-words/0° display 67% of the time,  $BF_{Alt} = 25.05$ , MAP = 66%, 95% HDI [56%, 76%],  $p < .01$  Binomial test. In contrast, for time the dimension individuals chose the 1-words/110° display as having a larger time requirement relative to the 3-words/0° display only 38% of the time (i.e., individuals more often chose the 3-words/0° display as more time demanding),  $BF_{Alt} = 3.18$ , MAP = 39%, 95% HDI [29%, 48%],  $p = .03$ . Results supported the null hypothesis when comparing the effort and error-



likelihood dimensions,  $BF_{Null} = 5.56$ ,  $\chi^2(1) < .1$ ,  $p > .1$ , suggesting that the choices for the 1-word/110° display were the same whether asking about effort or error-likelihood.

Comparisons of the effort and accuracy dimensions against the time dimension demonstrated at least very strong evidence for the alternative in both cases,  $BF_{Alt} = 278.94$ ,  $\chi^2(1) = 13.53$ ,  $p < .001$ ,  $BF_{Alt} = 309.56$ ,  $\chi^2(1) = 13.66$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time respectively, meaning that evaluations of time requirements differed markedly from evaluations of effort and error likelihood. In short, individuals similarly chose the 1-word/110° display as the more effortful and less accurate option, but the 3-words/0° display as the more time demanding option.

*1-word/110° vs. 4-words/0°*. A BF computed demonstrated strong evidence for the alternative that each column of the data had different frequencies,  $BF_{Alt} = 22.35$ ,  $\chi^2(2) = 13.02$ ,  $p < .001$ . For the effort dimension, individuals chose the 1-word/110° display as the more effortful option relative to the 4-words/0° display 50% of the time,  $BF_{Null} = 3.85$ ,  $MAP = 50\%$ , 95% HDI [40%, 60%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as having a higher error-likelihood relative to the 4-words/0° display 56% of the time,  $BF_{Alt} = .43$ ,  $MAP = 55\%$ , 95% HDI [45%, 66%],  $p > .1$  Binomial test. In contrast to effort and errors, individuals chose the 1-words/110° display as having a larger time requirement relative to the 4-words/0° display only 30% of the time (i.e., individuals more often chose the 4-words/0° display as more time demanding),  $BF_{Alt} = 296.17$ ,  $MAP = 31\%$ , 95% HDI [22%, 41%],  $p < .001$  Binomial test. Results moderately supported the null hypothesis when comparing the effort and error-likelihood,  $BF_{Null} = 4.00$ ,  $\chi^2(1) = .33$ ,  $p > .1$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated moderate and very strong

evidence for the alternative,  $BF_{Alt} = 8.90$ ,  $\chi^2(1) = 6.77$ ,  $p < .01$ ,  $BF_{Alt} = 64.69$ ,  $\chi^2(1) = 10.64$ ,  $p = .001$ , for effort vs. time and error-likelihood vs. time respectively. Thus, individuals similarly chose the 1-word/110° display at near chance levels for the effort and error-likelihood dimensions, but the 4-words/0° display as the more time demanding option. As with the 3-words/0° vs. 1-word/110° displays, ratings of effort and error-likelihood tracked one another, and differed markedly from ratings of time requirements.

*1-word/110° vs. 5-words/0°*. A BF demonstrated anecdotal evidence for the alternative that each column of the data has different frequencies,  $BF_{Alt} = 2.02$ ,  $\chi^2(2) = 8.33$ ,  $p = .02$ . For the effort dimension, individuals chose the 1-word/110° display as the more effortful option relative to the 5-words/0° display 34% of the time,  $BF_{Alt} = 11.59$ ,  $MAP = 36\%$ , 95% HDI [25%, 45%],  $p < .01$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as having a higher error-likelihood relative to the 5-words/0° display 36% of the time,  $BF_{Alt} = 4.91$ ,  $MAP = 36\%$ , 95% HDI [26%, 47%],  $p = .02$  Binomial test. For time, individuals chose the 1-words/110° display as having a larger time requirement relative to the 5-words/0° display 18% of the time,  $BF_{Alt} > 1,000,000$ , 95% HDI [13%, 29%],  $p < .001$  Binomial test. Results supported the null hypothesis when comparing the effort and error-likelihood dimensions,  $BF_{Null} = 4.00$ ,  $\chi^2(1) < .1$ ,  $p > .1$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated moderate evidence for the alternative in both cases,  $BF_{Alt} = 3.74$ ,  $\chi^2(1) = 5.40$ ,  $p = .02$ ,  $BF_{Alt} = 5.45$ ,  $\chi^2(1) = 6.07$ ,  $p = .01$ , for effort vs. time and error-likelihood vs. time respectively. Thus, individuals chose the 5-words/0° display as the more effortful, less accurate, and more time demanding option relative to the 1-

word/110° display. Though, effort and error-likelihood choices were similar with both dimensions producing more 1-word/110° choices relative to the time dimension.

### **Experiment 2b**

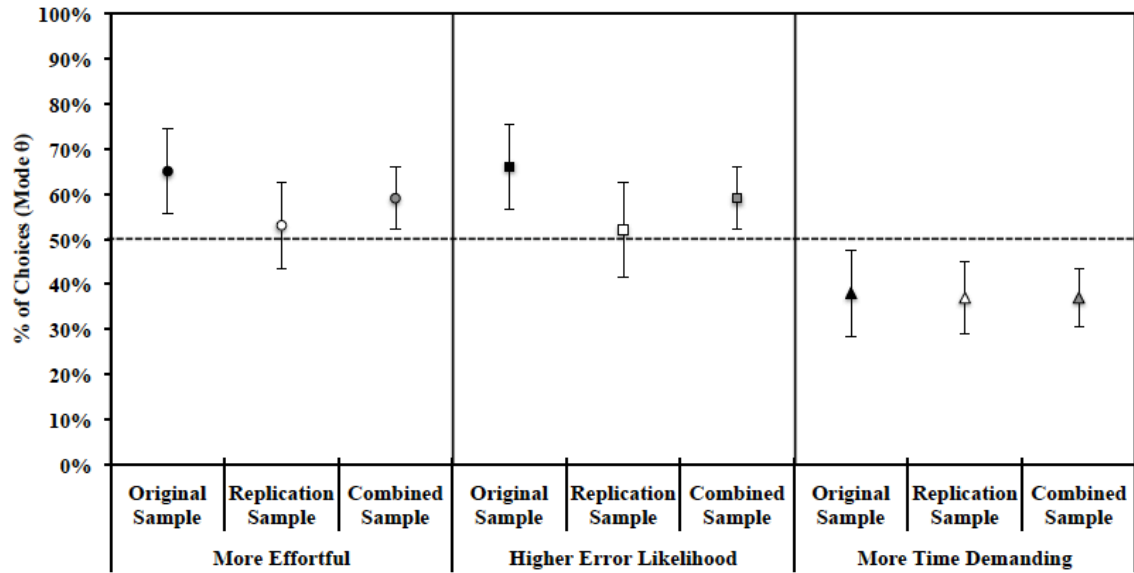
*1-word/110° vs. 3-words/0°*. A BF demonstrated only anecdotal evidence for the alternative that each column of the data had different frequencies,  $BF_{Alt} = 1.26$ ,  $\chi^2(2) = 7.63$ ,  $p = .02$ . For the effort dimension, individuals chose the 1-word/110° display as the more effortful option relative to the 3-words/0° display 53% of the time,  $BF_{Null} = 3.57$ ,  $MAP = 52\%$ , 95% HDI [43%, 63%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as having a higher error-likelihood relative to the 3-words/0° display 52% of the time,  $BF_{Null} = 3.57$ ,  $MAP = 52\%$ , 95% HDI [41%, 62%],  $p > .1$  Binomial test. For the time dimension, individuals chose the 1-word/110° display as having a larger time requirement relative to the 3-words/0° display 36% of the time (i.e., individuals more often chose the 3-words/0° display as more time demanding),  $BF_{Alt} = 20.84$ ,  $MAP = 37\%$ , 95% HDI [29%, 45%],  $p < .001$  Binomial test. Results demonstrated evidence for the null hypothesis when comparing the effort and error-likelihood dimensions,  $BF_{Null} = 5.26$ ,  $\chi^2(1) < .1$ ,  $p > .1$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated moderate and anecdotal evidence for the alternative,  $BF_{Alt} = 3.23$ ,  $\chi^2(1) = 5.28$ ,  $p = .02$ ,  $BF_{Alt} = 2.06$ ,  $\chi^2(1) = 4.34$ ,  $p = .04$ , for effort vs. time and error-likelihood vs. time respectively.

Given the replication sample used the exact methods and procedures relative to Experiment 2a, we tested for differences in choices for each comparison conditions across the initial and replication samples. Bayes Factors demonstrated only anecdotal evidence at best that choices differed across the original and replication samples for all

the rating dimensions,  $BF_{Alt} < 1.25$  for all,  $p > .07$  for all chi-square tests. Therefore, the initial sample and replication sample were combined for the following unregistered analyses to provide a clearer estimate of the choices for each rating dimension in the 1-word/110° vs. 3-words/0° condition ( $N = 567$ ).

For the effort dimension, individuals chose the 1-word/110° display as the more effortful option relative to the 3-words/0° display 59% of the time,  $BF_{Alt} = 4.32$ ,  $MAP = 59\%$ , 95% HDI [52%, 66%],  $p = .01$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as having a higher error-likelihood relative to the 3-words/0° display 59% of the time,  $BF_{Alt} = 3.60$ ,  $MAP = 59\%$ , 95% HDI [52%, 66%],  $p = .02$  Binomial test. For the time dimension, individuals chose the 1-word/110° display as having a larger time requirement relative to the 3-words/0° display 37% of the time (i.e., individuals more often chose the 3-words/0° display as more time demanding),  $BF_{Alt} = 238.66$ ,  $MAP = 37\%$ , 95% HDI [31%, 44%],  $p < .001$  Binomial test. Results demonstrated evidence for the null hypothesis when comparing the effort and error-likelihood dimensions,  $BF_{Null} = 7.69$ ,  $\chi^2(1) < .1$ ,  $p > .1$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated extreme evidence for the alternative in both cases,  $BF_{Alt} > 1,000$ ,  $\chi^2(1) = 19.39$ ,  $p < .001$ ,  $BF_{Alt} = 2185.93$ ,  $\chi^2(1) = 18.55$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time respectively. This combined analysis confirms the pattern that individuals similarly chose the 1-word/110° display as the more effortful and greater error-likelihood, but the 3-words/0° display as the more time demanding option (see Figure 3).

Figure 3. Individuals' More Effortful, Higher Error-Likelihood, and More Time Demanding Choices for the 1-word/110° display in Experiment 2, the Replication Sample, and Combined Samples.



*Note:* all presented data points are for choices of the 1-word/110° display. The mode percentage of choices is based on the posterior  $\theta$  distribution. Error bars represent 95% Highest Density Intervals (HDI).

**Discussion**

Additional associations between effort and error judgments that differed from time judgments were observed in Experiments 2a and 2b. The 1-word/110° versus 3-words/0° comparison produced the strongest dissociation between the dimensions. Individuals more often choose the 1-word/110° display as more effortful and error prone, but the 3-words/0° display as more time demanding. Furthermore, the replication sample provided strong evidence for choices for the time dimension coinciding with the original sample, though choices for the effort and error dimensions were somewhat lower relative to the original sample. Nonetheless, when taking the original and replication samples into account the full dissociation persisted.

Results from the 1-word/110° versus 4-words/0° comparison suggests that though participants overwhelmingly chose the 4-words/0° option as more time demanding, this difference in perceived time demand (and objective time demand based on Experiment 1a) did not translate to ratings of effort where individuals' choices were at chance. The objective difference in time requirements based on the performance estimates from Experiment 1a are very large (1340 ms;  $d = 3.28$ ), though this time cost did not appear to largely factor into anticipated effort judgments for half of the individuals choosing in the comparison condition.

All dimensions did eventually favor the larger set size as being more effortful, time demanding, and having a higher error-likelihood in the 1-word/110° versus 5-words/0° comparison condition. Nonetheless, time judgments favored the 5-words/0° option with effort and error judgments being closer to chance. Importantly, though, in all cases evidence favored effort and error-likelihood choices being similar with both being markedly different relative to time choices (i.e., effort = errors > time). Thus, Experiment 2 further demonstrates a clear association between and effort and error judgments that differed from time judgments.

### **Experiment 3**

The previous experiments have relied on manipulations of stimulus rotation and set size to generate a direct trade-off between time and errors. One could argue, however, that closely associated effort and error judgments are not being driven by the evaluation of errors as costly, but rather based on low perceptual fluency related to a rotated display. Indeed, manipulations of perceptual fluency have been shown to influence a wide range of judgments and performance (e.g., Dunn et al., 2016; Reber, Winkielman, & Schwarz,

1998; Winkielman, Schwarz, Fazendeiro, & Reber, 2003). To rule out perceptual fluency, then, options in Experiment 3 were changed from stimulus rotation and set size to different types of math problems. Specifically, we contrasted multiplication problems with addition problems in the same trade-off context as used above, and had individuals make judgments about effort, error-likelihood, and time requirements in a between-subject design.

Based on previous research (e.g., Ashcraft & Faust, 1994; Siegler & Lemaire, 1997, Walsh & Anderson, 2009) one  $N \times NN$  multiplication problem would be expected to generate more errors relative to six simple addition problems, whereas solving six addition problems would be expected to take longer relative to a single multiplication problem. Hence, again individuals were faced with a direct trade-off between time and errors. Based on Experiments 1 and 2, if judgments of effort and errors are closely associated, then we would expect individuals to choose the one multiplication problem as the more effortful alternative relative to six addition problems. In addition, choices for the error dimension should track with effort choices, whereas choices for the time dimension should favor the six addition problems.

## **Method**

### **Participants**

*Initial Sample Size Determination.* The sample sizes of  $n = 96$  for each rating dimension (i.e., effort, accuracy, and time) was carried over from the previous experiments.

*Current Sample.* Two-hundred and eighty-eight MTurk workers were recruited for the online study and compensated \$1 USD for participating. Seven percent of

individuals failed the attention check embedded in the survey resulting in a final  $N$  of 268 ( $Median_{Age} = 33$  years,  $Min_{Age} = 18$  years,  $Max_{Age} = 71$  years, 43% male participants, and 50% reported completing a Bachelor's degree or higher).

### **Design**

A one-factor (Rating Dimension: Effort, Time, Accuracy) between-subject design was employed, where all individuals made a choice between one  $N \times NN$  multiplication problem and six simple single digit addition problems.

### **Stimuli**

The general choice screen was similar to the previous experiments. To attempt to control for simple retrieval-based strategies for multiplication problems, all  $N$  digits ranged from five to nine and for the  $NN$  problems all digits ranged from 12 to 19. Seven unique problems were randomly generated:  $15 \times 9$ ,  $17 \times 6$ ,  $13 \times 7$ ,  $12 \times 7$ ,  $16 \times 8$ ,  $19 \times 5$ , and  $14 \times 8$ . For the simple addition problems, digits ranged from one to four. Seven unique problems were randomly generated:  $2 + 4$ ,  $4 + 3$ ,  $1 + 1$ ,  $1 + 2$ ,  $3 + 1$ ,  $2 + 3$ , and  $4 + 1$ . The order of the addition problems presented within the choice screen was counterbalanced along with the multiplication problems resulting in seven unique choice screens.

### **Procedure**

All portions of the procedure followed the previous experiments.

### **Results**

All of the following analyses utilized priors as described in the previous experiments. All results in the following are expressed as the overall proportion of



choices for the one multiplication problem relative to the six addition problems alternative.

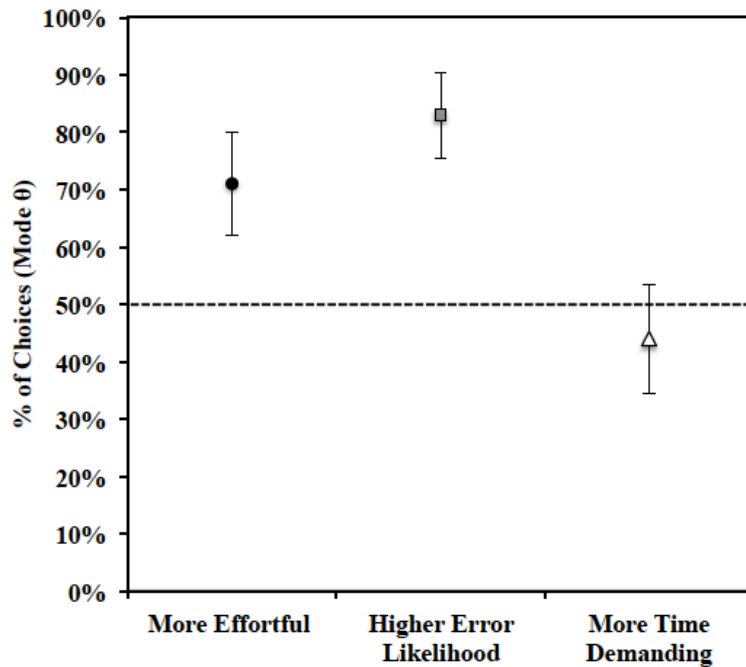
First, a BF demonstrated extreme evidence for the alternative that each column of the data has different probabilities,  $BF_{Alt} > 1,000,000$ ,  $\chi^2(2) = 36.45$ ,  $p < .001$ . For effort, individuals chose the one multiplication problem as the more effortful option relative to the six addition problems 72% of the time,  $BF_{Alt} = 1344.85$ ,  $MAP = 71\%$ , 95% HDI [61%, 79%],  $p < .001$  Binomial test. For error-likelihood, individuals chose the one multiplication problem as the less accurate option relative to the six addition problems 85% of the time,  $BF_{Alt} > 1,000,000$ ,  $MAP = .83$ , 95% HDI [75%, 90%],  $p < .001$  Binomial test. Last, for time, individuals chose the one multiplication as the more time demanding option relative to the six addition problems only 44% of the time,  $BF_{Alt} = .48$ ,  $MAP = 44\%$ , 95% HDI [35%, 54%],  $p > .1$  Binomial test. Results reveal only anecdotal evidence for the alternative hypothesis when comparing the effort and error-likelihood dimensions,  $BF_{Alt} = 1.51$ ,  $\chi^2(1) = 3.89$ ,  $p = .05$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated extreme evidence for the alternative in both cases,  $BF_{Alt} = 310.66$ ,  $\chi^2(1) = 13.68$ ,  $p < .001$ ,  $BF_{Alt} > 1,000,000$ ,  $\chi^2(1) = 31.84$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time respectively (see Figure 4).

## Discussion

Consistent with the previous experiments, the option associated with a higher likelihood of an error generated more effortful and higher error likelihood evaluations that were both similarly above chance, with choices for time requirements differing markedly from these and being slightly below chance. Therefore, the association between

effort and error judgments is not driven merely by the low fluency associated with the stimulus rotation manipulation in Experiments 1 and 2. Rather, the association between judgments of effort and errors that differs from judgments of time generalizes to an additional trade-off context using math problems.

Figure 4. Individuals' More Effortful, Less Accurate, and More Time Demanding Choices for the One Multiplication Problem in Experiment 3.



*Note:* All presented data points are for choices of the one multiplication problem. The mode  $\theta$  values are based on the posterior distribution. Error bars represent 95% Highest Density Intervals (HDI).

#### Experiment 4

Experiments 1 through 3 demonstrated a clear association between individuals' effort and error judgments that differed from time judgments across two different error/time trade-off contexts (i.e., stimulus rotation vs. set size; a multiplication problem vs. addition problems). To this point we have used a between-subject design across these

experiments to provide strong test of heuristic reasoning in judgment. In Experiment 4 we move to a within-subject design to examine whether this association persists when there is explicit information (i.e., self-knowledge) about making judgments for the effort, error-likelihood, and time dimensions. Here, participants provided judgments for all three dimensions for one of the comparison conditions again using stimulus rotation and set size. Furthermore, the within-subject design provides the opportunity to directly predict individuals' effort judgments using a logistic regression approach. In addition, given the potential issues of carry-over effects with within-subject designs, we can use the first judgments that individuals complete as a type of replication benchmark with larger sample sizes against the within-subject results, as well as the original between-subject results presented in Experiments 1 and 2 (see Figure 6 and c.f. Figures 2 and 3).

Following from Experiments 2a and 2b, we would expect effort and error-likelihood judgments to be closely associated across all of the comparison conditions, and these judgments should differ from time judgments. Specifically, for the 1-word/110° versus 2-words/0° comparison we hypothesize that individuals will choose the 1-word/110° display as the more effortful and higher error-likelihood option, but the time requirement choices being at chance. For the 1-word/110° versus 3-words/0° comparison we hypothesize that individuals will choose the 1-word/110° display as the more effortful and higher error-likelihood option, but the 3-words/0° display as the option associated with a larger time requirement. For the 1-word/110° versus 4-words/0° comparison we hypothesize that individuals will choose the two options at chance for the effort and error-likelihood dimensions, but the 4-words/0° display as the option associated with a larger time requirement. These patterns are hypothesized for both the within-subject and

between-subject analyses. Furthermore, we hypothesize that error-likelihood judgments relative to time judgments will be a better predictor of effort judgments.

## Method

### Participants

*Initial Sample Size Determination.* Sample sizes were increased to  $n = 420$  for each dimension (i.e., effort, error-likelihood, and time) to ensure full counterbalancing of the order that each dimension was presented, as well as having sufficient power to adequately estimate the odds ratios associated with error and time choices predicting effort choices.

*Current Sample.* Twelve-hundred and sixty total MTurk workers were recruited and compensated \$1 USD for participating. This is an increase in pay relative to the previous experiments given the increased time requirement needed to complete choices for all the rating dimensions and filler questionnaires (see below for details). Individuals were randomly assigned to one of the comparison condition dimensions

For the 1-word/110° versus 2-words/0° comparison, four percent of individuals failed the attention check embedded in the survey resulting in a final  $n$  of 402 ( $Median_{Age} = 33$  years,  $Min_{Age} = 19$  years,  $Max_{Age} = 77$  years, 44% male participants, and 54% reported completing a Bachelor's degree or higher). For the 1-word/110° versus 3-words/0° comparison, five percent of individuals failed the attention check embedded in the survey resulting in a final  $n$  of 400 ( $Median_{Age} = 32$  years,  $Min_{Age} = 19$  years,  $Max_{Age} = 70$  years, 49% male participants, and 51% reported completing a Bachelor's degree or higher). Last, for the 1-word/110° versus 4-words/0° comparison, five percent of individuals failed the attention check embedded in the survey resulting in a final  $n$  of 400

( $Median_{Age} = 33$  years,  $Min_{Age} = 18$  years,  $Max_{Age} = 79$  years, 52% male participants, and 51% reported completing a Bachelor's degree or higher). We did not include the 1-word/110° versus the 5-words/0° comparison given all judgments favored the 5-words/0° option, therefore we would not expect error and time judgments to differentially predict effort judgments.

### **Design**

A 3 (Rating Dimension: Effort, Time, Accuracy) x 3 (Comparison Condition: 1-word/110° vs. 2-words/0°, 1-word/110° vs. 3-words/0°, 1-word/110° vs. 4-words/0°) mixed design was employed. Rating dimension was manipulated within-subjects, where individuals made three judgments for one of the comparison conditions.

### **Stimuli**

All stimuli for the comparison conditions followed Experiments 2a and 2b.

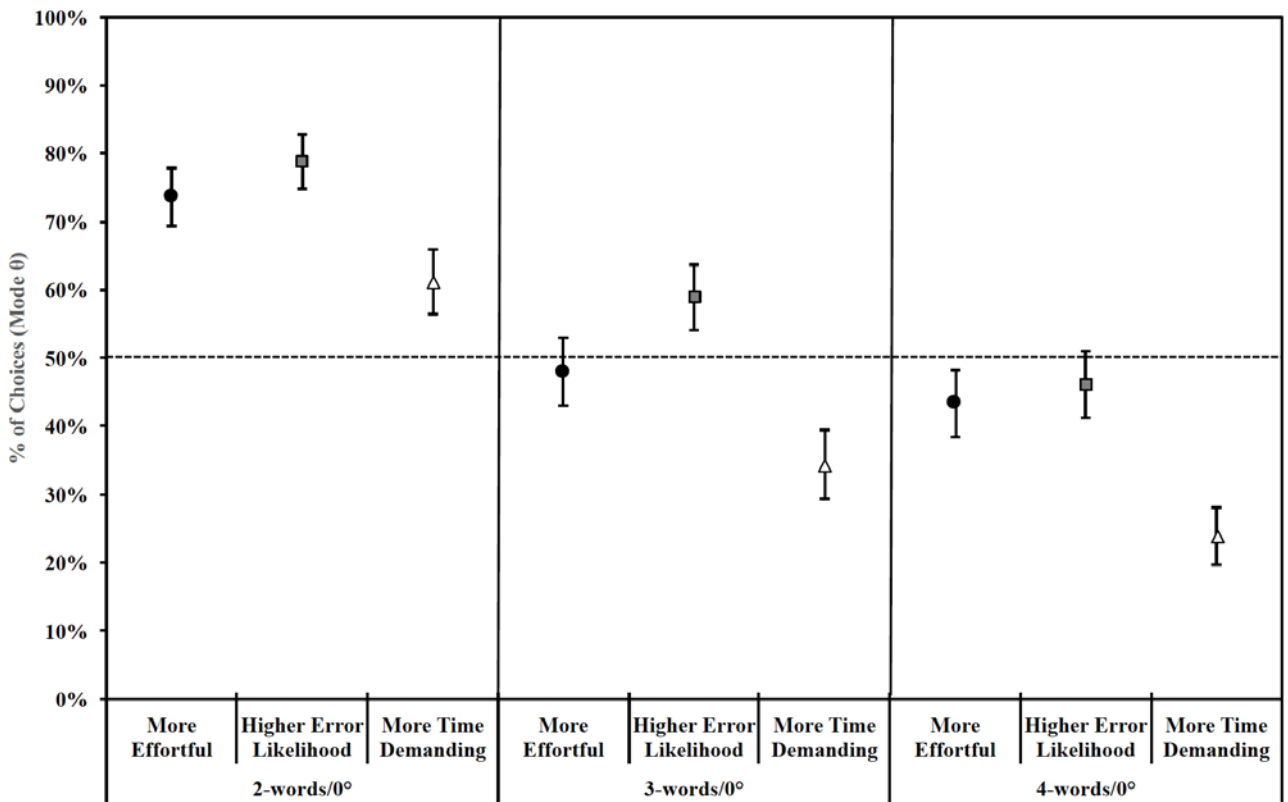
### **Procedure**

All parts of the procedure followed the previous experiments with the exception of individuals making all three dimension judgments within an assigned comparison condition. To attempt to prevent carryover over of judgments across the three dimensions, individuals completed the Metacognitive Awareness Inventory (52 items, Schraw & Dennison, 1994) and the (44 items, John & Srivastava, 1999) after their first and second judgments (i.e., 1<sup>st</sup> judgment – distractor questionnaire – 2<sup>nd</sup> judgment – distractor questionnaire – 3<sup>rd</sup> judgment). The order of the dimensions and distractor questionnaires was counterbalanced across participants.

**Results**

All of the following followed the same methods as described in the previous experiments. All results in the following are expressed as the overall proportion of choices for the 1-word/110° display relative to the described alternative.

Figure 5. More Effortful, High Error-likelihood, and More Time Demanding Choices for the 1-word/110° Display for the Within-Subject Analyses.



*Note:* All presented data points are for choices of the 1-word/110° display; the alternative choice relative to the 1-word/110° display is plotted on the X-axis. That is, choices below chance (50%) would reflect a tendency to more often choose the alternative choice denoted on the X-axis. The mode  $\theta$  values are based on the posterior distribution. Error bars represent 95% Highest Density Intervals (HDI).

**Within-subject Analyses**

*1-word/110° vs. 2-words/0°*. Within-subject choices across the three dimensions were first submitted to a Cochran's  $Q$  test. Three individuals did not offer choices for all three dimensions and are thus excluded in the following analyses. Results demonstrated that the difference in probabilities across the three dimensions was different than zero,  $Q(2) = 44.09, p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 2-words/0° display 74% of the time,  $BF_{Alt} > 1,000,000$ ,  $MAP = 74\%$ , 95% HDI [69%, 78%],  $p < .001$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as the less accurate option relative to the 2-words/0° display 79% of the time,  $BF_{Alt} > 1,000,000$ ,  $MAP = 79\%$ , 95% HDI [75%, 83%],  $p < .001$  Binomial test. For time, individuals chose the 1-word/110° display as the more time demanding option relative to the 2-words/0° display 61% of the time,  $BF_{Alt} > 1,000$ ,  $MAP = 61\%$ , 95% HDI [56%, 66%],  $p < .001$  Binomial test. Specific comparisons across the dimensions using Wilcoxon sign tests revealed a higher magnitude of 1-word/110° choices for the effort and error likelihood dimensions relative to the time dimension,  $p < .001$  for both, and a small though significant difference between the effort and error dimensions,  $p = .042$ . Therefore, effort and error likelihood 1-word/110° choices closely followed with both being greater than the time dimension. Though in contrast with Experiment 1, the choices for the time dimension were greater than chance (see Figure 5).

*1-word/110° vs. 3-words/0°*. Four individuals did not offer choices for all three dimensions. Results demonstrated that the difference in probabilities across the three dimensions was different than zero,  $Q(2) = 66.33, p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 3-words/0° display 48% of

the time,  $BF_{Null} = 6.05$ ,  $MAP = 48\%$ , 95% HDI [43%, 53%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as the less accurate option relative to the 3-words/0° display 59% of the time,  $BF_{Alt} = 68.43$ ,  $MAP = 59\%$ , 95% HDI [54%, 64%],  $p < .001$  Binomial test. For time, individuals chose the 1-words/110° display as the more time demanding option relative to the 3-words/0° display 34% of the time (i.e., individuals more often chose the 3-words/0° display as more time demanding),  $BF_{Alt} > 1,000,000$ ,  $MAP = 34\%$ , 95% HDI [29%, 39%],  $p < .001$  Binomial test. Comparisons across the dimensions revealed a higher magnitude of 1-word/110° choices for the error likelihood dimension relative to the effort and time dimensions,  $p < .001$  for both. In addition, there was a higher magnitude of 1-word/110° choices for the effort dimension relative to the time dimensions,  $p < .001$ . Thus, when contrasting to the Experiment 2 between-subject choices, error likelihood and time choices closely followed the patterns in Experiment 2. Effort choices matched the pattern in the Experiment 2 replication where choices were near chance (see Figure 5).

*1-word/110° vs. 4-words/0°.* Three individuals did not offer choices for all three dimensions. Results demonstrated that the difference in probabilities across the three dimensions was different than zero,  $Q(2) = 66.01$ ,  $p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 4-words/0° display 43% of the time,  $BF_{Alt} = 4.51$ ,  $MAP = 43\%$ , 95% HDI [38%, 48%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as the less accurate option relative to the 4-words/0° display 46% of the time,  $BF_{Null} = 2.34$ ,  $MAP = 46\%$ , 95% HDI [41%, 51%],  $p > .1$  Binomial test. For time, individuals chose the 1-words/110° display as the more time demanding option relative to the 4-words/0° display 24% of the time



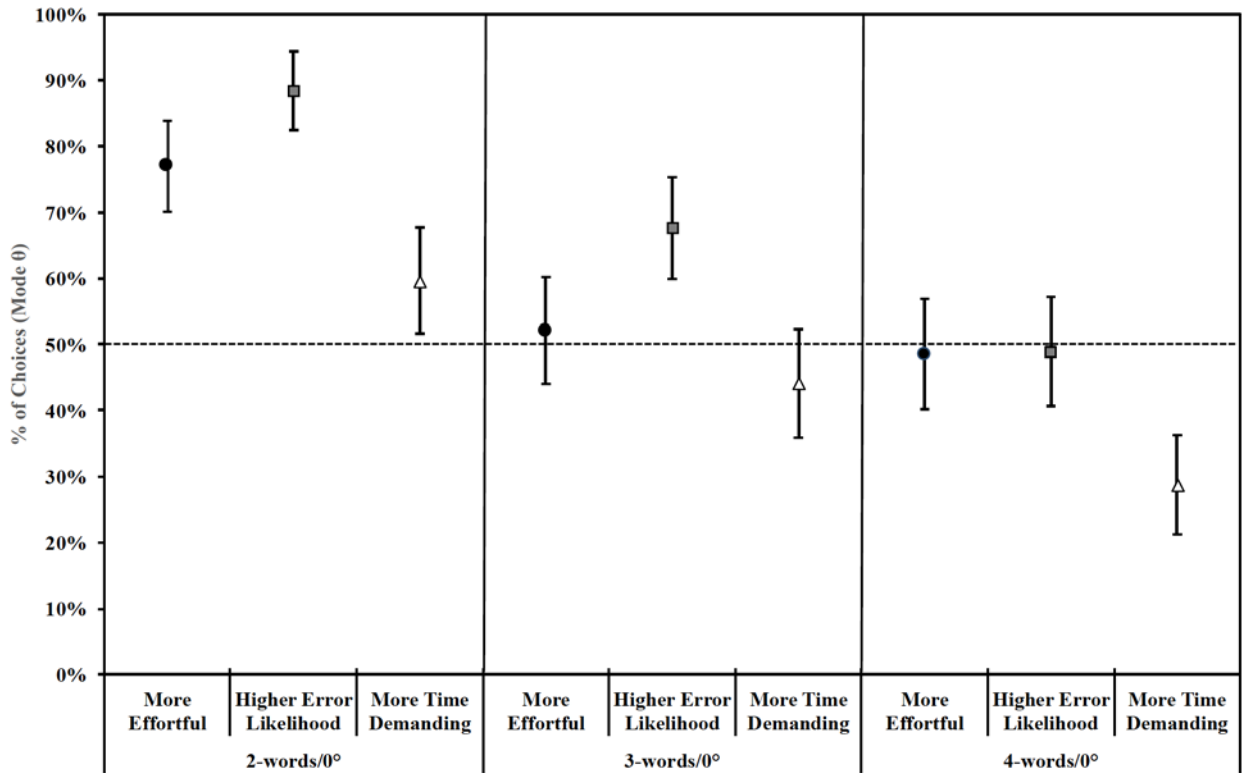
(i.e., individuals more often chose the 4-words/0° display as more time demanding),  $BF_{Alt} > 1,000,000$ , MAP = 34%, 95% HDI [29%, 39%],  $p < .001$  Binomial test. Comparisons across the dimensions revealed a similar magnitude of 1-word/110° choices for the error likelihood and effort dimensions,  $p > .1$ , with both demonstrating a larger magnitude of 1-word/110° choices relative to the time dimension,  $p < .001$ . Thus, when contrasting to the Experiment 2 between-subject choices, error likelihood and time choices closely followed the patterns in Experiment 2, though effort choices were below chance in this instance (see Figure 5).

*Logistic regressions.* In the following analyses, individuals' error likelihood and time judgments were permitted to interact allowing for each individual's pattern of choices across the dimensions to be modeled. All following estimates are derived from these interaction models. The interaction between the error-likelihood and time judgments were not significant in any of the three comparison condition models,  $p > .1$  for all models. Removing the interaction term from the following models does not change the overall results reported below.

For the 1-word/110° vs. 2-words/0° comparison condition, error likelihood was a significant predictor of effort ratings,  $b = 1.44$ ,  $SE = .37$ ,  $Z = 3.84$ ,  $p < .001$ , OR = 4.21, 95% CI [2.05, 8.49]. Time judgments were also a significant predictor of effort judgments to a similar extent as error-likelihood judgments,  $b = 1.32$ ,  $SE = .47$ ,  $Z = 2.84$ ,  $p = .004$ , OR = 3.77, 95% CI [1.54, 9.68]. Similar to the above comparison condition, both error-likelihood and time judgments predicted effort judgments in the 1-word/110° vs. 3-words/0° comparison condition,  $b = .77$ ,  $SE = .26$ ,  $Z = 2.94$ ,  $p = .003$ , OR = 2.15, 95% CI [1.3, 3.61],  $b = .96$ ,  $SE = .4$ ,  $Z = 2.53$ ,  $p = .011$ , OR = 2.61, 95% CI [1.24, 5.55],

for the error-likelihood and time dimensions respectively. Last, for the 1-word/110° vs. 4-words/0° comparison condition, again both error-likelihood and time judgments predicted effort judgments,  $b = .9$ ,  $SE = .25$ ,  $Z = 3.56$ ,  $p < .001$ ,  $OR = 2.47$ , 95% CI [1.5, 4.07],  $b = 1.78$ ,  $SE = .42$ ,  $Z = 4.21$ ,  $p < .001$ ,  $OR = 5.96$ , 95% CI [2.65, 14.17], for the error-likelihood and time dimensions respectively. Although this is the biggest difference of the estimates observed between error-likelihood and time judgments as predictors, a test of the difference in ORs was not significant,  $Z = 1.78$ ,  $p = .074$ . Thus, in all comparison conditions both error-likelihood and time judgments were significant predictors of effort judgments.

Figure 6. More Effortful, High Error-likelihood, and More Time Demanding Choices for the 1-word/110° Display for the Between-Subject Analyses.



Note: All presented data points are for choices of the 1-word/110° display; the alternative choice relative to the 1-word/110° display is plotted on the X-axis. That is, choices below chance (50%) would reflect a tendency to more often choose the alternative choice denoted on the X-axis. The mode  $\theta$  values are based on the posterior distribution. Error bars represent 95% Highest Density Intervals (HDI).

**Between Subject Analyses**

In the following we report the results from the between-subject analyses where only the first judgments that individuals completed were used.

*1-word/110° vs. 2-words/0°.* A BF computed for the 2 x 3 data demonstrated extreme evidence for the alternative that each column of the data had different probabilities,  $BF_{Alt} > 1,000$ ,  $\chi^2(2) = 25.38$ ,  $p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 2-words/0° display 80% of the time,

$BF_{Alt} > 1,000,000$ ,  $MAP = 77\%$ , 95% HDI [70%, 84%],  $p < .001$  Binomial test. This result held for the error dimension where the 1-word/110° was chosen as the less accurate option relative to the 2-words/0° display 86% of the time,  $BF_{Alt} > 1,000,000$ ,  $MAP = 84\%$ , 95% HDI [78%, 90%],  $p < .001$  Binomial test. For time, individuals chose the 1-words/110° display as the more time demanding option relative to the 2-words/0° display 60% of the time,  $BF_{Alt} = 2.92$ ,  $MAP = 60\%$ , 95% HDI [29%, 68%],  $p = .02$ . Results more favored the null hypothesis when specifically comparing the effort and error-likelihood dimensions,  $BF_{Null} = 2.83$ ,  $\chi^2(1) < .1$ ,  $p > .1$ . Thus, coinciding with Experiment 1, choices for the 1-word/110° display were similar for effort and error-likelihood. Comparisons of the effort and accuracy dimensions against the time dimension demonstrated at least very strong evidence for the alternative in both cases,  $BF_{Alt} = 36.37$ ,  $\chi^2(1) = 10.23$ ,  $p = .001$ ,  $BF_{Alt} > 1,000$ ,  $\chi^2(1) = 21.36$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time respectively, again mirroring the pattern of results from Experiment 1 (see Figure 6).

*1-word/110° vs. 3-words/0°*. Results demonstrated very strong evidence for the alternative that each column of the data had different probabilities,  $BF_{Alt} = 70.71$ ,  $\chi^2(2) = 16.21$ ,  $p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 3-words/0° display 52% of the time,  $BF_{Null} = 4.21$ ,  $MAP = 52\%$ , 95% HDI [44%, 60%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as the less accurate option relative to the 3-words/0° display 68% of the time,  $BF_{Alt} > 1,000$ ,  $MAP = 68\%$ , 95% HDI [60%, 75%],  $p < .001$  Binomial test. For time, individuals chose the 1-words/110° display as the more time demanding option relative to the 3-words/0° display 44% of the time,  $BF_{Null} = 1.83$ ,  $MAP = 44\%$ , 95% HDI [36%, 53%],  $p > .1$ . In contrast to Experiment 2, results supported the alternative hypothesis

when comparing the effort and error-likelihood dimensions where choices were higher for the 1-words/110° display for the error-likelihood dimension,  $BF_{Alt} = 5.89$ ,  $\chi^2(1) = 6.71$ ,  $p = .01$ . Again, contrasting with Experiment 2, comparison of the effort dimension against the time dimension demonstrated evidence favoring the null where choices for the 1-words/110° display were similar across the dimensions,  $BF_{Null} = 2.81$ ,  $\chi^2(1) = 1.39$ ,  $p > .1$ . Last, comparison of the error-likelihood dimension against the time dimension demonstrated extreme evidence for the alternative,  $BF_{Alt} = 387.5$ ,  $\chi^2(1) = 14.68$ ,  $p < .001$ . Thus, individuals similarly chose the 1-word/110° display as the more effortful and more time demanding option near chance, whereas error-likelihood choices for the 1-word/110° were much greater than these later two options (see Figure 6).

*1-word/110° vs. 4-words/0°*. Strong evidence for the alternative that each column of the data has different probabilities,  $BF_{Alt} = 58.58$ ,  $\chi^2(2) = 15.65$ ,  $p < .001$ . Individuals chose the 1-word/110° display as the more effortful option relative to the 4-words/0° display 48% of the time,  $BF_{Null} = 4.37$ , MAP = 49%, 95% HDI [40%, 57%],  $p > .1$  Binomial test. For error-likelihood, individuals chose the 1-word/110° display as the less accurate option relative to the 4-words/0° display 49% of the time,  $BF_{Null} = 4.54$ , MAP = 49%, 95% HDI [40%, 57%],  $p > .1$  Binomial test. For time, individuals chose the 1-word/110° display as the more time demanding option relative to the 4-words/0° display 28% of the time (i.e., individuals more often chose the 4-words/0° display as more time demanding),  $BF_{Alt} > 1,000$ , MAP = 29%, 95% HDI [22%, 37%],  $p < .001$  Binomial test. Results demonstrated positive evidence for the null hypothesis when comparing the effort and error-likelihood,  $BF_{Null} = 6.46$ ,  $\chi^2(1) < .1$ ,  $p > .1$ . Comparisons of the effort and error-likelihood dimensions against the time dimension demonstrated very strong evidence for

the alternative,  $BF_{Alt} = 54.49$ ,  $\chi^2(1) = 10.92$ ,  $p < .001$ ,  $BF_{Alt} = 73.69$ ,  $\chi^2(1) = 11.49$ ,  $p < .001$ , for effort vs. time and error-likelihood vs. time respectively. These results nicely replicate the pattern of results reported in Experiment 2, where individuals chose the 1-word/110° display as the more effortful and higher error likelihood option at chance, whereas the 1-word/110° display was chosen well below chance for the time dimension (see Figure 6).

### **Discussion**

Overall, patterns of within-subject choices fell within expected values observed in Experiments 2a and 2b, though there were some variations. For the 1-word/110° vs. 2-words/0° condition time judgments more favored the 1-word/110° display contrasting with Experiment 1. Again, this is interesting given the performance estimates derived in Experiment 1a that demonstrated a medium-sized effect in RTs where the 2-words/0° condition took longer to read aloud. Nonetheless, we still observed a similarity in choices between effort and error-likelihood that differed from time judgments. In the 1-word/110° vs. 3-words/0° condition, choices for the effort dimension were at chance, though error-likelihood judgements favored the 1-word/110° and time judgments favored the 3-words/0° condition replicating Experiments 2a and 2b. In contrast, however, all three dimensions in this case were statistically different. The between-subject judgments (taken from the first dimension an individual made a choice) demonstrated these same inconsistencies in the 1-word/110° vs. 2-words/0° and somewhat in the 1-word/110° vs. 3-words/0° conditions relative to Experiments 1 and 2, though effort and error-likelihood judgments did not statistically differ in the latter case.

The conclusions that can be drawn from the logistic regression analyses are less clear. Individuals' error-likelihood and time judgements predicted their effort choices to a similar extent in each of the comparison conditions. This can be interpreted in two potential ways. First, individuals may indeed take both dimensions into account in the present experimental context when generating how effortful they feel a task may be in both the between- and within-subject designs. For example, when contrasting 1-word/110° vs. 2-words/0°, individuals may be evaluating errors in a way that outweighs or dominates time when judging effort. This is not to say that time *does not* play any role in the evaluation, rather that error-likelihood may be the more salient dimension when the evaluation takes place. Alternatively, demonstrating that both error-likelihood and time similarly predict effort may simply be an artifact of using a within-subject design. As noted in the introduction, within-subject designs provide information to individuals that is potentially not available to those individuals in the between-subject design. Though we attempted to minimize carry-over effects by distractor questionnaires between choices, individuals presumably had information about what their choices were in all three dimensions. Therefore, they may have utilized this information to remain consistent in their choices across the dimensions (Kahneman & Tversky, 1996) leading to both error-likelihood and time being significant predictors of effort. Nonetheless, some evidence does exist demonstrating striking consistencies in effort ratings across within- and between-subject designs (Dunn, Koehler, & Risko, 2017).

### **General Discussion**

Here, we investigated the influence of two potential perceived cognitive costs often associated with effortfulness: error-likelihood and time requirements. Experiments

1b and 1c provided evidence of differences between effort, error-likelihood, and time. The option associated with the higher error-likelihood generated higher effort and error choices while time choices were equivalent across the options. Experiment 2 further demonstrated clear dissociations between effort, errors, and time choices, with effort and error choices closely tracking one another. This was the case even in light of the objective performance estimates derived in Experiment 1a that would not predict the specific pattern of results across the two experiments. To generalize the findings from Experiments 1 and 2, Experiment 3 utilized different task conditions in the same trade-off context. Again, individuals chose the option associated with a higher likelihood of an error as more effortful, with effort and error judgments tracking closely, but not time choices. Last Experiment 4 looked to test the strength of the effort and error-likelihood association by turning to a within-subject design. Overall, patterns of choices were similar to Experiments 1 and 2. Nonetheless, both error-likelihood and time judgments predicted effort judgments to a similar extent across all comparison conditions. In the following we discuss why error-likelihood and time may differentially affect individuals' evaluations of anticipated effort, and suggest avenues for future research.

#### *Associates of Anticipated Cognitive Effort*

Why are some cognitive tasks evaluated as being effortful? Cognitive effort, as a psychological construct, is defined in various ways such as a mediating process (Shenhav et al., 2017) or an inferential metacognitive evaluation (Dunn et al., 2016). We have shown here that in several contexts, lines of action associated with a higher anticipated likelihood of an error tracks closely with judgments of anticipated effort. Utilizing error-likelihood as a type of cue when generating judgments of effort can provide a low-cost



approximation of the potential cognitive demand associated with a line of action. As reviewed above, several accounts suggest that error commission signals the need to engage demanding control over behavior; for example, attempting to correct a deviation from intended behavior to be in line with expected rewards, avoiding cross-talk situations that quickly lead to capacity limits, or configuring behavior in ways that avoid danger and threats to safety. Under all of these accounts, errors can be considered aversive as they signal the potential for engaging demanding control processes that are intimately linked to increased cognitive work across the executive control network (for reviews see Botvinick & Braver, 2015; Inzlicht et al., 2015; Shenhav et al., 2017). As such, evaluating lines of actions that are associated with an increased error-likelihood as effortful would be expected to be adaptive to the organism. Therefore, to the extent that effort is to be avoided during action selection (e.g., it is not offset by some reward; Botvinick & Braver, 2015), examination of cues can be thus limited to only those coinciding with perceived anticipated errors.

From a metacognitive standpoint (Dunn et al., 2016; Dunn & Risko, submitted), then, individuals may potentially utilize and weight the likelihood of an error over salient available cues associated with an action, such as stimulus rotation in Experiments 1, 2, and 4 or math problem type in Experiment 3 to generate their judgment of anticipated effort. This can serve as a type of inferential heuristic in guiding initial action selection by generating satisfactory solutions for action selection while costing only modest amounts of cognitive work (Gigerenzer, 2008; Gigerenzer, Todd, & ABC Research Group, 1999; Gigerenzer & Goldstein, 1996; Simon, 1982; 1990). Following Shah and Oppenheimer's (2008) framework, this type of inferential metacognitive evaluation of

cognitive effort can predictably reduce cognitive work by (1) simplifying the weighting principle for cues, (2) allowing for the examination of fewer cues, (3) reducing the work associating with storing and retrieving specific values, (4) requiring less information to be integrated, and (5) potentially leading to examining fewer alternatives, as we detail below.

While simplifying weighting principles and cue examination, utilizing error-likelihood to determine anticipated effort can also circumvent the issue of potential increased cognitive demands associated with using more complex processes during evaluation. Recently, competing cost/benefit accounts of how control should be deployed while situated within a task have been proposed (Gershman, Horvitz, Tenenbaum, 2015; Griffiths, et al., 2015). For example, the Expected Value of Control account (EVC; Shenhav, Botvinick, & Cohen, 2013; Shenhav, Cohen, & Botvinick, 2016) proposes that the allocation of control processes is driven by the computation of the expected gains and costs associated with the intensity of a given configuration of control signals and is contingent on continuous monitoring of present state cost information (e.g., conflict, errors, time delay, or negative feedback) through the dorsal anterior cingulate cortex (dACC). A comparison between this mechanism and the utilization of error-likelihood heuristically provides an important contrast with regards to perceived effort. For example, extending the error-likelihood idea to *experienced* effort, it would not be expected to be dependent on demanding online monitoring of information, and would thus be expected to require less information relative to more complex alternatives.

In contrast to error-likelihood, time demands were not as closely associated with effort judgments across the current experiments. This claim dovetails with recent works

that have demonstrated dissociations between effort-based decisions and the time costs associated with a task (Dixon & Christoff, 2012; Dunn et al., 2016; Kool et al., 2010; Westbrook et al., 2013), but diverges from models that suggest that processes that require more time will be perceived as more effortful (e.g., Gray et al., 2006; Inzlicht et al., 2014). It is important to note, however, that we do not make the claim that time costs play *no role* in judging effort. Rather, error-likelihood appears to show a closer association to anticipated cognitive effort. One potential explanation of this divergence is that committing an error arguably generates a more immediate call for demanding control to the system relative to increased time requirements. Classically, post-error slowing in speeded tasks (Rabbitt, 1966) has been conceptualized as a compensatory process tuned to improve performance on subsequent trials (Gehring & Fencsik, 2001), and this slowing has been shown to correlate positively with the likelihood of success on a following trial (i.e., minimizing errors; Hajcak, McDonald, & Simons, 2003). Hence, the system takes on a time cost to account for an error and minimizes the future likelihood of more errors.

When considering time as related to opportunity costs, these signals are hypothesized to tally and accrue over periods of time producing aversive signals used by the system to exert control and move behaviors to a more rewarding alternative (Inzlicht et al., 2014; Westbrook & Braver, 2016). Thus, one could reasonably conceive of increased opportunity costs requiring longer time scales to signal a need for control, relative to the more immediate timescale associated with errors as discussed above. To nicely demonstrate the relation between errors and time at a relatively longer timescale, recent work by Blain and colleagues (2016) demonstrated that aversive signals related to cognitive fatigue only affected individuals' propensity to engage in impulsive choices

after very long periods of time of engaging in a demanding task. A significant increase of impulse related choices was only observed after four-and-a-half hours where accuracy in the tasks remained relatively constant across the six-hour session. A time-based opportunity cost perspective would suggest that the opportunity cost of engaging in the control demanding task took hours before the aversive signal (i.e., fatigue) was great enough to divide capacity across different processes then potentially leading to more impulsive choices (i.e., divided capacity led to decreased ability to engage more analytical processing during choice; Evans & Stanovich, 2013). Thus, fatigue required a relatively long amount of time to presumably become aversive enough to initiate control, where resources were split to other processes leading to more impulsive choices.

A comparable conclusion can be drawn currently from the observation that *More Effortful* choices only began to favor the option associated with the higher objective time cost when this cost was very large across the options (i.e., in the 1-word/110° vs. 5-words/0° comparison,  $d = 3.60$  from Experiment 1a; c.f. Experiment 4 within-subject analysis,  $BF_{Alt} = 4.51$  favoring the 5-words/0° condition). Nonetheless, judgments of error-likelihood also similarly followed. Furthermore, we demonstrated in Experiment 4 that both error-likelihood and time significantly predicted individuals' effort judgments across three comparison conditions in a within-subject design (c.f., Experiment 4 Discussion). Therefore, it is again important to note that time and errors are likely not mutually exclusive determinants of cognitive effort. Indeed, these two are intrinsically connected (e.g., taking too long in a task can be considered an error and errors sometimes lead to slowing down on tasks).

One potentially interesting avenue to pursue with regard to error-likelihood and time demands working conjunctively to produce evaluations of effort comes from the EVC theory of control allocation (Shenhav et al., 2013; Shenhav et al., 2016). Within the EVC theory, the intensity of a control signal reflects the level of activation of the task units required and how the signal impacts information processes. Higher required intensity may lead to increased performance costs such as errors. In relation to the data presented here, then, individuals may estimate the intensity of the control signal(s) needed for a task and how long that level of intensity would be required when generating a judgment of anticipated effort. This account can be used to explain the results of Experiment 2a and 4, where we increased time requirements by increasing set size. When set size increased, effort judgments began to favor the option associated with larger time requirements. Therefore, high-intensity over a shorter term was anticipated to be more effortful (i.e., the rotated word condition) which may have been better captured by the error-likelihood framing, until time requirements became large and a relatively lower-intensity over a longer term became more effortful. Approaching cognitive effort through the lens of the EVC theory may prove fruitful in disentangling the basic constituents of anticipated cognitive effort both experimentally and through formal modelling methods.

#### *Limitations and Future Avenues for the Study of Cognitive Effort*

Though we directly manipulated specific trade-off contexts between errors and time, the current set of studies demonstrates only a stronger association between effort and error-likelihood that differed from time judgments in several contexts. Given the methods used we cannot claim a causal role of error-likelihood or time requirements playing a direct causal role in generating an evaluation of anticipated effort. Individuals

may have utilized something beyond these dimensions to determine which option they believed would be more effortful. Though this point highlights a critical issue to consider in studying cognitive effort. Indeed, there are numerous aversive elements of tasks that are potentially confounded with cognitive effort in experimental settings, for example: task difficulty (Vassena et al., 2014), conflict (Botvinick, 2007), perceived risk (Brown & Braver, 2007), the volatility of outcomes (Behrens et al., 2007), uncertainty (Shenhav et al., 2017), surprise (O'Reilly et al., 2013), boredom (Danckert & Allman, 2005; Milyavskaya, Inzlicht, Johnson, & Larson, 2017), negative affect (Inzlicht et al., 2015), fatigue (Hockey, 2011), and disfluency (Dreisbach & Fischer, 2011; also see Westbrook & Braver, 2015 for a brief review on the relation between several types of aversive signals and cognitive effort). Indeed, many of these signals originate in the ACC, a critical component of the executive control network (Kolling, Behrens, Wittman, & Rushworth, 2016; Vassena, Holroyd, & Alexander, 2017). As an interesting example, recent evidence suggests that processing affective stimuli in an empathic manner can be perceived as being effortful (Cameron, Hutcherson, Ferguson, Scheffers, & Inzlicht, 2017). Thus, attempting to control for these signals in trade-off contexts is an important issue to consider moving forward. Furthermore, identifying a potential general aversive (or costly) element common amongst these signals can help to shed light on the basic determinant of cognitive effortfulness in trade-off situations where effort is evaluated.

An additional limitation of the current set of studies concerns the role of motivation and reward in generating judgments of anticipated effort. Critically, these factors are known to affect effort-based decision-making (for a review see Botvinick & Braver, 2015) and were not specifically manipulated here. Explicit reward is well known

to discount the level of cognitive demand when individuals make effort-based decisions (e.g., Apps, Grima, Manohar, & 2015; Chong et al., 2017; Klein-Flügge, Kennerley, Firston, & Bestmann, 2016; Nishiyama, 2014; 2016; Phillips, Walton, & Jhou, 2007; Westbrook et al., 2013), and this devaluation process has been proposed to be specifically modulated in the nucleus accumbens (NAcc; Botvinick, Huffstetler, McGuire, 2009). As an example, the assumption that demanding (and in many cases effortful) tasks are aversive is violated if participants are highly motivated or rewarded to counteract boredom rather than to avoid cognitive demand (Milyavskaya et al., 2017; Westbrook & Braver, 2015). Following from this, effortful lines of action can be considered intrinsically rewarding for some individuals. Classically, Cacioppo and Petty (1982) demonstrated individual differences in *effort-seeking* behavior where engaging in such demanding tasks was associated with being rewarding. Recent endeavors have focused on the neural processes associated with how engaging in effortful actions modulates reward (e.g., Boehler et al., 2011; Hernandez-Lallement et al., 2014; Ma, Meng, Wang, & Shen, 2014) For example, Wang and colleagues (2017) recently demonstrated more pronounced FRN/P300 signals upon positive feedback in a high-effort task, suggesting that successes in this condition might include added subjective value. This finding suggests that engaging in high-effort tasks may carry intrinsic reward. Accounting for baseline motivations and expected reward (whether extrinsic or intrinsic) in judgments of anticipated effort thus represents an important consideration moving forward.

One interesting potential avenue for assessing the use of error-likelihood in generating evaluations of effort is through considering individual differences in personality and clinical contexts. For example, hyperactive error sensitivity has been

demonstrated in clinical populations with obsessive-compulsive disorder (Gehring, Himle, & Nisenson, 2000), and also in healthy samples of individuals high in negative affect (Hajack et al., 2004; Pailing & Segalowitz, 2004). Hence, a straightforward prediction is that these individuals may show deficiencies in attempting to override strong error biases in making effort judgments. In addition, several clinical disorders such as alexithymia (Maier et al., 2016) and Schizophrenia (Alain, McNeely, He, Christensen, & West, 2002; Bates, Kiehl, Laurens, & Liddle, 2002) demonstrate hypoactive error processing in affected individuals. As an interesting example, Gold and colleagues (2014) recently demonstrated that Schizophrenic patient samples were unable to avoid courses of action associated with high levels of cognitive effort. The authors attributed this failure to avoid effortful courses of action to deficits in the monitoring of control costs. A complementary explanation based on the error account proposed here then would suggest that insensitivity to errors may have caused the patient sample to inadequately map differential error-likelihoods, and thus differential effort-likelihoods to the options.

### **Conclusion**

The observation that humans are predisposed to avoid effortful actions is a cornerstone of theories of human behavior (e.g., Zipf, 1949). A clear assumption of this claim is that effortful actions are inherently costly and evoke a need to be avoided. Here, we demonstrate a strong association between judgments of effort and judgments of likelihood of errors of tasks, which differed in many cases from judgments of time requirements. Though we cannot completely rule out time as an important factor in effort judgments, the effort/error correspondence provides several interesting avenues to pursue moving forward to disentangle the key constituents of cognitive effort.



**Compliance with Ethical Standards**

**Funding:** This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) and funding from the Canada Research Chairs program to Evan F. Risko.

**Conflict of Interest:** All authors declare that they have no conflicts of interest pertaining to the development or submission of this manuscript.

**Ethical approval:** All procedures performed in these studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards, and were approved by the University of Waterloo Office of Research Ethics.

**Informed consent:** Informed consent was obtained from all individual participants included in the study.

### References

- Apps, M. A., Grima, L. L., Manohar, S., & Husain, M. (2015). The role of cognitive effort in subjective reward devaluation and risky decision-making. *Scientific Reports*, 5.
- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607-617.
- Akçay, Ç., & Hazeltine, E. (2007). Conflict monitoring and feature overlap: Two sources of sequential modulations. *Psychonomic Bulletin & Review*, 14(4), 742-748.
- Alain, C., McNeely, H. E., He, Y., Christensen, B. K., & West, R. (2002). Neurophysiological evidence of error-monitoring deficits in patients with schizophrenia. *Cerebral Cortex*, 12(8), 840-846.
- Apps, M. A., Grima, L. L., Manohar, S., & Husain, M. (2015). The role of cognitive effort in subjective reward devaluation and risky decision-making. *Scientific Reports*, 5.
- Ashcraft, M. H., & Faust, M. W. (1994). Mathematics anxiety and mental arithmetic performance: An exploratory investigation. *Cognition & Emotion*, 8(2), 97-125.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8, 47-89.
- Bates, A. T., Kiehl, K. A., Laurens, K. R., & Liddle, P. F. (2002). Error-related negativity and correct response negativity in schizophrenia. *Clinical Neurophysiology*, 113(9), 1454-1463.

- Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214-1221.
- Bijleveld, E., Custers, R., & Aarts, H. (2009). The unconscious eye opener: Pupil dilation reveals strategic recruitment of resources upon presentation of subliminal reward cues. *Psychological Science*, *20*(11), 1313-1315.
- Blain, B., Hollard, G., & Pessiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences*, *113*(25), 6967-6972.
- Boehler, C. N., Hopf, J. M., Krebs, R. M., Stoppel, C. M., Schoenfeld, M. A., Heinze, H. J., & Noesselt, T. (2011). Task-load-dependent activation of dopaminergic midbrain areas in the absence of reward. *Journal of Neuroscience*, *31*(13), 4955-4961.
- Boksem, M. A., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain Research Reviews*, *59*(1), 125-139.
- Botvinick, M. M., Braver, T. S. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83-113.
- Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 16-27.
- Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research PRPF*, *73*(6), 835-842.

- Boureau, Y. L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self control and meta-decision making. *Trends in Cognitive Sciences*, *19*(11), 700-710.
- Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(5712), 1118-1121.
- Bryce, D., & Bratzke, D. (2014). Introspective reports on reaction times in dual-tasks reflect experienced difficulty rather than the timing of cognitive processes. *Consciousness & Cognition*, *27*, 254-267.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3-5.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116-131.
- Cameron, D., Hutcherson, C., Ferguson, A. M., Scheffer, J. A., & Inzlicht, M. (2017, July 28). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. Retrieved from [psyarxiv.com/jkc4n](https://psyarxiv.com/jkc4n)
- Chong, T. T. J., Apps, M., Giehl, K., Sillence, A., Grima, L. L., & Husain, M. (2017). Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS biology*, *15*(2), e1002598.
- Danckert, J. A., & Allman, A. A. A. (2005). Time flies when you're having fun: Temporal estimation and the experience of boredom. *Brain and Cognition*, *59*(3), 236-245.
- Davenport, H. J. (1911). Cost and its significance. *The American Economic Review*, *1*(4), 724- 752.

- Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science, 5*(5), 303-305.
- Desender, K., Buc Calderon, C., Van Opstal, F., & Van den Bussche, E. (2017). Avoiding the conflict: Metacognitive awareness drives the selection of low-demand contexts. *Journal of Experimental Psychology: Human Perception and Performance, 43*(7), 1397-1410.
- Desender, K., Van Opstal, F., & Van den Bussche, E. (2017). Subjective experience of difficulty depends on multiple cues. *Scientific Reports, 7*, 44222. Doi: 10.1038/srep44222
- Diede, N. T., & Bugg, J. M. (2017). Cognitive effort is modulated outside of the explicit awareness of conflict frequency: Evidence from pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(5), 824-835.
- Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: Neural and behavioral evidence. *PLoS One, 7*(12), e51637.
- Dreisbach, G., & Fischer, R. (2012). Conflicts as aversive signals. *Brain and Cognition, 78*(2), 94-98.
- Dunn, T. L., Koehler, D. J., & Risko, E. F. (2017). Evaluating effort: Influences of evaluation mode on judgments of task-specific efforts. *Journal of Behavioral Decision Making*. Advanced online publication. Doi: 10.1002/bdm.2018
- Dunn, T. L., Lutes, D. J. C., & Risko, E. F. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1372-1387.
- Dunn, T. L., & Risko, E. F. (2016). Toward a metacognitive account of cognitive offloading. *Cognitive Science, 40*(5), 1080-1127.

- Eriksen, C. W. (1995). The flankers task and response competition: a useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2-3, 101-118.
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspective on Psychological Science*, 8(3), 223-241.
- Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, 51(2), 87-107.
- Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitation in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, 14(1), 129-146.
- Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116-124.
- Frank, M. J., Worocho, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, 47(4), 495-501.
- Gehring, W. J., & Fencsik, D. E. (2001). Functions of the medial frontal cortex in the processing of conflict and errors. *Journal of Neuroscience*, 21(23), 9430-9437.
- Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, 4(6), 385-390.
- Gehring, W. J., Himle, J., & Nisenson, L. G. (2000). Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological Science*, 11(1), 1-6.

- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278.
- Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20-29.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650-669.
- Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that makes us smart*. New York, NY: Oxford University Press.
- Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, *19*(2), 483-495.
- Gold, J. M., Kool, W., Botvinick, M. M., Hubzin, L., August, S., & Waltz, J. A. (2014). Cognitive effort avoidance and detection in people with schizophrenia. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(1), 145-154.
- Gray, W. D., Sims, C. R., Fu, W-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461-482.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217-229.
- Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *Journal of chemical neuroanatomy*, *26*(4), 317-330.

- Hajack, G., & Foti, D. (2008). Errors are aversive: Defensive motivation and the error related negativity. *Psychological Science, 19*(2), 103-108.
- Hajack, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology, 40*(6), 895-903.
- Hajack, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology, 42*(2), 151-160.
- Hernandez-Lallement, J., van Wingerden, M., Marx, C., Srejic, M., & Kalenscher, T. (2014). Rats prefer mutual rewards in a prosocial choice task. *Frontiers in Neuroscience, 8*.
- Hockey, G. R. J. (2011). *A motivational control theory of cognitive fatigue*. In P.L. Ackerman (Ed.), *Cognitive fatigue: multidisciplinary perspectives on current research and future applications* (pp. 167-188). Washington, DC: American Psychological Association.
- Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679-709.
- Inzlicht, M., Bartholow, B. D., & Hirsh, J. B. (2015). Emotional foundations of cognitive control. *Trends in Cognitive Sciences, 19*(3), 126-132.
- Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences, 18*(3), 127-133.
- Jeffreys, H. (1961). *Theory of probability* (3<sup>rd</sup> ed.). Oxford, England: Oxford University Press.



- John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York: Guilford Press.
- Jordan, K., & Huntsman, L. A. (1990). Image rotation of misoriented letter strings: Effects of orientation cuing and repetition. *Perception & Psychophysics*, *48*(4), 363-374.
- Kaas, R. E., & Raferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.
- Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: consensus and controversy. *Neuron*, *63*(6), 733-745.
- Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.
- Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583-1585.
- Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, *79*(1, Pt.1), 164-167
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582-591.
- Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*(5660), 1023-1026.

- Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort-reward trade-off. *Journal of Neuroscience*, *36*(39), 10002-10015.
- Kolling, N., Behrens, T. E. J., Wittmann, M. K., & Rushworth, M. F. S. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology*, *37*, 36-43.
- Kool, W., & Botvinick, M. M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 131-141.
- Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665-682.
- Koriat, A., & Norman, J. (1984). What is rotated in mental rotation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 421-434.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573-603.
- Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology*, *7*, 67-70.
- Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model subjective effort and task performance. *Behavioral and Brain Sciences*, *36*, 661-726.
- Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences*, *111*(6), 2343-2348.

- Lawrence, M. A. (2015). ez: Easy analysis and visualization of factorial experiments. R package version 4.3. Retrieved from <http://CRAN.Rproject.org/package=ez>
- Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York, NY: Cambridge University Press.
- Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: a review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, 2(2), 174-207.
- Luu, P., Collins, P., & Tucker, D. M. (2000). Mood, personality, and self-monitoring: Negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring. *Journal of Experimental Psychology: General*, 129(1), 43-60.
- Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science*, 14(1), 47-53.
- Ma, Q., Meng, L., Wang, L., & Shen, Q. (2014). I endeavor to make it: Effort increases valuation of subsequent monetary reward. *Behavioural Brain Research*, 261, 1-7.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2), 163-203.
- Maier, M. E., Scarpazza, C., Starita, F., Filogamo, R., & Làdavas, E. (2016). Error monitoring is related to processing internal affective states. *Cognitive, Affective, and Behavioral Neuroscience*, 16(6), 1050-1062.
- Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition*, 115(2), 303-313.

- McGuire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences, 107*(17), 7922-7926.
- Miller, J., Vieweg, P., Kruize, N., & McLea, B. (2010). Subjective reports of stimulus, response, and decision times in speeded tasks: How accurate are decision time reports? *Consciousness & Cognition, 19*(4), 1013-1036.
- Milyavskaya, M., Inzlicht, M., Johnson, T., & Larson, M. J. (2017, August 16). Reward sensitivity following boredom and cognitive effort: A high-powered neurophysiological investigation. Retrieved from [psyarxiv.com/2czjv](https://psyarxiv.com/2czjv). DOI 10.17605/OSF.IO/B3MYA
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134-140.
- Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.11-1. Retrieved from <http://CRAN.R-project.org/package=BayesFactor>
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16*(5), 1936-1947.
- Naccache, L., Dehaene, S., Cohen, L., Habert, M. O., Guichart-Gomez, E., Galanaud, D., & Willer, J. C. (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia, 43*(9), 1318-1328.
- Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review, 86*(3), 214-255.

- Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752-760.
- Nishiyama, R. (2014). Response effort discounts the subjective value of rewards. *Behavioural Processes*, *107*, 175-177.
- Nishiyama, R. (2016). Physical, emotional, and cognitive effort discounting in gain and loss situations. *Behavioural Processes*, *125*, 72-75.
- Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520.
- O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660-E3669.
- Pailing, P. E., & Segalowitz, S. J. (2004). The error-related negativity as a state and trait measure: Motivation, personality, and ERPs in response to errors. *Psychophysiology*, *41*(1), 84-95.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York City, NY: Cambridge University Press.
- Phillips, P. E., Walton, M. E., & Jhou, T. C. (2007). Calculating utility: Preclinical evidence for cost-benefit analysis by mesolimbic dopamine. *Psychopharmacology*, *191*(3), 483-495.

- Protopapas, A. (2007). CheckVocal: a program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*, 859-862.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org>
- Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264-272.
- Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45-48.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomics Bulletin & Review*, *21*(2), 301-308.
- Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.
- Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460-475.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.
- Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin*, *134*(2), 207-222.
- Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

- Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (in press). Toward a rational and mechanistic account of cognitive effort. *Annual Review of Neuroscience*.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127-190.
- Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: testing predictions of ASCM using the choice/no-choice method. *Journal of experimental psychology: General*, *126*(1), 71-92.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution. *Dialogue, The Official Newsletter of the Society for Personality and Social Psychology*, *26*(2), 4-7.
- Simon, H. A. (1982). *Models of bounded rationality (Vol. 3): Empirically grounded economic reason*. Cambridge, MA: MIT Press.
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1-20.
- Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: Recent finding and theoretical perspectives. *The Neuroscientist*, *13*(2), 160-172.
- Van Steenbergen, H., & Band, G. P. H. (2013). Pupil dilation in the Simon task as a marker of conflict processing. *Frontiers in Human Neuroscience*, *7*, 215.  
<http://doi.org/10.3389/fnhum.2013.00215>

- Vassena, E., Holroyd, C. B., & Alexander, W. H. (2017). Computational models of anterior cingulate cortex: At the crossroads between prediction and effort. *Frontiers in Neuroscience, 11*.
- Vassena, E., Silvetti, M., Boehler, C. N., Achten, E., Fias, W., & Verguts, T. (2014). Overlapping neural systems represent cognitive effort and reward anticipation. *PLoS One, 9*(3), e91008.
- Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: a neurocomputational model. *Frontiers in Behavioral Neuroscience, 9*(57). doi: 10.3389/fnbeh.2015.00057
- Walsh, M. M., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology, 58*(3), 416-440.
- Wang, L., Zheng, J., & Meng, L. (2017). Effort provides its own reward: Endeavors reinforce subjective expectation and evaluation of task performance. *Experimental Brain Research, 235*(4), 1107-1118.
- Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron, 89*(4), 695-710.
- Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS One, 8*(7), e68210.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.
- Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C.



- Klauer (eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189-217). Mahwah, NJ: Erlbaum.
- Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, *111*(4), 931-959.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.