On the Determinants of Effort

by

Timothy Lee Dunn

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirements for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Maggie Toplak Professor
Supervisor	Dr. Evan Risko Assistant Professor
Internal Member	Dr. Colin MacLeod Professor
Internal Member	Dr. Daniel Smilek Professor
Internal-external Member	Dr. Selcuk Onay Associate Professor

Author's Declaration

This thesis consists of material all of which I authored or co-authored: See Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I would like to acknowledge the names of my co-authors who contributed to the research described in this dissertation. These include:

- David Lutes
- Dr. Derek Koehler
- Dr. Michael Inzlicht
- Dr. Evan Risko

Abstract

That humans configure their behaviors in ways to avoid effortful actions is undoubtedly one of the most pervasive hypotheses put forth to account for a wide range of human behaviors. This dissertation describes a series of experiments aimed at testing accounts of how individuals make effort-based decisions, and why actions may be evaluated as effortful. In Chapter 1, I contrasted the hypothesis that individuals' effort avoidance behaviors and conceptions of effort are driven by the performance associated with different lines of actions, versus the hypothesis that individuals generate a kind of metacognitive evaluation of effort that can be dissociated from performance. I found that individuals' choices were not associated with their performance or with a physiological measure of demand, but rather tracked closely with subjective perceptions of effort associated with embedded cues supporting the latter hypothesis. Chapter 2 extended the idea that individuals make their effort-based decisions by utilizing cues through pitting options against one another that vary on cue saliency, demands on executive control, and performance. Chapter 3 looked to test how individuals evaluate expected effort across a range of tasks through manipulations of evaluation mode (i.e., whether options are evaluated comparatively or in isolation). Following from these results, I propose that evaluable forms of effort are driven by the presence of a failure point associated with the task. In Chapter 4, specific determinants of effort were examined by crossing anticipated time demands and error likelihood across different choice options. Data supported the notion that judgments of effort are closely related to the perceived likelihood of an error associated with a task, but not to the time demands. I conclude by proposing that cognitive effort can best be conceived of as a type of inferential metacognitive evaluation made over available cues that are weighed on the perceived likelihood of making errors.

Acknowledgements

First and foremost, I am forever grateful for the unconditional support given to me by my mother, father, and brother throughout my time as a student. Their selflessness is beyond comparison. This dissertation is just as much a result of their work as it is my own.

Second, I am extremely thankful for the extraordinary investment of my advisor Evan Risko in my graduate career. His continual guidance over the years has helped me grow not only intellectually but personally as well. I feel that our relationship has developed beyond just that of advisor and student and into a true friendship between colleagues.

Third, I am grateful for the members of the Cognition and Natural Behavior Laboratory. Without their dedication to research none of this work would have been possible. It has been a pleasure to work, mentor, and build relationships with the group over the years.

Last, I would like to thank Colin MacLeod and Dan Smilek for their thoughtful and constructive comments on this thesis.

Table of Contents

List of Tables

List of Figures

Introduction

"…the entire behavior of an individual is at all times motivated by the urge to minimize effort." (Zipf, 1949, p. 3).

The above statement by G. K. Zipf in his seminal work *Human Behavior and the Principle of Least Effort* effectively captures the ubiquity of the idea of effort minimization: the effort associated with some line of action is most often conceptualized as a cost to be minimized in decision-making processes (Botvinick, 2007; Botvinick & Braver, 2015; Inzlicht, Schmeichel, & Macrae, 2013; Kool, McGuire, Rosen, & Botvinick, 2010; Kool & Botvinick, 2014; Kurzban, Duckworth, Kable, & Myers, 2013; Shenhav, Botvinick, & Cohen, 2013; Shenhav et al., 2017; Westbrook & Braver, 2015; 2016). This proposed *principle* (Clark, 2010; Zipf, 1949) or *law* (Hull, 1949) of human behavior has been applied as an explanatory tool to a wide range of behaviors ranging from attempting to suppress inappropriate motor responses (e.g., the flanker task, Eriksen & Eriksen, 1974), to engaging in shallow intuitive thinking (Pennycook, Fugelsang, & Koehler, 2015), to issues with resisting temptations and regulating emotions (see de Ridder et al., 2012 for a review). Although much work has recently focused on cognitive effort, how individuals make decisions based on effort and how effort is evaluated are still open questions in need of addressing. Thus, the present work aimed to test accounts of how individuals make effort-based decisions, and why some actions may be evaluated as effortful.

Chapter 1 tests the hypothesis that individuals avoid courses of effortful action based on their associated performance with each action (i.e., response times and errors). This is contrasted with the hypothesis that individuals avoid such actions based on an inferential metacognitive evaluation of demand. Across two experiments individuals completed free-choice Demand Selection Tasks (DST) where individuals were free to choose alternative stimuli to read on a trial-by-trail basis that used conditions that dissociate performance from perceived effort. Overall patterns of preference choices in the DST followed those of perceived demand rather than performance. Furthermore, choices were dissociated from a peripheral physiological measure of demand as indexed by blink rates. A final experiment utilized a forced-choice version of a DST that asked individuals to make choices based on effort, time, and errors. Patterns of choices were similar across the three rating dimensions, supporting the notion that a general metacognitive evaluation of demand is based on cues that drive choices.

Chapter 2 extends the cue utilization account of how individuals make effort-based decisions. Specifically, the influences of time and demands on executive control are contrasted with the influence of an available effort cue across three experiments. Using a forced-choice version of a DST where effort minimization was instructed, evidence is demonstrated for avoidance behaviors dissociating from both time and demands on executive control in a manner predicted by a cue-utilization account. Furthermore, assessments of cue awareness were collected through self-report. Interestingly, awareness of the specific high-demand cues associated with the options (i.e., more stimulus rotation and more task switching) tracked closely with rates of avoidance of those high-demand conditions.

Chapter 3 investigates how individuals evaluate effort. Here, the evaluability of effort is examined through methods described by the General Evaluability Theory (GET; Hsee & Zhang, 2010). GET argues that evaluable attributes (here effort) should be consistently judged across joint (i.e., judged comparatively) and single (i.e., judged in isolation) evaluation modes. Individuals judged the anticipated effort of four task-specific efforts indexed by stimulus rotation, items to-be-remembered, weight to-be-lifted, and stimulus degradation. Across six experiments, I demonstrate that effort judgments associated with items to-be-remembered, weight to-be-lifted, and stimulus degradation can be considered relatively evaluable, while the effort associated with stimulus rotation may be relatively inevaluable. I argue that evaluability is driven by reference information specifically pertaining to when an expected error or failure may occur as perceived effort increases.

Chapter 4 addresses the question of why some actions are evaluated as effortful. Individuals' perceptions of anticipated effort were collected in contexts where two basic determinants of effort were traded off: time and errors. Across three experiments, I demonstrate that individuals perceive options that are associated with low time requirement but high error-likelihood as more effortful than options that are more time consuming but have a low error-likelihood. Furthermore, effort-based and error-based choices closely tracked one another, whereas time-based choices did not demonstrate this same pattern. Results support the conclusion that the likelihood of error commission drives effort-based choices in this trade-off context.

I conclude this thesis by integrating the evidence from these four chapters into a novel account of cognitive effort. Specifically, I argue that cognitive effort is best conceptualized as an inferential evaluation of demand based on available cues, and that these cues are weighted based on the likelihood of error commission associated with a line of action. Future implications for this account are additionally discussed.

Chapter 1

The following work has been published in the *Journal of Experimental Psychology: Human Perception and Performance* (Dunn, Lutes, & Risko, 2016).

Changes have been introduced to improve the flow of the dissertation.

In the mid-18th century, French mathematician and philosopher Pierre-Louis Moreau de Maupertuis formulated one of the most pervasive hypotheses with regard to the physical world– the Principle of Least Action–according to which all phenomena occurring in nature minimize some form of "action" (Maupertuis, 1750). Given the sweeping implications of Maupertuis' assertion, the notion that physical systems act in a way that minimizes action began to be applied across a range of human behaviors. In the case of human systems, the "action" that was thought to be minimized or avoided was some form of effort or demand, whether it be classified as physical or cognitive (e.g., Allport, 1954; Clark, 2010; Hull, 1943; Rosch, 1998; Solomon, 1948; Zipf, 1949). Interestingly, not until recently have systematic attempts been made to examine, experimentally, the latter type of demand avoidance behavior (Gold et al., 2014; Kool, McGuire, Rosen, & Botvinick, 2010; Westbrook, Kester, & Braver, 2013). This research has mostly relied on assessing individuals' preference (as indexed via free-choice) for selecting particular courses of action that vary in cognitive demand. In the present investigation, I extend this line of research by examining the role of metacognitive evaluation in these types of decisions.

Avoiding Cognitive Demand

Possibly one of the most widespread hypotheses put forward regarding demand avoidance is likening strategies that optimize performance, such as minimizing response times or errors (Maglio, Wenger, & Copeland, 2008; Payne, Bettman, & Johnson, 1988; 1992; Siegler & Lamaire, 1997; Walsh & Anderson, 2009), to minimizing demand. One prevalent example of this notion, the soft-constraints hypothesis (Gray & Fu, 2004; Gray et al., 2006), states that the selection of interactive behaviors (i.e., strategy selection involving external resources) tends to minimize objective costs in terms of time (Gray & Boehm-Davis, 2000). Gray and colleagues (2006) suggest that individuals possess implicit knowledge pertaining to the costs (in terms of time) associated with the activation of an item to be retrieved from memory, actual retrieval of that item, and the probability that the item will be recalled despite decay and noise. It is these objective, time-oriented costs that the system takes into account when selecting a demand-minimizing action. Such a view suggests the cognitive system maintains a type of "direct-access" to differences in time in guiding action selection.

An alternative to a performance-driven account of demand avoidance, the notion that cognitive effort is equated with the amount of controlled or executive processing required for a task, has received much attention. Research investigating the avoidance of demand within the context of executive control has used variants of what are referred to as demand selection tasks (DST; Botvinick & Rosen, 2009; Gold et al., 2014; Kool et al., 2010; McGuire & Botvinick, 2010). In these tasks, individuals are faced with a repeated free-choice between two alternative courses of action. For example, to examine demand avoidance Kool et al. (2010) manipulated, across the two tasks, the amount of demand placed on the executive control system, operationalized as the likelihood of a task switch (see Monsell, 2003). Critically, across several experiments, Kool et al. (2010) demonstrated that individuals readily avoided selecting courses of action associated with a greater likelihood of a task switch (i.e., greater cognitive effort).

An important feature of the design of Kool et al.'s (2010) demand selection experiments was the attempt to control "time-on-task" across the two choices. This is important because making a task objectively more cognitively demanding (e.g., through a high proportion of task switches) typically increases the amount of time it takes to complete the task and, consequently, participants' choices could reflect a desire to minimize time as mentioned above (e.g., Gray et al., 2006; Payne, Bettman, & Johnson, 1988; Siegler & Lamaire, 1997) rather than to minimize executive processing demands. Kool and colleagues (2010) sought to address this issue by creating a DST that involved a high-demand task-switching strategy that would result in minimizing the objective time required to complete the goal of the task, and a low-demand strategy that would avoid the need to engage in a task switch but result in longer times to complete the goal (i.e., a global cost). Individuals most often selected the low-demand strategy, thus avoiding the task switch even in light of a global cost in terms of time. Westbrook and colleagues (2013) additionally demonstrated that individuals forego awards to conserve cognitive effort while controlling for time-on-task. Such findings have led to the conclusion that a desire to minimize objective costs (in terms of performance) cannot fully account for demand avoidance behavior.

Subjectivity and Explicitness

Given the claim that demand avoidance can be observed even when controlling for objective costs (e.g., time-on-task), subjective accounts of demand avoidance have recently come into focus theoretically. Westbrook and Braver (2015; see also Kool & Botvinick, 2013) suggest that cognitive effort carries a subjective cost that may covary with objective costs (i.e., time, accuracy), but nonetheless cannot be described strictly in these objective terms. Such subjective costs are argued to be generated in the lateral prefrontal cortex (lPFC) and are independent of signals associated with objective costs (McGuire & Botvinick, 2010; Naccache, et al. 2005). Recently, Kurzban and colleagues (2013) have specifically classified this hypothesized subjective cost in terms of an opportunity cost. That is, "…the costs of performing task X include the potential benefit of doing those other tasks (A, B, C, etc.) that are precluded because the systems required for the task X cannot be used for alternatives A, B, or C" (p. 664). Thus, tasks that utilize a range of systems required for many other tasks (e.g., executive processes) carry large opportunity costs. Subjective effort costs under this framework are the output from mechanisms used to measure the opportunity costs of engaging in a task and, critically, this explicit consciously experienced phenomenon serves as an adaptive signal as action selection unfolds (Kurzban et al., 2013)

The idea that subjective costs and some degree of explicit awareness of said costs are tightly coupled in driving action selection represents an intriguing view with direct consequences for extant theories. Gold et al. (2014) recently demonstrated this notion utilizing the same DST method as in Kool et al. (2010). An initial experiment failed to replicate avoidance of cognitive demand (i.e., participants did not avoid the choices associated with more task switches). Interestingly, only when participants were explicitly instructed of differences in demand across conditions did Gold et al. (2014) find the pattern of expected demand avoidance (i.e., participants avoided the choices associated with more task switches). The authors interpreted the inability to observe demand avoidance in the free-choice DST as a failure within the sample to monitor the objective task-switching costs present in the task. Alternatively, the need to explicitly instruct participants in this manner suggests the possibility that individuals, in the context of the DST, are selecting tasks based on a determinant beyond the putative demands placed on the executive control system by task-switching.

A Metacognitive Framework for Demand Avoidance

Taken together, the foregoing evidence suggests that cognitive demand avoidance is driven by subjective costs, potentially correlated with, but not necessarily driven by objective task demands, and these costs may need to be explicit to the individual for avoidance behaviors to manifest (cf. General Discussion). Thus, one potentially productive avenue to pursue with regard to action selection within the context of demand avoidance is to view these decisions as a kind of metacognitive control (Dunn & Risko, 2016). Metacognitive control refers to the regulation of cognitive processes or behavior dynamically guided by an individual's subjective monitoring of their own cognitive processes (Koriat, Ma'ayan, & Nussinson, 2006; Nelson & Narens, 1990; Son & Schwartz, 2002). Critically, the output of this monitoring process that is individuals' metacognitive experiences (e.g., feeling of knowing, Hart, 1965; Koriat & Goldsmith, 1996; judgments of learning, Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002; confidence judgments, Koriat & Goldsmith, 1996; Robinson, Johnson, & Herndon, 1997) have been demonstrated to play a causal role in controlling behavior (Koriat & Goldsmith, 1996; Koriat et al., 2006; Nelson, 1996).

While authors typically postulate specific types of metacognitive experiences (e.g., feeling of knowing), others have hypothesized that a possible function of subjective experience may be to serve as a type of "summary" signal or global index of processing quality that reduces various processes into a single condensed experience (Epstein, 2000; Mangan, 2001), for example a subjective assessment of expected processing fluency (Arango-Muñoz, 2014; Reber, Fazendiero, & Winkleman, 2002; Reber & Schwarz, 2001). Similarly, recent work by Dunn and Risko (2016) suggests that a useful way of conceptualizing effort is to liken it to a type of general metacognitive evaluation of perceived task demand. Specifically, the authors found analogous patterns of perceived ratings of effort, time, and accuracy in a perceptual task. Hence, these metacognitive evaluations may incorporate several types of information.

In this vein, individuals' metacognitive evaluations of demand can be considered as akin to a type of metacognitive experience (e.g., perceived effort, Efklides, 2001; 2002; Efklides, Kourkoulou, Mitsiou, & Ziliakopoulou, 2006; Efklides & Petkaki, 2005; Koriat & Levy-Sadot, 2001) and selecting a particular course of action (e.g., in a DST) as the behavior being regulated. From this metacognitive perspective, the experiences and the judgments driving control can be considered largely inferential and, consequently, sensitive to factors such as preconceived biases, beliefs, or intuitive theories (Nelson & Narens, 1990; Koriat, 2007). Thus, it would be expected that individuals rely on multiple cues when making selections in a DST, only some of which will be directly observable in terms of objective costs (e.g., performance). Indeed, numerous dissociations between types of objective costs and metacognitive evaluations exist across a wide range of domains such as spatial attention (Wilimzig, Tsuchiya, Fahle, Einhäuser, & Koch, 2008), metamemory (Castel, Rhodes, & Friedman, 2013; Logan, Castel, Haber, & Viehman, 2012; Metcalfe, Schwartz, & Joaquin, 1993), blind insight (Scott, Dienes, Barrett, Bor, & Seth, 2014; Timmermans, Schilbach, Pasquali, & Cleeremans, 2012), and cognitive offloading (Gilbert, 2015; Risko & Dunn, 2015).

Present Investigation

In the present investigation, I examine this metacognitive framework through a test of the hypothesis that individuals avoid courses of action based on a metacognitive evaluation of demand. As discussed above, I consider perceived demand to be a general-purpose signal that incorporates several types of information. To examine the contribution of perceived demand to free-choice in a demand selection task, I took advantage of a recent demonstration of stimulus conditions that yielded a dissociation between performance (i.e., time and errors) and ratings of perceived time demand, effort, and likelihood of error (i.e., what is being referred to here as demand more generally; see Dunn & Risko, 2016). Critically, this dissociation allows a strong test of the hypothesis that choices in a demand selection task are based on the metacognitive evaluation of demand, as different patterns of choices would be expected based on objective performance (i.e., time, accuracy) and perceived demand.

The dissociation between perceived demand and performance was observed in the context of a reading task that involved the factorial crossing of word and frame rotation in multi-element arrays: Upright Words-Upright frame (UW-UF), Upright Words-Rotated Frame (UW-RF), Rotated Words-Upright Frame (RW-UF), and Rotated Words-Rotated Frame (RW-RF; see Figure 1). Critically, Dunn and Risko (2016) demonstrated that performance in the RW-RF condition was equivalent to performance in the RW-UF condition with both producing longer RTs and more errors than the UW-RF and UW-UF conditions (i.e., RW-RF = RW-UF > UW-RF > UW-UF). Individuals, however, perceived reading the RW-RF array to be more effortful, time demanding, and less accurate relative to the RW-UF array, followed by the UW-RF and UW-UF arrays (i.e., RW-RF > RW-UF > UW-RF > UW-UF). The authors suggested that this dissociation was due to individuals' preconceived (and mostly accurate) notion that "disoriented" visual stimuli are difficult to process and that the RW-RF display is *more* disoriented (i.e., both words and reading direction are rotated) than the RW-UF display (i.e., only the words are rotated). Based on this notion, participants infer based on the explicit cues available (i.e., the number of dimensions rotated in the display) that the RW-RF display will be associated with greater effort, more time, and more errors (i.e., greater perceived demand). Across three experiments (and 2 more reported here), however, there was no evidence of any performance differences between the RW-RF and RW-UF conditions.

As discussed above, previous research has suggested that demand avoidance behavior can be dissociated from objective demand (e.g., time demands). However, this work is not without potential shortcomings that the present investigation directly addresses. For example, with regard to the Kool et al. (2010) experiment highlighted above, in choosing the low-demand strategy and avoiding the cost associated with a task switch, individuals were able to minimize a local time cost in that maintaining an established strategy was faster at the trial level (i.e., they avoided switching between trials but ended up performing a larger number of trials and hence taking longer). The notion that individuals may focus on local optimization while demonstrating suboptimal performance at the global task level is a common theme within the strategy selection literature (Anderson, 1990; Fu & Gray, 2004, Gray et al., 2006). In this sense, the demand avoidance behavior in Kool et al. (2010) could have reflected the avoidance of local objective time-based demands. Moreover, in Westbrook et al. (2013), individuals completed a discounting task utilizing varying objective demands in terms of *N*-back levels. Although performance diminished as *N*-back levels increased, regression results suggested unique contributions of demand beyond performance on individuals' wagers. Importantly, however, in each of these cases, performance differences were present and it is unclear how the system may exploit these signals during demand avoidance. Therefore, the demonstration of demand avoidance in the absence of an objective performance cost, as methods in the present investigation allows, represents a theoretically important hypothesis to test.

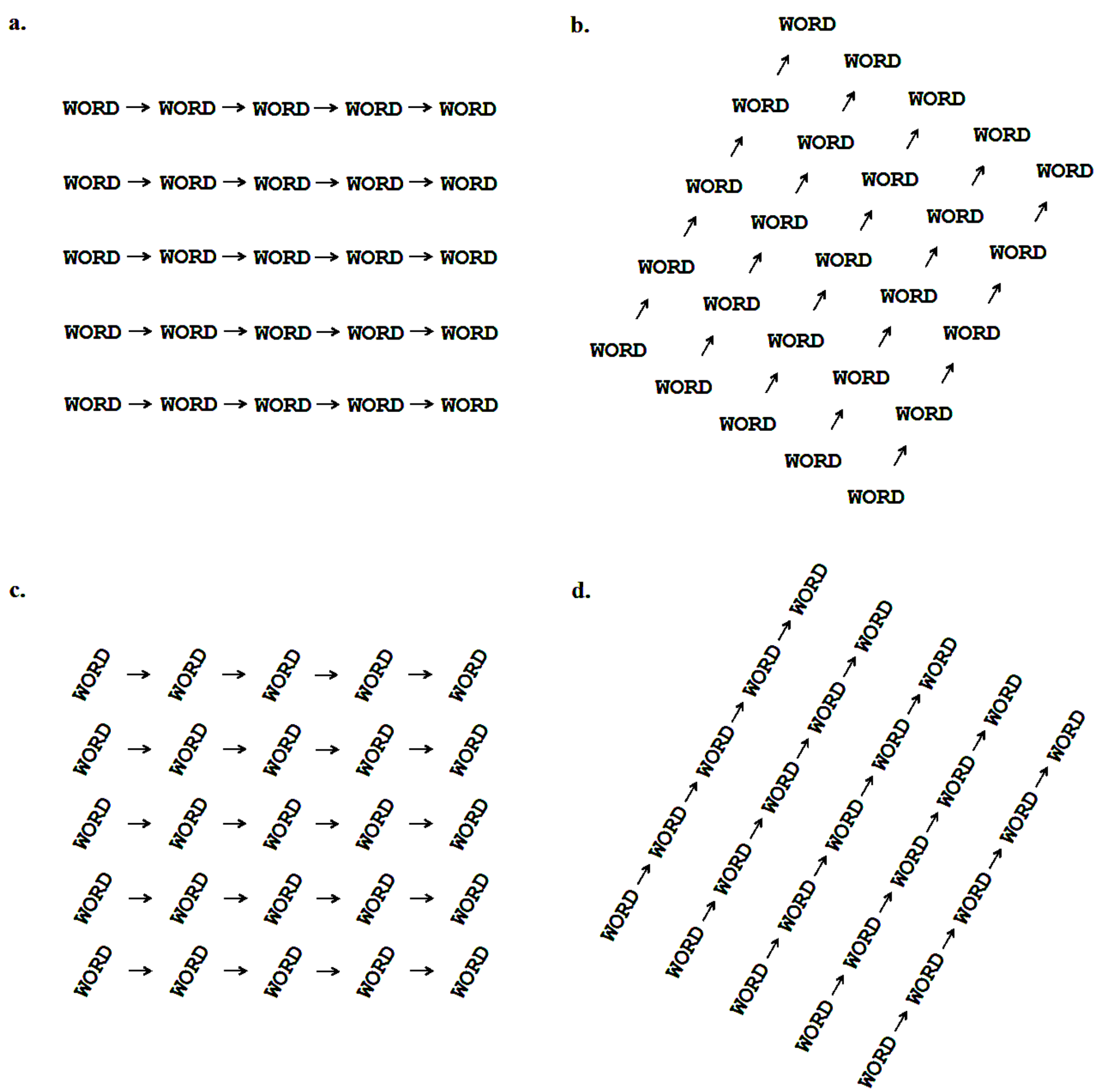
Experiment 1

In Experiment 1, participants took part in a variant of a free-choice DST that involved choosing which of two displays of words they would prefer to read. The methodological move to a DST utilizing the different arrays described above provides several important extensions to the initial work by Dunn and Risko (2016). First, Dunn and Risko (2016) examined the role of perceived demand within the context of cognitive offloading (i.e., integrating an external strategy into a cognitive act; Risko, Medimorec, Chisholm, & Kingstone, 2014). Although the potential role of perceived demand as a determinant in deciding to integrate an external strategy and making selections in a DST context is an intuitive assumption, it remains a possibility that such decision-making contexts differ in the role that perceived demand may play. Thus, providing evidence for a metacognitive account across unique decision-making paradigms represents an important finding to establish. Second, perceived demand was assessed in Dunn and Risko (2016) by way of individuals' self-reported ratings on a likert-type scale. Recently, however, assessing demand in terms of decision-making (i.e., selections in the DST) rather than self-report has been argued to be a more fruitful evaluation of demand avoidance (Westbrook et al., 2013; Westbrook & Braver, 2015). Furthermore, additional issues with self-reported values of demand are examined in the General Discussion.

In the DST, each trial consisted of a selection between two arrays of words that participants would then have to read. Each array consisted of 25 words arranged into the four different array configurations highlighted above (see Figure 1). If individuals avoid reading arrays based on a metacognitive evaluation of perceived task demand, then individuals should be more likely to select to read the RW-UF arrays than the RW-RF arrays, with both being selected less than the UW-RF and UW-UF arrays. Alternatively, if individuals avoid reading arrays based on objective performance demands, then I would expect no difference in the likelihood that individuals select the RW-UF and RW-RF arrays. In addition to the critical DST, individuals also completed a reading task utilizing the same arrays as described above in order to replicate the critical patterns in performance (i.e., the equivalence in terms of reading time and accuracy across the RW-RF and RW-UF displays).

Figure 1.

Examples of Array Types



Note: (a) Upright Word-Upright Frame (UW-UF); (b) Upright Word-Rotated Frame (UW-RF); (c) Rotated Word-Upright Frame (RW-UF); (d) Rotated Word-Rotated Frame (RW-RF). Each array was presented using 25 unique words configured into 5 x 5 displays rotated at both ±60° from upright (excluding the UW-UF array). Arrows were included in the arrays to ensure natural reading direction.

Method

Participants

Thirty-six University of Waterloo undergraduate students participated in the study in exchange for research credit. Sample size was estimated assuming a moderate effect size for the critical RW-RF and RW-UF comparison group (i.e., Cohen's *d* of .48) with an 80% probability to detect the effect for a within-subjects design at $\alpha = .05$. This effect size was computed using the perceived effort data from Experiment 3 in Dunn and Risko (2016). Individuals provided self-report ratings of perceived effort after completing a reading task using the same stimuli used in the current study. The *d* value reported here is derived from the pairwise comparison between the RW-RF and RW-UF arrays (i.e., SD_{Diff} as the standardizer). In addition, obtaining a large enough sample size to complete full counterbalancing of conditions was taken into consideration.

Design

For both the demand selection task (DST) and the reading task, a one factor (Array Type: Upright Word-Upright Frame: UW-UF, Upright Word-Rotated Frame: UW-RF, Rotated Word-Upright Frame: RW-UF, Rotated Word-Rotated Frame: RW-RF; see Fig. 1) within-subject design was employed.

Apparatus

Presentation of the stimuli and collection of vocal and manual button-press responses were handled by DMDX software (Forster & Forster, 2003). The stimuli were presented on a 24" LCD monitor and participants sat approximately 70 cm away from the monitor while loosely maintaining an upright head position in a headrest. Individuals were not asked to completely place their chin in the headrest so they could comfortably respond aloud. Participants used a standard QWERTY keyboard to enter their manual responses.

Stimuli

Array stimuli were taken from the experiments reported in Dunn and Risko (2016). Arrays consisted of 25 random nouns and verbs, 4 to 5 letters per word (mean written word frequency = 58.96 per million), arranged into 5 x 5 displays (see Figure 1). All words were presented in 18 point Courier New font. Arrays for the DST only consisted of 5-letter words. The RW-RF and UW-RF array types subtended approximately 15° x 14° (H x W), while the RW-UF and UW-UF array types subtended approximately 9.5° x 11.5°. The first word in each array was colored red and arrows were included between words to ensure natural reading. Each disoriented array type (RW-RF, RW-UF, and UW-RF) was presented to the right of upright (0°) and to the left of upright (i.e., + and - 60°) an equal amount of times, while the UW-UF array type was only presented at 0°. Twelve unique lists were created and counterbalanced such that each 25-word set appeared an equal number of times in each condition.

Stimuli for the array selection portion of the DST consisted of two arrays presented on the left and right sides of the screen separated by a black line. Each array was structured in the same manner as mentioned above, but "WORD" was presented in each position rather than a unique word to ensure that individuals were selecting based on the explicit structure of the array (i.e., rotated words, rotated frame) rather than the individual words within the array. Thus, in contrast to DSTs using a sampling methodology, that is using variations across decks that require monitoring of objective costs at the block-level, all cues were explicitly presented at the trial-level. Instructions on the top of the selection screen stated to, "*Please* c*hoose the type of array you would prefer to read."* The combination of all pairwise comparisons of the arrays resulted in six unique comparison groups: UW-UF | UW-RF, UW-UF | RW-UF, UW-UF | RW-RF, UW-RF | RW-UF, UW-RF | RW-RF, and RW-UF | RW-RF. Each comparison group was randomly presented 8 times during the task using 12 unique lists so that each array type appeared an equal number of times in each position in the selection screen (i.e., left or right) and each direction (i.e., 0° and ±60°).

Procedure

Participants read instructions based on the task that they were completing first (i.e., reading task or DST). Instructions for the reading task stated to read each word in the presented array aloud as quickly and accurately as possible and to press the *B* button once they had finished reading every word. Instructions for the DST stated that individuals were to choose as quickly as possible which of the two arrays they would prefer to read by pressing the *Z* button to choose the array presented on the left side of the screen or the *M* button to choose the array on the right side of the screen. Once the preferred array was chosen, participants then read the selected array type aloud. Emphasis was added in the instructions for both tasks to not hit the *B* button prematurely to avoid incomplete responses. Participants completed 32 randomized trials of the reading task and 48 randomized trials of the DST. Order of task was counterbalanced across participants. The entire study took approximately 45 minutes to complete.

Results

Results are reported first for the reading task followed by the DST, followed by Bayesian analyses of effect sizes. All analyses employed repeated measures ANOVA. Effect sizes are presented with all ANOVA results (eta-squared for one-way models) and pairwise comparisons (Cohen's *d* using SD_{avg} as the standardizer term; see Cumming, 2012, p. 291). In addition, 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals (DiCiccio & Efron, 1996) computed for the mean difference across groups are presented for all pairwise comparisons. Greenhouse-Geisser corrections are used where applicable.

Reading Task

Response times (RT) were calculated as the amount of time between stimulus onset and the vocal onset of the last word in the array (i.e., the 25th word) using CheckVocal software (Protopapas, 2007). Grand mean outlier and by-subject within condition outlier analyses were conducted on raw RTs using a 2.5 standard deviation cut-off in both cases (Van Selst & Jolicoeur, 1994), resulting in the removal of 2% of the trials. In addition, spoiled trials in which participants cut their vocal responses off early (i.e., pressing the *B* button before they had finished responding) were removed (8% of trials). Errors were considered anytime a participant misread a word, skipped reading a word, repeated a correctly read word, or navigated the array in an irregular fashion. Trials in which one or more errors occurred were not removed from RT analysis because removal of such trials would result in a large proportion of trials being removed given that each array type affords 25 individual chances to make an error. Furthermore, previous studies using large multi-element displays in reading tasks have not found differences in RT analyses when trials with errors are included or removed (e.g., Kolers, 1975).

Response Time (RT). A one-way repeated measures ANOVA demonstrated a significant effect of array type on participants' response times, $F(3,105) = 13.15$, $MSE = 421247.87$, $\eta^2 = .27$, $p < .001$. Critically, comparisons did not demonstrate a significant difference in RT for the RW-RF condition ($M = 16551$ms, $SD = 3189$) relative to the RW-UF condition ($M = 16419$ ms, $SD = 3270$), $M_{Diff} = 132$ ms, $t(35) = 0.97$, 95% BCa CI [-109.85, 411.81], $d = .04$, $p = .34$. Comparisons for all arrays relative to the UW-UF condition ($M = 15698$ ms, $SD = 2989$) demonstrated at least marginal differences in RT, *Min* $d = .10$, *Max* $d = .28$, all p's < .07. Both the RW-RF and RW-UF conditions demonstrated significant differences in RT relative to the UW-RF condition ($M = 15994$ ms, $SD = 3202$), $d = .13$, $p < .001$, and $d = .17$, $p < .01$, respectively (see Figure 2). Thus, similar to Dunn and Risko (2016), results demonstrated that the RTs associated with the RW-RF condition are similar to the RW-UF condition, with both of these RTs being longer than in the UW-RF condition, and all being longer than in the UW-UF condition.

Accuracy. A one-way repeated measures ANOVA did not demonstrate a significant effect of array type on errors, $F(1,105) = .14$, $MSE = .12$, $\eta^2 = .004$, $p = .94$. Similar to RT, comparisons did not demonstrate a significant difference in errors for the RW-RF condition ($M = 1.01$, $SD = .69$) relative to the RW-UF condition ($M = 1.06$, $SD = .73$), $M_{Diff} = -.05$, $t(35) = -.61$, 95% BCa CI [.09, -.20], $d = .06$, $p = .54$. Qualitatively, errors per trial did not vary as a function of array type, with individuals making approximately one error per trial for all arrays (UW-UF, $M = 1.02$, $SD = .75$; UW-RF array, $M = 1.01$, $SD = .77$)

Demand Selection Task

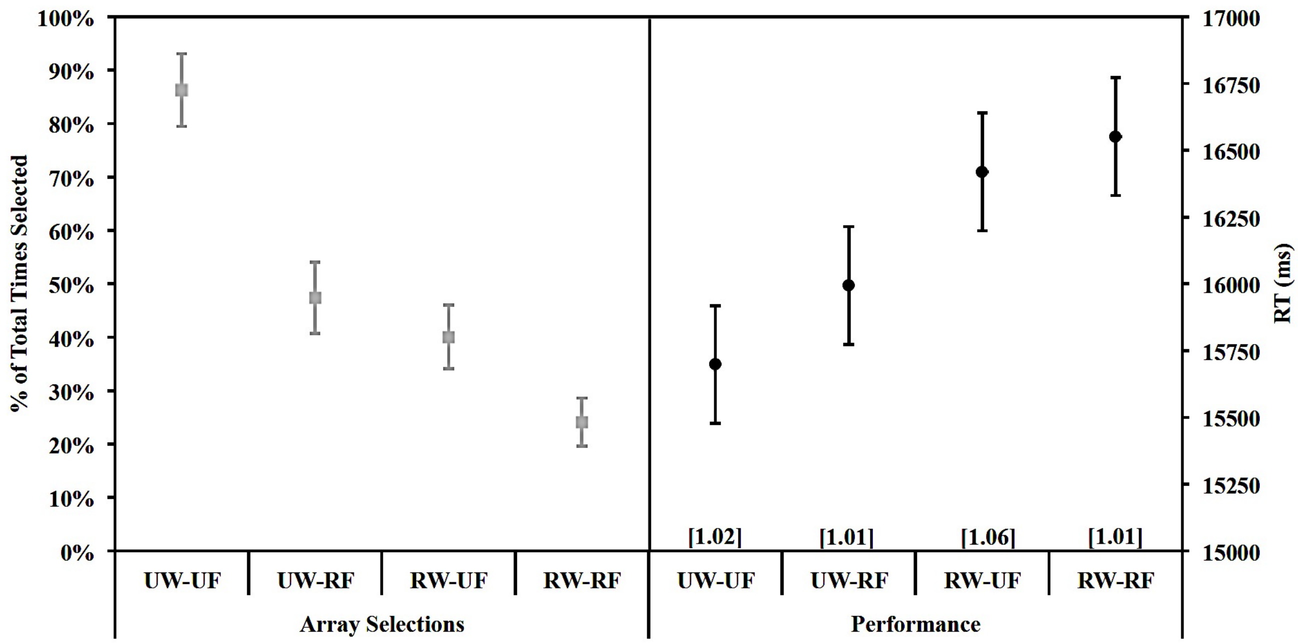
Responses where individuals mistakenly hit the *B* button rather than selected an array were removed for all analyses (hitting the *B* button at this point would simply abort the trial). Array selection times were calculated as the time from the selection screen onset to when the participant made their manual response to select an array. These selection times were highly skewed (*Min* = 8.29 ms, *Max* = 44135.78 ms, Skewness = 14.23), and were thus trimmed using a trial-level grand mean outlier procedure (i.e., trimming values above and below 2.5 SD units; Van Selst & Jolicoeur, 1994) to exclude extremely short and long selection times (approximately 2% of all trials). Proportions of array selections were computed using the trimmed data set.

First, I examined the effect of each array type on selections relative to the total number of times the array appeared within the experiment. A one-way repeated measures ANOVA demonstrated a significant effect of array type on participants' selections, $F(3, 105) = 55.67$,

$MSE = .05$, $\eta^2 = .61$, $p < .001$. Importantly, a pairwise comparison demonstrated a large difference in selections across the RW-RF condition ($M = 24.12\%$, $SD = 14.17\%$) relative to the RW-UF condition ($M = 40.05\%$, $SD = 19.16\%$), $M_{Diff} = 15.93\%$, $t(35) = 3.91$, 95% BCa CI [8.33%, 24.20%], $d = .96$, $p < .001$. Pairwise comparisons also demonstrated a significant difference in selections for the UW-RF ($M = 47.36\%$, $SD = 20.56\%$) relative to the RW-RF condition, $M_{Diff} = 23.24\%$, $t(35) = 5.38$, 95% BCa CI [22.07%, 23.27%], $d = 1.34$, $p < .001$, but not relative to the RW-UF condition, $M_{Diff} = 7.31\%$, $t(35) = 1.23$, 95% BCa CI [5.6%, 61.1%], $d = .37$, $p > .1$. Similar to RT, comparisons for all arrays relative to the UW-UF condition ($M = 86.28\%$, $SD = 19.18\%$) demonstrated significant differences in array selections, *Min* $d = 1.96$, *Max* $d = 3.73$, all p's < .001 (see Figure 2). Moreover, an analysis examining array selection within each of the six comparison groups yielded results consistent with the overall analysis reported above (see Table 1).

Figure 2.

Mean Overall Array Selection and Performance in Experiment 1



Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame. Results displayed are overall array selection and reading task response times and errors (in brackets) in the left and right panels respectively. Error bars for the left panel represent 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals. Error bars for the right panel represent 95% within-subjects confidence intervals (Masson & Loftus, 2003).

The current experimental design allows for the examination of effects of intra-experimental experience on array selections. Specifically, if individuals are monitoring and exploiting performance information during their selections, then I would expect those individuals that completed the reading task prior to the DST to show patterns of selections similar to the patterns of performance reported above. However, the between-subjects effect of task order (i.e., DST 1^{st} vs. reading task 1^{st}) did not demonstrate an effect on array selection, $F(1, 34) = 2.60$, $MSE = .001$, $\eta^2_p = .07$, $p > .1$. Thus, experience with the objective costs (as derived from performance) associated with processing the arrays prior to completing the DST did not seem to influence how individuals selected arrays to read.

Table 1.

Array Selection as a Function of Comparison Group for Experiment 1 and 2

		Experiment 1			Experiment 2		
Comparison Group		*M*	*d*	95% BCa CI	*M*	*d*	95% BCa CI
Group 1	*UW-UF* *UW-RF*	85% 14%	1.57	[78%, 93%]	86% 14%	1.50	[78%, 93%]
Group 2	*UW-UF* *RW-UF*	85% 15%	1.41	[76%, 93%]	89% 11%	1.59	[81%, 95%]
Group 3	*UW-UF* *RW-RF*	88% 12%	2.28	[82%, 94%]	92% 8%	2.02	[85%, 97%]
Group 4	*UW-RF* *RW-UF*	55% 45%	.14	[45%, 65%]	63% 37%	.41	[55%, 71%]
Group 5	*UW-RF* *RW-RF*	74% 26%	.87	[65%, 82%]	80% 20%	1.09	[72%, 88%]
Group 6	*RW-UF* *RW-RF*	64% 36%	.53	[56%, 72%]	60% 40%	.32	[50%, 69%]

Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame; *d* = Cohen's *d;* BCa CI = bias-corrected and accelerated (BCa) confidence intervals. Cohen's *d* is calculated based on a one-sample test against chance for the highest proportion value in each group. Confidence intervals are computed for the same highest proportion value.

Bayesian Analyses

The two critical conditions at the center of the argument in favor of metacognitive evaluations of perceived task demand playing an important role in selections are the RW-RF and RW-UF conditions, specifically, the notion that the two arrays are matched on performance measures (i.e., a null finding across performance variables). Thus, to further highlight the relative effects of each array within performance and selections, Bayesian estimation analyses (Kruschke, 2014) were conducted on response times, errors, and overall array selections using the BEST package (Kruschke, 2014) in R (R Core Team, 2014). For each dependent variable, 100,000 estimates of the effect size (calculated for all analyses as RW-RF minus RW-UF) using Markov Chain Monte Carlo (MCMC) sampling were simulated. 95% Highest Density Intervals (HDI), as well as the percentage of the effect size distribution contained within the Region of Practical Interest (ROPE) are presented (ROPE = -.1, .1; Kruschke, 2013). In addition, Bayes Factors (BF) computed using the BayesFactor package (Morey & Rouder, 2015) in R are presented. Interpretations of Bayes Factors follows the criteria outlined by Kass and Raftery (1995). The inclusion of both effect size estimates and Bayes Factors allows for a more robust picture of the effects on performance and array selection between the RW-RF and RW-UF arrays.

First, estimation of the effect size (ES) in response times demonstrated that 42% of the distribution fell within the ROPE, Mode ES = .04, 95% HDI [-.31, .38]. In addition, the computed Bayes Factor for RT across the two arrays demonstrated positive evidence for the null, $BF_{Null} = 3.61$. Analyses of errors produced similar results to RT, with 41% of the distribution falling within the ROPE, Mode ES = -.07, 95% HDI [-.41, .31], $BF_{Null} = 4.69$. Last, the estimation of effect sizes in array selection demonstrated that 0% of the distribution fell within the ROPE, Mode ES = -.69, 95% HDI [-1.1, -.32.]. In this case, the computed Bayes Factor demonstrated strong evidence for the alternative, $BF_{Alt} = 66.67$. Thus, considering the RW-RF and RW-UF arrays, the effect sizes reported for performance can be considered negligible, whereas the effect size reported for individuals' selections could be considered medium in size and reliable.

Discussion

Experiment 1 confirmed the hypothesis that individuals would select arrays based on the previously reported patterns of perceived demand (Dunn & Risko, 2016) associated with processing the arrays despite there being no time- or accuracy-based reason to do so. Individuals most often avoided reading the RW-RF array, followed by the RW-UF array, UW-RF array, and lastly the UW-UF array. However, individuals' performance in terms of RT and errors did not play this pattern out. Effect sizes across the RW-RF and RW-UF arrays for both performance measures were negligible as demonstrated by both inferential statistics and Bayesian analyses. Thus, while performance across the two conditions is nearly equivalent, there was a marked difference in individuals' selections in the DST across the two conditions. Based on previous research demonstrating a robust difference between these two conditions, as well as the difference between UW-RF and these arrays, in terms of perceived demand (Dunn & Risko, 2016), it can be concluded that the perceived demand represents a critical factor driving selections in the DST.

Although Experiment 1 confirmed the major hypothesis put forth above that individuals will select lines of action based on perceived demand, several potential shortcomings exist. First, while the results based on inferential statistics and Bayesian analyses suggest that performance across the RW-RF and RW-UF arrays is equivalent, the RW-RF array did demonstrate slightly slower reading times (by 132 ms). Despite this difference being statistically indistinguishable from zero, there exist several demonstrations arguing that the cognitive system can be sensitive to similarly small time costs (e.g., Gray et al., 2006; Gray & Boehm-Davis, 2000). Therefore, it remains a possibility that individuals are integrating this cost into the array selection process. Second, although the pattern of array selections followed the predicted pattern based on previous research (Dunn & Risko, 2016), perceived demand was not directly measured in the sample. Thus, it remains a possibility that self-reported patterns of perceived task demand may dissociate from the patterns of array selection within the current experimental context, although the pattern here matches that previously reported. I looked to address these concerns in Experiment 2.

Experiment 2

Experiment 2 followed a similar procedure to that of Experiment 1, with several critical changes. First, all individuals completed the reading task prior to completing the DST. Thus, as opposed to Experiment 1, all individuals received experience with the objective costs associated with reading each type of array prior to making their selections in the DST. I expect the pattern of selections to mirror the pattern outlined above (i.e., UW-UF > UW-RF > RW-UF > RW-RF) given that selections for individuals that received experience from the reading task first in Experiment 1 did not differ from those for individuals that completed the DST first. In addition, upon completion of the DST, individuals provided perceived effort rankings of the four different arrays. The straightforward hypothesis is that rankings should closely track selections in the DST. That is, the RW-RF array should be ranked as the most effortful, followed by the RW-UF array, UW-RF array, and lastly the UW-UF array.

Moreover, an additional measure of objective demand was considered alongside RT and errors. Assessing how peripheral physiological measures track demand avoidance behaviors represents a theoretically important issue (Westbrook & Braver, 2015). Thus, individuals' eye blink data was collected during the reading task. Rates of eye blinks have been consistently shown to be a sensitive measure of demand during visual tasks (Drew, 1951; Fogarty & Stern, 1989; Recarte et al., 2008; Ryu & Myung, 2005; Stern et al., 1994; Stern, Walrath, & Goldstein; 1984; Veltman & Gaillard, 1998). Specifically, blink suppression (i.e., lower blink rates) is often observed for more demanding tasks. For example, Wilson (2002) found suppressed blink rates in professional pilots during flight segments that required more attention to cockpit instruments. Therefore, given the demonstration of equivalent objective demand (as indexed by performance) across the RW-RF and RW-UF arrays, I can derive a straightforward hypothesis regarding blink rates and the utilized arrays: I would expect the UW-UF array to show the highest blink rates (i.e., less blink suppression), followed by the UW-RF array, with blink rates being similar across the RW-UF and RW-RF arrays (i.e., RW-RF = RW-UF < UW-RF < UW-UF). That is, similar rates of blink suppression should be observed for the RW-RF and RW-UF arrays. Thus another dissociation is predicted, this time between a yet to be tested measure of demand (i.e., blink rates) and selections in the demand selection task. Observing this second dissociation would lend additional credence to the notion that individuals' selections are driven by a metacognitive evaluation of perceived task demand and, to our knowledge, would be the first such demonstration utilizing a peripheral physiological measure.

Method

Participants

Forty-eight University of Waterloo undergraduate students participated in the study in exchange for research credit. The decision was made to increase sample size to achieve more precise estimates of the effects across arrays specifically for performance measures relative to Experiment 1. One individual could not be included in the following analyses due to a program error, resulting in a final *N* of 47.

Design

For both the demand selection task (DST) and the reading task, a one factor (Array Type: Upright Word-Upright Frame: UW-UF, Upright Word-Rotated Frame: UW-RF, Rotated Word-

Upright Frame: RW-UF, Rotated Word-Rotated Frame: RW-RF; see Figure 1) within-subject design was employed.

Apparatus

All apparatus was the same as in Experiment 1. To record individuals' eye blinks, a webcam was placed on the top of the monitor facing the individual. No attempt was made to hide the camera.

Stimuli

Stimuli were the same as Experiment 1.

Procedure

The procedure for Experiment 2 was similar to that of Experiment 1. However, all individuals completed the tasks within the experiment in the same order. First, individuals completed the reading task, followed by the DST, and lastly the effort rankings. For effort rankings, individuals first were randomly presented each of the four array conditions at both rotation angles (i.e., ±60°, not including the UW-UF array) simultaneously. Each set of arrays was then randomly displayed at one of four quadrants on the screen labeled A through D (i.e., individuals saw all four array types together when producing their rankings). Individuals were then asked to rank each array type from most effortful (1) to least effortful (4) to read. The entire experimental session took approximately 45 minutes to complete.

Results

Results are reported first for the reading task (i.e., RT, accuracy, blink rate) followed by the DST, Bayesian analyses, and finally subjective effort. Perceived effort ratings are reported as reversed-scored ranks (i.e., initially lower ranks denoted higher effort in the task). All reporting procedures for results are similar to Experiment 1.

Reading Task

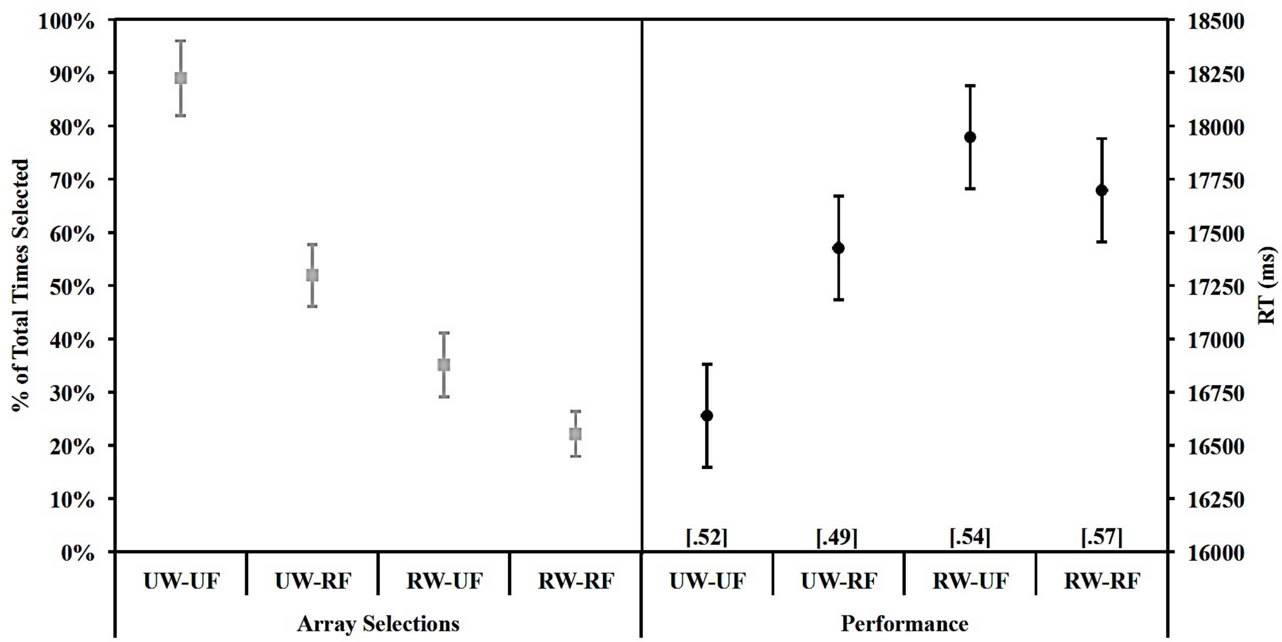
Outlier analyses resulted in the removal of approximately 2% of the trials. Spoiled trials in which participants cut their vocal responses off early were removed (approximately 4% of trials). Eye blinks were coded by individuals blind to the condition on a trial-by-trial basis. The original coder and an additional coder recoded 25% (12 individuals) of participants' eye blink data. Both inter- and intra-rater reliability was high, $K = .95$, for both respectively. To control for trial level time differences, blink rates were divided by the trial RT (i.e., blinks-per-second; BPS). Five individuals did not consent to be filmed during the experimental session and are thus not included in the blink rate analysis ($N = 42$).

Response Time (RT). Results demonstrated a significant effect of array type on participants' response times, $F(2.57, 118.08) = 22.41$, $MSE = 788971.1$, $\eta^2 = .33$, $p < .001$. Importantly, comparisons did not demonstrate a significant difference in RT for the RW-RF condition ($M = 17698$ ms, $SD = 3549$) relative to the RW-UF condition ($M = 17947$ ms, $SD = 3575$), $M_{Diff} = 249$ ms, $t(46) = 1.69$, 95% BCa CI [-30.19, 554.36], $d = .07$, $p = .09$, and critically, the direction of the difference between the arrays flipped relative to Experiment 1. Both the RW-RF and RW-UF conditions demonstrated significant differences in RT relative to the UW-RF condition ($M = 17426$ ms, $SD = 3383$), $d = .08$, $p = .048$, and $d = .15$, $p < .01$, respectively. Comparisons for all arrays relative to the UW-UF condition ($M = 16638$ ms, $SD = 3529$) demonstrated significant differences in RT, *Min* $d = .23$, *Max* $d = .37$, all p's $< .001$ (see Figure 3). Thus, the pattern of RT results replicated the pattern demonstrated in Experiment 1.

Accuracy. Results did not demonstrate a significant effect of array type on errors, $F(2.52,116.23) = .14$, $MSE = .09$, $\eta^2 = .01$, $p > .1$. Similar to Experiment 1, errors per trial did not vary as a function of array type, with individuals making less than one error per trial for all arrays (RW-RF, $M = .57$, $SD = .78$; RW-UF, $M = .54$, $SD = .92$; UW-RF, $M = .49$, $SD = .8$; UW-UF, $M = .52$, $SD = .8$)

Figure 3.

Mean Overall Array Selection and Performance in Experiment 2

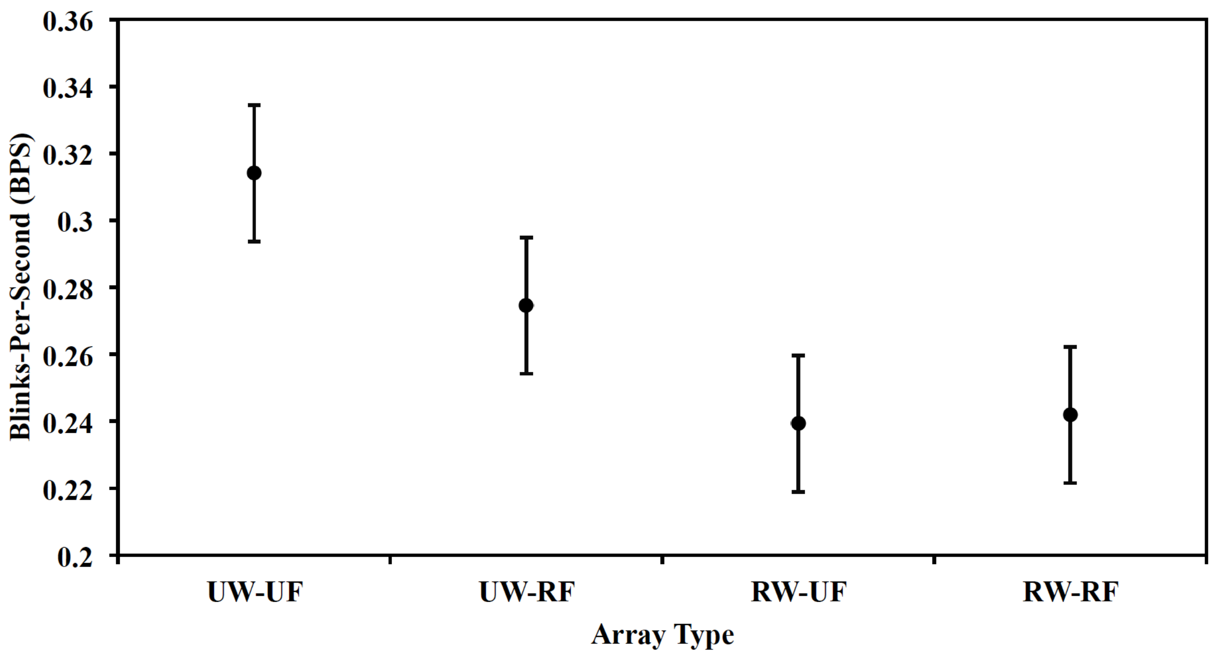


Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame. Results displayed are overall array selection and reading task response times and errors (in brackets) in the left and right panels respectively. Error bars for the left panel represent 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals. Error bars for the right panel represent 95% within-subjects confidence intervals (Masson & Loftus, 2003).

Eye-Blink Rates (BPS). Results demonstrated a significant effect of array type on BPS, $F(2.48, 101.68) = 26.93$, $MSE = .002$, $\eta^2 = .4$, $p < .001$. Again, comparisons did not demonstrate a significant difference in BPS for the RW-RF condition ($M = .24$, $SD = .15$) relative to the RW-UF condition ($M = .24$, $SD = .15$), $M_{Diff} = .003$, $t(41) = .35$, 95% BCa CI [-.01, .02], $d = .02$, $p > .1$. Both the RW-RF and RW-UF conditions demonstrated significant differences in RT relative to the UW-RF condition (*M* = .27, *SD* = .16), *d* = .21, *p* < .01, and *d* = .23, *p* < .01, respectively. Comparisons for all arrays relative to the UW-UF condition (*M* = .31, *SD* = .18) demonstrated significant differences in BPS, *Min d* = .23, *Max d* = .43, all *p*'s < .001 (see Figure 4). Therefore, BPS closely followed the hypothesized pattern above (i.e., UW-UF > UW-RF > RW-UF = RW-RF).

Figure 4.

Mean Blink Rates (Blinks-per-Second) in Experiment 2



Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame. Error bars represent 95% within-subjects confidence intervals (Masson & Loftus, 2003).

Demand Selection Task

Array selection times again showed signs of skewness (*Min* = 223.26, *Max* = 8677.9, Skewness = 2.31), and were trimmed to exclude extremely short and long selection times (approximately 2% of all trials). Proportions of array selections were computed using the trimmed data set.

Results demonstrated a significant effect of array type on participants' overall selections, $F(2.63, 120.84) = 76.43$, $MSE = .06$, $\eta^2 = .62$, $p < .001$. Similar to Experiment 1, a pairwise comparison demonstrated a large difference in selections across the RW-RF condition (M = 22.14%, SD = 15.7%) relative to the RW-UF condition (M = 35.11%, SD = 20.57%), M_{Diff} = 12.98%, $t(46) = 3.43$, 95% BCa CI [5.72%, 19.99%], $d = .72$, $p < .01$. Pairwise comparisons also demonstrated a significant difference in selections for the UW-RF (M = 51.92%, SD = 20.04%) relative to the RW-RF condition, M_{Diff} = 29.79%, $t(46) = 7.52$, 95% BCa CI [22.45%, 37.27%], $d = 1.67$, $p < .001$, and RW-UF condition, M_{Diff} = 16.81%, $t(46) = 3.21$, 95% BCa CI [6.67%, 26.07%], $d = .83$, $p < .01$. Comparisons for all arrays relative to the UW-UF condition (M = 88.97%, SD = 21.91%) demonstrated significant differences in array selections, *Min* $d = 1.77$, *Max* $d = 3.55$, all p's < .001 (see Figure 3). Array selection within each of the six comparison groups again was consistent with the patterns of overall array selection (see Table 1).

Bayesian Analyses

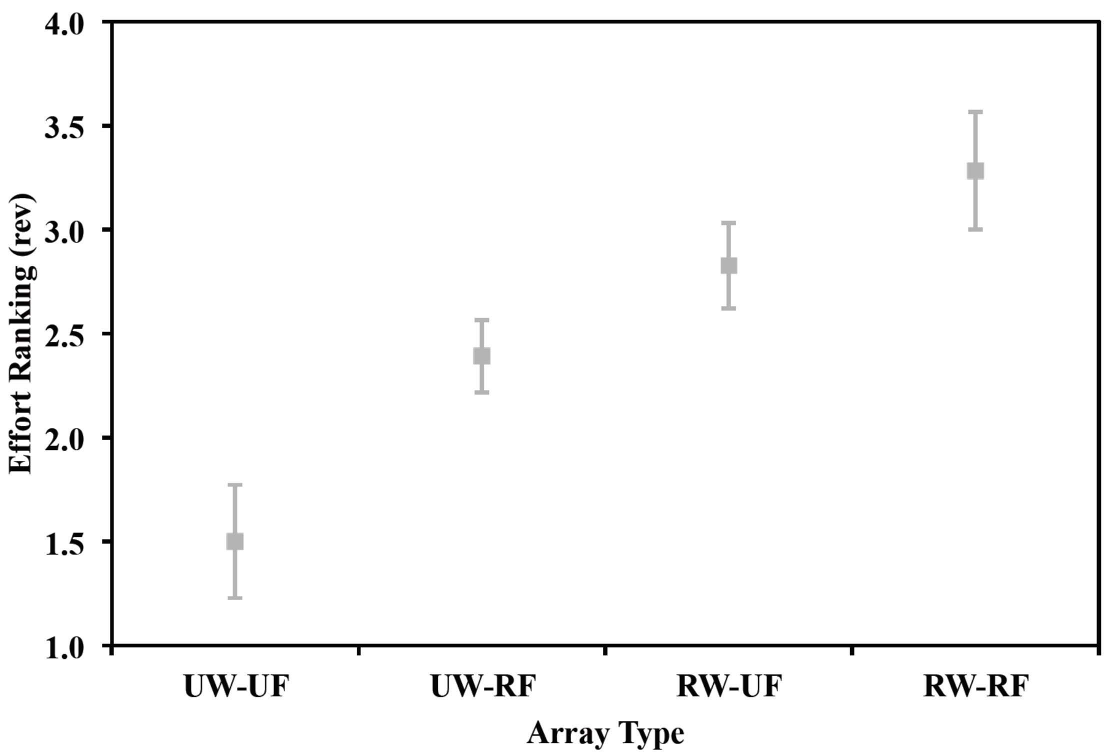
First, estimation of the effect size (ES) in response times demonstrated that 46% of the ES distribution fell within the ROPE, Mode ES = -.06, 95% HDI [-.36, .24]. In addition, the computed Bayes Factor for RT across the two arrays demonstrated no evidence for the alternative or the null, $BF_{Null} = 1.69$. Effect size analysis of errors produced similar results to RT, with 47% of the distribution falling within the ROPE, Mode ES = .03, 95% HDI [-.26, .35], but showed positive evidence for the null, $BF_{Null} = 5.79$. Analyses of BPS demonstrated that 16% of the distribution fell within the ROPE, Mode ES = -.25, 95% HDI [-.58, .06], $BF_{Null} = 5.67$. Last, the estimation of effect sizes in array selection demonstrated that 0% of the distribution fell within the ROPE, Mode ES = -.6, 95% HDI [-.99, -.25.]. The computed Bayes Factor demonstrated strong evidence for the alternative, $BF_{Alt} = 23.35$. Therefore, similar to Experiment 1, the effect sizes reported for performance as well blink rates can be considered negligible, whereas the effect size reported for individuals' selections can be considered medium in size and reliable.

Perceived Effort

Rank data were submitted to a Freidman's test. Results demonstrated a significant effect of array type on individuals' perceived effort rankings, $X^2(3) = 47.77$, $p < .001$. Wilcoxon rank-sum tests were conducted to assess differences in mean ranks across array types. First, the RW-RF array (*M = 3.28,* SD = 1.05) was ranked as significantly more effortful than the RW-UF array (*M = 2.83,* SD = .74), $Z = 2.22$, $r = .27$, $p = .03$, and the UW-RF array (*M = 2.4,* SD = 1.2), $Z = 3.52$, $r = .36$, $p < .001$. In addition, the RW-UF array was ranked as significantly more effortful than the UW-RF array, $Z = 2.21$, $r = .23$, $p = .03$. All disoriented arrays were ranked as more effortful relative to the UW-UF array (*M* = 1.5*,* SD = 1.09), all *Z'*s > 3.79, *r*'s > .39, *p'*s < .001. Thus, the RW-RF was ranked as the most effortful array to read followed by the RW-UF, UW-RF, and UW-UF arrays. Importantly, this pattern of results tracks the inverse pattern of individuals' array selections in Experiment 1 (see Figure 5).

Figure 5.

Mean Effort Rankings (reversed scored) in Experiment 2



Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame. Error bars represent 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals.

Discussion

First and foremost, Experiment 2 provides a clear replication of the major findings from Experiment 1. Individuals most often selected to read the UW-UF array followed by the UW-RF, RW-UF, and RW-RF arrays. Results from performance measures again demonstrated equivalent response times across the RW-RF and RW-UF arrays, with both producing longer RTs relative to the UW-RF and UW-UF arrays. Importantly, the direction of the relation between the RW-RF and RW-UF arrays flipped relative to Experiment 1. That is, the RW-UF array demonstrated slightly longer reading times relative to the RW-RF array (249 ms). If individuals were sensitive to this time cost, then I would have expected the RW-UF array to be chosen less often relative to the RW-RF array, though this notion was not supported. In addition, the added measure of demand (i.e., blink rates) supported the notion that the RW-RF and RW-UF arrays produce similar levels of demand, with both being greater than the UW-RF and UW-UF arrays. Thus, tracking performance, demand avoidance in terms of the DST was dissociated from this peripheral physiological measure. This is the first such dissociation of which I am aware. Last, individuals' perceived effort rankings closely matched array selections; the RW-RF array was most often ranked as the most effortful, followed by the RW-UF, UW-RF, and UW-UF arrays. This pattern replicated the pattern of perceived demand as well as perceived time and accuracy across three experiments reported in Dunn and Risko (2016). This general pattern has now been demonstrated across rankings, subjective self-report ratings, and selections in a DST, and dissociated from objective performance in terms of response times and errors and now also from blink rates. Thus, if selections during the DST can be viewed as a proxy for demand avoidance, then action selection in this context seems to be largely driven by a metacognitive evaluation of perceived task demand.

Experiment 3

Throughout the current chapter, I have put forth the argument that individuals make use of a general metacognitive evaluation of perceived task demand during action selection in demand avoidance. As highlighted above, recent work by Dunn and Risko (2016) found similar patterns of ratings of perceived effort, time, and accuracy for perceptual stimuli. Therefore, this account would predict that if individuals were forced to make selections based on a given performance dimension (i.e., effort, time, accuracy) rather than selecting based on a free-choice preference, then patterns of selections should be similar across all dimensions. Although the similar patterns across performance dimensions have been observed in self-report ratings of demand (Dunn & Risko, 2016), these patterns have not been demonstrated in a decision-making context (see above) and thus utilization of a forced-choice DST paradigm affords the opportunity to generalize the expected pattern to the perhaps more revealing decision-making context.

To address this decision-making situation, participants in Experiment 3 first completed a reading task similar to Experiments 1 and 2 online via Amazon Mechanical Turk. Amazon Turk workers have been shown to be much more demographically diverse than university samples and to provide reliable data when benchmarked against traditional experimental methods (Buhrmester, Kwang, & Gosling, 2011). At the conclusion of the reading task, individuals completed a forced-choice version of the DST where instructions explicitly stated to select on every trial, based on the condition to which they were assigned, which of the two presented arrays was the most effortful, time demanding, or least accurate to read. Following our assumption that individuals exploit a general metacognitive evaluation during selections, I do not expect selections to vary as a function of rating dimension (i.e., effort, time, and accuracy), but rather the RW-RF array will be *most* often selected, followed by the RW-UF, UW-RF, and UW-UF arrays. Importantly, however, if one dimension deviates from the patterns of selections observed in the free-choice DST, then it can be taken as evidence that individuals may not be attempting to integrate that dimension into their evaluations during free-choice.

Method

Participants

One hundred and eight individuals were recruited from Amazon Mechanical Turk. All participants were over the age of 18 and native English speakers. Participants received $2.00 as compensation upon completion of the task.

Design

For the reading task, a one factor (Array Type: Upright Word-Upright Frame: UW-UF, Upright Word-Rotated Frame: UW-RF, Rotated Word-Upright Frame: RW-UF, Rotated Word-Rotated Frame: RW-RF; see Figure 1) within-subject design was employed. For the DST, a 3 (Dimension: Effort, Time, Accuracy) by 4 (Array Type: UW-UF, UW-RF, RW-UF, and RW-RF) mixed design was employed.

Stimuli

Stimuli for the two tasks were the same as in Experiments 1 and 2.

Procedure

Participants chose and accepted the human intelligence task (HIT) on Amazon Mechanical Turk and electronically provided informed consent. Participants first completed 16 randomly presented trials of the reading task with each array type occurring four times. Instructions for the reading task stated that participants were to read through each array as quickly and accurately as they were able to themselves, and to try their best to keep their head upright while reading. In addition, in attempt to ensure that some time was spent actually reading the arrays participants were told that they would be timed while reading [1]. Individuals then completed 36 trials of the DST similar to Experiments 1 and 2, however choice instructions varied based on the rating dimension (i.e., Effort, Time, Accuracy) to which they were randomly assigned. For the effort rating condition, individuals were asked, "*Which array do you believe would be more effortful in reading aloud?*" For the time rating condition, individuals were asked, "*Which array do you believe would be more time demanding in reading aloud?*" For the accuracy rating condition, individuals were asked, "*Which array do you believe would be less accurate in reading aloud?*" Upon completion of the DST, all individuals were debriefed and required to enter a randomized code into Amazon Mechanical Turk to receive remuneration.

Results

Demand Selection Task

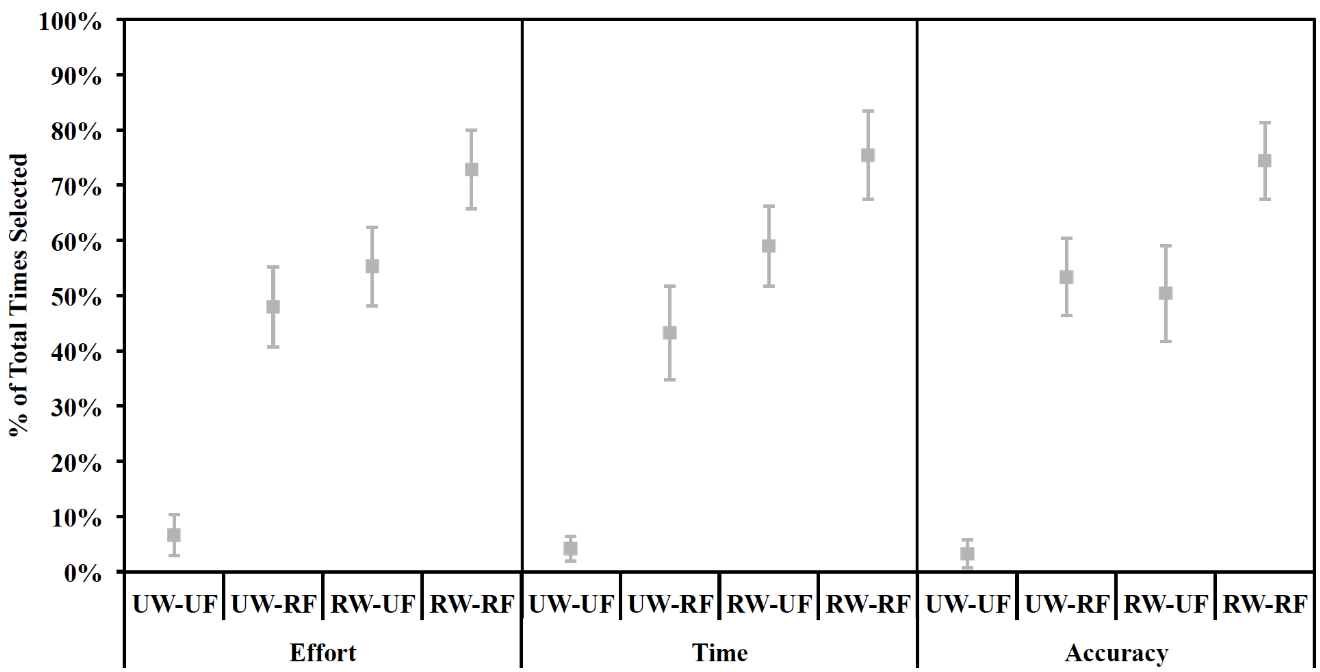
First, results did not demonstrate a significant a main effect of Dimension, $F(2, 105) = .88$, $MSE = .0004$, $\eta^2_p = .02$, $p > .1$ or a Dimension x Array interaction, $F(2.23, 237.29) = .99$, $MSE = .08$, $\eta^2_p = .02$, $p > .1$. However, results did demonstrate a significant main effect of array type on selections, $F(2.26, 237.29) = 157.82$, $MSE = .0004$, $\eta^2_p = .6$, $p < .001$ (see Figure 6). Collapsing across rating dimensions, a pairwise comparison demonstrated a large difference in selections across the RW-RF condition ($M = 74.22\%$, $SD = 22.99\%$) relative to the RW-UF condition ($M = 54.87\%$, $SD = 23.35\%$), $M_{Diff} = 19.25\%$, $t(107) = 5.1$, 95% BCa CI [11.95%, 27.9%], $d = .84$, $p < .001$. Pairwise comparisons also demonstrated a significant difference in

[1] There is considerable lack of control with regard to participants' RTs from the online task. For example, differences in Internet speed could cause stimuli to load slower, or individuals could simply click through the arrays without responding. Even after aggressive outlier trimming, mean RTs for each array type were much faster and produced much larger SDs than the two controlled reading tasks reported in Experiments 1 and 2 (RW-RF, M =11676 ms, SD = 6496; RW-UF, M =12565 ms, SD = 7294; UW-RF, M =12395 ms SD = 7468; UW-UF, M =11204 ms, SD = 7054). Thus, I do not present inferential statistics for RTs in the reading task.

selections for the UW-RF (M = 48.16%, SD = 24.37%) relative to the RW-RF condition, M_{Diff} = 26.06%, t(107) = 6.62, 95% BCa CI [19.29%, 33.26%], d = 1.1, p < .001, but not relative to the RW-UF condition, M_{Diff} = 6.71%, t(107) = 1.67, 95% BCa CI [.6%, 14.23%], d = .28, p = .1. Comparisons for all arrays relative to the UW-UF condition (M = 4.67%, SD = 10.35%) demonstrated significant differences in selections, $Min\ d$ = 2.51, $Max\ d$ = 4.18, all p's < .001.

Figure 6.

Mean Overall Array Selections in Experiment 3



Note: UW-UF = Upright Word-Upright Frame; UW-RF = Upright Word-Rotated Frame; RW-UF = Rotated Word-Upright Frame; RW-RF = Rotated Word-Rotated Frame. Error bars represent 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals.

Discussion

Experiment 3 looked to directly address the assumption that individuals make use of a general metacognitive evaluation of perceived task demand during action selection in a forced-choice DST where they were asked to select the more effortful, the more time demanding, or the least accurate array to read. Results supported this notion and replicated the pattern of self-report in Dunn and Risko (2016) in a decision-making context. Patterns of selections did not vary as a function of rating dimension. Individuals selected the RW-RF array to be more effortful, time demanding, and less accurate to read relative to the RW-UF array. Thus, forcing individuals to choose based on a specific conceptualization of demand did not change patterns of selections and, interestingly, demonstrated a similar pattern of results relative to individuals' selections in the free-choice DST. Within the context of DSTs, individuals may indeed be attempting to integrate some type of performance information into their selections. However, the addition of ancillary metacognitive information (e.g., explicit cues) may lead to an "error" in selections as defined by objective demands. In a similar vein, theories incorporating the notion that metacognitive experiences are based on an evaluation of one's performance in light of some metacognitive theory can be found within the literature (Whittlesea, 1997; 2003). It is important to note, though, that interpreting the similar patterns of selections across dimensions as evidence for a general evaluation of demand, does not support a claim that *all* dimensions are necessarily included in the evaluation all the time.

General Discussion

Chapter 1 looked to closely examine the role of metacognitive evaluations of task demand in a demand selection task (DST). To achieve this, a novel DST was employed utilizing stimuli that yield equal objective costs but varying levels of perceived task demand. Importantly, individuals actively avoided selecting to read arrays based not on objective costs but on metacognitive evaluations of task demand. Specifically, individuals less often selected to read the RW-RF array over the RW-UF array, followed by the UW-RF and UW-UF arrays.

Performance in the reading task, however, revealed that RT and errors for the RW-RF and RW-UF conditions were similar. Thus, the pattern of array selections was dissociated from the pattern of objective performance as indexed by the reading task. Experiment 2 provided a straightforward replication of the critical findings demonstrated in Experiment 1. Moreover, an additional peripheral physiological measure of demand (i.e., eye blinks) nicely dovetailed with the notion that objective demand was equivalent across the RW-RF and RW-UF arrays. In addition, individuals' perceived effort rankings closely tracked selections in the DST. Experiment 3 looked to address the notion that individuals base their selections on a general metacognitive evaluation of perceived demand by forcing individuals to make selections based on effort, time, and accuracy. Patterns of selections were similar across all dimensions and matched patterns observed in the free-choice DSTs suggesting that a type of general metacognitive evaluation is being exploited during the selection process. In the following, I consider the above results within a metacognitive framework of demand avoidance, provide connections to extant theories, and provide testable predictions for future work.

Toward A Metacognition of Demand Avoidance

The experiments in Chapter 1 provide evidence consistent with the notion that individuals avoid perceived demand independently of "demand" as indexed by time-on-task or performance more generally in a DST[2]. In addition, I have provided early evidence of demand avoidance being dissociated from a physiological measure of demand. I suggest that demand avoidance is driven by an individual's metacognitive evaluation of perceived task demand elicited by available cues. As highlighted in the introduction, these results suggest that one potentially productive avenue to pursue with regard to demand avoidance behavior is under a metacognitive control framework (Dunn & Risko, 2016). From this framework, selecting courses of action based on demand can be considered largely inferential. As such, individuals may rely on multiple cues during the action selection process (e.g., Payne, Bettman, & Johnson, 1992), and evaluations are influenced by factors such as preconceived biases, intuitive theories, and past experience (Dunn & Risko, 2016; Koriat, 2007). Furthermore, dissociations across objective demands and demand avoidance behavior could be considered the rule rather than an exception (cf. Koriat et al., 2006).

The results in Chapter 1 are consistent with such an account. The RW-RF and RW-UF arrays are read in about the same amount of time, generate about the same number of errors, and generate similar responses using a physiological measure of demand (i.e., blink rates). It would be difficult to make the case that the two conditions differ markedly in any kind of objective

[2] Following Kool et al. (2010), I conducted an individual differences analysis on individuals' selections in the DST and, in this case, rotation costs. Performance from the reading task and DST selections was pooled across Experiments 1 and 2 ($N = 83$). Correlations were conducted within each of the comparison groups incorporating disoriented arrays using a change score for both arrays (i.e., RW-RF – RW-UF; RW-RF – UWRF; RW-UF – UW-RF) for each variable of interest. We would expect moderate negative correlations between performance and selections for those individuals that showed larger rotation costs for one array relative to the other within specific comparison groups. However, all correlations were extremely small, *Min* $r = -.01$, *Max* $r = -.13$, all p's $> .1$. Blink rates in Experiment 2 showed similar results, *Min* $r = -.13$, *Max* $r = -.24$, all p's $> .1$. Although these findings strengthen the overall argument presented here, they must be interpreted with caution given major issues concerning small sample size correlation analyses (Schönbrodt & Perugini, 2013) and utilizing change scores to index costs (Peter, Churchill, & Brown, 1993).

costs. Critically, though, participants report that reading the RW-RF display is more effortful, more time consuming, and more error prone than reading the RW-UF displays. This belief likely derives from the intuitive notion that stimulus rotation impairs performance, and that the RW-RF array contains two visually salient rotations. Importantly, when faced with a free decision about which display to read, participants overwhelmingly avoid the RW-RF display. In this vein, the fact that the RW-RF condition is not associated with more time or errors is evidence for the contribution of some extra-experimental factor (e.g., preconceived bias, theory, past experience).

This explanation above also goes some way to explaining the fact that there was a relatively small difference in the selections between the RW-UF and UW-RF displays despite large differences in performance. Critically, both displays feature the disorientation of a single dimension and on that standard alone could be perceived as approximately equally effortful, time consuming, and error prone. The objectively greater cost associated with rotating words relative to rotating the frame is to some extent "lost" on participants.

What are the implications of adopting a metacognitive approach for the idea that individuals avoid demands based on executive control (Botvinick & Rosen, 2009; Gold et al., 2014; Kool et al., 2010; McGuire & Botvinick, 2010)? One perspective is that a kind of sheer effort cost signal associated with demand on executive control exists and that this is exploited as a type of cue in action selection during the course of a DST. Importantly, a metacognitive control account of demand avoidance, specifically a bi-directional and sequential model of monitoring and control (Koriat & Levy-Sadot, 2001; Son & Metcalfe, 2005; Vernon & Usher, 2003), allows for the updating of an existing theory based on feedback from control. In those cases where variations in performance costs become explicit, feedback from the ultimate selection of a condition could subsequently inform further monitoring and theories about said condition. As an example, Kool et al. (2010) demonstrated that individuals who exhibited larger task switching costs subsequently exhibited more extreme effort avoidance behavior.

Alternatively, avoidance of task switching in previous experiments may have been driven by a preexisting theory that task switching is demanding, and individuals thus avoid the option associated with more switches. Gold et al.'s (2014) demonstration that avoiding task switching was contingent on providing instructions about differences in effort across two options may reflect those instructions sensitizing individuals to variations in the probability of a task switch across decks. Importantly, although I have argued here that demand avoidance is driven by an individual's metacognitive evaluation of perceived demand, such a view does not preclude the notion of individuals avoiding cognitive effort as defined by variation in executive control or some other conceptualization of effort more generally. Rather, and critically, my argument is that the driving factor of avoidance is the metacognitive evaluation.

Subjective Control Costs as Metacognitive Evaluations

An interesting question to address moving forward is what comprises a subjective control cost signal? As highlighted in the introduction, such subjective decision costs have been associated with increased activity in the lateral prefrontal cortex (PFC; Mcguire & Botvinick, 2010). I have suggested above that individuals utilize a type of general metacognitive evaluation of task demand comprised of several types of objective information influenced by cues, theories, and biases, leading to perceived effort. Comparably, recent neuropsychological models of metacognition have also identified areas of the PFC as critical in metacognitive monitoring and control. For example, the rostrolateral PFC is hypothesized to receive input from areas of the cortex involved in a "closed-loop" of monitoring and control (i.e., the anterior cingulate cortex, ACC; see Botvinick, Cohen, & Carter, 2004, for a review), generating a metacognitive evaluation of the state of the system that can be deployed or reported beyond the local task (Fleck et al., 2006; Fleming & Dolan, 2012; Medalla & Barbas, 2010). Although substantial overlap of processes involved in the ACC (i.e., executive control) and PFC (i.e., metacognition) exists (Fernandez-Duque, Baird, & Posner, 2000), they can be dissociated (Naccache, et al. 2005). In this vein, Hebscher et al. (2015) recently demonstrated that deficits in metamemory monitoring were associated with damage to the ventromedial PFC, whereas deficits in control (e.g., withholding retrieved responses) were associated with damage to the orbitofrontal cortex (OFC). Moreover, factors such as environmental cues have been posited to activate representations within the PFC that can drive action selection (see Miller & Cohen, 2001, for a review). Therefore, unique contributions of the PFC may be an essential determinant in controlling behavior within a metacognitive account of demand avoidance. Empirically exploring this notion represents a theoretically important issue for future investigations.

A Role for Implicit Costs?

Though much of the current discussion has focused on explicit demand costs in demand avoidance, one theoretical issue in need of addressing is the role of implicit costs in driving the behavior. Westbrook et al. (2015) note that, “Cost signals need not always be conscious to influence behavior, but they may become conscious when the costs are sufficiently high” (p. 399). For example, Kool et al. (2010) reported effort avoidance for individuals who did not self-report a difference between decks in the DST (though this proportion of individuals was relatively small). The opportunity cost model (Kurzban et al., 2013) would seem to suggest that such costs would not be expected to influence effort avoidance behaviors. That is, costs that are not “felt” would not be expected to affect decision-making. The influences of such implicit costs within competing frameworks are less clear. The metacognitive account put forth here suggests that the source of the cost signal would not necessarily need to be identifiable to drive demand avoidance. For example, if an individual perceives some difference in demand across conditions but cannot necessarily identify the source of the difference (e.g., differences in task switching), that individual may simply infer one alternative is more demanding relative to the other. Therefore, if cognitive demand carries a subjective cost (Kool & Botvinick, 2013; Westbrook & Braver, 2015), then robustly detailing what entails an implicit cost under this framework constitutes a significant issue to address.

Variations in Judgments across Evaluation Modes

One potentially interesting finding emerging from the DST results in Chapter 1 is that, although selections in the task showed a qualitatively similar pattern of results relative to the subjective demand ratings reported in Experiment 3 in Dunn and Risko (2016), the effect of a rotated frame seemed to have a larger influence on selections relative to self-report ratings for arrays containing rotated frames (i.e., the UW-RF and RW-RF arrays). For example, in Experiment 3 the difference in the number of selections between the UW-RF array and the RW-UF array was smaller than the reported effect sizes for the same arrays in self-reported subjective ratings[3]. That is, the difference in selections between the UW-RF and RW-UF arrays seems less than subjective ratings would predict. In addition, when considering the RW-RF and RW-UF comparison, the difference in selections was larger in Experiment 3 relative to the average effect for this comparison reported in self-reported subjective ratings. Again, within this comparison group individuals selected the array with the rotated frame (RW-RF) more often than what

[3] Experiment 1 shows a similar pattern for this comparison group (d = .37) relative to Experiment 3, however selections decrease in this context given selections in the free-choice DST represents an *avoidance* of demand as opposed to selecting which option is more demanding. The same holds true for the RW-RF and RW-UF comparison in Experiments 1 (d = .96) and 2 (d = .72).

subjective ratings would predict. In both cases, the existence of an array with a rotated frame increased the likelihood of selections of that array relative to other arrays.

Why might the influence of a rotated frame have a larger effect on selections in the DST than it does on subjective ratings? General Evaluability Theory (GET; Hsee et al., 1999; Hsee & Zhang, 2010; Hsee, 1996) hypothesizes important differences in reported subjective value across single and joint evaluation modes. The joint evaluation mode represents instances where items are juxtaposed and evaluated comparatively. Joint evaluation is most associated with subjective values generated during the decision-making stage of choice. In contrast, single evaluation mode represents instances where items are evaluated in isolation and are associated with the experience stage of a decision (Hsee et al., 2013). Critically, GET predicts differences in evaluations (e.g., ratings or choices) across the two modes when a level of one factor is inherently inevaluable relative to another level. This "low-evaluability" level holds little prominence (e.g., low ratings) within single evaluation contexts, but becomes more evaluable (e.g., high ratings) during joint evaluation when contrasted with another level of a factor. For example, Hsee (1996) demonstrated that individuals evaluating two candidates for a computer programmer position favored the candidate with a higher GPA in single evaluation, but in joint evaluation reversed their preferences opting for the individual with more experience. Hsee (1996) concludes that the participants offering evaluations possessed richer knowledge concerning GPA relative to programming experience.

Within the current context, one can conceive of rotated frames in multi-element arrays as inherently inevaluable relative to rotated words. Specifically, in single evaluation (i.e., subjective ratings on a likert-type scale), arrays containing rotated words (RW-RF and RW-UF) produced higher ratings of effort relative to the array with just a rotated frame (UW-RF). Thus, in single evaluation, rotated words are more salient. In joint evaluation, however, rotated frames become more evaluable when contrasted with rotated words, and selections increase for those arrays (i.e., RW-RF and UW-RF) relative to rotated words only (RW-UF). Hence, a type of quantitative shift was observed for rotated frames across single and joint evaluations for subjective judgments.

This distinction between single and joint evaluation modes, together with the associated possible variations in judgments and choices, represent potentially important issues moving forward. First, individuals in trade-off contexts may more often select option X in a free-choice context suggesting that option X is the least demanding of the alternatives, but they may subsequently self-report (i.e., using likert-type scales) selections from X as being more demanding relative to the available alternative Y, thus demonstrating a type of demand *seeking* behavior (i.e., mischoice; Hsee & Zhang, 2010). Second, this approach may provide useful insight into the inherent evaluability of specific types of perceived demand (i.e., perception, memory). Specifically, investigating differences across evaluation modes may provide a clearer picture of how domain-specific demands are weighted and contribute to general metacognitive evaluations.

Conclusion

The framework developed in Chapter 1 makes clear that understanding demand-based decision-making will be critically dependent on understanding metacognitive evaluations of demand. In particular, on a metacognitive account, these evaluations play a causal role in the control of effort avoidance behaviors. Despite the importance of subjective demand to such decisions gaining in prominence (e.g., Kool & Botvinick, 2013; Westbrook & Braver, 2015; Westbrook, Kester, & Braver, 2013) there remains much work to be done. The present work has provided a step in this direction.

Chapter 2

Results from Chapter 1 are consistent with the hypothesis that individuals utilize available cues to generate their perceptions of effort and to guide choices in a DST, as opposed to indexing their performance. The aim of Chapter 2 is to extend this cue utilization account of how individuals generate their notion of effort. Specifically, I introduced demands on executive control, alongside demands on performance, as an additional source of information that individuals may use to generate their perceptions of effort. These two potential determinants were contrasted with the potential contribution of a salient effort cue that demands little in terms of executive control and is associated with relatively low performance demands, to determine which form of information drives decisions in an effort minimization context.

The following work is currently under revision at the *Journal of Experimental Psychology: Human Perception and Performance* (Dunn & Risko, submitted).
Changes have been introduced to improve the flow of the dissertation.

The idea that humans adapt their behavior in an attempt to avoid effort is indeed pervasive (e.g., Clark, 2010; Rosch, 1998; Solomon, 1948; Zipf, 1949). Given the driving role that effort avoidance is purported to play in our day-to-day lives, understanding *how* individuals decide which course of action is the least effortful – a seeming pre-requisite for successful effort avoidance – is critical. To this end, in the present chapter I consider three theoretical proposals regarding the information on which effort-based decisions may be made: time, demands placed on the executive control system, and available effort cues.

Avoiding Cognitive Effort

Understanding what individuals are minimizing when they attempt to minimize effort has been arguably dominated by two influential ideas. The first is that individuals look to minimize time (e.g., Gray & Boehm-Davis, 2000; Gray & Fu, 2004; Gray, Sims, Fu, & Schoelles, 2006; Maglio, Wenger, & Copeland, 2008; Morgan, Patrick, Waldron, King, & Patrick, 2009; Siegler & Lemaire, 1997), and the second is that individuals minimize the level of demand placed on the executive control system (e.g., Botvinick & Braver, 2015; Kool et al., 2010; Kurzban, Duckworth, Kable, & Myers, 2013; Westbrook & Braver, 2016; Yeung & Monsell, 2003).

According to accounts arguing that time acts as a key determinant in effort-based decisions, individuals will minimize effort by selecting the course of action that will take the least amount of time (while maintaining an acceptable level of performance/reward). For example, according to the Soft-constraints Hypothesis (Gray et al., 2006) individuals select courses of action that tend to minimize performance costs in terms of time. On this account, the cognitive system assigns no privileged status amongst different types of effort, such as perceptual-motor effort (e.g., making eye movements) or memorial effort (e.g., retrieving an item from memory), rather "...only milliseconds matter" (Gray et al., 2006, p. 463). Consistent with this idea, Gray and Fu (2004) demonstrated that individuals become increasingly likely to opt for a less accurate memory-based strategy as the time costs associated with a more accurate perceptual-motor strategy increase. In a similar vein, Siegler and Lemaire (1997) demonstrated that when individuals had a free choice between different multiplication strategies, the relative speed of strategies was the best predictor of what strategy would be deployed.

In addition to time playing a determining role in effort avoidance, several accounts of behavioral control argue that, when faced with a choice between different courses of action, individuals will minimize their effort by selecting an alternative that will put fewest demands on the executive control system (while maintaining an acceptable level of performance/reward). For example, Botvinick and Braver (2015) tie effort to the engagement of the executive control system spanning the dorsolateral prefrontal cortex (dPFC), anterior cingulate cortex (ACC), and intraparietal cortex. Specifically, tasks that engage the processes associated with control (i.e., the sets of superordinate functions that encode and maintain representations and feed to subordinate processes) are closely tied with cognitive effort (Botvinick & Braver, 2015).

Within an opportunity cost framework of effort, Kurzban and colleagues (2013) have argued that tasks that are associated with utilizing a wide range of cognitive processes, such as those requiring high levels of executive control (e.g., the Stroop, Simon, or Flanker tasks), are associated with higher levels of subjective effort. Thus, such tasks would be expected to be avoided during decision-making when alternative lines of action are available. Consistent with this idea, when individuals were free to choose between alternative line of actions, Kool et al. (2010) found that individuals avoided options associated with a higher probability of a task switch (i.e., the option that placed higher demands on executive control) relative to an alternative associated with a lower probability of a task switch (i.e., the option that placed lower demands on executive control). Importantly, though, in many cases increasing demands on the executive control system also increases the time associated with completing the task. Kool et al. (2010) additionally demonstrated that individuals would avoid demands on executive control at the cost of greater time spent, a result consistent with demands on executive control making a unique contribution to effort avoidance behavior beyond time demands at least in that experimental context.

Beyond Time and Executive Control

While the influence of both time and executive control demands on effort avoidance have shared support from a number of domains, recent evidence has suggested the need for a broader framework when examining such decisions. Specifically, Chapter 1 demonstrated that effort-based preferences can be dissociated from effort measured both with traditional performance measures such as time and accuracy, and with a physiological measure of demand indexed by blink rate. I argued that this dissociation was inconsistent with time and likely demands on the executive control system as well (i.e., there was no a priori reason to expect that options varied on demands on executive control) being solely responsible for effort-based decisions. An account was suggested wherein choosing the least effortful course of action is conceived of as a kind of inference based on available effort cues (see similarly Payne, Bettman, & Johnson, 1993). This account draws largely from cue-utilization models of metamemory judgments that attribute judgments to inferential or heuristic processes based on a variety of cues (Koriat, 1997) rather than on direct access to memory traces (e.g., King, Zechmeister, & Shaughnessy, 1980; Schwartz, 1994). For example, Koriat (1997) argued that individuals' judgments of learning (JOLs) were based on a variety of extrinsic, intrinsic, and mnemonic cues (e.g., repetition, relatedness to other items, cue familiarity) rather than being based on direct access to the strength of memory traces. In support of this view, Castel (2008) demonstrated that individuals utilize cues pertaining to serial position of words in a memory task in generating their JOLs when such cues were made salient. Other cues purported to support metamemory judgments include, for example, font size (Rhodes & Castel, 2008), memorization effort (Koriat, Ma'ayan, & Nussinson, 2006), and presentation time (Mazzoni, Cornoldi, & Marchitelli, 1990). Thus, in the context of effort avoidance, individuals would be expected to utilize a variety of "effort cues" (e.g., the orientation of a stimulus to be named; Chapter 1; Dunn & Risko, 2015) rather than directly assessing levels of demand when attempting to minimize effort.

Although a cue-utilization account need not deny the influence of time or demands on executive control for effort-based decisions, the shift in perspective offered by a cue-utilization account moves to center stage the need to understand the cues that individuals use and the inferences/heuristics that they employ when making effort-based decisions. It also draws attention to the fact that effort-based decisions will be subject to systematic biases, for example, the influence of misleading cues or intuitive theories pulled from past-experience (Dunn & Risko, 2016; Koriat, 1997; 2007). Critically, it is proposed that these cues drive the evaluation of effort generally (Chapter 1, see also Epstein, 2000; Mangan, 2001), incorporating several types of information such as expected performance (Whittlesea, 2003) or processing fluency (Reber, Fazendeiro, & Winkielman, 2002). As an illustration, within the context of metamemory judgments, Rhodes and Castel (2008) found that individuals assigned higher JOLs to items that were presented in larger fonts although there was no evidence of a relation between font size and recall. Similarly, Castel and colleagues (2012) found that individuals assigned higher JOLs when asked to say "YES" to specific items that were to-be-remembered (i.e., a variant of the production effect) although this action, not being item-specific, afforded no benefit in recall.

Thus, in both cases, individuals were exploiting putatively misleading or superficial cues in generating their JOL rather than directly monitoring memory strength. Applying the cue-utilization account to effort avoidance then carries over the expectation that effort-based decisions will deviate from effort as indexed by performance (Chapter 1; Kool et al., 2010; Westbrook, Kester, & Braver, 2013) as well as from demands placed on the executive control system (McGuire & Botvinick, 2010).

Present Investigation

In Chapter 2, I contrasted the influence of time and demands on the executive control system against the influence of a salient effort cue using variants of the free-choice demand selection task often used in investigations of effort-based decision-making (DST; Botvinick & Rosen, 2009; Gold et al., 2014; Kool et al., 2010; McGuire & Botvinick, 2010). This particular strategy was selected to provide a strong test of a cue-utilization account. That is, by contrasting the contributions of two factors thought to heavily drive effort-based decision-making against the contribution of putatively misleading cues, I can determine the utility of more seriously considering the latter as a major contributor in this experimental context. That said, it is important to note that the goal here is not to falsify time or executive control as being important factors in effort-based decision-making. Rather, the goal in Chapter 2 is to provide positive evidence for cue utilization by generating directly competing alternatives hypotheses.

Across the three experiments in Chapter 2, in a given block of trials, individuals were presented with a pair of tasks consisting of a high- and a low-demand option, where one option took longer and placed more demands on the executive control system relative to the alternative option. Participants were first given experience with each option (i.e., they were forced to select that option a given number of times) and were subsequently asked to generate a least-effortful preference. Performance on these “forced” trials provides an index of each option’s effort. Importantly, “effort” here is defined as the demands associated with a line of action by way of a direct measure of time and an indirect measure of demands on executive control through the proportion of task switching (see below). Participants’ choices on the “free” trials provide an index of their effort avoidance. Unlike previous research using the demand selection paradigm, the addition of an initial forced familiarization stage with each option should provide individuals ample opportunity to differentiate the different options. Furthermore, given that individuals’ preferences in free-choice contexts can be considered to be constructed on-the-fly and labile (Payne et al., 1993; Lichtenstein & Slovic, 2006), directed instructions with regard to generating a less-effortful preference (see similarly Gold et al., 2014) gives a strong indication of how each manipulation may affect individuals’ effort-based decisions. Specifically, participants should be biased toward selecting the low demand option to the extent that they perceive a difference in effort between the two options. By explicitly instructing participants to make decisions based on effort, the potential noise associated with eliciting preference-based choices (which has been utilized in previous iterations of the DST, Gold et al., 2014, Experiment 1; Kool et al., 2010) can be minimized.

The critical manipulation in the present experiments is the nature of the factor that differentiates the low- and high-demand options in each pair presented to participants. In Experiment 4, two different pairs of options were used. The first pair of options differed in whether a to-be-identified stimulus was rotated or not (i.e., the rotation pair). The time costs of rotation on object identification are small (e.g., Jolicoeur, 1990) and arguably place minimal demands on executive control. Critically, stimulus rotation represents a salient effort cue. That is, it is both highly “available” and perceived as effortful (Chapter 1; Dunn & Risko, 2015). Here I define salience specifically as the extent to which individuals are aware of either the rotation or the switching cue. To specifically index awareness, individuals completed a self-report questionnaire (see below) aimed at gauging awareness of effort cues and explicit strategies employed (similar to that employed by Kool et al., 2010, Experiment 1).

In the second pair, the options differed in the probability of a task switch (i.e., the switching pair; 90% vs. 10%). Task switching is often associated with a large time cost (e.g., Monsell, 2003), and is a representative task that engages the putatively effortful processes associated with executive control (Botvinick & Braver, 2015; Botvinick & Rosen, 2009; Kool et al., 2010; McGuire & Botvinick, 2010). Nonetheless, recent research suggests that the probability of a task switch appears to be a relatively "poor" effort cue. Individuals will often not avoid options associated with more frequent switching even with large differences in the probability of a task switch (i.e., 90% vs. 10%) and in the associated time costs and demands on executive control. For example, although Kool et al. (2010) demonstrated effort avoidance for low- and high-demand options consisting of a 10% and a 90% probability of a task switch respectively, Gold and colleagues (2014) failed to replicate these results and were only able to demonstrate similar patterns of effort avoidance using the same probabilities when individuals were explicitly instructed that a difference existed across alternatives. That is, they tended not to notice unless told to look for a difference. This lack of awareness occurs despite the large observed costs of increases in the proportion of a task switch and arguably reflects (among other possibilities) the fact that (a) task switching consists of a relation between consecutive trials, and (b) individuals are often unaware of even rather large manipulations of the proportion of a given trial type (e.g., Risko & Stolz, 2014; Schmidt, Crump, Cheesman & Besner, 2007). While proportion of task switching manipulations might be opaque to participants, the costs are large on performance and theoretically can be tied directly to increased demands of the executive control system (Monsell, 2003). Indeed, in Kool et al.'s (2010) influential work demonstrating the avoidance of cognitive demand, proportion of switches was one of the central manipulations.

It is important to note that, given the specific task switching manipulation used in the present context, the time and demands on executive control accounts will largely make the same predictions. Contrasting these two contributions has been the focus of other work (Kool et al., 2010). That said, in addition to contrasting overall performance across different proportions of task switching, I also compare the magnitude of the switch cost across these conditions. In their Experiment 3, Kool and colleagues (2010) reported smaller switch cost in a condition with a higher probability of a task switch despite performance overall being poorer in that condition relative to a condition with a lower probability of a task switch. This likely represents a type of preparation effect (Monsell, 2003) provided the variation in the likelihood of a task switch. Importantly, participants avoided the option with a high proportion of switches and lower switch cost and not the option with a low proportion of switches and an associated larger switch cost. Thus, I report switch costs as another potential executive demand to further explore.

From the perspective of time and overall demands on executive control driving effort avoidance, I would expect to observe higher rates of effort avoidance (i.e., a stronger bias toward the low demand option) in the switching pair relative to the rotation pair. On the other hand, from a cue-utilization perspective, saliency of the effort cue should contribute to effort-based decisions and be signaled by rates of effort avoidance equivalent or greater for the rotation pair which features a salient effort cue, than for the switching pair which features a less salient effort cue.

Here the “deck” is strongly stacked against the rotation pair generating much in the way of effort avoidance behavior relative to the switching pair provided that the switching pair features two options that will differ strongly in terms of time (i.e., hundreds of milliseconds) and demands on executive control (i.e., almost no task switching vs. task switching nearly every trial). Based on time and demands on executive control being the driving determinant within this context, there should be high rates of effort avoidance particularly given that I asked individuals to make their choices based on effort. In the rotation pair, the tasks should differ little in terms of time and arguably will not differ in terms of demands on executive control, barring the claim that the processes involved in dealing with modest stimulus rotation places higher demands on the executive control system than the processes involved in task switching. Against this clear difference in relative time and executive control demands, the rotation pair features only a salient effort cue (i.e., stimulus rotation). Considering the strong evidence that time and demands on executive control can drive effort-based decisions, positive evidence for the cue-utilization account in this context would represent strong support for that approach to understanding effort avoidance: that is, similar or greater effort avoidance in the rotation pair relative to the switching pair.

Experiment 4

Method

Participants

Thirty-six University of Waterloo undergraduate students participated in the study in exchange for research credit. Sample size was estimated (Power = .8, $\alpha = .05$) from a pilot study suggesting a moderate effect in low-demand selections across the rotation and switching pairs ($d = .44$). In addition, completing full condition counterbalancing was also taken into consideration.

One participant was unable to satisfy the practice criterion outlined below and a program issue caused no switch trials to be delivered to another participant for the low-demand option, thus their data are not included in the following analyses (N = 34).

Design

A 2 (Pair of Alternatives: Rotation Pair, Switching Pair) x 2 (Demand Option: Low Demand, High Demand) within-subjects design was employed.

Apparatus

The DST was programmed in MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007). Stimuli were presented on a 24" LCD monitor approximately 70 cm away from the participant. Participants used a standard optical mouse to provide responses.

Stimuli

The stimuli employed followed Kool et al. (2010; Experiment 3) and Gold et al. (2014; see Figure 7)[4]. For each pair of alternatives, individuals were presented with two options, a low-demand and a high-demand option (the relative demand was not communicated to participants), to the left and right of the center of the screen. Once the mouse was placed over a target centered between the two options, both pairs were activated for selection. Individuals were then free to move the mouse over an option to reveal a colored digit, either blue or yellow, to which a response was to be made.

[4] I thank Wouter Kool for providing the base MATLAB code for the experiment.

When digits (1, 4, 7, & 8)[5] were colored blue, individuals were tasked with making a magnitude judgment (i.e., greater than or less than) with five as the reference number. When digits were colored yellow, individuals were tasked with making a parity judgment (i.e., odd or even). The left mouse button served as a less than five or odd response; the right mouse button served as a greater than five or even response. For the rotation pair, the low-demand option consisted of all upright digits (0°), whereas the high-demand option consisted of all digits rotated ±90°, with both options associated with a 50% probability of a task switch. For the switching pair, the low-demand option consisted of a 10% probability of a task switch whereas the high-demand option consisted of a 90% probability of a task switch, with digits always presented upright for both options.

Procedure

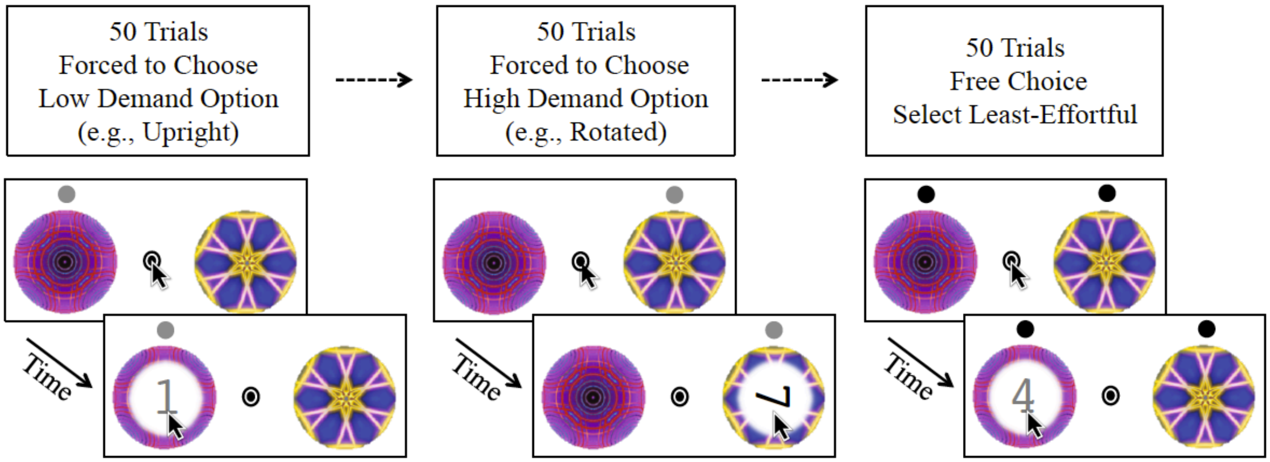
Individuals entered the testing room and provided informed consent. First, for individuals to gain sufficient experience with task response mappings, a practice session consisting of three blocks with feedback was completed. Individuals were required to reach a minimum accuracy of 90% in the last practice session prior to moving on to the main experiment. Instructions for the DST stated that individuals were to complete 3 blocks in the experiment for each pair of alternatives. To provide experience with the relative demand within each pair, the first two blocks constituted forced-choice instructions where a small cue displayed above one of the two options signaled that the individual was to choose only from that cued option. The starting option to be sampled from was randomized and counterbalanced. Once the first cued option was

[5] In both Kool et al. (2010) and Gold et al. (2014), the digit set included 1, 2, 3, 4, 6, 7, 8, and 9. However, during debriefing sessions of the pilot study, individuals often stated that the 6 and 9 digits were indistinguishable in the rotated conditions, therefore they were removed from the digit set for the current experiments. Removal of these digits also caused a need to remove an odd and even digit below 5 as well to keep the digit set balanced across both the parity and magnitude tasks.

finished being sampled from for 50 trials, the cue would then move to the other option where individuals would sample for 50 trials. Critically, the third block constituted the free-choice portion of the experiment. Individuals were instructed that that they may have noticed differences between the options after the first two blocks, and that some individuals begin to feel that one option is more effortful than the other. For this block, individuals were explicitly told to attempt to develop a preference about which option they felt was the *least effortful* of the two and to continue to choose that option until the block was complete (see Gold et al., 2014 for a similar approach). Individuals were randomly assigned to an order (i.e., rotation pair first, switching pair first) and order was counterbalanced across participants.

Figure 7.

Example DST in the Rotation Pair Condition



Note: In the first block, participants were cued to select one option by a small circle placed above the to-be-selected option. In the second block, the cue moved to the opposing option where individuals completed an additional 50 trials. The order of the first two blocks was randomized. In the third block, a circle was placed above both options indicating that the participant was to attempt to generate a least-effortful preference for one of the options and to continue to select that option until the block was complete. Each of the options was either a low-demand or a high-demand option. The rotation pair used in Experiments 4 and 5 is displayed in the figure where the low demand option consisted of upright digits and the high demand option consisted of rotated digits. In the switching pair (not shown), the low demand option consisted of a lower likelihood of switching (i.e., 10% Experiment 4; 30% Experiment 5) and the high demand option consisted of a higher likelihood of switching (i.e., 90% Experiment 4; 70% Experiment 5). In Experiment 6, the low demand option consisted of digits rotated 90° with a 30% probability of switching whereas the high demand option consisted of digits rotated 15° with a 70% probability of switching.

At the end of the DST, individuals completed a self-report questionnaire similar to that reported in Kool et al. 2010 (pp. 668) which included questions regarding whether individuals felt that they had constructed a less-effortful preference, what that preference was based on, whether a noticeable difference was apparent across options, and whether individuals were explicitly aware of a difference in the probabilities of a task switch. In addition, a question was added assessing how confident (from 0% to 100%) individuals were in the belief that there was some difference across options (see Appendix A). The entire experiment took approximately 45 minutes to complete.

Results

Results are reported first for RTs for task responses, switch costs (following Kool et al., 2010), and accuracy, followed by the DST (see Table 2) and self-report data. All effect sizes associated with within-subjects comparisons are Cohen's *d* using SD_{avg} as the standardizer term (see Cumming, 2012, p. 291) and all reported 95% confidence intervals (CI) are bootstrapped bias-corrected and accelerated (BCa) intervals (DiCiccio & Efron, 1996). To supplement inferential statistics, Bayesian estimation analyses were conducted on selections in the DST and RT using the BEST package (Kruschke, 2014) in R (R Core Team, 2014) using 100,000 estimates of the effect size across groups using Markov Chain Monte Carlo (MCMC) sampling. Ninety-five percent Highest Density Intervals (HDI), as well as the simulated mode effect size (i.e., the maximum *a posteriori* estimate; MAP) are presented. In addition, Bayes Factors (BF) computed using the BayesFactor package (Morey & Rouder, 2015) in R are presented. Interpretation of BFs follows the criteria outlined by Kass and Raftery (1995). In addition, any manual response times faster than 200 ms would be removed for all RT analyses.

Performance

No manual RTs were faster than 200 ms. Grand mean outlier and by-subject within condition outlier analyses were conducted on raw RTs using a 2.5 standard deviation cut-off VanSelst & Jolicoeur, 1994) in both cases and resulted in the removal of approximately 6% of the trials.

Performance. Accuracy did not vary as a function of pairs or demand option (all *F's* < 1.33). With respect to task RTs at the block level, a 2 (Pair of Alternatives) x 2 (Demand Option) repeated measures ANOVA demonstrated a significant main effect of pair of alternatives, $F(1, 33) = 7.77$, $MSE = 29979.34$, $\eta^2_p = .19$, $p < .001$, and demand option, $F(1, 33) = 73.98$, $MSE = 14453.64$, $\eta^2_p = .69$, $p < .001$, as well as a pair x demand option interaction, $F(1, 33) = 20.96$, $MSE = 18496.28$, $\eta^2_p = .39$, $p < .001$. The difference in response time between the low and high demand option (i.e., the demand effect) for the rotation pair (M = 71 ms, SD = 155) was significantly smaller than for the switching pair, (M = 284 ms, SD = 204), M_{Diff}= 213 ms, $t(33) = 4.56$, $SE = 45.36$, $d = 1.19$, 95% BCa CI [121 ms, 304 ms], $p < .001$. Aligned with the reported inferential statistics, Bayesian analyses demonstrated strong evidence for the alternative (i.e., that there is a difference between means), MAP = .79, 95% HDI [.38, 1.19], BF_{ALT} = 380.87.

Next for switch costs, a 2 (Pair of Alternatives) x 2 (Demand Option) x 2 (Switch Trial) repeated measures ANOVA demonstrated a significant main effect of demand option, $F(1, 33) = 24.91$, $MSE = 31387.63$, $\eta^2_p = .43$, $p < .001$, a significant main effect of switch trial, $F(1, 33) = 26.41$, $MSE = 51295.65$, $\eta^2_p = .45$, $p < .001$, a significant demand option x switch trial interaction, $F(1, 33) = 6.46$, $MSE = 17906.98$, $\eta^2_p = .16$, $p = .02$, and a significant pair of alternatives x demand option x switch trial interaction, $F(1, 33) = 9.57$, $MSE = 27693.48$, $\eta^2_p = .23$, $p < .01$. To further examine this three-way interaction, switch costs (i.e., switch trials – no switch trials) were computed for each demand option within the rotation and switching pairs. Pairwise comparisons demonstrated that switch costs within the rotation pair did not differ across the low demand (M = 110 ms, SD = 184) and high demand (M = 152 ms, SD = 174) options, M_{Diff}= 42 ms, $t(33) = 1.09$, $SE = 39.74$, $d = .24$, 95% BCa CI [-37ms, 122ms], $p > .1$, MAP =

.24, 95% HDI [-.13, .69], BF_{NULL} = 3.17. In contrast, higher switch costs for the low demand (*M* = 255ms, SD = 319) relative to the high demand (*M* = 48 ms, SD = 174) option, were demonstrated within the switching pair, M_{Diff} = 42 ms, $t(33) = 3.35$, *SE* = 61.49, *d* = .67, 95% BCa CI [82 ms, 333 ms], $p < .01$, MAP = .58, 95% HDI [.20, .96], BF_{ALT} = 16.96.

Thus, the difference in task RTs between the low and high demand options in the switching pair was significantly larger (about 3x) than in the rotation pair. In contrast, for the rotation pair the average switch costs for the high-demand option was similar to that for the low-demand option, whereas for the switching pair, the low-demand option produced higher average switch costs relative to the high-demand option. The latter interaction likely reflects the fact that the high-demand option in the switching pair featured frequent switching whereas switching was relatively rare in the low demand condition. Indeed, Kool and colleagues (2010) reported a similar finding in their Experiment 3, and smaller switch costs associated with a higher probability of a switch trial may therefore represent a type of preparation effect (Monsell, 2003) given the experience afforded to individuals for each block.

If individuals are utilizing the salient effort cue associated with stimulus rotation, then I should observe higher rates of low-demand options for the rotation pair relative to the switching pair. Alternatively, if time and overall demands on executive control are driving effort-based decisions in the DST, then there should be a corresponding difference in terms of the selection of the low and high demand options in the DST. That is, there should be a stronger bias toward the low demand option in the switching pair than in the rotation pair. Last, if relative switch costs are driving selections in the DST, then I should observe similar frequencies of selections across options for the rotation pair, and higher frequencies of selections of the *high demand* option for the switching pair (i.e., individuals should avoid the low-demand option associated with high average switch costs).

Table 2.

Demand Selection Task and Mean Performance Results for Experiments 4 and 5

	Rotation Pair		Switching Pair	
	Low Demand Option	High Demand Option	Low Demand Option	High Demand Option
Experiment 4				
Low Demand Selections	67% (37%)	-	62% (39%)	-
Block RT (ms)	1121 (241)	1201 (278)	927 (196)	1216 (279)
Switch Cost (ms)	110 (184)	152 (174)	255 (319)	47 (297)
Repeat Trials	1070 (244)	1115 (249)	923 (225)	1175 (327)
Switch Trials	1180 (290)	1267 (307)	1178 (339)	1222 (291)
Accuracy	97% (4%)	98% (2%)	98% (2%)	98% (2%)
Experiment 5				
Low Demand Selections	72% (35%)	-	47% (37%)	-
Block RT (ms)	1041 (251)	1097 (250)	1120 (286)	1212 (285)
Switch Cost (ms)	141 (181)	181 (200)	154 (217)	80 (185)
Repeat Trials	1002 (252)	1046 (248)	1074 (290)	1160 (320)
Switch Trials	1143 (281)	1227 (289)	1228 (338)	1240 (296)
Accuracy	98% (4%)	98% (2%)	98% (2%)	98% (3%)

Note: Standard deviations are presented in parentheses.

Demand Selection Task

Individuals chose the low-demand option in the rotation pair 67% of the time (SD = 37%). This value significantly differed from chance and demonstrated positive evidence for the alternative (i.e., low-demand selections were greater than chance), $t(33) = 2.62$, $SE = .06$, $d = .45$, 95% BCa CI [53%, 70%], $p = .013$, MAP = .47*,* 95% HDI [.09, .90], $BF_{ALT} = 3.44$. Individuals chose the low-demand option for the switching pair 62% of the time (SD = 39%). This value did not significantly differ from chance and did not demonstrate evidence for either the null or the alternative, $t(33) = 1.71$, $SE = .10$, $d = .29$, 95% BCa CI [49%, 74%], $p = .1$, MAP = .30*,* 95% HDI [-.05, .70], $BF_{ALT} = .68$. Low-demand choices across pairs did not significantly differ, $M_{Diff} = 5\%$, $t(33) = .64$, $SE = .08$, $d = .14$, 95% BCa CI [-9%, 2%], $p > .1$, and provided evidence for the null, MAP $d = .07$*,* 95% HDI [-.30, .44], $BF_{NULL} = 4.51$.

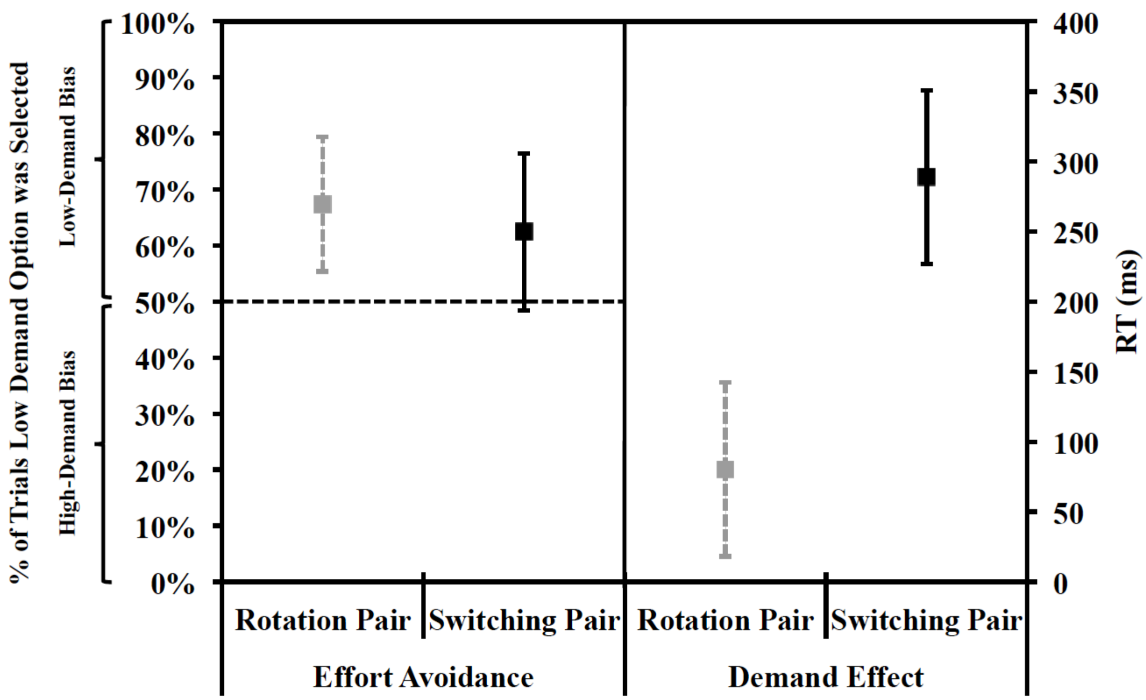
Thus, effort avoidance was similar across the rotation and switching pairs, despite the fact that the difference in terms of time and overall demand on executive control was much greater in the switching pair than in the rotation pair (see Figure 8)[6]. Furthermore, if I redefine demands on executive control as the magnitude of the switch cost across options, this does not change the conclusion. The differences in switch cost between options was much greater in the switching pair (larger switch cost in the low demand option with infrequent switching) than in

[6] In addition to these analyses, non-parametric tests were conducted on rates of low-demand selections. Wilcoxon Signed-rank tests confirmed the parametric and Bayesian analyses in that low-demand selections for the rotation decks differed from chance, $p = .02$, whereas low-demand selections for the switching decks did not, $p > .1$. Furthermore, treating individuals dichotomously as "avoiders" within each deck type (i.e., greater than 50% low-demand selections) demonstrated that 68% of individuals fell into the "avoider" category for the rotation decks, binomial test against chance, $p = .058$, whereas 66% of individuals fell into the category for the switching decks, $p = .12$. Importantly, however, dichotomizing individuals into these categories can be argued to be too conservative given that individuals with selections near chance are treated the same as individuals with high frequencies of low-demand selections.

the rotation pair, and effort avoidance went in the direction opposite that predicted pattern in the former pair (i.e., participants selected the low demand option/high switch cost option as the least effortful more often). Furthermore, in the rotation pair there was no difference in the magnitude of the switch cost but there was a marked difference in low effort selections.

Figure 8.

Percentage of Low-Demand Selections and Demand Effect for Rotation and Switching Pairs in Experiment 4



Note: The left panel consists of the percentage of selections of the low-demand options for each pair. The right panel consists of the demand effect in RT (i.e., high-demand option minus low-demand option) for each pair. Error bars in the left panel represent 95% Bias-corrected and accelerated (BCa) confidence intervals. Error bars in the right panel represent 95% within-subject confidence intervals (Masson & Loftus, 2003).

Self-Report

One individual was unable to complete the self-report portion of the experiment (N = 33 for the following analyses). Considering all individuals, 79% (26) reported noticing some difference across options for the rotation pair and 70% (23) reported noticing some difference for the switching pair, Sign Test $p > .1$. Likewise, self-reported confidence in reporting noticing a difference across options was similar for the rotation pair (M = 82%, SD = 20%) and the switching pair (M = 76%, SD = 21%), M_{Diff} = 6%, $t(31) = 1.52$, $SE = .04$, $d = .28$, 95% BCa CI [-1%, 13%], $p > .1$. Sixty-one percent of individuals (20) reported that they became explicitly aware of a difference in switching across the options in the switching pair, Binomial Test, $p > .1$ (participants were not explicitly asked about stimulus rotation). Sixty-four percent of individuals (21) explicitly reported stimulus rotation as the determinant of their preference for the rotation pair relative to only 33% (11) of individuals explicitly reporting differences in the probability of switching as the determinant of their preference for the switching pair, Sign Test $p = .031$. Consistent with effort avoidance rates being similar in Experiment 1, individuals reported noticing “some” difference (and confidence in that difference) in effort across the low and high demand options at about the same rate across the rotation and switching pairs. Interestingly, in the rotation pair, individuals were more likely to say that stimulus rotation was the basis of their preference than individuals in the switching pair were to say that the high rate of task switching was the basis of their preference (see Table 3).

Table 3.

Self-report Data for Experiment 4

	Rotation Pair	Switching Pair
Reported that there was a difference between the two options (i.e., low- and high-demand).	79%	70%
Confidence that there was a difference between the options.	82%	76%
Generated a less-effortful preference for one of the options in the final (free choice) block.	85%	76%
Explicitly reported preference was based on stimulus rotation (rotation pair) / % of switching (switching pair).	64%	33%
Explicitly aware of a difference in % of switching for the switching pair.		61%

Discussion

In Experiment 4, the low- and high-demand options in the switching pair were associated with a greater difference in task RTs within block and theoretically a greater difference in demand on executive control than in the rotation pair. Nevertheless, individuals avoided the high demand option in the rotation pair (i.e., the 90° rotated stimulus) at a rate similar to the high demand option in the switching pair (i.e., a 90% chance of a task switch; see Figure 8). From a cue-utilization perspective, this result can be seen to reflect the influence of the salient effort cue in the rotation pair (i.e., stimulus rotation). That is, the presence of this effort cue led to rates of effort avoidance equivalent to that produced by a large difference in time and demands on executive control. Given the central role that the latter two contributions have played in theoretical discussion of effort-based decision-making, this result provides strong support for a cue-based contribution. It is also important to note that while switch costs were similar across options for the rotation pair, the low-demand option produced larger switch costs relative to the high-demand option for the switching pair. Yet in neither case did it appear that effort-based selection followed relative switch costs.

Self-report data demonstrated that a 90° rotated stimulus and a high-probability of a task switch (i.e., 90% vs. 10%) evoked a similar level of subjective experience of relative effortfulness, consistent with the DST results. Again this is despite the large differences in time costs and demands on executive control. While individuals were equally likely to report "some" difference in effort between the pairs, more individuals reported stimulus rotation than task switching as the determinant of their choice. It may be the case that, although becoming aware of the switching cue, individuals did not endorse task switching as a source of effort to-be-avoided (e.g., Michaelian, 2012). For example, some individuals may not have garnered sufficient intra- or extra-experimental experience necessary to generate the belief that task switching is difficult. Alternatively, individuals may have experienced differences in effort across the options (e.g., Kurzban et al., 2013), but were unable to explicitly identify switching as the source of the difference.

Experiment 5

Experiment 5 looked to extend the results of Experiment 4 by reducing the difference in the probabilities of a task switch in the switching pair to 70% and 30% (from 90% vs. 10%) for the high- and low-demand options respectively. In Experiment 4, rates of effort avoidance for the switching pair were statistically similar to rates for the rotation pair. Critically, according to a cue-utilization account, if I were to reduce the putative salience of the effort cue in the switching pair (i.e., by reducing the difference in the proportion of switching), then low-demand selections should be reduced as well. While I would also expect the difference in time and demand on executive control to decrease, given how large the difference was in Experiment 4 it would be unlikely to become less than that observed between options in the rotation pair. If I assume that the low-demand selections will remain similar in the rotation pair, then this raises the interesting possibility that individuals will avoid the high demand option in the rotation pair more than the high demand option in the switching pair, despite the difference in time and demand on executive control not being smaller in the latter than the former.

Last, in Experiment 5, I looked to further gauge individuals' explicit awareness of each of the effort cues. To do so, I added a self-report question allowing individuals to express what they believed the specific difference was (if any) across the options for both of the pairs (this was only done for the switching pair in Experiment 4).

Method

Participants

Thirty-six University of Waterloo undergraduate students participated in the study in exchange for research credit.

Design

A 2 (Pair of Alternatives: Rotation Pair, Switching Pair) x 2 (Demand Option: Low Demand, High Demand) within-subjects design was employed.

Apparatus

All apparatus were the same as Experiment 4.

Stimuli

For the rotation pair, the low-demand option consisted of all upright digits, whereas the high-demand option consisted of all digits rotated ±90°, and the probability of a task switch was reduced to 30% for both options. For the switching pair, the low-demand option consisted of a 30% probability of a task switch, whereas the high-demand option consisted of a 70%

probability of a task switch, with digits always presented at ±15° for both options. The move to add 15° of rotation to the high-demand option was to ensure that both manipulations were crossed to some extent across the options (i.e., high-and-low degree of stimulus rotation, high-and-low probability of switching).

Procedure

The procedure was the same as in Experiment 4. However, the self-report questionnaire completed by individuals at the end of the DST was changed to include a free-response question asking what attribute of the task individuals felt the difference across options for each pair was (see Appendix B).

Results

All reporting procedures follow Experiment 4 (see Table 2).

Performance

One observation had a response faster than 200 ms and was removed. Grand mean outlier and by-subject within condition outlier analyses were conducted on raw RTs using a 2.5 standard deviation cut-off in both cases, resulted in the removal of approximately 6% of the trials.

Performance. Again, accuracy did not vary as a function of pairs or demand option (all *F's* < 1.15). With respect to block response time, a 2 x 2 repeated measures ANOVA demonstrated a significant main effect of pairs of alternatives, $F(1, 35) = 10.39$, $MSE = 32520.3$, $\eta^2_p = .23$, $p < .01$, and demand option, $F(1, 35) = 9.04$, $MSE = 21922.92$, $\eta^2_p = .21$, $p < .01$, but not a significant pair x demand option interaction, $F(1, 35) = 1.25$, $MSE = 9896.01$, $\eta^2_p = .04$, $p > .1$. The difference between the low and high demand option for the rotation pair ($M = 56$ ms, SD = 158) was statistically similar to the difference for the switching pair, ($M = 92$ ms, SD = 196),

M_{Diff} = 37 ms, $t(35) = 1.12$, $SE = 32.05$, $d = .19$, 95% BCa CI [-30 ms, 102 ms], $p > .1$. Similarly, Bayesian analyses demonstrated positive evidence for the null (i.e., the two demand differences are equivalent), MAP = .21, 95% HDI [-.15, .54], $BF_{NULL} = 3.13$.

Next for switch costs, a 2 (Pair of Alternatives) x 2 (Demand Option) x 2 (Switch Trial) repeated measures ANOVA demonstrated a significant main effect of pair, $F(1, 35) = 5.30$, $MSE = 68533.83$, $\eta^2_p = .13$, $p = .027$, a significant main effect of demand option, $F(1, 35) = 4.59$, $MSE = 49135.36$, $\eta^2_p = .12$, $p = .039$, and a significant main effect of switch trial, $F(1, 35) = 51.62$, $MSE = 27007.58$, $\eta^2_p = .59$, $p < .001$. Furthermore, a marginally significant pair x demand option x switch trial interaction was demonstrated, $F(1, 35) = 4.06$, $MSE = 14357.99$, $\eta^2_p = .10$, $p = .052$. Following Experiment 4, to examine this three-way interaction switch costs were computed for each demand option within the rotation and switching pairs. Pairwise comparisons demonstrated that switch costs within the rotation pair were similar across the low demand (M = 141 ms, SD = 181) and high demand options (M = 181 ms, SD = 200), M_{Diff}= 40 ms, $t(33) = .97$, $SE = 39.35$, $d = .21$, 95% BCa CI [-37 ms, 105 ms], $p > .1$, MAP = .24, 95% HDI [-.15, .60], $BF_{NULL} = 3.63$. Similar switch costs for the low demand (M = 154 ms, SD = 217) relative to the high demand options (M = 80 ms, SD = 185), were demonstrated within the switching pair as well, M_{Diff}= 74 ms, $t(33) = 1.64$, $SE = 44.25$, $d = .37$, 95% BCa CI [-3 ms, 159 ms], $p > .1$, MAP = .27, 95% HDI [.08, .63], $BF_{NULL} = 1.67$.

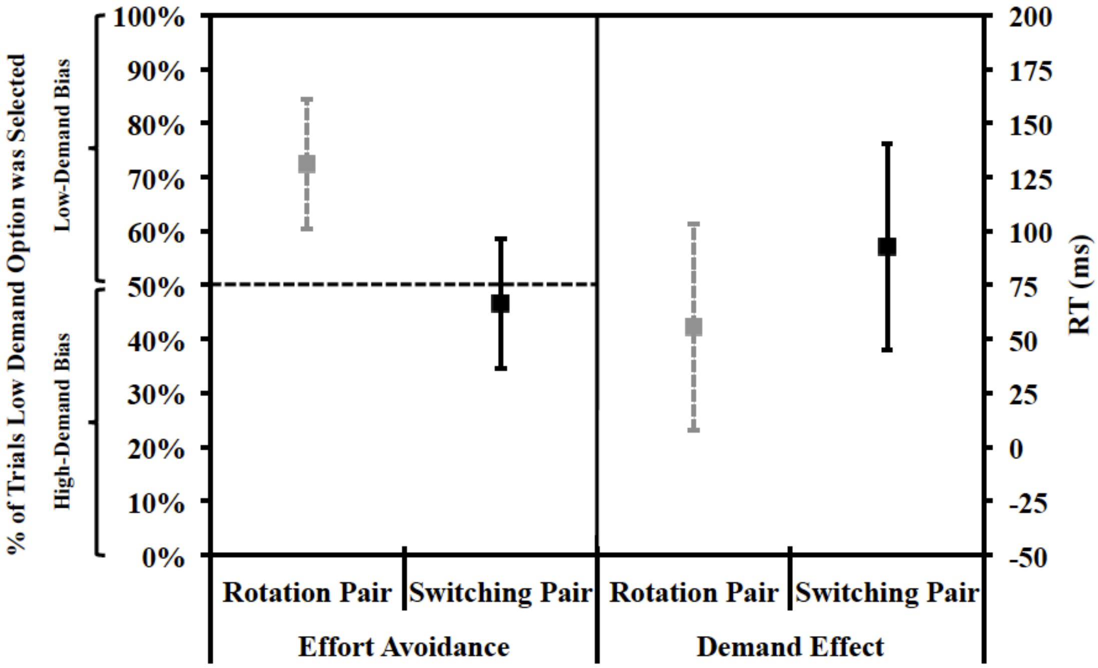
In contrast to Experiment 4, both the rotation and switching pairs demonstrated statistically similar differences between the low- and high-demand options. Furthermore, switch costs again were similar across options for the rotation pair, and larger in the low-demand option relative to the high-demand option for the switching pair. Again, if individuals are utilizing the salient effort cue associated with stimulus rotation, then I should observe higher rates of low-demand options for the rotation pair relative to the switching pair. Alternatively, if time and overall demands on executive control are the determinant of choices in the DST, then there should be no difference in participants' selections of the low demand option across the two pairs. Last, if switch costs are being avoided then I should see low-demand selections near chance for the rotation pair and higher *high-demand* selections for the switching pair.

Demand Selection Task

Individuals chose the low-demand option in the rotation pair 72% of the time (SD = 35%), and this value significantly differed from chance and demonstrated strong evidence for the alternative hypothesis (i.e., selections are greater than chance), $t(35) = 3.85$, $SE = .06$, $d = .64$, 95% BCa CI [61%, 84%], $p < .001$, MAP = .57, 95% HDI [.36, 1], $BF_{ALT} = 61.11$. Individuals chose the low-demand option for the switching pair only 47% of the time (SD = 37%). This value did not significantly differ from chance and demonstrated positive evidence for the null (i.e., selections are equivalent to chance), $t(35) = -.55$, $SE = .06$, $d = -.09$, 95% BCa CI [35%, 58%], $p > .1$, MAP = -.1, 95% HDI [-.43, .25], $BF_{NULL} = 4.86$. In addition, low-demand choices across deck type differed significantly, $M_{Diff} = 25\%$, $t(35) = 3.14$, $SE = .08$, $d = .72$, 95% BCa CI [8%, 41%], $p > .01$, and provided positive evidence for the alternative, MAP = .51, 95% HDI [.17, .92], $BF_{ALT} = 10.73$. Therefore, despite equivalent demand effects in terms of time across the rotation and switching pairs, avoidance rates were demonstrably higher for the rotation option than for the switching options[7] (see Figure 9). Furthermore, selections did not follow the pattern hypothesized by switch costs alone driving selections.

Figure 9.

Percentage of Low-Demand Selections and Demand Effect for Rotation and Switching Pairs in Experiment 5



Note: The left panel consists of the percentage of selections of the low-demand options for each pair. The right panel consists of the demand effect in RT (i.e., high-demand option minus low-demand option) for each pair. Error bars in the left panel represent 95% Bias-corrected and accelerated (BCa) confidence intervals. Error bars in the right panel represent 95% within-subject confidence intervals (Masson & Loftus, 2003).

[7] Wilcoxon Signed-rank tests confirmed the parametric and Bayesian analyses in that low-demand selections for the rotation decks differed from chance, $p = .01$, whereas low-demand selections for the switching decks did not, $p > .1$. Furthermore, 75% of individuals fell into the "avoider" category for the rotation decks, $p < .01$, whereas only 50% of individuals fell into the category for the switching decks, $p > .1$.

Self-Report

Considering all individuals, 92% (33) reported noticing some difference across options for the rotation pair whereas only 47% (17) reported noticing some difference for the switching pair, Sign Test $p < .001$. Fifty-six (20) percent of individuals explicitly reported that the difference across options for the rotation pair was stimulus rotation, whereas only 22% (8) explicitly reported switching as the difference across options for the switching pair, Sign Test $p < .05$. Last, 64% of individuals (23) explicitly reported stimulus rotation as the determinant of their preference for the rotation pair relative to only 22% (8) of individuals explicitly reporting differences in the probability of switching as the determinant of their preference for the switching pair, Sign Test $p = .001$. Thus, as opposed to Experiment 1, more individuals reported being aware of and utilizing the stimulus rotation cue relative to switching (see Table 4).

Table 4.

Self-report Data for Experiment 5

	Rotation Pair	Switching Pair
Reported that there was a difference between the two options (i.e., low- and high-demand).	92%	47%
Confidence that there was a difference between the options.	89%	77%
Difference between options was stimulus rotation (rotation pair) / % of switching (switching pair).	56%	22%
Generated a less-effortful preference for one of the options in the final (free choice) block.	81%	72%
Preference was based on stimulus rotation (rotation pair) / % of switching (switching pair).	64%	22%

Combined Experiments 1 and 2 Analyses

In the following, I directly compare demand selection, demand costs, and self-reports in the switching pairs across the different probabilities of a task switch (i.e., Experiment 4: 90% vs. 10% vs. Experiment 5: 70% vs. 30%). For the switching pair, low-demand selections significantly declined from Experiment 4 (M = 69%, SD = 47%) to Experiment 5 (M = 47%, SD = 37%), M_{Diff} = 22%, $t(64.62) = 2.18$, SE = .1, d = .52, 95% BCa CI [2%, 42%], p = .03. In addition, the demand effect across options significantly reduced from Experiment 4 (M = 289 ms, SD = 218) to Experiment 5 (M = 93 ms, SD = 196), M_{Diff} = 196 ms, $t(69) = 3.98$, SE = 49.24, d = .94, 95% BCa CI [98ms, 294ms], $p < .001$. Last, considering individuals that reported becoming aware of some difference across options (this question was consistent across the experiments) 71% of individuals reported a difference in Experiment 4 relative to only 47% in Experiment 5, $\chi^2(1) = 3.93$, p = .05. Thus, reducing the difference in the probability of a task switch across experiments led to a smaller difference in demand across the high and low options, a reduced likelihood of noticing "some" difference in effort and a reduced likelihood of selecting the low demand option.

Discussion

In Experiment 5, the difference in the proportion of switching across the options in the switching pair was reduced. This led to a reduction in the difference in time across the low- and high-demand options such that the difference between the two options was statistically similar to the difference across the options in the rotation pair. Nonetheless, effort avoidance rates for the rotation pair were much higher than in the switching pair (see Figure 9). This dissociation was predicted based on a cue utilization account. Despite the equivalence in demand, the salient effort cue in the rotation pair led to higher rates of effort avoidance. Furthermore, the self-report data from Experiment 2 are consistent with the interpretation provided of the dissociation observed. Individuals more frequently reported being aware of, and making use of, the rotation cue for their less-effortful choices even though differences in RTs between the low and high demand options across the switching and rotation pairs and demands on executive control in the former pair were similar. Overall, the results of Experiment 5 provide further support for cue-utilization being a critical determinant in effort avoidance in situations where these three types of information (i.e., cues, time, and demands on EC) are available.

Experiment 6

To this point, I have relied on the relative rates of effort avoidance across the rotation and switching pairs to test the hypotheses derived at the outset. In Experiment 6, I pit the high-demand options from each pair in Experiment 5 directly against one another. Participants selected between two alternatives: (1) a low probability of a task switch with the stimulus rotated 90°, and (2) a high probability of a task switch with the stimulus rotated 15°. The first option should be both faster and less demanding on executive control than the second option, hence I refer to (1) as the low-demand option and (2) as the high-demand option. Thus, in this context, there is a direct trade-off between avoiding a high-demand option associated with higher time demands and demands on executive control, and avoiding a relatively low-demand option associated with a salient effort cue. Again, from a time and overall demands on executive control perspective, I would expect individuals to avoid the high-demand option. Alternatively, from a cue-utilization perspective, the effort cue associated with the low-demand option (i.e., stimulus rotation) should counteract this effect given the higher rates of awareness relative to the switching cue as demonstrated in Experiment 5. Strong evidence for cue utilization in this trade-off situation would be signaled by equivalent selections across the two options or a bias to select the high-demand option.

Method

Participants

Thirty-six University of Waterloo undergraduate students participated in the study in exchange for research credit.

Design

A one-factor (Demand Option: Low Demand, High Demand) within-subjects design was employed.

Apparatus, Stimuli, and Procedure

All apparatus and stimuli were the same as Experiment 5.

Results

All reporting procedures follow Experiments 4 and 5 (see Table 5).

Table 5.

Demand Selection Task and Mean Performance Results for Experiment 6

	Low Demand Option	High Demand Option
Low Demand Selections	42% (37%)	-
Block RT (ms)	1266 (327)	1390 (312)
Switch Costs (ms)	221 (254)	106 (258)
Repeat Trials	1208 (323)	1324 (344)
Switch Trials	1429 (402)	1430 (333)
Accuracy	98% (3%)	98% (3%)

Note: Standard deviations are presented in parentheses.

Performance

No RTs were faster than 200 ms. Grand mean outlier and by-subject within condition outlier analyses were conducted on raw RTs using a 2.5 standard deviation cut-off in both cases; this resulted in the removal of approximately 6% of the trials.

Accuracy did not vary as a function of demand option (t = .42). RTs for the high-demand option (M = 1390 ms, SD = 312) were slower relative to the low-demand option, (M = 1266 ms, SD = 327), M_{Diff}= 124 ms, t(35) = 2.85, SE = 41.9, d = .38, 95% BCa CI [45 ms, 205 ms], p < .01. Similarly, Bayesian analyses demonstrated positive evidence for the alternative, MAP = .49, 95% HDI [.12, .91], BF_{ALT} = 5.57.

To examine switch costs, a 2 (Demand Option) x 2 (Switch Trial) repeated measures ANOVA was conducted. Results demonstrated a significant main effect of switch trial, F(1, 35) = 29.57, MSE = 32597.23, η^2_p = .46, p < .001. The switch cost for the low-demand option (M = 221 ms, SD = 254) was larger than the switch cost for the high demand option (M = 106 ms, SD = 258) producing a marginally significant demand option x switch trial interaction, F(1, 35) = 3.62, MSE = 32789.90, d = .45, p = .066, MAP = .33, 95% HDI [-.03, .68], BF_{NULL} = 1.11.

Therefore, as expected, the option associated with a higher probability of a task switch was significantly slower than the option associated with a higher degree of stimulus rotation. If individuals are utilizing the salient effort cue associated with stimulus rotation, then I should observe higher rates of choices for the high-demand option (i.e., individuals will avoid the rotated stimulus). Alternatively, according to time and overall demands on executive control driving avoidance, individuals should avoid the high-demand option in the DST.

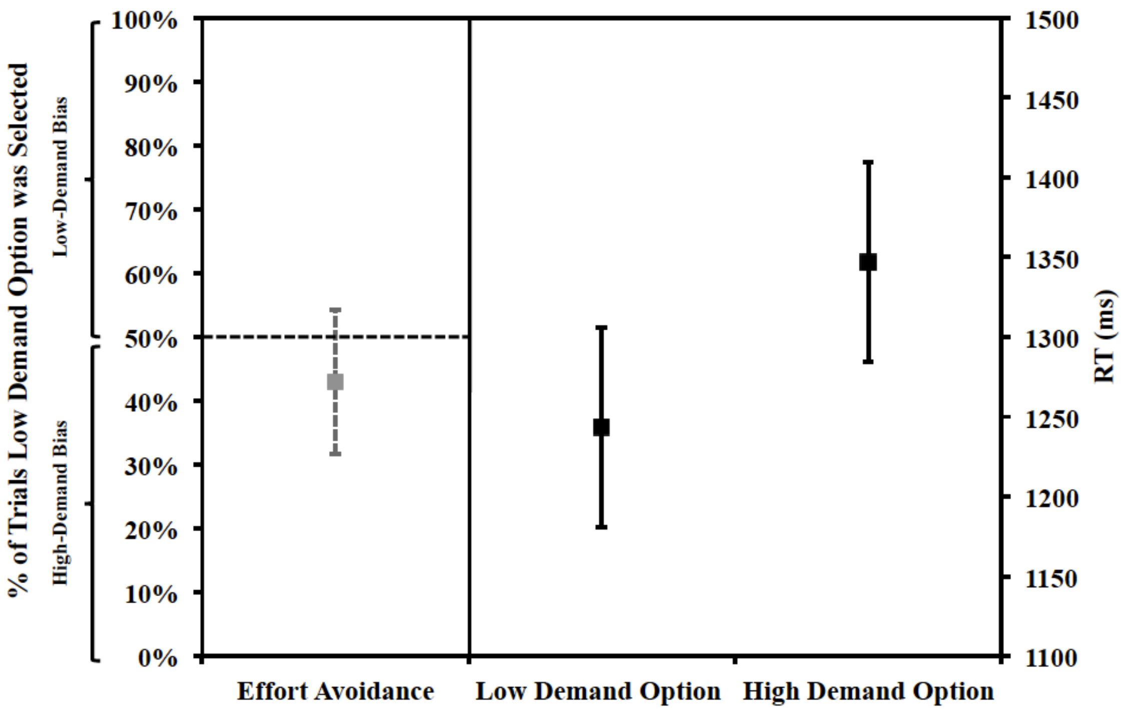
Demand Selection Task

Individuals chose the low-demand option in the rotation decks only 42% of the time (SD = 37%); this value did not significantly differ from chance, $t(35) = -1.36$, $SE = .06$, $d = -.23$, 95% BCa CI [30%, 54%], $p > .1$, Bayesian analysis, MAP = -.24, 95% HDI [-.6, 1], $BF_{ALT} = .41$. In addition, I could test the hypothesis that low-demand selections were at or below chance (i.e., the negative interval $-\infty < d < 0$, where 0 = chance, i.e., 50%, and d = the low-demand selection effect) relative to above chance (i.e., the positive interval $0 < d < +\infty$). Results demonstrated positive evidence for the alternative $BF_{ALT} = 9.07$. That is, selections occurred in the negative interval at or below chance. Indeed, the 95% HDI reported above [-.6, 1] lends additional evidence to this notion given that the majority of the simulated d distribution was negative. Thus, individuals did not avoid the high-demand option[8] (see Figure 10).

[8] Again, Wilcoxon Signed-rank tests confirmed the parametric and Bayesian analyses in that low-demand selections did not differ from chance, $p > .1$, and only 39% of individuals fell into the “avoider” category, $p > .1$.

Figure 10.

Percentage of Low-Demand Selections and Demand Effect in Experiment 6



Note: The left panel consists of the percentage of selections of the low-demand options for each pair. The right panel consists of the demand effect in RT (i.e., high-demand option minus low-demand option) for each pair. Error bars in the left panel represent 95% Bias-corrected and accelerated (BCa) confidence intervals. Error bars in the right panel represent 95% within-subject confidence intervals (Masson & Loftus, 2003).

Self-Report

Sixty-nine (25) percent of all individuals reported noticing some difference across the low- and high-demand options, Binomial Test $p = .029$. Forty-four percent (16) of individuals explicitly reported that the difference across options was stimulus rotation, whereas only 8% (3) explicitly reported switching as the difference, Sign Test $p < .01$. Fifty percent of individuals (18) explicitly reported stimulus rotation as the determinant of their preference relative to only

6% (2) of individuals explicitly reporting differences in the probability of switching as the determinant of their preference, Sign Test $p < .001$ (see Table 6).

Table 6.

Self-report Data for Experiment 6

Reported that there was a difference between the two options (i.e., low- and high-demand).	69%
Difference between options was stimulus rotation / % of switching.	44% / 8%
Confidence that there was a difference between the options (rotation / % switching).	84% / 83%
Generated a preference for one of the options in the final (free choice) block.	92%
Preference was based on stimulus rotation / % of switching.	55% / 6%

Discussion

In Experiment 6, participants chose between a low-demand option consisting of a low probability of a task switch with the stimulus rotated 90° and a high-demand option (i.e., significantly slower and more demand on executive control) consisting of a high probability of a task switch with the stimulus rotated only 15°. Critically, results demonstrated that individuals did not avoid the high-demand option (see Figure 10). Experiment 6 therefore demonstrates the importance of effort cues in effort-based decision-making. That is, individuals did not avoid an option associated with greater time costs and greater demands on executive control when it was pitted against an option with a salient effort cue. Interestingly, the switch cost was higher for the low demand option as has been observed in previous experiments, thus one might suggest that this executive cost was driving effort-based selections. However, the results from Experiments 4 and 5 together (and Kool et al., 2010; Experiment 3) suggest that this cost is unlikely to be the driver in selections.

General Discussion

Chapter 2 examined how individuals decide which course of action is the least effortful. To achieve this, using variants of the DST, I contrasted the influence of time and demands on the executive control system against the influence of a salient effort cue not associated with much in the way of either time or executive demands. All three experiments demonstrated clear evidence for the influence of a kind of cue utilization process in effort-based decision-making in the DST. In Experiment 4, individuals avoided the high-demand option associated with stimulus rotation at a rate similar to the high-demand option associated with a high probability of a task switch, even though the performance costs and demands on executive control in the switching pair were much greater than in the rotation pair. In Experiment 5, individuals avoided the high-demand option associated with stimulus rotation at a higher rate than the high-demand option associated with a high probability of a task switch, even though the performance costs across the pairs were similar (though the difference in demands on executive control likely still favored the switching pair). Last, in Experiment 6, individuals did not avoid the high-demand option associated with a high probability of a task switch when pitted against a low-demand option associated with a salient effort cue, again despite the former inducing significant performance costs and demands on executive control. The results observed here are all consistent with cue-utilization being an important determinant in effort avoidance behaviors. As such, the present results provide strong evidence that this theoretical approach to effort avoidance is viable.

Cue-Utilization in Effort Avoidance

A critical variable in effort-based decisions according to a cue-utilization account is the availability of cues and the inferences/heuristics applied to those cues when generating a subjective evaluation of effort. While Chapter 2 has focused on contrasting cue-utilization with time and overall demands on executive control, it is important to note that the current proposal does not make the claim that the latter two factors do not contribute to the control of behavior in effort avoidance. For example, if the option associated with a high proportion of switches was indicated to participants (or discovered spontaneously), then individuals should avoid that option assuming that they held or generated the belief that switching was effortful. The results of Experiments 4 and 5 are consistent with this notion. When the difference in the proportion of switching between options was 90%/10%, participants avoided the high demand option somewhat more than chance, but when it was reduced to 70%/30% they did not. That is, in certain situations, high demands on executive control (or time) can serve as salient cues exploited in the evaluation process. Thus, the cue-utilization approach does not preclude time or demands on executive control playing a role in effort-based decisions. Rather, cue-utilization simply predicts that individuals exploit available effort cues with the strongest (or most salient) cue being utilized to generate which option is more effortful.

While the theoretical proposal that I offer here might appear counterintuitive in that individuals could often end up making non-optimal effort-based decisions (as they did for example in Experiment 6; also see below), it is important to note that effort avoidance is itself a potentially effortful task (Boureau, Sokol-Hessner, & Daw, 2015; Payne et al., 1993). That is, monitoring some veridical effort cost (i.e., which option takes longer, which option makes greater demands on executive control), assuming that one can do so, with the goal of optimally minimizing effort may place additional demands on the system beyond that of the task at hand. Both the time- and executive control-based accounts imply what could be argued to be a demanding monitoring process (e.g., Shenhav, Botvinick, & Cohen, 2013) to arrive at a least "effortful" solution. Exploiting a salient effort cue on the other hand arguably requires little in terms of processing. Similarly, Payne et al. (1993; see also Gigerenzer & Goldstein, 1996) note that the deployment of a heuristic strategy such as cue utilization would be expected to be associated with a large savings in effort when contrasted with more normative strategies, especially in decision environments where effort minimization is emphasized. Reliance then on effort cues that might be detached from objective demand (i.e., the *cognitive work* taken on by the system) could itself be conceptualized as a least effort solution to the problem of selecting a least effort solution[9].

The results demonstrated across Chapter 2 additionally highlight the importance of the level of explicit availability of an effort cue in effort-based decision-making (Dunn et al., 2016). Cue-utilization predicts that optimal effort avoidance would only be expected in situations where cues associated with variations in objective effort (e.g., time, demands on executive control) across alternatives become available to the individual. That is, a demand cue (or cues) becomes explicit to the point that a qualitative judgment can be made in the decision-making process (i.e., *"Option A is more effortful than Option B"*). Effort avoidance would be expected to often follow objective demand in situations where tasks are structured in ways to facilitate such evaluations (e.g., explicit feedback, a large number of trials) or where a difference in some form of objective demand across alternatives is large. Dissociations between objective demand and effort avoidance, on the other hand, should occur in situations where differences in objective demand

[9] I thank Gordon Pennycook for useful discussions on this point.

are present, but there is a lack of awareness of a cue highlighting such a difference (e.g., results from the switching pair in Experiment 2 and to some extent Experiment 1). Such a prediction is seemingly at odds with accounts suggesting that implicit costs or processes can influence control of behavior (e.g., Payne et al., 1993; Reber, 1989; Westbrook & Braver, 2015). Moreover, dissociations would be expected in situations where there are no differences in objective demands, however a salient cue leads to the inference that one option is more effortful than the other (see Chapter 1).

Furthermore, situations can undoubtedly arise where the deployment of cue-utilization is overridden with other decision strategies (e.g., strategies that reduce objective demands). For example, within cost-benefit accounts of control, motivation plays a critical role in configuring behaviors (e.g., Botvinick & Braver, 2015; Braver et al., 2014, Inzlicht & Schmeichel, 2012). Inzlicht and colleagues (2014) argue that effortful tasks become aversive when motivation shifts away from goal-directed behavior associated with a task to engaging in cognitive leisure (i.e., disengaging; see also Kool & Botvinick, 2014). As an example, if motivation is specifically manipulated, perhaps through the use of incentives to monitor, identify, and complete a task at some level of accuracy, then deployment of a cue-utilization strategy in generating less-effortful preferences may vary relative to what was observed in the current experiments. Here, arguably the only extrinsic motivation present was the explicit instructions to participants to attempt to generate a less-effortful preference; a situation in which cue-utilization drove choices in the DST. Testing how manipulations of motivation and incentive modulate less-effortful choices (e.g., Kool et al., 2010) represents an important future empirical endeavor.

Cue-Utilization as a Non-compensatory Strategy in Effort Avoidance

The claim that effort-based decision making relies on effort cues to form a kind of metacognitive evaluation of effort raises important theoretical questions about potential competition between cues. One interesting issue to consider is whether effort-based decision-making is best described as non-compensatory or compensatory (Gigerenzer & Goldstein, 1996; Payne et al., 1993; Tversky & Kahneman, 1974). Non-compensatory decision strategies do not make trade-offs across attributes (i.e., these strategies do not incorporate an ability of a good value on one attribute to make up for a bad value on another attribute) whereas compensatory strategies make such trade-offs possible (Payne et al., 1993). In Experiment 6, two potential effort cues (i.e., stimulus rotation and probability of a task switch) were pitted against one another. There was some evidence that the stimulus rotation cue actually overshadowed the probability of a task switch consistent with a non-compensatory strategy. In particular, the self-report data revealed that awareness and utilization of the switching cue reduced from Experiment 5 (22%) to Experiment 6 (6%), $\chi^2(1) = 4.18$, $p = .04$. That said, if individuals were relying on the stimulus rotation cue to the complete exclusion of the probability of the task switch, I would have expected many more selections of the high demand option (i.e., less rotated option) given the high rates of avoiding the rotated option in Experiments 4 and 5 (where rotation was nominally the only available effort cue). Future work designed to test between non-compensatory and compensatory versions of the cue-utilization account is needed.

Awareness and Utilization of Effort Cues

While the experiments in Chapter 2 were not directly designed to assess the relation between the explicit awareness of effort cues and effort avoidance, a number of notable patterns emerged across experiments. For example, in Experiment 4, rates of effort avoidance across pairs were similar, as were the frequencies of individuals self-reporting some difference across options in each of the pairs. In addition, when the probability of switching was reduced in Experiment 5, reported awareness and effort avoidance similarly decreased for the switching pair. Thus, it appears that explicit awareness of effort cues and effort avoidance might be tightly coupled (see Chapter 1). Furthermore, reported utilization of the cues was higher for stimulus rotation relative to switching in Experiment 4 even in light of similar rates of effort avoidance suggesting that, in addition to requiring awareness, cues seem to need to be endorsed (e.g., Michaelian, 2012) as a source of effort for a least-effortful preference to be generated.

Indexing Demands on Executive Control

In Chapter 2, I relied throughout on an indirect measure of demands on executive control (i.e., the probability of a task switch) to test the hypothesis that such demands will drive effort avoidance behavior. Furthermore, I relied on response times to bolster this inference. Of course, given that demands on executive control cannot be measured directly, both of these assumptions should be assessed critically. For example, the implications of the present results would change if stimulus rotation produced greater demands on executive control than task switching. While *a priori* this idea seems inconsistent with most theoretical conceptions of executive control, it is nonetheless worth considering. Importantly, any test of the relation between demands on executive control and effort-based decision-making will necessarily rely on assumptions about the nature of that demand. Being clear about the nature of this assumption is important to avoid the circular argument that the act of avoidance in and of itself constitutes evidence of a difference in demands on executive control. For example, if one was to make the argument that the equal rates of avoidance in Experiment 4 are a result of the difference between upright and rotated stimuli and a 90% and 10% probability of switching are equally demanding on executive control. Along similar lines, it is important that I note that previous work has demonstrated that RTs can be dissociated from these demands on executive control (e.g., Kool et al., 2010; McGuire & Botvinick, 2010) and, thus, represents only a noisy index of such demands. Thus, again it is important to rely on our theoretical understanding of the types of tasks or manipulations that put demands on the executive control system and not to rely solely on measures of time.

The Probability of a Switch-by-Switch Cost Interaction

With respect to the notion that individuals avoid demands on executive control, the consistent observation of an interaction between the probability of switch trials and switch costs provides a potentially important observation. Specifically, this interaction can be interpreted in the context of Braver's (2012) dual mechanism framework that distinguishes between proactive and reactive control. Proactive control consists of a form of sustained, prepatory control whereas reactive control consists as a type of "as-needed" control recruited when an event demanding control is detected. The reduction in the switch cost in a context with a high switch probability can be interpreted as the engagement of a more proactive mode of control relative to a more reactive mode of control adopted when switching occurs less frequently. Proactive control in this context could, for example, consist of maintaining both tasks sets in a partially activated state. Consistent with such a mechanism is the observation that the probability of switching by switch cost interaction seems largely due to differences in repeat trial response times (see Tables 2 and 5).

Against this background, two points are worth noting. First, in Experiment 4, where this interaction was particularly large, effort-based choices demonstrated an avoidance of the high probability switch option, which can be taken as being associated with more of a proactive mode of control, rather than the avoidance of the low probability switch option, which can be taken as being associated with a more reactive mode of control. This suggests that avoidance of demands on executive control might be more strongly tuned to avoiding proactive rather than reactive control. The second observation is that the presence of the interaction between the probability of switching and the switch costs suggests, at least on some level, a response to the probability of a switch manipulation. Thus, in Experiment 5, while there was no evidence of differential avoidance of the options observed at the level of effort-based decisions, there was at the level of the cognitive systems' response to the proportion switching (e.g., a switch between more proactive and reactive modes of control). While not the aim of the present research, further examination of the proportion switching by switch cost interaction might promise a more nuanced understanding of the relation between different types of control and effort based decisions.

Conclusion

One of the fundamental principles of human psychology *principle* (e.g., Clark, 2010; Zipf, 1949) is that humans tend to avoid effort. The experiments reported here contribute to our understanding of how it is that individuals achieve this goal. Future work examining the novel theoretical account provided here promises further insights into the inner workings of the cognitive miser.

Chapter 3

Chapters 1 and 2 provided evidence in support of the hypothesis that the utilization of available cues drives perceptions of effort (i.e., in self-report) and effort-based decision (i.e., in a DST). That is, individuals appear to infer effort through a type of metacognitive evaluation made over available cues. However, this cue utilization account does not address *why* some cues would signal that a task may be effortful. Chapter 3 begins to address this question by investigating how individuals evaluate effortful tasks through the lens of the General Evaluability Theory (GET; Hsee & Zhang, 2010).

The following work has been accepted for publication and is currently in press at the *Journal of Behavioral Decision Making* (Dunn, Koehler, & Risko, in press).

Changes have been introduced to improve the flow of the dissertation.

Recently, a focus on effort-based decision-making has pushed the hypothesis of effort minimization to the fore as a critical behavior to understand empirically in human decision-making (e.g., Botvinick & Braver, 2015; Dunn, Lutes, & Risko, 2016; Dunn & Risko, 2016; Kool, McGuire, Rosen, & Botvinick, 2010; Kool & Botvinick, 2014; Kurniawan, Guitart-Masip, & Dolan, 2011; McGuire & Botvinick, 2010; Westbrook & Braver, 2015; Westbrook, Kester, & Braver, 2013). Although many of these endeavors have focused on the factors that drive effort-avoidance (e.g., cues, see Chapters 1 and 2; demands on executive control, Kool et al., 2010), less attention has been paid specifically to individuals' subjective evaluation of efforts (Clithero & Rangel, 2014; Bartra, McGuire, & Kable, 2013; Kable & Glimcher, 2009; Otto, Zijlstra, & Goebel, 2014). That is, how do individuals appraise the level of anticipated effort associated with some action within the decision-making process? In the current set of experiments, I examined subjective evaluations of task-specific efforts in three domains through the scope of General Evaluability Theory (Hsee & Zhang, 2010).

General Evaluability Theory

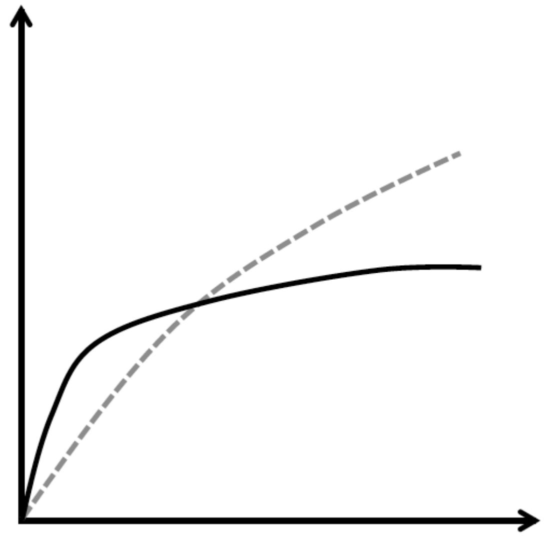
It is proposed that humans integrate various dimensions of an option (e.g., type and quantity of some reward) into a singular abstract measure of subjective value during evaluation and decision-making processes (Hsee & Zhang, 2010; Kable & Glimcher, 2009; Kahneman & Tversky, 1979). A critical theoretical issue in understanding individuals' subjective evaluations of a given attribute (here effort) is their *value sensitivity* (i.e., the responsiveness of evaluations to changes in the value of the attribute). One influential theory of value sensitivity is the General Evaluability Theory (GET; Hsee & Zhang, 2010; also see Hsee, Loewenstein, Blount, & Bazerman, 1999; Hsee, 1996). According to GET, value sensitivity is dependent on the *evaluability* of an attribute value, defined as the extent to which a person possesses relevant reference information needed to gauge values and map them onto a subjective evaluation (Hsee & Zhang, 2010). In a typical university student sample, an example of an evaluable attribute might be a students' GPA. In making a judgment about job prospects, students would be expected to be sensitive to a difference between candidates with a 1.9 GPA and a 3.5 GPA. An example of an inevaluable attribute (for students) might be the value of a diamond based on size. In estimating the value of a diamond, students might not be sensitive to a difference between a 10 karat and a 15 karat diamond.

Within GET, evaluability is dependent on three types of reference information: *mode*, *knowledge*, and *nature*. First, *mode* refers to the evaluation mode that a person is placed in. Any evaluation takes place in either a (1) Joint Evaluation (JE) mode in which two or more alternatives are explicitly juxtaposed and evaluated comparatively, or a (2) Single Evaluation (SE) mode where evaluation of a single value takes place in isolation. As an example, individuals may make a choice between renting one of two apartments (i.e., JE mode), or individuals can provide a willingness-to-pay estimate of only one of the options without knowledge of the other (i.e., SE mode). *Knowledge* refers to the distributional information, such as the variability and average, about some target attribute that is gained through experience. For example, college students have greater knowledge about Grade Point Average (GPA) than diamond size. Last, *nature* refers to whether individuals possess a stable physiological or psychological reference system to evaluate some value, for example, temperature. Those values that have such a scale are considered *inherently evaluable.* Importantly, these three types of reference information are conjunctive in determining sensitivity, or consistency, of valuation. That is, the manipulation of one type of information such as *knowledge* would be expected to modulate the influence of another type of information such as *mode*.

GET provides a theoretically motivated means of assessing the evaluability of a given attribute as it posits a straightforward relation between level of evaluability and mode: low-sensitivity attributes will produce less evaluability in the SE mode relative to the JE mode. As Hsee et al. (1999) note, difficult-to-evaluate attributes have little influence in differentiating the evaluations of values in SE, whereas in the comparative JE mode difficult-to-evaluate attributes become easier to evaluate and hence exert a greater influence. By contrast, easier-to-evaluate attributes will have similar impact in SE and JE. The effect of mode (SE vs. JE) should then be observable in individuals' subjective value functions (i.e., the functions that demonstrate how objective quantities map onto subjective evaluations; Kahneman & Tversky, 1979). GET proposes that subjective value functions for relatively inevaluable attributes are more linear in JE relative to SE (see Figure 11) by virtue of individuals being similarly sensitive to qualitative/categorical (i.e., from some meaningful reference point) and quantitative/continuous differences (i.e., judgments not made to the reference point) across incremental values. In the SE mode, in contrast, individuals are expected to be more sensitive to qualitative differences from a reference point and to show less sensitivity across quantitative values thereafter similar to a log function. Examining the relation between levels of some value, such as effort, and subjective values across evaluation modes thus provides a useful tool in determining the evaluability of a given value.

Figure 11.

Subjective Value Functions for a Relatively Inevaluable Attribute under Single and Joint Evaluation Modes



Note: The subjective value function for the Single Evaluation (SE) mode is presented as the solid line, whereas the function for the Joint Evaluation (JE) mode is presented as the dashed line. General Evaluability Theory proposes that subjective value functions are more linear in JE than in SE across values (adapted from Hsee & Zhang, 2010).

The Evaluability of Effort

Although the evaluability of effort has yet to receive considerable attention, it would seem plausible to suggest that effort should be a highly evaluable dimension. That is, effort may be both high in knowledge, where individuals putatively have a swath of experience making effort-based decisions, and in nature where effort may be inherently evaluable. Consistent with these ideas, there have been a number of demonstrations that individuals can behave in a manner that would suggest the ability to accurately evaluate effort (e.g., Bitgood & Dukes, 2006; Gray, Sims, Fu, & Schoelles, 2006; Kool et al., 2010; Kurniawan et al., 2010; Siegler & Lemaire, 1997; Walsh & Anderson, 2009). As an example, utilizing a free-choice task where participants made decisions between simply holding a grip device and engaging in effortful gripping (i.e., squeezing of air compressed cylinders), Kurniawan and colleagues (2010) demonstrated less frequent engagement in the high-effort gripping option. In a similar vein, Walsh and Anderson (2009) demonstrated that as the effort to successfully compute a solution to a multiplication problem increased, and consequently performance decreased, reliance on an external strategy of using a calculator increased. Similar findings consistent with a least effort principle have additionally been demonstrated in a variety of animal behaviors, for example, foraging (Marsh, Schuck-Paim, & Kacelnik, 2004; Stephens & Krebs, 1986).

While the evaluability of effort appears intuitive, it might be prudent to consider whether different types of effort may be more or less evaluable. One critical distinction to make may thus be between more physical and more cognitive (mental) forms of effort. With regard to physical effort, one could hypothesize that individuals possess a clear signal (e.g., energetic costs) needed to accurately evaluate the costs of expending physical effort. However, this does not appear to be the case with regard to cognitive effort. Cognitive effort is a far more challenging construct to define and study empirically (Botvinick & Braver, 2015; Dunn & Risko, 2016; Kurzban, 2016; Westbrook & Braver, 2015). Although many accounts have attempted to generalize the energetic aspect of physical effort accounts to cognitive effort accounts (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Boksem & Tops, 2008; Gailliot & Baumeister, 2007), this premise has been met with a large amount of skepticism based on theoretical and empirical grounds (e.g., Botvinick & Braver, 2015; Carter & McCullough, 2013; Gibson, 2007; Hockey, 2011; Inzlicht & Schmeichel, 2013; 2012; Job, Walton, Bernecker, & Dweck, 2013; Kelly, Sunram-Lea, & Crawford, 2015; Lange & Eggert, 2014; Lurquin et al., 2016; Kurzban, 2010; Kurzban, Duckworth, Kable, & Myers, 2013; Raichle & Mintun, 2006; Vadillo, Gold, & Osman, 2016;

Westbrook & Braver, 2015). Furthermore, whether similar systems underlie evaluation of both physical and cognitive effort remains to be determined (Westbrook & Braver, 2015; cf. Boksem & Tops, 2008). Therefore, the information (e.g., the reference information) available to individuals while attempting to evaluate physical versus cognitive effort may indeed differ.

Present Investigation

In Chapter 3, I use the GET framework to examine the evaluability of task-specific efforts via manipulations of evaluation mode in the context of effort judgments across perceptual, memorial, and motor tasks. Here I focus specifically on evaluations of *anticipated* effort as opposed to *experienced* effort. Payne and colleagues (1993) make a critical distinction between the two arguing that it is the evaluations and judgments of *anticipated* (or perceived) effort that play the primary role in strategy selection and decision-making, for example, deciding to not engage in a task outright. Nonetheless, experienced effort can indeed theoretically become anticipated effort given some hypothesized monitoring mechanism (e.g., Anzai & Simon, 1979; Koriat & Levy-Sadot, 2001; Payne et al., 1993; Son & Metcalfe, 2005; Vernon & Usher, 2003). Several recent proposals have argued that perceived effort is subjective in nature (Dunn & Risko, 2016; Kool et al., 2010; Westbrook & Braver, 2015) serving as a type of "summary signal" used to select lines of action (see Chapter 1). It is important to note, though, that while effort is often closely coupled with other potential determinants of behavior (e.g., task difficulty) and often covaries with similar signals (e.g., fatigue or arousal), effort can be understood as a unique cost considered in the decision-making process (for reviews on these issues see Kurzban, et al., 2013, and Westbrook & Braver, 2015).

The application of the GET framework to effort affords an examination of whether different task-specific efforts can be considered evaluable. The extent to which effort judgments may vary across JE and SE modes can provide evidence concerning the evaluability of a given type of task-specific effort. Critically, attributes that are high in evaluability do not demonstrate increased sensitivity in the JE mode because they are evaluable in the SE mode as well. Highly evaluable attributes should thus be consistently evaluated (i.e., demonstrate similar patterns of judgments) across the two modes. If a given task-specific effort does not show a susceptibility to evaluation mode, then I would expect subjective rating functions in both SE and JE to be similar (cf. Figure 11). Such a pattern, if observed consistently across different types or manipulations of effort, would provide initial evidence that effort and its determinants are evaluable. That being said, given the partial exploratory nature of Experiment 7, no specific hypothesis is offered concerning whether one specific type of effort is expected to be evaluable or not.

To foreshadow, in Chapter 3, six studies were carried out utilizing the GET framework for examining mode by value effects (Hsee & Zhang, 2010). I examined evaluations of effort where individuals rated perceived effort related to stimulus rotation and stimulus degradation in a reading task, set size pertaining to a short-term memory task, and lifting various degrees of weight. Of these four specific tasks, results suggested that perceived effort related to stimulus rotation is the least evaluable, whereas the perceived effort associated with the other three tasks could be considered relatively evaluable.

Experiment 7

To examine the influence of evaluation mode on task-specific efforts, evaluation mode (i.e., JE and SE) was manipulated between subjects. Individuals assigned to the JE mode were presented with all values of a task to be evaluated together, whereas individuals assigned to the SE mode evaluated only one value of a task in isolation. I employed three types of task-specific efforts to be evaluated by individuals across three domains: perceptual, memorial, and motor. For the perceptual domain, effort was manipulated by stimulus rotation, for the memorial domain by the number of to-be-remembered items, and for the motor domain by weight to-be-lifted (see below for more details). I chose these rather simple tasks in order to isolate specific types of effort as they allowed straightforward (and empirically confirmed) parametric increases in the putative effort within each task. As an example, increasing the number of items to-be-remembered increases the amount of perceived effort associated with engaging in the task (e.g., Risko & Dunn, 2015). Using a more complex "everyday" task (e.g., math), at least at this stage, could complicate the inferences that could be drawn about the specific type of effort being indexed.

Method

Participants

Five hundred and forty Amazon Mechanical Turk (MTurk) workers participated in the online study (see Buhrmester, Kwang, & Gosling, 2011) for compensation of $1 USD. One hundred and eighty individuals were assigned to each of the effort dimensions, 30 in each of the six nested effort manipulation judgment groups (five in SE and one in JE; see below). Twenty-four percent of individuals failed at least one of three attention checks embedded in the survey (see below) resulting in a final N of 411 (M_{Age} = 34 years, 47% female participants, 57% reported completing a Bachelor's degree or higher).

Design

A 3 (Effort Dimension: Perceptual Task, Memorial Task, Motor Task) x 5 (Effort Level: Stimulus Rotation – 0°, 45°, 90°, 135°, 180°; Set Size – 2-letters, 4-letters, 6-letters, 8-letters, 10-letters; Weight – 5 lbs., 10 lbs., 15 lbs., 20 lbs. 25 lbs.) x 2 (Evaluation Mode: Joint Evaluation, Single Evaluation) design was employed.

Stimuli

For stimulus rotation (perceptual), individuals judged visual displays consisting of the stimulus “WORD” rotated from 0° to 180° in 45° increments that were displayed on the screen to the participants. The use of “WORD” rather than an actual word was to tune individuals to the perceptual manipulation, rather than incorporating a potential confound of an actual word (e.g., a word randomly high in concreteness) driving judgments. Stimuli for items to-be-remembered (memorial) consisted of randomized letter strings presented in audio form through the survey software. Individuals were required to hit “Play” for each one of the set sizes to hear the specific stimuli over their headphones or speakers (the requirement of having working speakers or headphones was outlined to participants prior to consent). Set sizes ranged from 2-letters to 10-letters in increments of two, with each letter presented at 1-second intervals. For weight-to-be lifted (motor), a visual diagram featuring an individual lifting an unlabeled bag from the ground in three steps with the weight to be judged labeled beneath the diagram was presented to individuals on their screen. Presented weight ranged from 5 lbs. to 25 lbs. in five-pound increments. For perceptual effort conditions, individuals were asked “*How effortful would it be to read this word aloud?* For memorial effort conditions, individuals were asked “*How effortful would it be to recall all X letters immediately in the order that they are presented?”* For motor effort conditions, individuals were asked “*How effortful would it be to lift X lbs. starting from the ground?”* The evaluation scale was kept consistent across modes (i.e., a sliding 0-100 scale; see below).

Procedure

MTurk workers selected and accepted the Human Intelligence Task (HIT) and provided informed consent electronically. All participants first read instructions outlining the rating scale to be used in the study. Instructions stated that individuals were to make their judgments on a scale ranging from "*0 – No Action Taken"* to "*100 – Full Effort*". A rating of "*0 – No Action Taken*" was explained as entailing not engaging in the task outlined to the participant. For example, if an individual was assigned to the motor effort group, then instructions stated that a rating of "*0 – No Action Taken*" would entail *not* attempting to lift the amount of weight specified in the question. This lower anchor was chosen in an attempt to keep ratings off the floor, as well as to encourage individuals to imagine at least attempting the presented task when generating their effort rating (e.g., theoretically there should not be any "0" ratings if individuals are following instructions).

Individuals were then asked to move the rating scale to "*0 – No Action Taken*" and move onto the next portion of instructions. The next section outlined the "*100 – Full Effort*" rating. Individuals were instructed that a "*100 – Full Effort*" rating would entail "*…a mental or physical act that would require all of your effort to complete successfully (i.e., if it was any more effortful you would not have been able to complete it successfully)*". Furthermore, individuals were asked to freely respond in a text box with a description of "*one mental or physical act that they had completed in the past that required all of their effort to complete successfully*" and instructed that this act (i.e., the one self-reported) would be equivalent to a "*100 – Full Effort*" rating. Individuals were then asked to move the rating scale to "*100 – Full Effort*". The movement of the scale to "*0 – No Action Taken*" and to "*100 – Full Effort*", as well as the free response regarding an act entailing a *"Full Effort"* action, served as attention checks. All instructions were then briefly reiterated before individuals moved on to the judgment portion of the survey.

Evaluation mode was manipulated between subjects. Individuals assigned to the JE mode were presented with all values to be evaluated together, whereas individuals assigned to the SE mode evaluated only one value in isolation. Individuals assigned to the SE conditions received only one randomly assigned effort manipulation (e.g., for stimulus rotation, either 0°, 45°, 90°, 135°, or 180°). Individuals in the JE conditions were presented with all five effort levels sequentially from the lowest effort manipulation to the highest (see Hsee & Zhang, 2004, for a similar approach). Upon completion of the judgment portion of the survey, individuals were asked to complete three short demographic questions about their age, sex, and highest level of education completed. Individuals were then given the option to provide any feedback to the researchers and debriefed electronically.

Results

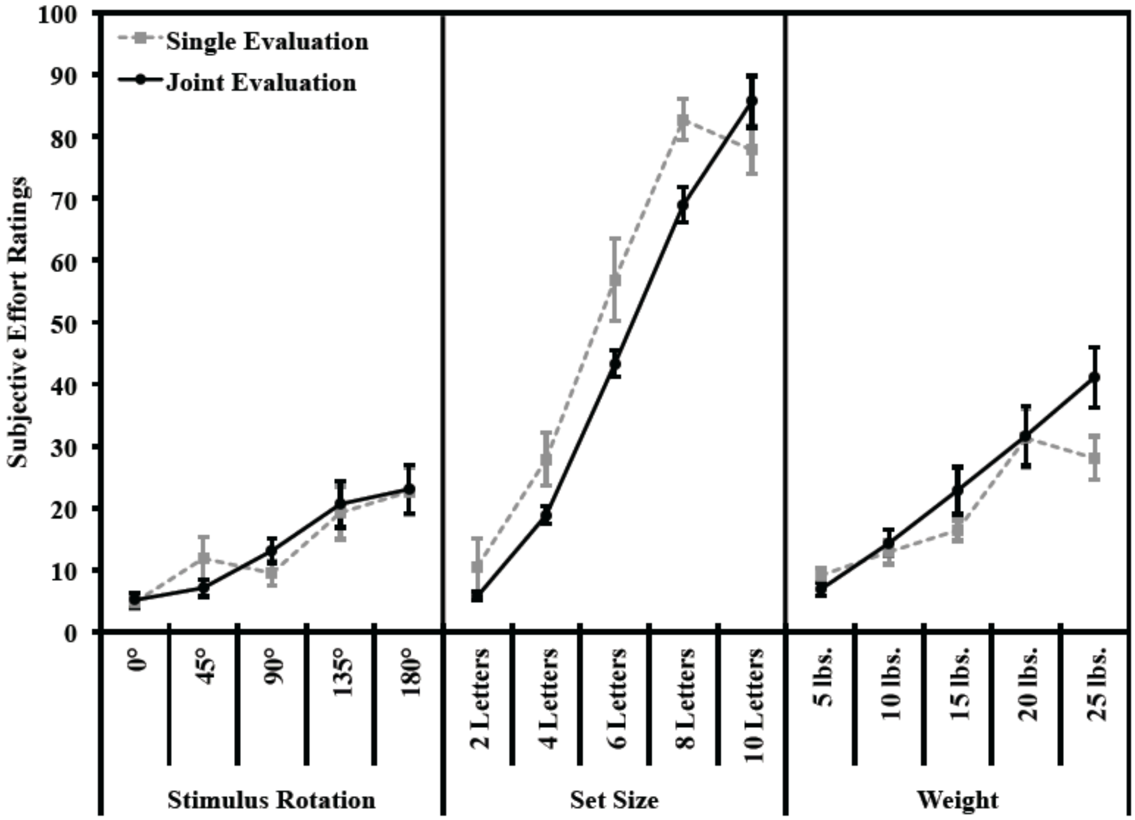
All reported analyses were conducted using R statistical software (R Development Core Team, 2015). Results are reported first for JE judgments followed by SE judgments for each effort dimension (see Figures 12 and 13). For JE judgments (i.e., within-subject judgments), Linear Mixed Models (LMM) were constructed using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015). All models incorporated a crossed random effect structure including random subject slopes, and slope-by-intercept correlations[10] (Baayen, Davidson, & Bates, 2008). In addition, the *RePsychLing* package (Bates, Kliegl, Vasishth, & Baayen, 2015) was employed

[10] All random effect correlations produced across efforts were relatively negative. These results suggest a type of "fanning in" pattern of individuals' within-subject ratings. Specifically, individuals who started their ratings lower on the scale generated more positive slopes, whereas individuals who started higher on the scale generated less positive slopes, reflecting in the latter case a type of ceiling effect for ratings.

to ensure that random effect structures were not over-fitted (cf. Barr, Levy, Scheepers, & Tily, 2013). Significance criterion for slope terms was set as $| t | > 2$ following Baayen et al. (2008). Model assumptions were assessed using visual depictions of residuals plots using the *car* package (Fox & Weisberg, 2011). In addition, influential case analysis (i.e., Cook's distance) was conducted using the *influence.ME* package (Nieuwenhuis, Grotenhuis, & Pelzer, 2012). To test slope model goodness-of-fit, the LMM containing the slope term (i.e., effort level for a particular dimension) was compared against an intercept-only model using a Log Likelihood test. Single evaluation judgments were analyzed using linear regression models (LM). Model assumptions and influential case analyses followed a similar procedure to LMMs. Removed cases (e.g., trials for LMMs and subjects for LMs) are reported at the start of each effort dimension section for both LMMs and LMs with all procedures following an iterative process for removal. Standardized beta values (β) and bootstrapped confidence intervals (CIs) are provided for all slope estimates. Last, all models were visually compared to loess fits to ensure that linear models relative to non-linear models were the most appropriate fit to the data post hoc.

Figure 12.

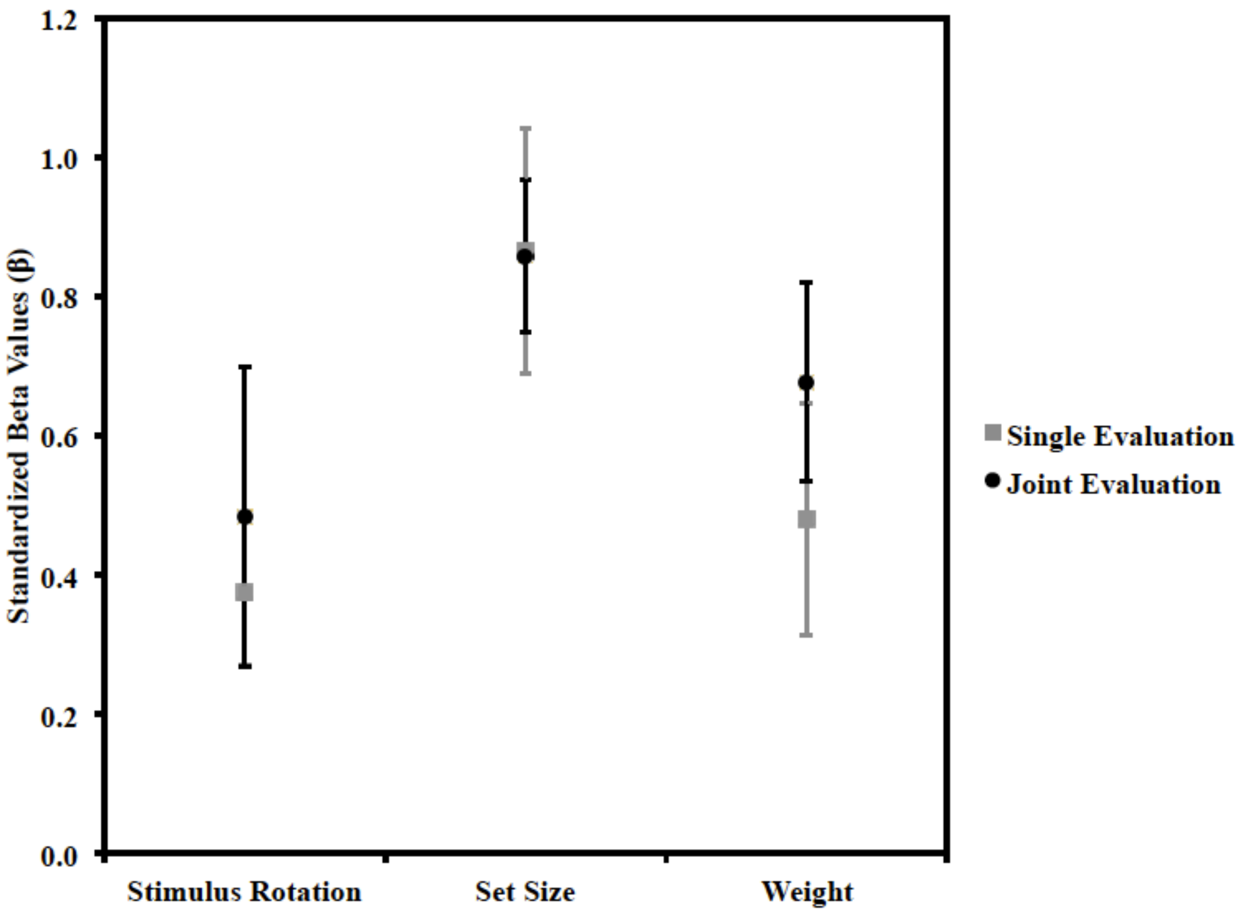
Single and Joint Evaluation Subjective Effort Rating Results in Experiment 7



Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings. Error bars represent ±1 SEM.

Figure 13.

Standardized Beta Values for Slopes in Experiment 7



Note: Single Evaluation (SE) represents between-subject ratings, whereas Joint Evaluation represents within-subject ratings. Error bars represent bootstrapped 95% confidence intervals.

Stimulus Rotation (Perceptual)

First, approximately 2% of cases were removed for joint evaluation. LMM results demonstrated a significant positive slope associated with increased degree of stimulus rotation, $b = 4.71$, $SE = 1.07$, $t = 4.42$, 95% CI [2.61, 6.69], $\beta = .48$, β 95% CI [.27, .69]. The slope model significantly improved model fit relative to the intercept-only model, $\chi2\,(1) = 15.01$, $p < .001$.

For single evaluation, approximately 8% of cases were removed. Similar to LMM results, LM results demonstrated a significant positive slope associated with increased stimulus rotation, b = .07, SE = .02, t = 4.16, p < .001, 95% CI [.04, .1], R^2 = .14, β = .37, β 95% CI [.2, .55]. The slope model significantly improved model fit relative to the intercept-only model, $F(1, 106) = 17.29$, residual SE = 10.47, p < .001.

Set Size (Memorial)

Approximately 4% of cases were removed in joint evaluation. LMM results demonstrated a significant positive slope associated with increased set size, b = 21.17, SE = 1.38, t = 15.37, 95% CI [18.21, 23.98], β = .86, β 95% CI [.75, .98]. The slope model significantly improved model fit relative to the intercept-only model, $\chi2\,(1) = 59.6$, p < .001. For single evaluation, approximately 5% of cases were removed. LM results demonstrated a significant positive slope associated with increased set size, b = 10.9, SE = .67, t = 16.35, p < .001, 95% CI [9.87, 11.91], R^2 = .75, β = .87, β 95% CI [.76, .97]. The slope model significantly improved model fit relative to the intercept only model, $F(1, 89) = 267.3$, residual SE = 17.93, p < .001.

Weight (Motor)

Approximately 2% of cases were removed in joint evaluation. LMM results demonstrated a significant positive slope associated with increases in weight, b = 8.01, SE = .86, t = 9.29, 95% CI [6.44, 9.69], β = .68, β 95% CI [.53, .82]. The slope model significantly improved model fit relative to the intercept-only model, $\chi2\,(1) = 40.21$, p < .001. Approximately 8% of cases were removed in single evaluation. LM results demonstrated a significant positive slope associated with increased weight, b = .7, SE = .12, t = 5.64, p < .001, 95% CI [.48, .98], R^2 = .23, β = .48, β

95% CI [.31, .65]. The slope model significantly improved model fit relative to the intercept only model, $F(1, 106) = 31.78$, residual $SE = 8.99$, $p < .001$.

Discussion

Experiment 7 demonstrated several interesting findings with regard to the evaluability of task-specific efforts. First, effort judgments for the memory task produced very similar positive slopes across both the JE and SE evaluation modes. This finding provides initial evidence that effort associated with items to-be-remembered may be highly evaluable. Both stimulus rotation and weight to-be-lifted showed similar patterns of perceived effort judgments across evaluation modes, but not to the same extent as the memory task. Slopes were somewhat more positive in JE relative to SE, though both dimensions did produce significant linear functions (i.e., all effort dimensions remained relatively linear in the SE mode). That is, for all dimensions across between-subject raters only rating a single effort level in isolation, appeared to demonstrate value sensitivity in a similar fashion to individuals rating in JE where all alternatives were explicitly present. Such correspondence suggests that each task-specific effort dimension may have high levels of both *knowledge* and/or *nature* according to GET. Examination of the stimulus rotation effort slopes, and to some extent the weight effort slopes, though, suggests that the judgments were relatively constrained toward the bottom of the rating scale. Therefore, it is possible that the overall low perceived effort for these dimensions did not allow the shape of the function to be fully demonstrated across effort manipulations. Experiment 8 sought to address this issue.

Experiment 8

Individuals in Experiment 8 performed a task similar to Experiment 7. To increase effort ratings across the evaluation scale, and to provide a clearer picture of the functions associated with the stimulus rotation and weight dimensions, effort levels were increased while effort associated with the memory task was kept consistent with Experiment 7. For stimulus rotation, displays were increased from a one-word display to a nine-word 3 x 3 display. The addition of items in a rotated display is known to increase performance costs and, thus, the putative expected effort required to read the display (Risko, Medimorec, Chisholm, & Kingstone, 2014). For weight, effort levels were doubled relative to Experiment 1. Effort levels were presented in 10 lbs increments starting at 10 lbs and ending at 50 lbs. Experiment 8 additionally affords the opportunity to replicate the main findings from Experiment 7. If all task-specific effort dimensions are evaluable, then I would expect to observe similar positive linear functions across evaluation modes for all dimensions.

Method

Participants

Seven hundred and twenty MTurk workers participated in the online study. Two hundred and forty individuals were assigned to each of the effort dimensions, 40 in each of the six nested judgments groups (i.e., five in SE and one in JE). Eight percent of individuals failed at least one of three attention checks embedded in the survey resulting in a final N of 663 (M_{Age} = 33 years, 42% female participants, 50% reported completing a Bachelor's degree or higher).

Design

A 3 (Effort Dimension: Perceptual Task, Memorial Task, Motor Task) x 5 (Effort Manipulation: Stimulus Rotation – 0°, 45°, 90°, 135°, 180°; Set Size – 2 letters, 4 letters, 6 letters, 8 letters, 10, letters; Weight – 10 lbs., 20 lbs., 30 lbs., 40 lbs., 50 lbs.) x 2 (Evaluation Mode: Joint Evaluation, Single Evaluation) design was employed.

Stimuli

Stimuli for the memorial task manipulations were kept the same as in Experiment 7. For the perceptual task manipulations, stimulus rotation was kept constant, but set size was increased to nine words presented in a 3 x 3 display. Motor task manipulations were increased to 10 lbs. to 50 lbs. in 10 lbs. increments.

Procedure

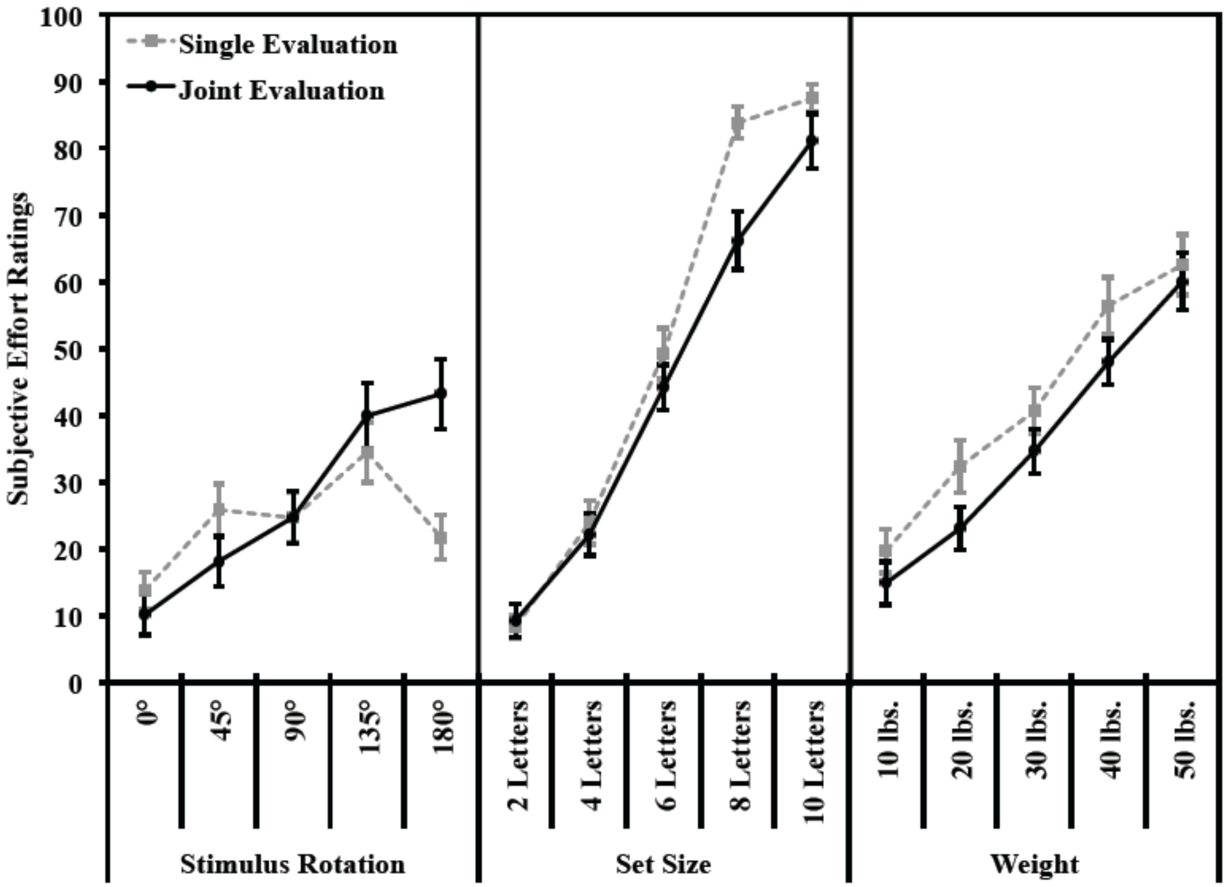
All procedures followed Experiment 7.

Results

All data analyses and reporting procedures followed Experiment 7 (see Figures 14 and 15).

Figure 14.

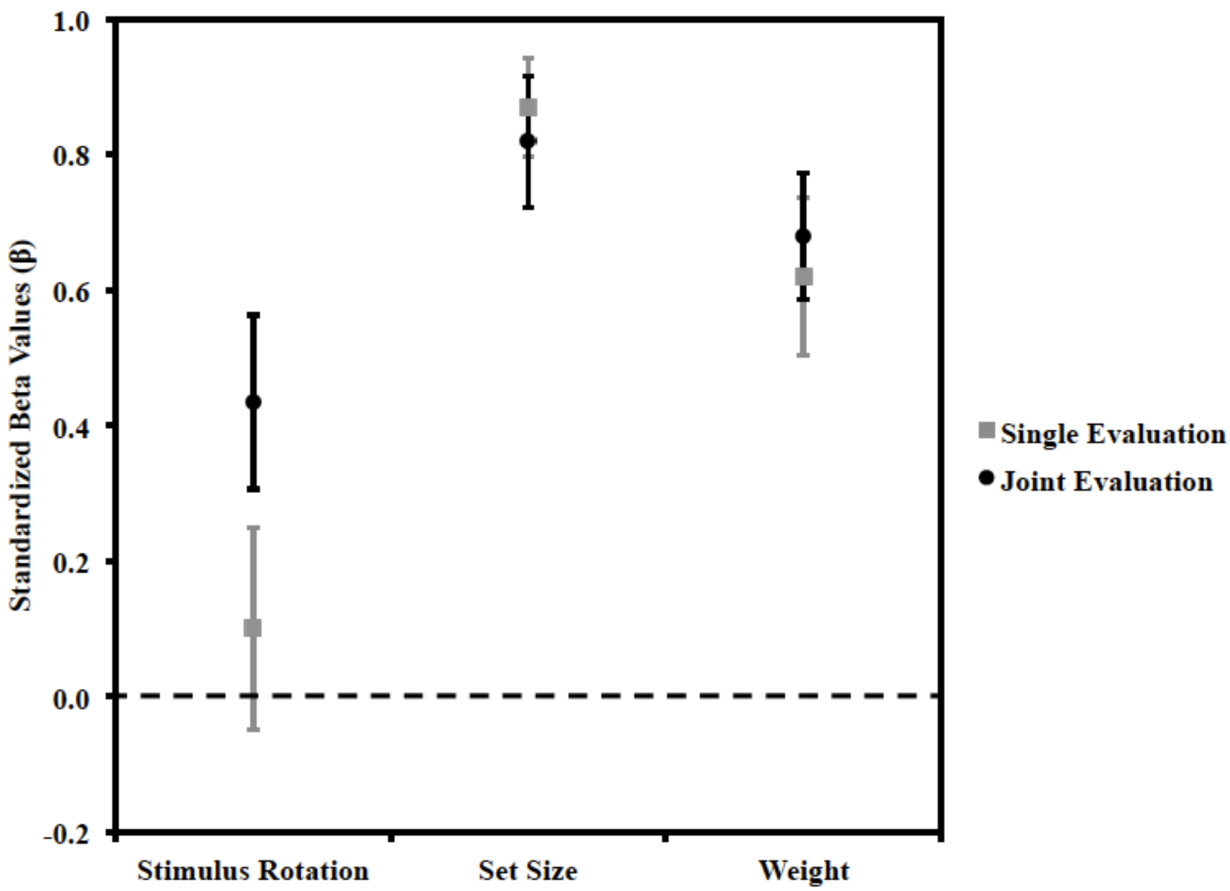
Single and Joint Evaluation Subjective Effort Rating Results in Experiment 8



Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings. Error bars represent ±1 SEM.

Figure 15.

Standardized Beta Values for Slopes in Experiment 8



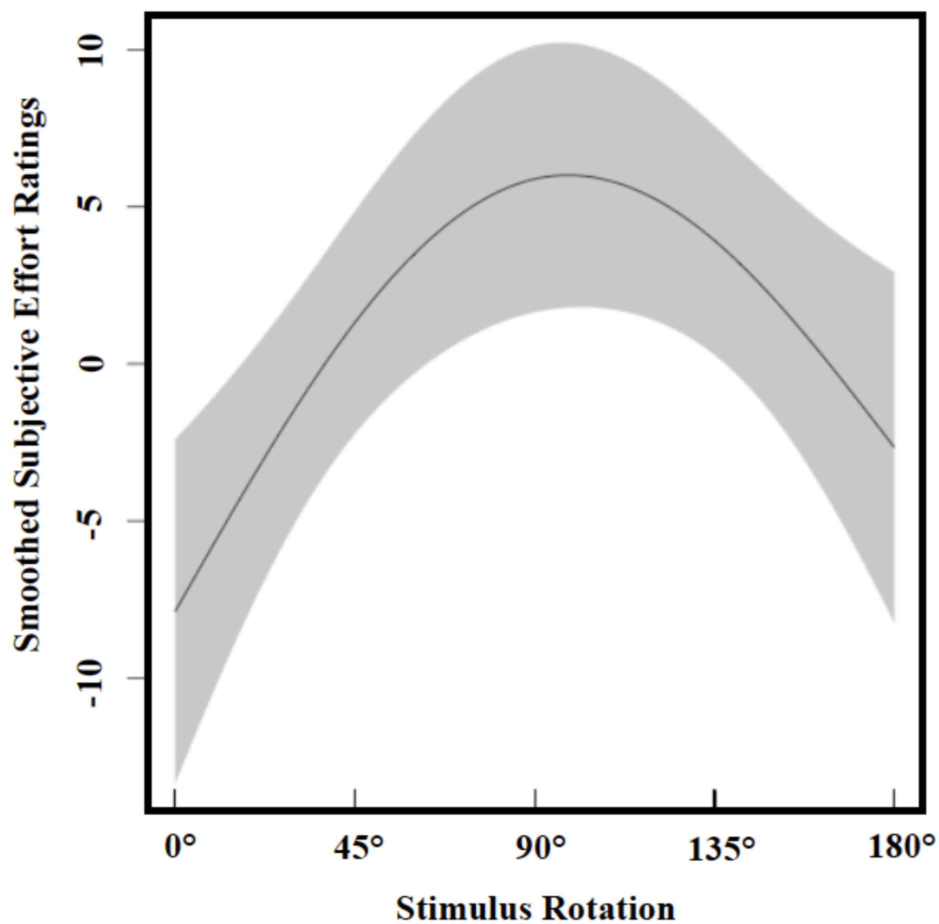
Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings Error bars represent bootstrapped 95% confidence intervals.

Stimulus Rotation

First**,** approximately 4% of cases were removed in JE. LMM results demonstrated a significant positive slope associated with increased degree of stimulus rotation, b = 8.33, SE = 1.26, t = 6.62, 95% CI [5.92, 10.91], β = .43, β 95% CI [.31, .56]. The slope model significantly improved model fit relative to the intercept-only model, $\chi2\,(1) = 29.48$, $p < .001$. For single evaluation approximately 6% of cases were removed. Conflicting with the JE results, LM results for SE judgments did not demonstrate a significant positive slope of stimulus rotation, b = .03, SE = .02, t = 1.32, 95% CI [-.01, .7], R^2 = .14, β = .1, β 95% CI [-.05 .25].

Figure 16.

Generalized Additive Model (GAM) Predicted Fit for Single Evaluation Stimulus Rotation Effort Ratings in Experiment 8



Note: Shaded region represents estimated SE of the smoothed estimate.

In addition to a demonstrated non-significant positive slope, visual inspection of loess fit to the data suggested a nonlinear model would perhaps best describe the SE data. Therefore, to test a non-linear fit, a generalized additive model (GAM) was constructed using the *mgcv* package (Wood, 2006). A cubic spline smoothing term was applied to degree of stimulus rotation utilizing three knots. Results demonstrated a significant smooth term, edf = 1.92, $F = 7.05$, approximate $p < .001$[11]. Furthermore, a deviance test demonstrated that the GAM model produced a better fit to the SE data relative to the LM model, $p < .001$, as well as a smaller Akaike information criterion (i.e., better goodness of fit; AIC; see Burnham, Anderson, & Huyvaert, 2011) value relative to the LM model, AIC = 1490, AIC = 1501.71, for the GAM and LM models respectively. Thus, GAM results suggest that the nonlinear model provided a better fit to the SE data relative to the LM model (see Figure 16). I return to the importance of this pattern in the discussion of this chapter.

Set Size

Approximately 4% of cases were removed in joint evaluation. LMM results demonstrated a significant positive slope associated with increased set size, $b = 11.18$, $SE = .48$, $t = 23.41$, 95% CI [17.36, 21.95], $\beta = .82$, β 95% CI [.72, .92]. The slope model significantly improved model fit relative to the intercept-only model, $\chi 2\,(1) = 78.09$, $p < .001$. Approximately 2% of cases were removed for single evaluation. LM results demonstrated a significant positive slope associated with increased set size, $b = 10.9$, $SE = .67$, $t = 16.35$, $p < .001$, 95% CI [10.51, 11.93], $R^2 = .75$, $\beta = .87$, β 95% CI [.8, .94]. The slope model significantly improved model fit relative to the intercept-only model, $F(1, 178) = 548$, residual $SE = 17.77$, $p < .001$.

[11] See Wood (2013) regarding issues associated with p-values and GAMs.

Weight

Approximately 6% of cases were removed in joint evaluation. LMM results demonstrated a significant positive slope associated with increases in weight, b = 12.22, SE = .86, t = 14.15, 95% CI [10.33, 13.99], β = .68, β 95% CI [.58, .77]. The slope model significantly improved model fit relative to the intercept-only model, $\chi 2\,(1) = 67.15$, $p < .001$. Approximately 3% cases were removed in single evaluation. LM results demonstrated a significant positive slope associated with increased weight, b = 1.24, SE = .12, t = 10.43, $p < .001$, 95% CI [1.01, 1.45], R^2 = .38, β = .62, β 95% CI [.5, .74]. The slope model significantly improved model fit relative to the intercept-only model, $F(1, 175) = 108.8$, residual SE = 22.1, $p < .001$.

Discussion

Experiment 8 replicated the main finding of Experiment 7 for the memorial and motor tasks: Both SE and JE produced highly similar positive linear functions of perceived effort ratings for the dimensions. Thus, Experiments 7 and 8 together suggest that the perceived effort associated with the specific memorial and motor tasks used here are evaluable. I return to potential explanations of why this may be the case for these dimensions in Experiment 11 and in the General Discussion of this chapter. In contrast to these tasks, stimulus rotation produced a flat slope in SE (as well as a nonlinear fit), whereas a positive linear pattern was observed in JE. As highlighted in the introduction, GET proposes that subjective value functions are more linear across values in JE relative to SE for inevaluable attributes (see Figure 11) due to individuals being similarly sensitive to quantitative and qualitative differences in JE, but only being sensitive to qualitative differences (i.e., from some meaningful reference point) with ratings asymptoting thereafter in SE. Individuals in SE seemed to be insensitive to the incremental differences in stimulus rotation past the 0° - 45° comparison. That is, individuals' effort judgments were sensitive to the qualitative shift from the 0° reference point, but not sensitive to the incremental differences between 45° - 90°, 90° - 135°, and 135° - 180°. Therefore, as opposed to Experiment 7 when the putative perceived effort was increased for stimulus rotation, the perceived effort associated with the task appeared to be relatively inevaluable.

Experiment 9

One potential explanation for the pattern of results for effort associated with stimulus rotation in Experiment 8 is that, in addition to 0° (i.e., upright), 180° stimulus rotation (i.e., upside down) may also serve as a meaningful reference point in judgments. This can be seen in the large drop in effort ratings from 135° to 180° (see Figure 14). Therefore, the nonlinear pattern in SE and, thus, the presumed inevaluability of the expected effort associated with stimulus rotation, may have been driven by the inclusion of two potential reference points within the effort level manipulation. Interestingly, the pattern of ratings produced in SE does somewhat follow patterns of response times (i.e., a performance index of effort) for reading simple rotated words aloud (see Koriat & Norman, 1985). Thus, an alternative account is that individuals are sensitive to the effort (in terms of response times) associated with stimulus rotation in SE, but having several alternatives present in JE modulates the pattern to be more positive. To address this, Experiment 9 manipulated stimulus rotation effort at 15° increments ranging from 0° to 90°, thus excluding the potential 180° reference point. If perceived effort associated with stimulus rotation is inevaluable, then I would expect to find similar results as Experiment 8: Individuals in SE should demonstrate greater sensitivity to the qualitative difference between 0° and 15°, but not demonstrate the same sensitivity to all differences thereafter. Judgments in JE should demonstrate a positive linear function across all differences. Alternatively, if individuals are sensitive to the response times associated with reading rotated words, then I would expect similar flat (or non-linear) functions in SE and JE.

Method

Participants

Four hundred MTurk workers participated. Fifty individuals were assigned to each of the eight nested judgment groups (i.e., seven in SE and one in JE). Eight percent of individuals failed at least one of three attention checks embedded in the survey resulting in a final *N* of 368 (M_{Age} = 34 years, 49% female participants, 50% reported completing a Bachelor's degree or higher).

Design

A 2 (Evaluation Mode: Joint Evaluation, Single Evaluation) x 7 (Stimulus Rotation: 0°, 15°, 30°, 45°, 60°, 75°, 90°) design was employed.

Stimuli

Set size was reduced to two-word displays and rotated from 0° to 90° in 15° increments.

Procedure

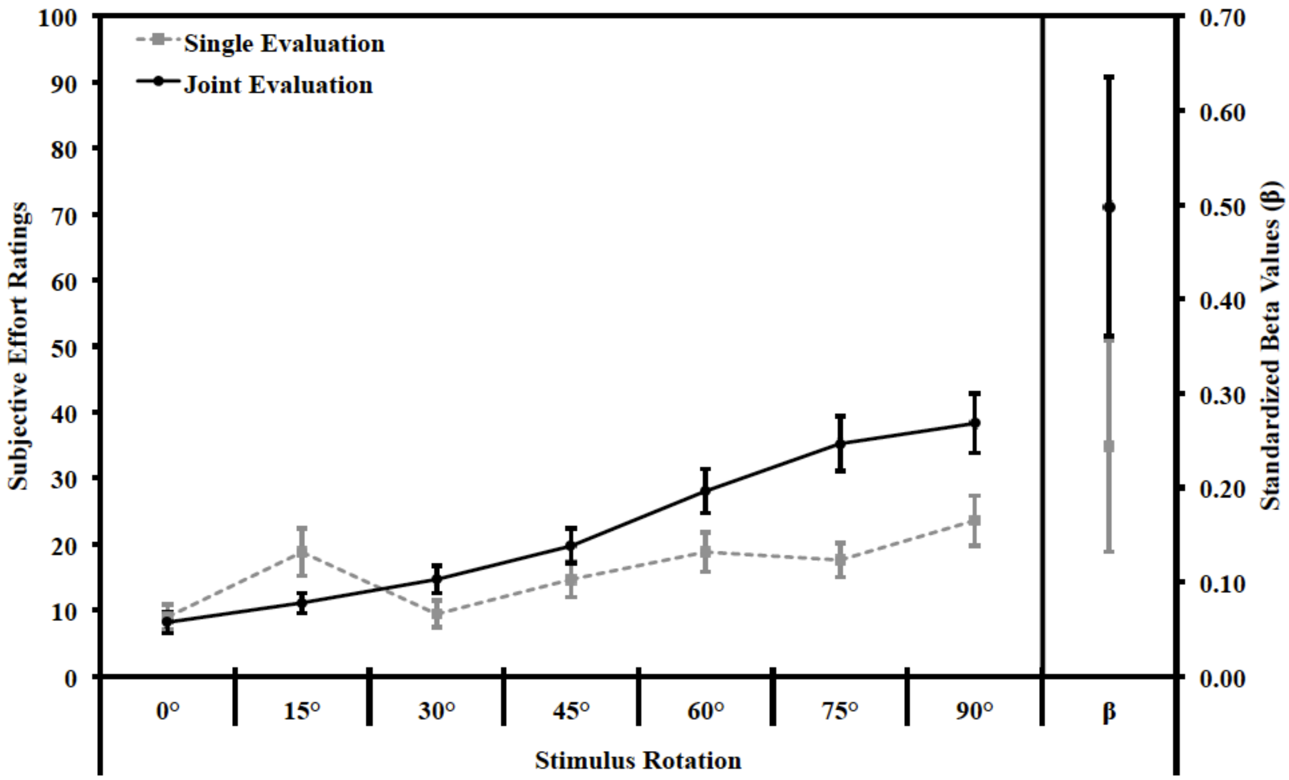
All procedures followed Experiments 7 and 8.

Results

All data analyses and reporting procedures followed Experiments 7 and 8.

Figure 17.

Single and Joint Evaluation Subjective Effort Rating Results (Left Panel) and Standardized Beta Values for Slopes (Right Panel) in Experiment 9



Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings. Error bars for the left panel represent ±1 SEM. Error bars for the right panel represent bootstrapped 95% confidence intervals.

Stimulus Rotation

Approximately 6% of cases were removed in JE. LMM results revealed a significant positive slope associated with increased degree of stimulus rotation, b = .36, SE = .05, t = 7.1, 95% CI [.27, .46], β = .5, β 95% CI [.36, .64]. The slope model significantly improved model fit relative to the intercept-only model, $\chi 2$ (1) = 35.02, p < .001. For single evaluation approximately 10% of cases were removed. Linear model results for SE ratings revealed a significant positive slope of stimulus rotation, $b = .09$, $SE = .02$, $t = 4.26$, $p < .001$, 95% CI [.05, .14], $R^2 = .06$, $\beta =$ .24, β 95% CI [.13, .36]. Thus, both evaluation modes produced significant positive slopes, however, comparing β values across the modes reveals that the slope fit in JE is more positive than the slope fit in SE (see Figure 17).

Discussion

Results in Experiment 9 demonstrated that individuals in SE produced only a slightly positive linear function across the effort levels, however, the function for JE was much more positive suggesting that individuals are not sensitive to response times associated with reading rotated words. Thus, similar to Experiment 8, there is a discrepancy between judgments made in SE relative to JE. Examination of the SE function demonstrates the large increase in ratings from the 0° reference point to the first effort level (15°), as predicted by GET. That is, individuals demonstrated sensitivity to the reference point and less sensitivity to incremental differences thereafter. Experiment 9 lends further evidence to the claim that the perceived effort associated with stimulus rotation may be relatively less evaluable.

Experiment 10

To this point in Chapter 3, I have focused on the relation between SE and JE to examine the evaluability of task-specific efforts through individuals' judgments of effort while keeping the evaluation scale consistent (i.e., a 0-100 scale). In addition to this method of examining evaluability, GET additionally predicts specific patterns of results across ratings and choices (e.g., would you buy A or B?) for low evaluability attributes. For instance, in GET, preference reversals across evaluation modes are credited to low evaluability values becoming more evaluable in JE choice relative to SE ratings (see similarly the *Prominence Effect*; Tversky, Sattath, & Slovic, 1988). Preference reversals (see Lichtenstein & Slovic, 2006, for a review) occur when a systematic change in preference order between normatively equivalent conditions is observed (Slovic & Litchenstein, 1983) and, as such, represents an internal inconsistency in judgment (Hsee, et al., 2004; Kahneman, 1994). For example, Hsee (1996) had individuals evaluate job candidates for a programming position on two attributes: GPA and experience. Hsee (1996) hypothesized that GPA would be the more evaluable attribute of the two given students' putative knowledge about GPA. In SE, the candidate with the higher GPA was favored, whereas in JE the candidate with more programming experience was favored, demonstrating a preference reversal across evaluation modes. The authors thus argued that GPA was more evaluable to students relative to experience. Hence, examining inconsistencies across ratings and choice represents an additional gauge of value sensitivity.

In Experiment 10, I utilized the stimulus rotation manipulation from Experiments 7-9 crossed with a set size manipulation (i.e., number of words in the display). Experiments 7-9 revealed that set size, in terms of number of items in the display, may be an evaluable attribute. Ratings in the 0° condition (i.e., upright) across the 1-word (Experiment 7), 2- word (Experiment 9), and 9-word (Experiment 8) set sizes demonstrated a relatively linear increase in ratings as set size increased (see Figures 12, 14, and 17). Therefore, similar to the example above, a relatively evaluable attribute is pit against a relatively inevaluable one. In SE, individuals rated either a one-word display rotated 90° or a two-word display rotated only 15° on the same 0-100 scale used previously. Individuals in JE were presented with both displays and were asked to make a choice about which of the two displays would be more effortful to read aloud.

I would expect individuals to rate the two-word display rotated 15° as more effortful relative to the one-word display rotated 90° in SE. This prediction is derived from SE results from Experiments 7 and 9. Individuals assigned to SE in Experiment 7 rated the one-word display rotated 90° as less-effortful to read than individuals assigned to SE in Experiment 9 rated the two-word display rotated 15° (see Figures 12 and 17). If the perceived effort associated with stimulus rotation is relatively inevaluable, then stimulus rotation should exert a greater influence in JE choice relative to SE ratings. That is, individuals should choose the one-word display rotated 90° as the more effortful alternative. Such a prediction for JE choice is counterintuitive given the clear difference in objective effort (e.g., as indexed by performance) between processing one word relative to two words (e.g., Dunn & Risko, 2016).

Method

Participants

Three hundred MTurk workers participated in the online study. Fifty individuals were assigned to the SE group and 100 to the JE group. Five percent of individuals failed one attention check embedded in the survey resulting in a final N of 368 (M_{Age} = 34 years, 49% female participants, 50% reported completing a Bachelor's degree or higher).

Design

A 2 (Evaluation Mode: Joint Evaluation, Single Evaluation) x 2 (Stimulus Rotation: 15°**,** 90°) x 2 (Set Size: 1 Word, 2 Words) design was employed.

Stimuli

Stimuli were similar to the previous experiments.

Procedure

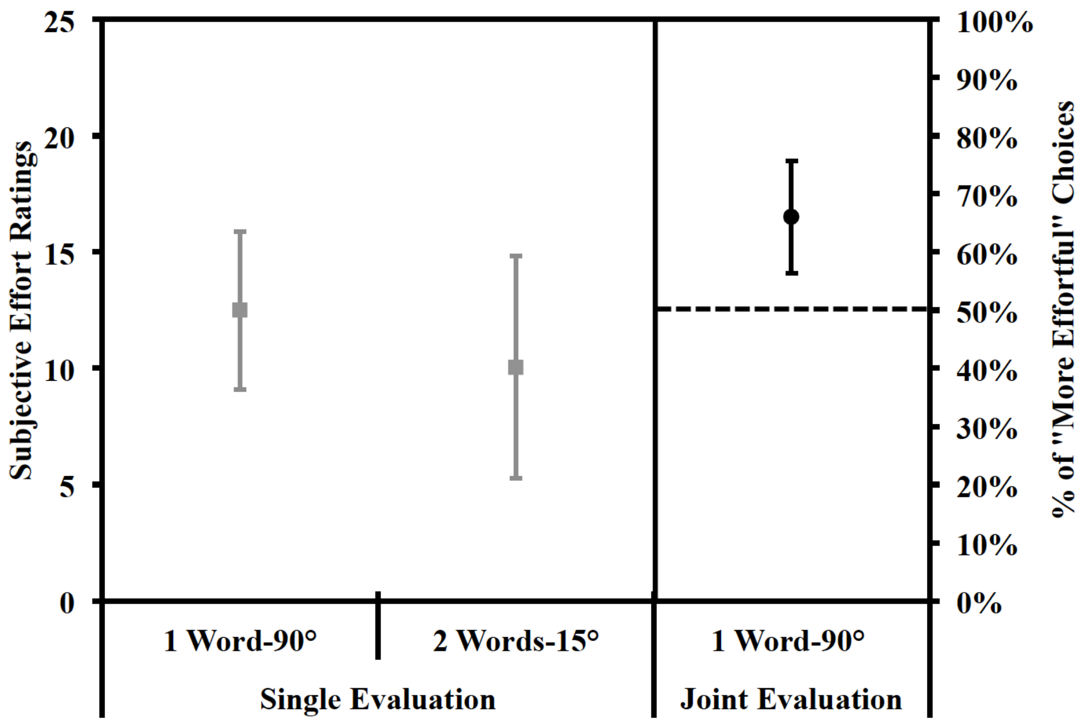
The procedure for SE was similar to the previous experiments. For JE, rather than providing judgments on the 0-to-100 scale, individuals were asked "*Which of the two displays above do you feel would be more effortful to read aloud?*" Individuals responded by either selecting "*Display A.*" or "*Display B". *Due to a program error, the position of the stimuli for JE (i.e., top or bottom position) was not counterbalanced across participants, and thus an additional 50 individuals were run through the JE condition (see above) to ensure equal presentation at the positions. Results were statistically similar across the runs therefore the aggregated data are reported for JE. The attention check for JE asked individuals to "*Please click on the Display A. button"* during the instruction phase.

Results

For SE, inferential statistics (i.e., between-group t-test) as well as Bayesian analyses were conducted on judgments with the *BEST* package (Kruschke, 2013) utilizing 100,000 estimates of the effect size (i.e., Cohen's *d)* by way of Markov Chain Monte Carlo (MCMC) sampling. In addition, 95% Highest Density Intervals (HDI) are presented, as well as Bayes Factors (BF) computed using the *BayesFactor* package (Morey & Rouder, 2015). Bayes Factor interpretation follows the criteria outlined by Kass and Raftery (1995). Furthermore, visual inspection of the SE judgment data revealed signs of outliers (skewness = 2.66), thus a grand-mean outlier cut was employed using a 2.5 SD cut-off criterion. This procedure resulted in the removal of approximately 3% of cases. For JE data, a Binomial Test was conducted on individuals' "more effortful" choices as well as a Bayes Factor test for binomial data.

Figure 18.

Single and Joint Evaluation Results in Experiment 10



Note: The left panel represents between-subject SE ratings; the right panel represents within-subject JE choices. Error bars in both panels represent bootstrapped 95% confidence intervals.

First, in SE ratings, individuals judged the 15°/2-Word display (M = 10.04, SD = 14.03) to be similarly effortful to read as the 90°/1-Word display (M = 12.94, SD = 11.99), $t(86) = -.87$, 95% CI [-7.99, 3.1], p = .38. Bayesian analyses revealed a simulated mode effect size of d = -.16, 95% HDI [-.58, .24], and positive evidence for the null, BF_{NULL} = 3.24. For JE choice, individuals selected the 90°/1-Word display as more effortful (66%, 95% CI [56%, 75%]) relative to the 15°/2-Word display, $p < .01$. Furthermore, a Bayes Factor computed for the binomial data demonstrated strong evidence for the alternative (i.e., that the proportion of choices is different from chance), BF_{Alt} = 31.95. Thus, individuals in SE rated the two displays as similarly effortful to read, whereas in JE, individuals selected the 90°/1-Word display as more effortful than the 15°/2-Word display (see Figure 18).

Discussion

Experiment 10 further examined the potential inevaluability of stimulus rotation by crossing stimulus rotation and a set size manipulation across a rating and choice context. In SE, individuals rated the 15°/2-Word display as similarly effortful to read aloud relative to the 90°/1-Word display (though the pattern was very slightly in the direction opposite to that predicted). In JE, however, individuals more often chose the 90°/1-Word display as the more effortful of the two alternatives. Although not a complete preference reversal, these results demonstrate an inconsistency across ratings and choice as predicted by GET if one inevauable attribute is included as an alternative alongside an evaluable attribute. That is, stimulus rotation exerted a greater influence in JE relative to SE. This was the case even in light of clear differences in objective effort across the two options. Stimulus rotation costs for single items are relatively small (e.g., Jolicoeur, 1990; Risko et al., 2014) and would not be expected to be larger than the cost of processing an additional item (i.e., reading a one-word display relative to reading a two-word display). Therefore, the results from Experiment 10 suggest, consistent with Experiments 7-9, that perceptual effort as indexed by stimulus rotation is weakly evaluable.

Experiments 11a and 11b

Experiments 7 through 10 have demonstrated, through the application of GET, that the perceived effort associated with stimulus rotation is relatively inevaluable in contrast to a memorial or motor task. One clear limitation is the constrained task-specific efforts used to this point (i.e., one task from the perceptual, motor, and memorial domains). From this, a clear question arising from Experiments 7-10 is whether the inevaluability evident with stimulus rotation is a product of perceptual effort being inherently inevaluable, or a product of stimulus rotation in-and-of-itself being relatively inevaluable.

To examine this possibility, in Experiment 11a individuals provided effort ratings based on an additional perceptual task: identifying degraded stimuli ranging from 0% pixel removal to 88% pixel removal in intervals of 22% pixels. Although stimulus degradation and stimulus rotation can both be considered perceptual manipulations, degradation arguably has associated with it a much clearer "failure point" (i.e., an upper limit in which an action cannot proceed without failure; see similarly Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). For example, such a failure point arguably exists for both set size and weight but not for the perceived effort associated with stimulus rotation (i.e., humans can process fully rotated stimuli at relatively minimal cost). Furthermore, this failure point *is* present for stimulus degradation as degradation can reach levels to the point where an item is unidentifiable. Therefore, following Experiments 7-9, if perceptual tasks are generally inevaluable, then I would expect individuals in SE to demonstrate greater sensitivity to the qualitative difference between 0% degradation and 22% degradation, but not to demonstrate the same sensitivity to all differences thereafter. Judgments in JE should demonstrate a relatively positive linear function across all differences. Alternatively, if evaluability is not contingent on domain-specific tasks but rather is driven by an inherent failure point, then I would expect similar subjective effort functions for both the SE and JE evaluation modes.

In addition, Experiment 11b looked to extend the general results from hypothetical perceived effort, as has been investigated thus far, to perceived effort where individuals were aware that they would engage in the task that they provide effort ratings for. Here, individuals were instructed that they would provide an effort rating (SE) or ratings (JE) for the levels of stimulus degradation and then be asked to attempt to read words that were presented to them on their screen. I offer no *a priori* hypothesis for why the pattern of ratings would differ if awareness of engaging in the task would affect perceived effort ratings relative to not having to engage in the task. However, if awareness that engaging in a task does not affect perceived effort ratings, then the patterns of subjective effort ratings should closely match those demonstrated in Experiment 11b.

Method

Participants

For Experiments 11a and 11b, 300 MTurk workers participated in the online study for compensation of .50¢ USD. Fifty individuals were assigned to each of the stimulus degradation dimensions in SE (i.e., 250 total for SE), and fifty were assigned to the JE group. In Experiment 11a, approximately eight percent of individuals failed at least one of three attention checks embedded in the survey resulting in a final N of 253 (M_{Age} = 35.5 years, 47% female participants, 61% reported completing a Bachelor's degree or higher). In Experiment 11b, approximately 14% of individuals failed at least one of three attention checks embedded in the survey resulting in a final N of 258 (M_{Age} = 32 years, 49% female participants, 53% reported completing a Bachelor's degree or higher).

Design

A 5 (Effort Level: Stimulus Degradation – 0%, 22%, 44%, 66%, 88%) x 2 (Evaluation Mode: Joint Evaluation, Single Evaluation) design was employed for both Experiments 11a and 11b.

Stimuli

Individuals judged displays consisting of the stimulus "WORD" degraded from 0% to 88% of pixel removal in 22% increments using a diagonal grating. Individuals were asked "*How effortful would it be to read this word aloud?*. For Experiment 11b, five high frequency nouns ($M_{WrittenFreq}$ = 326.80) were generated ("FELT", "WANT", "MIND", "DOOR", "HELP") and counterbalanced across five lists such that each word was presented equally across the five levels of stimulus degradation.

Procedure

The procedure for JE and SE ratings in Experiment 11a followed Experiments 7-9. The procedure for Experiment 11b closely followed the aforementioned experiments. However, individuals were instructed prior to completing their effort ratings that, upon completion of the ratings, they would be asked to attempt to read a word (for SE) or words (for JE) and to enter the word presented into a text box. In SE, individuals only received one word to attempt to read corresponding to the degradation level they rated. In JE, individuals received five words to attempt to read corresponding to all levels of the stimulus degradation manipulation.

Results

All data analyses and reporting procedures followed Experiments 7-9. Subjective effort rating results are first presented for Experiment 11a followed by Experiment 11b (see Figure 19 and Figure 20). Generalized additive mixed models (GAMM; see below) were constructed using the *mgcv* package (Wood, 2006).

Experiment 11a

Effort Ratings

Joint Evaluation Mode. No cases were removed for JE ratings. LMM results revealed a significant positive slope associated with increased degree of stimulus degradation, b = 16.11, SE = 1.82, t = 8.87, 95% CI [12.56, 19.77], β = .63, β 95% CI [.49, .77]. The slope model significantly improved model fit relative to the intercept-only model, $\chi 2\,(1) = 44.98$, $p < .001$. Furthermore, visual inspection of loess fit to the data suggested a nonlinear model would best fit the JE data (i.e., an exponential fit). A GAMM model demonstrated a significant smooth term, edf = 2.55, F = 671.60, approximate $p < .001$. A deviance test demonstrated that the GAMM model produced a better fit to the JE data relative to the LMM model, $p < .001$, as well as a smaller AIC value relative to LMM model, AIC = 1795.15, AIC = 1853.32, for the GAMM and LMM models, respectively.

Single Evaluation Mode. Approximately 5% of cases were removed. Linear model results for SE ratings revealed a significant positive slope associated with increased degree of stimulus degradation, b = 12.37, SE = 1.39, t = 8.90, $p < .001$, 95% CI [9.63, 15.10], R^2 = .28, β = .53, β 95% CI [.41, .64]. Similarly to JE ratings, inspection of a loess fit to the data suggested a nonlinear model may best fit the SE data (i.e., an exponential fit). A GAM model demonstrated a significant smooth term, edf = 1.86, F = 44.45, approximate $p < .001$. The GAM model produced a better fit to the SE data relative to the LM model, p = .003, as well as a smaller AIC value relative to LM model, AIC = 1981.10, AIC = 1987.76, for the GAM and LM models, respectively.

Experiment 11b

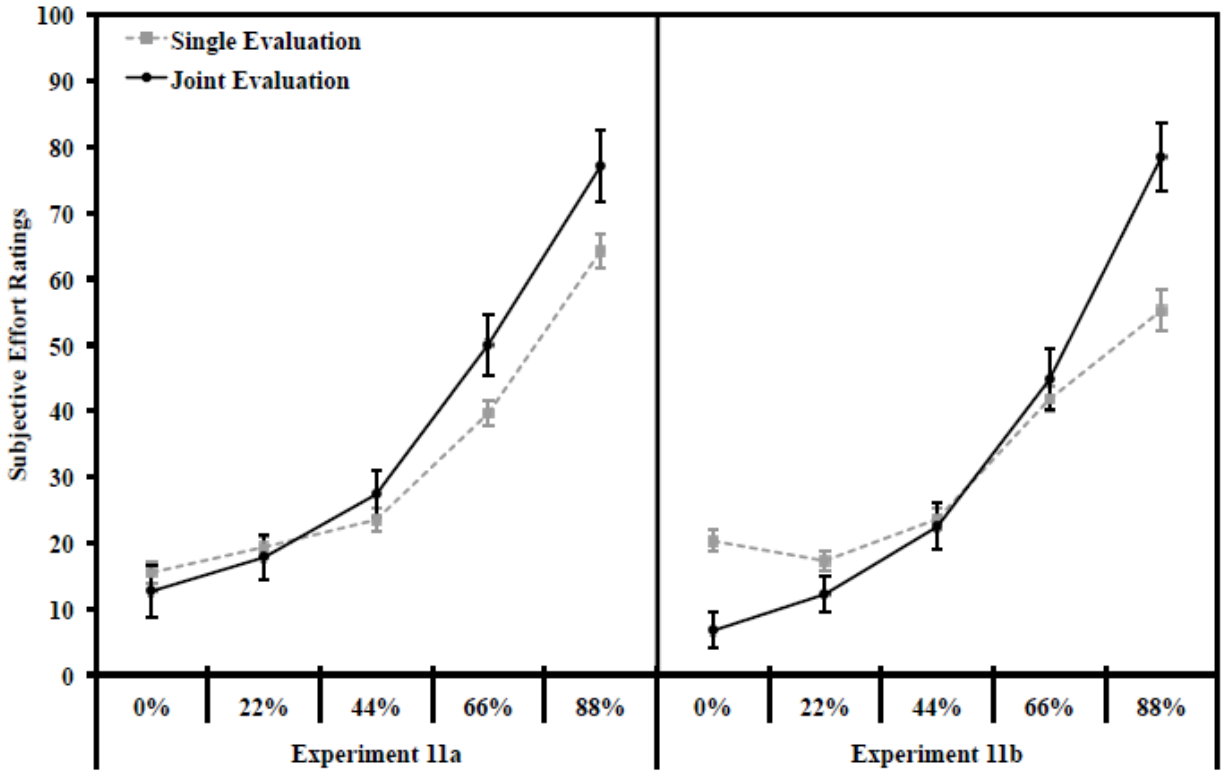
Effort Ratings

Joint Evaluation Mode. Less than 1% of cases were removed for JE ratings. LMM results revealed a significant positive slope associated with increased degree of stimulus degradation, $b = 18.00$, $SE = 1.44$, $t = 12.48$, 95% CI [15.10, 20.74], $\beta = .69$, β 95% CI [.59, .81]. The slope model significantly improved model fit relative to the intercept-only model, $\chi 2\,(1) = 64.82$, $p < .001$. Furthermore, visual inspection of loess fit to the data suggested a nonlinear model may best fit the JE data (i.e., an exponential fit). A GAMM model demonstrated a significant smooth term, edf = 2.79, $F = 1133.20$, approximate $p < .001$. Furthermore, a deviance test demonstrated that the GAMM model produced a better fit to the JE data relative to the LMM model, $p < .001$, as well as a smaller AIC value relative to LMM model, AIC = 1796.43, AIC = 1911.77, for the GAMM and LMM models, respectively.

Single Evaluation Mode. Approximately 10% of cases were removed. Linear model results for SE ratings revealed a significant positive slope associated with increased degree of stimulus degradation, $b = 10.01$, $SE = 1.42$, $t = 7.03$, $p < .001$, 95% CI [7.20, 12.82], $R^2 = .19$, $\beta = .44$, β 95% CI [.32, .56]. Similarly to JE ratings, inspection of a loess fit to the data suggested a nonlinear model may best fit the SE data (i.e., an exponential fit). A GAM model demonstrated a significant smooth term, edf = 1.99, $F = 13.45$, approximate $p < .001$. The GAM model produced a better fit to the SE data relative to the LM model, $p = .02$, as well as a smaller AIC value relative to LM model, AIC = 1991.20, AIC = 1994.81, for the GAM and LM models, respectively.

Figure 19.

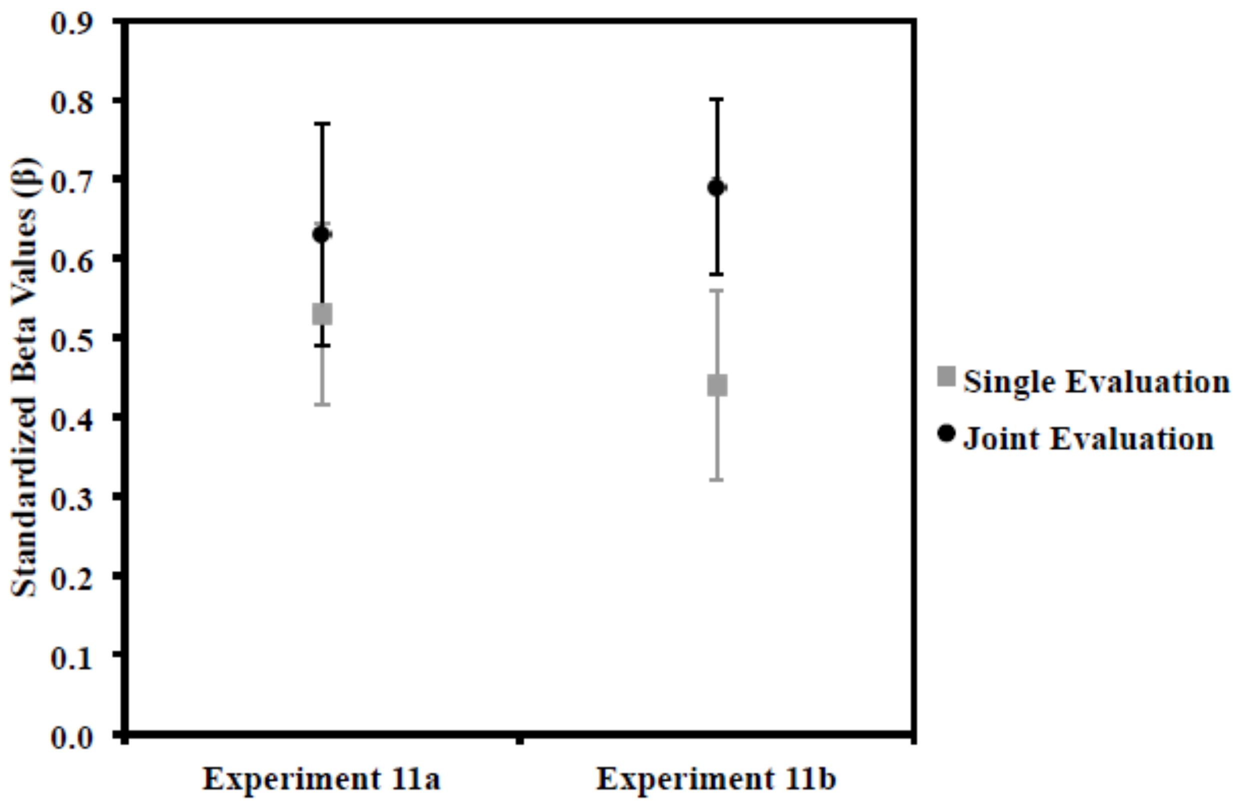
Single and Joint Evaluation Subjective Effort Rating Results for Stimulus Degradation in Experiments 11a (Left Panel) and 11b (Right Panel).



Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings. Error bars for the left panel represent ±1 SEM.

Figure 20.

Standardized Beta Values for Slopes in Experiments 11a and 11b.



Note: Single Evaluation (SE) represents between-subject ratings; Joint Evaluation represents within-subject ratings Error bars represent bootstrapped 95% confidence intervals.

Reading Accuracy. Given the lack of variability in accuracy across several of the stimulus degradation conditions, performance is reported here only qualitatively (see Table 7). First for JE, accuracy was at ceiling (i.e., 100%) for the 0%, 22%, and 44% degradation conditions. Accuracy fell to 83% for the 66% degradation condition and to floor (0%) for the 88% degradation condition. For SE, accuracy was similarly at ceiling for the 0%, 22%, 44% degradation conditions, then falling to 82% accuracy for the 66% degradation condition, and finally to floor (0%) for the 88% degradation condition.

Table 7.

Accuracy for Word Reading as a Function of Stimulus Degradation in Experiment 11b

	Stimulus Degradation				
Evaluation Mode	0%	22%	44%	66%	88%
Single Evaluation	94% (24%)	100% (0%)	98% (14%)	82% (34%)	0% (0%)
Joint Evaluation	100% (0%)	100% (0%)	100% (0%)	88% (32%)	0% (0%)

Note: Stimulus degradation was scaled as pixels removed from the original stimulus. Standard deviations are presented in parentheses.

Discussion

Experiment 11a provides evidence that evaluability is not domain-specific; rather, it may be task-specific and contingent on a failure point. In contrast to the stimulus rotation manipulation used in Experiments 7-10, subjective effort ratings for the stimulus degradation manipulation used in Experiments 11a and 11b can be considered relatively evaluable. Ratings for both the SE and JE modes produced similar positive linear slopes in Experiment 11a. In addition, both functions were best fit with a non-linear exponential function relative to the linear functions. Results for Experiment 11b, where individuals had awareness that they were required to complete the task that they would rate, demonstrated a less positive slope for SE than the slope for JE. This suggests that having such awareness may affect ratings in SE but not JE, given the JE slope for Experiment 11b was extremely similar to Experiment 11a. To confirm that the flatness of the SE slope relative to the JE slope was not a result of noise, additional samples were collected for SE ratings mirroring the procedures for 11a and 11b. Results demonstrated similar slopes across the new samples with both closely matching the SE slope from Experiment 11a[12]. Thus, the difference across JE and SE slopes in Experiment 11b may indeed have been due to random variability in ratings.

Critically, these findings contrast with Experiments 8 and 9 for the divergent patterns of effort ratings of stimulus rotation where SE slopes were flat and JE slopes were positively sloped. Tasks possessing a failure point associated with some value appear to drive consistent ratings across modes given the correspondence in effort functions across modes for degradation, set size, and weight. This similarity is absent for stimulus rotation. I return to these findings in the General discussion of this chapter. Furthermore, given the similarity of functions across Experiments 11a and 11b and replication samples (see Footnote 12), it does not appear that possessing awareness of having to engage in the task alters patterns of effort ratings.

General Discussion

General Evaluability Theory has been applied to various domains within economic, business, and management contexts (e.g., Bazerman, Loewenstein, & White, 1992; Hsee, 1993; 1996; Nowlis & Simonson, 1997) in an attempt to gauge the putative sensitivity of a value of interest. Chapter 3 applied the GET logic to an additional determinant of decision-making: effort (see Table 8 for a review).

---

[12] For each SE rating condition, 250 participants were recruited through Mechanical Turk. In the Experiment 11a replication sample approximately 11% of individuals failed at least one attention check resulting in a final *N* of 222 (M_{Age} = 32 years, 42% female participants, 53% reported completing a Bachelor's degree or higher). In the Experiment 11b replication sample approximately 12% of individuals failed at least one attention check resulting in a final *N* of 219 (M_{Age} = 34 years, 52% female participants, 53% reported completing a Bachelor's degree or higher). The two samples produced extremely similar positive slopes, $\beta = .51$, β 95% CI [.39, .63], $\beta = .54$, β 95% CI [.42, .65], for the Experiment 11a and Experiment 11b replication samples respectively, and are thus very comparable to the slope for SE in the original Experiment 11a experiment, $\beta = .53$, β 95% CI [.41, .64].

Table 8.

Summary of the Current Experiments in Chapter 3

	Task-Specific Effort(s)	JE-SE Relation	Slope Terms Mode	β	95% CI
Experiment 7	Stimulus Rotation	Similar positive slopes	*JE*	.48	[.27, .69]
			SE	.37	[.20, .55]
	Set Size	Similar positive slopes	*JE*	.86	[.75, .98]
			SE	.87	[.76, .97]
	Weight	Similar positive slopes	*JE*	.68	[.53, .82]
			SE	.48	[.31, .65]
Experiment 8	Stimulus Rotation	Positive slope in JE, flat slope in SE	*JE*	.43	[.31, .56]
			SE	.10*	[-.05, .25]
	Set Size	Similar positive slopes	*JE*	.82	[.72, .92]
			SE	.87	[.80, .94]
	Weight	Similar positive slopes	*JE*	.68	[.58, .77]
			SE	.62	[.50, .74]
Experiment 9	Stimulus Rotation	More positive slope in JE relative to SE	*JE*	.50	[.36, .64]
			SE	.24	[.13, .36]
Experiment 10	Stimulus Rotation & Set Size	Inconsistency across choices & ratings	-	-	-
			-	-	-
Experiment 11a	Stimulus Degradation	Similar positive slopes	*JE*	.63*	[.49, .77]
			SE	.53*	[.41, .64]
Experiment 11b	Stimulus Degradation	More positive slope in JE relative to SE	*JE*	.69*	[.59, .81]
			SE	.44*	[.32, .56]
Experiments 11a & 11b SE Mode Replications	Stimulus Degradation	-	*11a*	.51	[.39, .63]
			11b	.54	[.42, .65]

Note: The replications of the SE mode slopes for Experiments 11a and 11b are reported in Footnote 12.
* The *non-linear model (GAM or GAMM) produced a better fit to the data relative to the linear model*

In Experiment 7, individuals provided judgments of expected effort in either the SE or JE mode across three task-specific efforts: stimulus rotation, set size, and weight. Results demonstrated that judgments of perceived effort for set size produced similar positive slopes across evaluation modes. The perceived effort associated with stimulus rotation and weight demonstrated similar patterns across modes, however not to the same degree as set size, suggesting that each type of task-specific effort may be evaluable. Experiment 8 addressed a potential concern of a floor effect in Experiment 7 by increasing the putative levels of effort to-be-judged for the stimulus rotation and weight conditions. Results replicated the patterns found across modes for set size and weight, with both the SE and JE functions being very similar. Stimulus rotation, however, produced a positive linear function in JE and a flatter function in SE suggesting that perceptual effort in terms of stimulus rotation may not be highly evaluable. Experiment 9 looked to further examine this notion by focusing solely on judgments of stimulus rotations at finer rotation increments relative to Experiment 8. Again, JE produced a positive linear function across increments, whereas the linear function in SE was less positive. Taken together, Experiments 8 and 9 suggested that perceptual effort in terms of stimulus rotation may not be highly evaluable. That is, individuals do not appear to show value sensitivity to increasing effort in the task involving stimulus rotation.

In Experiment 10, I further examined this possibility by crossing the relatively inevaluable stimulus rotation with a manipulation of set size (i.e., number of items in the display) across both ratings and choice (i.e., "*Which is more effortful?*"). Within this context, GET hypothesizes that a low-evaluability value will exert greater influence in JE (choice) relative to SE (ratings). Individuals more often selected the condition associated with greater levels of stimulus rotation as the more effortful in JE, while both stimulus rotation and set size were rated as similarly effortful in SE. Experiments 11a and 11b looked to test the hypothesis that evaluability is not domain-specific, but rather specific to tasks that are associated with a failure point, and whether awareness of completing the task that is to-be-rated affects subjective effort functions. Specifically, using stimulus degradation, both SE and JE produced similar positive slopes across effort levels, as well as exponential fits to the ratings data. In the following, I discuss potential determinants of effort evaluability, highlight potential shortcomings of the current chapter, and suggest avenues for future research.

The Criticality of Reference Information

At a normative level, all decision-making can be thought of as occurring between alternatives even when alternatives are not explicit (Hsee et al., 1999). Thus, what type of reference information is available to individuals in SE that would lead to consistent judgments, when no explicit reference information is afforded? Within GET, both *knowledge* (i.e., distributional information, such as the variability and average, gained through experience) and *nature* (i.e., whether individuals possess a stable physiological or psychological reference system) reflect internal reference information available when generating judgments in the absence of explicit reference points. In contrast, *mode* as utilized in the present investigation can be considered *ad hoc* evaluability (Hsee & Zhang, 2010). Importantly, *knowledge* and *nature* both acting as internal forms of references may provide the necessary information needed to produce high levels of value sensitivity in the absence of explicit reference information with all factors conjunctively working toward evaluability.

Here I focus on potential candidates of what may constitute *knowledge-based* reference information and interact with *mode* potentially leading to, within the current experiments, some task-specific effort being relatively evaluable or inevaluable. I specifically focus on *knowledge* relative to *nature* given that *nature-based* reference information, according to GET, is explicitly tied to innate physiological or psychological scales not learned through experience. Aside from the clear issue of trying to generate testable predictions pertaining to proposed innate knowledge, there is no *a priori* reason to believe that humans possess innate psychological or physiological scales associated with remembering single items, lifting weights from the ground, or identifying degraded stimuli that would lead to heightened evaluability.

Following from the evidence presented here in Chapter 3, then, individuals may learn reference information over time specifically pertaining to when an expected error or failure will occur as putative effort increases. There are clear limits to the number of items that humans can hold in short-term memory at a given time (Atkinson & Shiffrin, 1971; Baddeley & Hitch, 1974; Cowan, 2001; Miller, 1956). Similarly, there are clear limits with respect to the amount of physical exertion that humans are able to invest at a given time (Poole, Ward, Garner, & Whipp, 1988; Suarez, 1996), and to the ability to identify degraded stimuli (Vokey, Baker, Hayman, & Jacoby, 1986). Individuals with good knowledge of these points may exploit this information to gauge their judgments from this point and, as such, benefit from increased evaluability across modes.

Interestingly, a failure point would seem to be absent in the context of stimulus rotation. Increasing stimulus rotation is finite; an object can only be rotated so many degrees before it returns to its canonical orientation. Moreover, though there are clear performance costs associated with stimulus rotation, individuals are very capable of processing a wide range of rotated items (e.g., Graf, 2006; Jolicoeur, 1985; Koriat & Norman, 1984; Risko et al., 2014; Tarr, 1995). That is, there does not seem to be a clear point of failure with regard to stimulus rotation in the current experimental context. Therefore, it would be expected that the lack of this information in SE would produce divergences in ratings across the modes. When this failure point information is present, however, individuals may exploit this reference information while generating their judgment of effort, leading to consistency in judgments across modes.

The proposal that successful evaluability of effort can be driven by a failure point shares similarities with a proposal by Tversky and Kahneman (1992, see also Kahneman & Tversky, 1979) concerning what they term *diminishing sensitivity* in S-shaped probability weighting functions. Diminishing sensitivity states that the impact of a change in probability diminishes as the distance increases from a reference point. Importantly, two natural reference points drive the S-shape of probability weighting functions: certainty and impossibility, with the former psychologically being analogous to "*certainly will happen*" (i.e., a 100% probability) and the latter being analogous to *"certainly will not happen"* (i.e., a 0% probability; Gonzalez & Wu, 1999). Here I have proposed that a failure point similar to impossibility potentially *increases* the evaluability of efforts. In contrast, following from GET, the existence of a failure point in probability weighting functions *decreases* the evaluability of the attribute. One clear reason for this discrepancy may be that perceived effort and perceived probability evoke different processes. Further considering how a failure point may foster evaluability in one case but inevaluability in another is an interesting question to pursue in the future.

Subjective versus Objective Effort Functions

The application of GET makes several assumptions with regard to value sensitivity that may pose problems for examining judgments of effort. First, there is a theoretically important distinction to make between an individual's *subjective* value functions with respect to effort and their *objective* functions with respect to effort. The notion of value sensitivity or evaluability can be seen as implying accuracy in the sense that an individual's evaluations of effortfulness would closely map onto some measure of the level of objective processing demand that an act places on the system. There are strong reasons, however, to doubt this assumption.

In Experiments 7-9, there was a close relation between JE and SE judgments of perceived effort associated with set size suggesting, according to GET, that this attribute is highly evaluable. However, if effort is measured through accuracy, then this is not the pattern expected. For example, Risko and Dunn (2015) demonstrated, using similar stimuli, that accuracy was near ceiling for smaller set sizes (2 and 4 letters), but fell radically at the medium set size (6 letters), and was near floor for the larger set sizes (8 and 10 letters). As would be expected from the experiments reported in this chapter, when individuals provided subjective ratings of accuracy and effort, results demonstrated relatively linear functions for both dimensions, where perceived effort increased and perceived accuracy decreased as the number of items increased. Therefore, the objective effort function (as indexed by accuracy) mimicked more of a decaying logistic function rather than a linear function as produced from individuals' subjective judgments.

A type of paradox then exists between subjective and objective effort functions: Based on the GET logic, a specific effort may be evaluable at the subjective level but not coincide with the associated objective function as indexed by some putative measure of objective effort (see also Hsee & Zhang, 2004, for a similar issue with regard to choice-experience inconsistencies). Resolving this apparent paradox requires taking seriously the notion that effort is a subjective phenomenon not necessarily tightly tied to objective processing demands (see Chapter 1; Kool & Botvinick, 2013; Westbrook & Braver, 2015). Indeed, several demonstrations exist of individuals' decisions and subjective evaluations based on effort being dissociated from objective demands as indexed by various indirect measures (e.g., Chapters 1 and 2; Westbrook, Kester, & Braver, 2013). From this perspective, evaluability would be defined by consistency in subjective evaluations (e.g., across modes) without any expectation that those evaluations accurately map onto processing demand per se (or indirect measures of it). Separating the concept of evaluability from "accuracy" in this manner does not diminish the importance of understanding the former. This is because it is individuals' subjective experience of effort that is proposed to drive behavior.

Evaluating "Apples and Oranges"

The current findings provide important insight into the relation between cognitive and physical judgments of effort. Kable and Glimcher (2009) note that subjective value allows for evaluation across options in such a way that decisions between "apples and oranges" are possible. Intriguingly, at specific effort levels in Experiments 7 and 8, individuals judged perceptual and memorial efforts as more effortful relative to physical effort and vice versa. For example, in Experiment 8, individuals' effort ratings demonstrated that attempting to hold 10 letters in memory and recall them accurately was perceived to be more effortful than lifting 50 lbs. of weight from the ground, whereas lifting the same amount of weight was perceived to be more effortful than attempting to hold 6 letters in memory. These situations represent interesting instances where, if placed in a decision-making context, individuals would be faced with trading-off one type of effort for another.

Examination of these situations has implications for the study of cognitive offloading where a decision to forego some form of internal processing (i.e., cognitive effort) is made in favor for external processing (e.g., physical effort; Dunn & Risko, 2016; Gilbert, 2015; Kirsh & Maglio; 1994; Martin & Schwartz, 2005; Risko & Dunn, 2015; Risko & Gilbert, 2016; Risko et al., 2014; Wilson, 2002). For example, external normalization (Dunn & Risko, 2016; Risko et al., 2014) represents an instance where individuals physically rotate their body to bring some disoriented display to its canonical orientation, instead of performing the analogous internal transformation. Here, individuals trade-off cognitive effort in the form of some type of internal transformation for physical effort in the form of moving one's body. Individuals more often choose to take on the physical effort associated with rotating the body as the putative cognitive effort of processing a rotated display increases (Risko et al., 2014). From the effort functions observed here, it can be argued that at some point the behavior would be expected to flip: Individuals would become more likely to take on the cognitive effort rather than the physical effort. Investigating whether a general bias exists toward avoidance of one form of effort, either cognitive or physical, in the context of offloading represents an intriguing avenue for future research.

Methodological Implications

With regard to evaluation mode, Hsee et al. (1999) note an important issue for researchers interested in subjective evaluations to consider: Which evaluation mode is "better", joint or separate? The authors note that often JE would be considered to be the better mode to place individuals in given that alternatives can be explicitly considered during judgment. However, in JE, individuals may be overly sensitive to a difference across options, thus inflating the within-subject effect when the effect may not even be detectable in SE (i.e., the between-subjects effect). Moreover, if the consumption of an option should theoretically take place in SE, then the judgments elicited in JE may show inconsistencies with an individual's actual consumption experience. Importantly, the design employed should match the researchers' specific question regarding the judgments of some value of interest: Are we as researchers interested in how individuals decide (i.e., a within-subject design)? Or are we interested in how individuals experience (i.e., a between-subject design)?

As an example, Schweitzer, Baker, and Risko (2013) examined the influence of including neuroimages (e.g., images of fMRI data) within scientific articles on individuals' favorability judgments. Across four experiments utilizing between-subjects designs no effects were found when images were included relative to control conditions without images suggesting no neuroimage bias. In a fifth experiment, however, an effect was found when a within-subject design was used. Is the neuroimage bias a real phenomenon then? As highlighted above, GET would suggest that the answer depends on the mode in which you would expect individuals to use the neuroimaging evidence. In particular, if jurors were asked to decide between an argument that included a neuroimgage and one that did not (i.e., JE), then you might expect to find a neuroimage bias in individuals verdicts. However, if an argument was presented in isolation either with or without a neuroimage (i.e., SE), then that image would not be expected to have a strong impact on the believability of the argument. Thus, this bias arguably is constrained to situations where individuals would be making decisions both with and without the neuroimages present.

Undoubtedly further examples exist where some effect is present in within-subjects designs but researchers fail to observe the same effect in between-subject designs (or vice versa), and further consideration of the issue of what evaluation mode to utilize would be of methodological importance across a wide range of contexts (e.g., Birnbaum, 1999). For example, within the context of heuristics and biases, Kahneman and Tversky (1996) argued that the two designs answer explicitly different questions. A between-subject design tests whether an individual relies on a given heuristic, whereas a within-subject design addresses how the conflict between heuristic use and some formal rule is resolved. That is, between-subject designs arguably offer a more realistic view of individuals' reasoning (Kahneman & Frederick, 2002; Tversky & Kahneman, 1983).

Conclusion

How individuals come to appraise the level of expected effort associated with some action is an essential question in the investigation of human decision-making. In this chapter, I have provided evidence with respect to how this is achieved through the lens of GET. Generally, both cognitive and physical effort as defined in the present investigation can be considered relatively evaluable, with increased evaluability being driven by exploiting a failure point associated with the attribute.

Chapter 4

Chapter 3 examined the evaluability of effort, I proposed a process whereby individuals utilize reference information pertaining to expected error or failure points to explain why some efforts are evaluable. The evidence in Chapter 3 thus provides useful insight into further addressing the question of why effortful actions are evaluated as being effortful. If effort and errors are closely associated, then I would expect to see strong contributions of error likelihood to judgments of effort. Chapter 4 looked to directly test this idea.

The following work is currently in review at *Psychological Research* (Dunn, Inzlicht, & Risko, submitted).

Changes have been introduced to improve the flow of the dissertation.

The question of what constitutes the cost of effort has recently come to the fore in research largely focused on the control of human behavior (e.g., Inzlicht, Bartholow, & Hirsh, 2015; Kurzban, 2016). Effortful actions are often attributed to instances where behavioral control is deployed, are generally thought to carry a cost in decision-making, to be aversive (Dreisbach & Fischer, 2012), and to 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 (Kool, McGuire, Rosen, & Botvinick, 2010) or requiring higher levels of reward to engage in such actions (Westbrook, Kester, & Braver, 2013).

Although the discussion of effort as a key determinant in decision-making and control is pervasive, what constitutes effort is still an open question. Thus, empirically testing what drives effort judgments remains an important empirical problem. To better understand the constituents of *effortfulness*, in this chapter, I pit two basic determinants of effort against one another: error-likelihood versus time-requirements. Specifically, I contrast individuals' perception of effort when faced with a trade-off between engaging in a small amount of hard work (high error-likelihood, low time requirements) versus a larger amount of easy work (low error-likelihood, high time requirements).

Constituents of Effortfulness

Time requirements. The claim that processes that take more time 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 work* (i.e., more time) being associated with effort costs is prevalent in several opportunity cost frameworks of behavioral control. An opportunity cost can generally be conceived of as 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, provide a useful construct in understanding cost-benefit analyses concerning behavioral control. For example, Niv and colleagues (2007) proposed that an average rate of reward serves as an 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) proposed that opportunity costs arise as a function of parallel-processing capacity being finite and thus scarce (Baddeley & Hitch, 1974; Kahneman, 1973; Kurzban et al., 2013; Navon & Gopher, 1979; Shiffrin & Schneider, 1977; Wickens, 2002), and dynamically allocated across processes. According to these approaches, the feeling of effort arises from an output of the mechanisms computing the costs and benefit of engaging in a task relative to alternative tasks to which the same processes could be applied. Costly, and therefore effortful tasks are those that engage work from multiple cognitive processes, such as control functions that encode and maintain representations of a task and marshal other functions such as attention, memory, and perception (Botvinick & Braver, 2015). Thus, increases in processing time prevent additional similar processes *X*, *Y*, *Z* from being carried out, given some amount of capacity is being held up by *Task A*. Furthermore, opportunity costs can accrue, be tallied, and tracked over time for use in the allocation of control (Westbrook & Braver, 2016)

Comparable to opportunity cost accounts, motivational accounts of self-control similarly assign a cost to increased time-on-task. Inzlicht, Schmeichel, and Macrae's (2014) shifting-priorities process model of self-control hypothesizes that, given time is limited, 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. Thus, *more work* (i.e., more time requirements) can be considered as a determinant of effortfulness.

Error-likelihood. Beyond time, errors 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., 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 the experienced reward functions as a signal in action and outcome learning that increases a behavior's 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 such as post error slowing (Rabbitt, 1966) or the reassessing 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 errors 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., in press). Recent work has suggested that the source of these capacity limitations lies 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 press). 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 configure processes through control to avoid situations of cross-talk.

Beyond response monitoring accounts of the error monitoring system (e.g., Dehaene, Posner, & Tucker, 1994; Holroyd & 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, & Làdavas, 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 as a basic motivational function. Situated within accounts of control, errors can then be considered to be particularly aversive, to generate strong emotional responses upon commission, and to require greater adjustments of effortful control to resolve (Inzlicht et al., 2015). Therefore, *hard work* (e.g., tasks associated with a higher likelihood of errors than an alternative course of action) can, in addition to *more work*, be considered a determinant of effortfulness.

Present Investigation

The set of experiments in Chapter 4 looked to highlight the association between effort, time, and errors in contexts where the latter two factors trade-off against one another.

Specifically, 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. Although many accounts of effort largely focus on why actions become subjectively effortful over time while situated within a task, relatively less attention has been paid to why some actions can be perceived as effortful at the point of initial evaluation, prior to engaging. Therefore, here I focus on individuals' prospective evaluations of *anticipated* or *perceived* effort, time, and errors.

Perceived effort, as opposed to experienced effort, has been argued to be crucial in the decision-making process (Dunn, Lutes, & Risko, 2016; Payne, Bettman, & Johnson, 1993). Dunn and colleagues (2016) recently likened perceived effort to a subjective metacognitive evaluation not necessarily tied to the cognitive work that a task requires (see also McGuire & Botvinick, 2010; Westbrook & Braver, 2015). Critically, the evaluation of effort is for guiding behavior whether situated in a task (e.g., disengaging from an action) or not (e.g., avoiding an action). Thus, investigating judgments of perceived effort, time, and errors affords 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 perceived effort and "online" evaluations of experienced effort both deserve researcher's 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).

Throughout this chapter, I looked to establish whether a stimulus associated with more time or more errors would be considered more effortful. To manipulate time and errors, I used stimulus rotation and set size. Specifically, 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. I also confirmed this general pattern of data in Experiment 12a. Thus, in Experiment 12b, 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 effort judgments. Experiment 12c utilized the same choice context and stimuli but manipulated the basis of individuals' choices. Specifically, 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 associated with more errors, 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 12a, 12b, and 12c

Method

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

Participants

Initial Sample Size Determination for Experiment 12a. Twenty University of Waterloo undergraduates participated for course credit. This sample size allows for the detection of at least a medium effect size across conditions for a within-subject design. No individuals were removed from the below analyses.

Initial Sample Size Determination for Experiments 12b and 12c. A pilot study was conducted based on Experiment 4 from Dunn et al. (submitted) 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, $g = .13$, $BF_{\text{Alt}} = 1.40$, suggesting a sample size of 93 was needed (based on null hypothesis significance testing; NHST). The initial sample size was set at $N = 96$ to ensure complete counterbalancing of the stimulus lists (see stimuli below). Given that the sizes of the effects for accuracy and time choices are unknown for Experiment 12b, the sample size from Experiment 12a was carried over for each of the dimensions.

Current Sample for Experiments 12b and 12c. Ninety-six Amazon Mechanical Turk (MTurk) workers were recruited in Experiment 12b 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 in Experiment 12c 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).

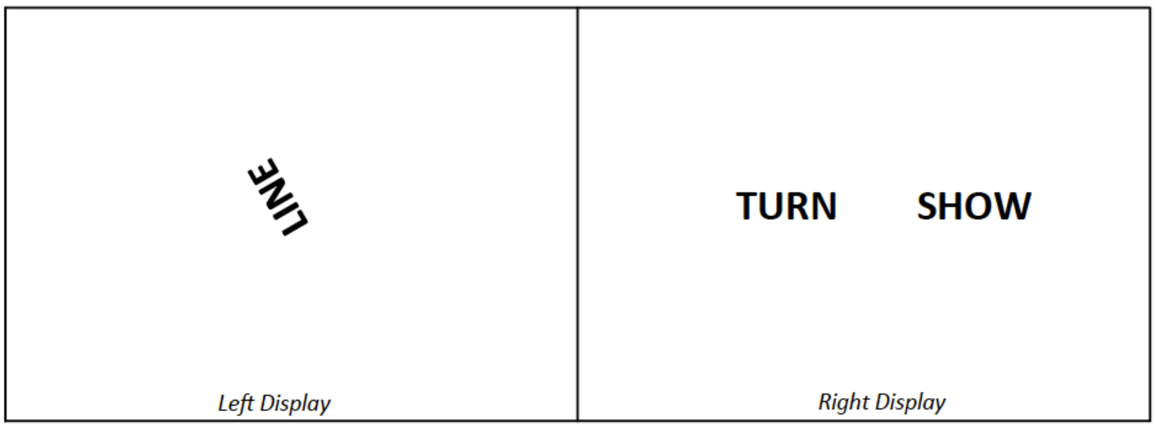
Design

For Experiment 12a, a one-factor (Choice Option: 1-word/110°; 2-words/0°**;** 3-words/0°**;** 4-words/0°**;** 5-words/0°) within-subject design was employed. For Experiment 12b, a one-factor

(Choice Option: 1-word/110°; 2-words/0°) within-subject design was employed. For Experiment 12c, a one factor (Rating Dimension: Time, Accuracy) x 2 (Choice Option: 1-word/110°; 2-words/0°) mixed design was employed. Rating dimension was manipulated between-subjects.

Figure 21.

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 12a and 13), more time demanding (Experiments 12b and 13), or less accurate (Experiments 12b and 13) to read aloud.

Apparatus

Experiment 12a was deployed using DMDX software (Forster & Forster, 2003). Stimuli were presented on a 24" LCD monitor with individuals sitting approximately 70 cm away. A standard QWERTY keyboard was used for manual responses.

Stimuli

Stimuli consisted of a single word presented at ±110° and two words presented at upright (0°). 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 21). All stimuli were similar in Experiment 12c, though the arrow was removed from the 2-word stimuli given that several participants in Experiment 12b 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 12a. 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 ensure 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 comfortable 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 12b and 12c. 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 produce 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 I 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 they confirmed 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 12c, 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)"* to complete.

Once participants made their choice, an attention check was presented displaying the same choice screen that the participant had 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

Experiment 12a

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 9).

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 no signs of extreme skewnesss (.58) or kurtosis (2.99). A one-way Bayes Factor (BF) ANOVA demonstrated positive evidence for the alternative (i.e., the choice option condition had an effect on RTs relative to an error-only model), $BF_{Alt} = 5.34$, $F(4, 76) = 296.33$, $MSE = 28595.49$, $p < .001$, $\eta^2 = .94$. Individual BFs were computed for the 1-word/110° condition relative to all other conditions. Extreme evidence for the alternative 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.

For accuracy, A one-way Bayes Factor (BF) ANOVA demonstrated strong evidence for the null, $BF_{Nullt} = 333.33$, $F(4, 76) = .91$, $MSE = .004$, $p > .1$, $\eta^2 = .05$, demonstrating that accuracy did not vary across the five choice option conditions, although 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 9.

Mean Performance Results for Experiment 12a

	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 12b and 12c

Bayesian analyses of the maximum a posteriori estimate (i.e., mode value; MAP) of the θ parameter (i.e., successes and failures for binomial data) and 95% Highest Density Intervals (HDI) around θ (Kruschke, 2013) were generated in R (R Core Team, 2014). Bayes Factors (BF) were computed using the BayesFactor package (Morey & Rouder, 2015) in R. Evidential strength categories for Bayes Factors (i.e., in favor of the alternative hypothesis) follow 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".* The determination of priors for the Beta distribution shape parameters β(a,b) (where a = successes and b = failures within the sample) used for HDIs and BFs is outlined preceding the reporting of results. Furthermore, Binomial and chi-square test results are presented alongside Bayesian analyses where applicable. For Experiments 12b and 12c, default priors were used for BFs (i.e., r scale = .707), and priors for estimation were set for β(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.

First, for Experiment 12b, 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} = 1569.77, MAP = 73%, 95% HDI [63%, 82%], $p < .001$ Binomial test. For the error-likelihood dimension in Experiment 12c, 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} = 457,656.10, 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_{Alt} = .26, 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., that each column of the data has different probabilities) 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. Comparisons of choices for the effort and error likelihood dimensions against the time dimension (50%) demonstrated very strong evidence for the alternative 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. This offers very strong support for the notion that ratings of effort do not track ratings of time-requirements, and also provides extreme support for the notion that ratings of error likelihood do not track ratings of time requirements.

In sum, individuals similarly chose the 1-word/110° display as the more effortful and more error-prone option relative to the 2-words/0° display. In contrast, by individuals showing no preference on the dimension of time requirement, they contrasted sharply with their very clear preference (for the 1-word/110° display) on the dimensions of effort and error likelihood (see Figure 22).

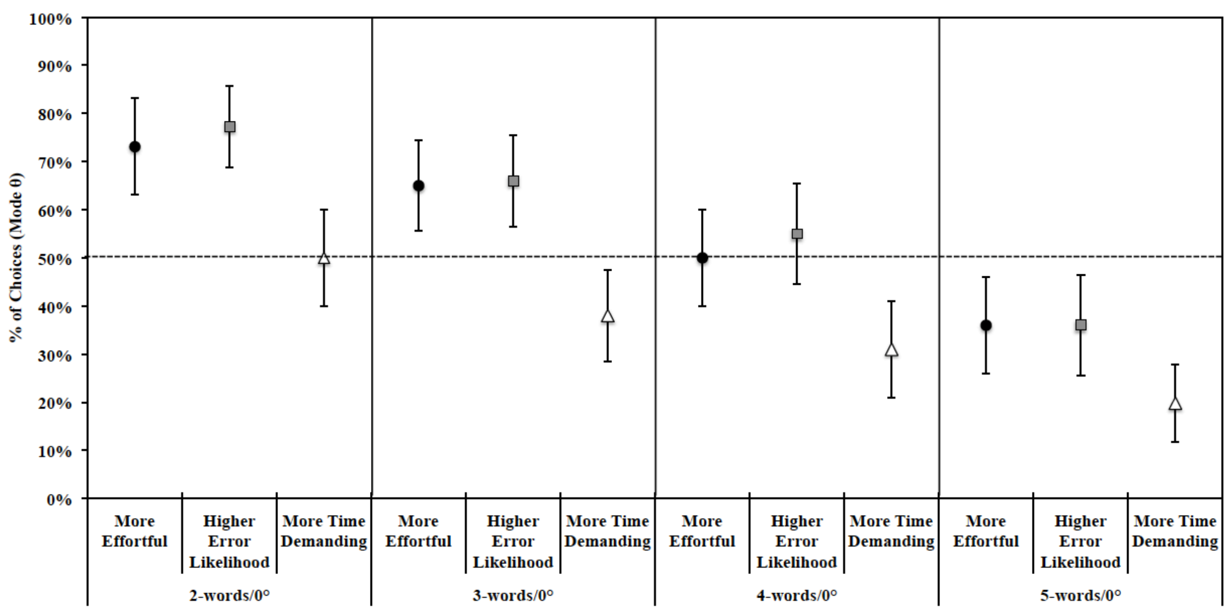
Discussion

Experiment 12 demonstrates a dissociation between effort, error, and time judgments. Individuals chose the 1-word/110° display as both more effortful and more error-prone relative to the 2-words/0° display despite accuracy being relatively equivalent across the options based on performance estimates. Individuals showed no clear preference for either display when evaluating the displays on time requirements. This was the case even though performance estimates demonstrated a large time cost for the 2-words/0° display relative to the 1-word/110°

display. Such 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, individuals' effort choices closely followed the option associated with a higher perceived error likelihood, but not the option associated with greater time demand.

Figure 22.

Individuals' More Effortful, Higher Error Likelihood, and More Time Demanding Choices for the 1-word/110° display in Experiments 12b, 12c, and 13a.



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 were collected in Experiments 12b and 12c. All other data were collected in Experiment 13a. The mode θ values are based on the posterior distribution. Error bars represent 95% Highest Density Intervals (HDI).

Experiments 13a and 13b

Experiment 13a aimed to further explore the observed dissociation between judgments (i.e., effort = errors > time). Individuals completed a judgment task similar to that in Experiments 12b and 12c. However, in Experiment 13a the set size manipulation ranged from 3 words to 5 words while keeping the 1-word/110° contrasting display constant. Thus, 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 the set size comparison in which this dissociation will occur, but rather by increasing set size, the likelihood that this dissociation will occur should increase accordingly. To foreshadow, Experiment 13a demonstrated a clear dissociation between effort and error-likelihood ratings relative to time when contrasting 1-word/110° vs. 3-words/0°**.** Given the importance of the clear dissociation between choices across the three dimensions, a registered replication was completed of the specific 1-word/110° versus 3-words/0° choice condition for all choice dimensions in Experiment 13b (please see https://osf.io/2szy3/registrations for the replication protocol).

Method

Participants

Initial Sample Size Determination for Experiment 13a. The sample sizes of $n = 96$ for each rating dimension (i.e., effort, accuracy, and time) from Experiments 12b and 12c were used in Experiment 2.

Current Sample for Experiment 13a. Eight hundred and sixty MTurk workers were recruited for the online study and compensated $1 USD for participating. Ten percent of individuals failed the attention check embedded in the survey, resulting in a final N of 771

($Median_{Age}$ = 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 13b. Three-hundred and twenty MTurk workers were recruited for the online study and compensated $1 USD for participating. The replication used optional stopping methods (Rouder, 2014) 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 tested until this cut-off was met (see below for final BFs).

Current Sample for Experiment 13b. Ninety-six individuals were tested in the effort and accuracy dimensions and 128 individuals were tested 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 (Choice Option: 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, Time, Accuracy) between-subjects design was employed for Experiment 13a. Experiment 13b utilized a 2 (Choice Option: 1-word/110° vs. 3-words/0) x 3 (Rating Dimension: Effort, Time, Accuracy) design.

Stimuli

The stimuli closely followed Experiments 12b and 12c, 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 procedures followed Experiments 12b and 12c.

Results

Results for Experiment 13a are presented (see Figure 22) followed by Experiment 13b. Following from Experiment 12, I present the results of each individual choice option condition separately to examine at what point judgments potentially become fully dissociated. All of the following analyses utilized priors as described in Experiments 12b and 12c.

Experiment 13a

1-word/110° vs. 3-words/0°. First, a BF computed for the 2 (choice) x 3 (rating dimension) data demonstrated extreme evidence for the alternative that each column of the data had different probabilities, $BF_{Alt} = 558.31$, $\chi^2(2) = 19.94$, $p < .001$. 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 the less accurate option 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. For time, individuals chose the 1-word/110° display as the more time demanding option relative to the 3-words/0° display 38% of the time, $BF_{Alt} = 3.18$, 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 for the 2 x 3 data demonstrated strong evidence for the alternative that each column of the data had different probabilities, $BF_{Alt} = 22.35$, $\chi^2(2) = 13.02$, $p < .001$. 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 the less accurate option relative to the 4-words/0° display 56% of the time, $BF_{Alt} = .43$, MAP = 55%, 95% HDI [45%, 66%], $p > .1$ Binomial test. For time, individuals chose the 1-words/110° display as the more time demanding option relative to the 4-words/0° display 30% of the time, $BF_{Alt} = 296.17$, 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 computed for the 2 x 3 data demonstrated anecdotal evidence for the alternative that each column of the data had different probabilities, $BF_{Alt} = 2.02$, $\chi^2(2) = 8.33$, $p = .02$. 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 the less accurate option 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 the more time demanding option 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. Nonetheless, effort and error likelihood choices were similar with both dimensions producing higher 1-word/110° choices relative to the time dimension.

Experiment 13b

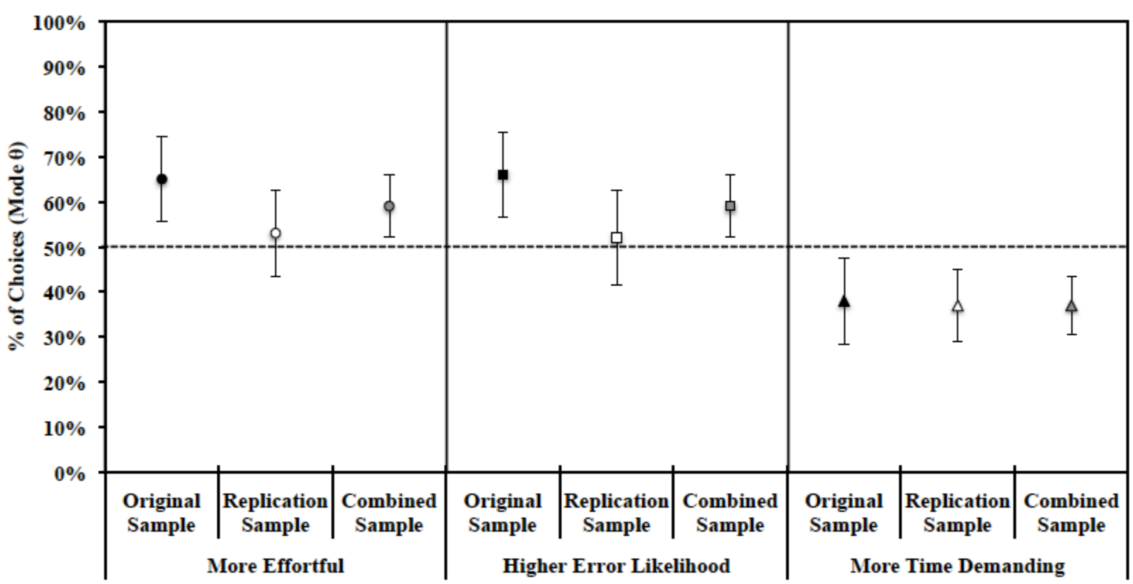
1-word/110° vs. 3-words/0°. A BF computed for the 2 x 3 data demonstrated only anecdotal evidence (based on the above evidential categories) for the alternative that each column of the data has different probabilities, $BF_{Alt} = 1.26$, $\chi^2(2) = 7.63$, $p = .02$. For the effort dimension, individuals 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 the error likelihood, individuals chose the 1-word/110° display as the less accurate option 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 the more time demanding option relative to the 3-words/0° display 36% of the time, $BF_{Alt} = 20.84$, 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 that the replication sample used the exact same methods and procedures relative to the original experiment, the initial sample and replication sample were combined to provide a clearer estimate of the choices for each rating dimension in the 1-word/110° vs. 3-words/0° condition ($N = 567$). The following analyses were unregistered. First, BFs 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. 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 the less accurate option 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 the more time demanding option relative to the 3-words/0° display 37% of the time, $BF_{Alt} = 238.66$, 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, $BF_{Alt} = 3278.02$, $\chi^2(1) = 19.39$, $p < .001$, $BF_{Alt} = 2185.93$, $\chi^2(1) = 18.55$, $p < .001$, for effort vs. time and error likelihood vs. time respectively. Therefore, the 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 23).

Figure 23.

Individuals' More Effortful, Higher Error Likelihood, and More Time Demanding Choices for the 1-word/110° display in Experiment 12 (Original Sample), 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 θ distribution. Error bars represent 95% Highest Density Intervals (HDI).

Discussion

Additional dissociations between effort, error, and time judgments were observed in Experiments 13a and 13b. 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 accuracy 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 suggest that although 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 12a) did not translate to ratings of effort, where individuals' choices were at chance. All dimensions did eventually favor the larger set size as being more effortful, time demanding, and less accurate at the 1-word/110° versus 5-words/0° comparison, though time judgments more favored the 5-words/0° option, with effort and accuracy judgments being closer to chance. Importantly, 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, Experiments 13a and 13b further demonstrate a strong association between effort and error judgments.

Experiment 14

The previous experiments have relied on manipulations of stimulus rotation and set size to generate a 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 aversive but rather are based on low perceptual fluency related to a rotated display being aversive. 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 as a determinant of effort judgments, in Experiment 14 options were changed from stimulus rotation and set size to math problems.

Based on previous research (e.g., Ashcraft & Faust, 1994; Siegler & Lemaire, 1997, Walsh & Anderson, 2009) one *N* x *NN* multiplication problem generates more errors relative to six simple addition problems, whereas solving six addition problems takes longer relative to a single multiplication problem. Hence, again individuals were faced with a trade-off between time and errors. If judgments of effort are driven by an evaluation of the likelihood of errors, then I 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 closely track effort choices, whereas choices for the time dimension should favor the six addition problems.

Method

Participants

Initial Sample Size Determination. The sample size 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 3 (Rating Dimension: Effort, Time, Accuracy) x 2 (Choice Option: 1 multiplication problem vs. 6 single digit addition problems) mixed design was employed. Rating dimension was manipulated between subjects.

Stimuli

The general choice screen was similar to the previous experiments. The two choice options consisted of one *N x NN* multiplication problem and six simple single digit addition problems. 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 x 9, 17 x 6, 13 x 7, 12 x 7, 16 x 8, 19 x 5, and 14 x 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 (i.e., the multiplication problem varied across individuals while the same addition problems were utilized for all individuals but shifted through the order that they were presented).

Procedure

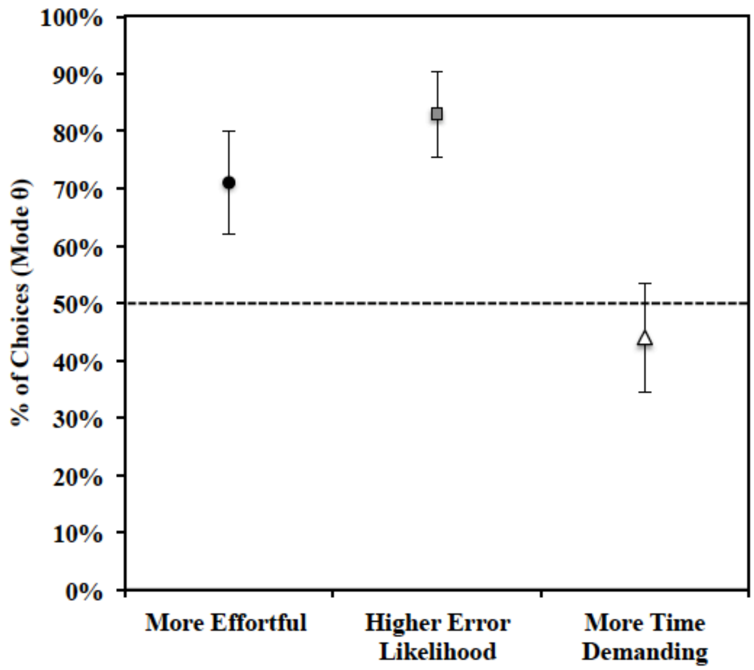
All portions of the procedure followed the previous experiments.

Results

All of the following analyses utilized priors as described in the previous experiments. A BF computed for the 3 x 2 data demonstrated extreme evidence for the alternative that each column of the data had 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 24).

Figure 24.

Individuals' More Effortful, Higher Error Likelihood, and More Time Demanding Choices for the One Multiplication Problem in Experiment 14



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

Discussion

Consistent with the previous experiments in this chapter, the multiplication option generated more effortful and higher error likelihood evaluations that were both similarly above chance, with evaluations of time requirements differing markedly from these and being slightly below chance. Therefore, the apparent error bias in effort judgments is not driven merely by the low fluency associated with the stimulus rotation manipulation in Experiments 12 and 13. Rather, this bias generalizes to additional trade-off contexts.

General Discussion

In Chapter 4, I investigated the influence of two potential determinants of effortfulness: time requirements and error likelihood. Experiments 12b and 12c provided evidence of a dissociation between effort, errors, 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 13 further demonstrated clear dissociations between effort, errors, and time choices, with effort and error choices closely tracking one another. To generalize the findings from Experiments 12 and 13, Experiment 14 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.

Perceived Error Likelihood as a Determinant of Effortfulness

Why are effortful actions evaluated as effortful? I propose that evaluations of effort are driven by a bias to more heavily weigh the perceived likelihood of error commission relative to time demands in explicit trade-off situations. Paired with the work outlined in Chapters 1 and 2, individuals can utilize and weight the likelihood of an error over salient available cues associated with an action, such as stimulus rotation in Experiments 12 and 13 or math problem type in Experiment 14, to generate their judgment of effortfulness. Such a bias 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 proposed bias predictably reduces 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 I detail below.

Utilizing error likelihood as a weighting principle across cues in judgments of effort can provide a low cost approximation of the potential effort associated with a line of action. As reviewed above, several accounts suggest that error commission signals the need to engage effortful 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; Shenav et al., in press). As such, evaluating lines of action that are associated with an increased error likelihood as effortful would be expected to be adaptive to the organism.

Time demands were not associated with effort judgments to the same extent as errors across the current experimental context. This claim dovetails with recent works that have demonstrated dissociations between effort-based decisions and the time costs associated with a task (see Chapters 1 and 2; Dixon & Christoff, 2012; Kool et al., 2010; Westbrook et al., 2013), but diverges from opportunity cost models that suggest processes that require more time will be perceived as more effortful (e.g., Kurzban et al., 2013). One potential explanation of this divergence is that, relative to taking more time on a task, committing an error arguably generates a more immediate call for effortful control to the system. 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 minimize the future likelihood of more errors.

In contrast, opportunity costs are hypothesized to tally and accrue over periods of time producing aversive signals used by the system to exert control and to move behaviors to a more rewarding alternative (Inzlicht et al., 2014; Kurzban et al., 2013). One could reasonably conceive of increased opportunity costs requiring longer timescales 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. An 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 potentially leading to more impulsive choices (e.g., divided capacity led to decreased ability to engage more analytical processing during choice; Evans & Stanovich, 2013). Thus, holding errors constant, fatigue required a relatively long time to initiate putative control 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 largely 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 12a), though judgments of error-likelihood also similarly followed. Therefore, it is important to note that time and errors should not be considered mutually exclusive determinants of judged effort. Indeed, these two are intrinsically connected and both may be integrated to some extent when generating evaluations of effort. For example, taking too long in a task can be considered an error and errors sometimes lead to slowing down on tasks. If time is expected to play a determining role in effort judgments, however, then future work should look to examine effort judgments across varying time scales where error likelihood and time costs are explicitly traded off.

While simplifying weighting principles and cue examination, utilizing error likelihood to determine prospective effortfulness can also circumvent the issue of the relatively increased cognitive demands associated with running more complex algorithms. 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) 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 information through the dorsal anterior cingulate cortex (dACC). A comparison between this mechanism and the bias proposed here provides an important contrast with regard to perceived effort. For example, using errors as a proxy to judge effort would not be expected to be dependent on demanding online monitoring of information. Thus, utilizing error likelihood would be expected to require less information relative to more complex alternatives.

A further potential avenue for assessing the pervasiveness of the proposed bias can be 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 that show high negative affect (Hajack et al., 2004) and are high in neuroticism (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. An alternative explanation based on the error account proposed here 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 evoke a need to be avoided. In this chapter, I have demonstrated a bias toward weighing the likelihood of errors in making these types of judgments. Time demands on the other hand demonstrated little association with effort judgments. Such findings pose challenges for opportunity cost accounts of decision-making that posit increased time as a key constituent of increased effort, and suggest that avoiding effortful actions may simply be avoiding actions associated with errors.

Concluding Remarks

In this series of studies, I have demonstrated that individuals utilize available cues during effort-based decision-making and that error likelihood appears to play an important role in the evaluation of effort.

Chapter 1 tested the hypothesis that individuals avoid courses of effortful action based on their associated performance with each action (i.e., response times and errors), versus the hypothesis that individuals avoid such actions based on an inferential metacognitive evaluation of demand. Individuals in Experiment 1 completed a free-choice Demand Selection Task (DST) that used stimuli known to yield a dissociation between performance and perceived effort. Patterns of choices followed that of perceived demand rather than performance. Experiment 2 provided a replication of this result, in addition to demonstrating a second dissociation between demand avoidance and a peripheral physiological measure of demand (i.e., blink rates), and a close correspondence between preference choices and self-reported ratings of effort. Experiment 3 directly tested the notion that a general metacognitive evaluation of demand drives selections. A DST was utilized in a forced-choice paradigm that asked individuals to select what they felt to be the most effortful, time demanding, or least accurate of two choices. Patterns of selections were similar across all of these rating dimensions, lending credence to the notion that a general metacognitive evaluation of demand based on cues drives choices.

Chapter 2 provided a further test of a cue utilization account of how individuals decide which course of action is the least effortful. Here, I contrasted the influences of time and demands on executive control with the influence of an available effort cue within the context of effort avoidance. Using a DST that focused on making effort-minimizing decisions, I provided evidence that avoidance behaviors can be dissociated from both time and demands on executive control in a manner predicted by a cue-utilization account. In Experiment 4, low- and high-demand options conditions were developed using manipulations of stimulus rotation and task switching. The former context can be argued to be associated with relatively low demands in terms of performance (i.e., be faster and cause less errors during processing) and executive control, whereas the latter context can be argued to be associated with relatively high demands on these dimensions. Interestingly, however, individuals avoided processing single rotated stimuli at a similar rate relative to more task switching. Experiment 5 specifically aimed to modulate avoidance behaviors by reducing the rate of task switching across options. In this context, individuals avoided processing single rotated stimuli at a greater rate relative to more task switching. Last, Experiment 6 directly pitted the stimulus rotation manipulation against the task switching manipulation in a DST. Again, individuals more often avoided the option associated with rotated stimuli. These results are consistent with individuals primarily utilizing the more salient effort cues in developing their effort preferences (i.e., individuals more often avoided rotated stimuli), rather than avoiding options associated with higher time demands and higher demands placed on the executive control system.

Chapters 1 and 2 provided evidence in support of the cue-utilization hypothesis as it pertains to the of perceived effort and effort-based decision, an account aimed at addressing *how* individuals generate a notion of effort across alternative lines of action. Chapter 3 aimed to begin to answer the question of *why* some cues would signal that a task may be effortful. Evaluations of effort were examined through methods described by the General Evaluability Theory (GET; Hsee & Zhang, 2010). Here, the evaluability of effort is determined by examining subjective value functions across different evaluation modes. Individuals judged the anticipated effort of four task-specific efforts indexed by stimulus rotation, items to-be-remembered, weight to-be-lifted, and stimulus degradation across joint (i.e., judged comparatively) and single (i.e., judged in isolation) evaluation modes. In the first three experiments (Experiments 7, 8, and 9), I demonstrated that the perceived effort associated with items to-be-remembered (memorial effort) and weight to-be-lifted (motor effort) can be considered relatively evaluable, whereas the perceived effort associated with stimulus rotation (perceptual effort) can be considered relatively inevaluable. Experiment 10 further established the inevauability of the perceived effort associated with stimulus rotation through a demonstration of inconsistencies in effort judgments across rating and choice contexts. Last, by examining an additional form of perceptual effort (i.e., stimulus degradation), Experiments 11a and 11b provided evidence that evaluability is not domain-specific; rather, it may be task-specific and contingent on a failure point. Following from the current evidence, I proposed that individuals utilize learned reference information specifically pertaining to when an expected error or failure may occur as perceived effort increases, and these points facilitate evaluability of effort values.

The experiments in Chapter 4 aimed to address the question of why some actions are evaluated as effortful. I examined individuals' perception of effort when they were faced with making judgments in a trade-off context between how much time a task takes versus how error-prone it is. That is, do people consider a small amount of hard work (i.e., low time requirement, but high error likelihood), or a large amount of easy work (i.e., high time requirement, but low error likelihood) to be more effortful? Across three experiments and two separate trade-off contexts, I demonstrated that individuals perceive options that are associated with low time requirement but high error-likelihood as more effortful than options that are more time consuming yet low in error-likelihood. Moreover, when individuals were asked to evaluate which of 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. These results support the conclusion that the likelihood of error commission, rather than time demands, appears to drive effort-based choices.

Effort as an Inferential Evaluation

Taken together, the evidence presented in the current thesis supports the hypothesis that perceived effort is the result of a general metacognitive evaluation made over available cues. To the extent that cues serve as a signal for effort is driven by an inferential approximation of the likelihood of error commission. That is, the higher the perceived error likelihood relative to alternative lines of action, the higher the level of perceived effort. This account is at odds with the view that individuals possess direct-access to the “demand” or “work” associated with executing some cognitive processes, with effort reflecting the cost of engaging in this work. Importantly, however, the account offered here can be situated within current frameworks of control and, specifically, action selection (i.e., “Should I chose to engage in this line of action?” or “Should I disengage from this action?”). As an example, several recent proposals that have focused on cost-benefit mechanisms have stressed that effort weighs as a subjective cost in the control of behavior (Kool & Botvinick, 2014; Kool et al., 2010; Kurzban et al., 2013; McGuire & Botvinick, 2010; Westbrook, Kester, & Braver, 2013). As developed here, the general evaluation based on available cues and weighted on errors represents a strong candidate process for this subjective cost. This account contrasts with recent claims that such costs arise from an approximation of the opportunity cost in terms of deploying limited resources (Kool & Botvinick, 2014). Indeed, the potential subjective cost arising from an evaluation based on cues need not be based specifically on perceived opportunity costs, but rather is approximated through errors. Furthermore, 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 a high likelihood of perceived errors relative to alternative lines of action.

Future Directions

Much future work is needed to flesh out the extent that the use of a loose approximation of errors over cues is employed in effort evaluation. Nonetheless, several clear avenues for extension of the account are apparent. For example, using errors to determine how effortful a line of action is would not necessarily be constrained to generating prospective effort judgments. Individuals may simply estimate perceived errors while situated within a task where experience with the putative cognitive work of engaging in a task (e.g., control costs) is afforded and avoidance of a line of action is possible. Importantly, examining how an error bias may contribute to the feeling of expected and experienced effort relative to other signals such as conflict, uncertainty, or risk (Shenhav et al., in press) will provide useful insight into how individuals carry out action selection (e.g., avoidance behaviors) while situated in a task. As an example, the concept of risk entails the product of the predicted likelihood of error commission (as highlighted here) and the predicted consequence of the action (i.e., the gain/loss associated with the action, Brown & Braver, 2007; 2008). Specifically, risk has been proposed to account for many of the findings previously proposed to be best explained by conflict signals associated with processes carried out by the executive control network (e.g., the conflict monitoring hypothesis, Botvinick, 2007). Thus, risk represents an additional potential aversive element (e.g., effort, conflict, uncertainty, etc.) that is common and confounded across many contexts where effort avoidance behaviors are examined. Critically, stringently testing these determinants against one another in decision and judgment contexts will allow for a clearer picture of effort as a unique driver of behavior.

Conclusion

Effort is undoubtedly a challenging construct to define theoretically and test empirically (Botvinick & Braver, 2015; Dunn & Risko, 2016; Kurzban, 2016; Westbrook & Braver, 2015). Nevertheless, empirically testing hypotheses aimed at addressing how individuals minimize effort and why effort is often considered aversive represents a critical theoretical undertaking in understanding human behavior. This is especially true given that, in its strongest form, effort avoidance is lauded as a *law* or *principle* of human behavior (Allport, 1954; Clark, 2010; Hull, 1943; Rosch, 1998; Solomon, 1948; Zipf, 1949). The cue-utilization account presented here provides insight into both of these questions. Specifically, I proposed that effort is a result of a metacognitive evaluation made over cues in which individuals largely weigh the likelihood of a committing an error relative to alternative lines of action. This evaluation can then be used to drive action selection (e.g., engage in a task or not). The account I have developed in this thesis provides an important move forward in the journey to not only understand effort, but also human behavior generally.

References

Allport, G. (1954). *The nature of prejudice.* Cambridge, MA: Perseus Books.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, *86*(2), 124-140.

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.

Arango-Muñoz, S. (2014). The nature of epistemic feelings. *Philosophical Psychology, 27*(2), 193-211.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126-131.

Ashcraft, M. H., & Faust, M. W. (1994). Mathematics anxiety and mental arithmetic performance: An exploratory investigation. *Cognition & Emotion*, *8*(2), 97-125.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*, 82-90.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390 412.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8,* 47-89.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412-427.

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.

Bates D., Maechler, M., Bolker, B. & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-8. Retrieved from http://CRAN.R project.org/package=lme4.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv,* arXiv:1506.04967.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252-1265.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly, 37,* 220-240.

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., & Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement Disorders, 12*(6), 1028-1034.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between subjects design. *Psychological Methods*, *4*(3), 243-249.

Bitgood, S., & Dukes, S. (2006). Not another step! Economy of movement and pedestrian choice point behavior in shopping malls. *Environment and Behavior*, *38*(3), 394 405.

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.

Boksem, M. A., & Tops, M. (2008). Mental fatigue: Costs and benefits. *Brain Research Reviews*, *59*(1), 125-139.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 356-366.

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., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*(12), 529-546.

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.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106-113.

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.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23-35.

Carter, E. C., & McCullough, M. E. (2013). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect. *Behavioral and Brain Sciences*, *36*(6), 683-684.

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*(2), 429-437.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41*(1), 28-35.

Clark, A. (2010). *Supersizing the mind.* Oxford: Oxford University Press.

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social, Cognitive, and Affective Neuroscience*, *9*(9), 1289-1302.

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, *24*(1), 154-176.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189-212.

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.

Drew, G. C. (1951). Variations in reflex blink-rate during visual motor tasks. *Quarterly Journal of Experimental Psychology, 3*(2), 73-88.

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2017, February 17). Determinants of effort: Effort-based choices are associated with error-likelihood, not time demand. Available at Research Gate: https://researchgate.net/publication/309386517_Determinants_of_Effort_Effort-based_Choices_are_Associated_with_Error-Likelihood_not_Time_Demand

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.

Efklides, A. (2001). Metacognitive experiences in problem solving: Metacognition, motivation, and self-regulation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 297-323). Dordrecht, The Netherlands: Kluwer.

Efklides, A. (2002). The systemic nature of metacognitive experiences: Feelings, judgments, and their interrelations. In M. Izaute, P. Chambres, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use. Dordrecht*, The Netherlands: Kluwer.

Efklides, A., Kourkoulou, A., Mitsiou, F., & Ziliaskopoulou, D. (2006). Metacognitive knowledge of effort, personality factors, and mood state: Their relationships with effort related metacognitive experiences. *Metacognition and Learning, 1,* 33-49.

Efklides, A., & Petkaki, C. (2005). Effects of mood on students' metacognitive experiences. *Learning and Instruction, 15,* 415-431.

Epstein, R. (2000). The neural-cognitive basis of the Jamesian stream of thought. *Consciousness and Cognition, 9*(4), 550-575.

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.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition, 9*(2), 288-307.

Fleck, M. S., Daselaar, S. M., Dobbins, I. G., & Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cerebral Cortex, 16*(11), 1623-1630.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 367*(1594), 1338 1349.

Fogarty, C., & Stern, J. A. (1989). Eye movements and blinks: their relationship to higher cognitive processes. *International Journal of Pscyhophysiology, 8*(1), 35-42.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116-124.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression.* Thousand Oaks, CA: Sage Publishing.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, *47*(4), 495-501.

Fu, W. T., & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science, 28*(6), 901-935.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review*, *11*(4), 303-327.

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.

Gibson, E. L. (2007). Carbohydrates and mental function: Feeding or impeding the brain? *Nutrition Bulletin*, *32*, 71-83.

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.

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition, 33,* 245-260.

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.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129-166.

Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin*, *132*(6), 920-945.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6*(4), 322-335.

Gray, W. D., & Fu, W. T. (2004). Soft-constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28*(3), 359-382.

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.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208-216.

Hebscher, M., Barkan-Abramski, M., Goldsmith, M., Aharon-Peretz, J., & Gilboa, A. (2016). Memory, decision-making, and the ventromedial perfrontal cortex (vmPFC): The roles of subcallosal and posterior orbitofrontal cortices in monitoring and control processes. *Cerebral Cortex, 26*(12), 4590-4601.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5,* 215-227.

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.

Hsee, C. K. (1993). When trend of monetary outcomes matter: Separate versus joint evaluation and judgments of feelings versus choice. Unpublished manuscript.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*(3), 247-257.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576-590.

Hsee, C. K., & Zhang, J. (2004). Distinction bias: misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*(5), 680-695.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5*(4), 343-355.

Hsee, C. K., Zhang, J., & Chen, J. (2004). Internal and substantive inconsistencies in decision-making. In D. Koehler M & N. Harvey (Eds.), *Blackwell handbook of judgment and decision-making* (pp. 360-378). Oxford, England: Blackwell Publishing.

Hsee, C. K., Zhang, J., Wang, L., & Zhang, S. (2013). Magnitude, time, and risk differ similarly between joint and single evaluations. *Journal of Consumer Research, 40*(1), 172-184.

Hull, C. L. (1943). *Principles of behavior*. New York, NY: Appleton–Century.

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. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, *7*(5), 450-463.

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* (3rd ed.). Oxford, England: Oxford University Press.

Job, V., Walton, G. M., Bernecker, K., & Dweck, C. S. (2013). Beliefs about willpower determine the impact of glucose on self-control. *Proceedings of the National Academy of Sciences*, *110*(37), 14837-14842.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition*, *13*(4), 289-303.

Jolicoeur, P. (1990). Identification of disoriented objects: A dual-systems Theory. *Mind & Language*, *5*(4), 387-410.

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., & Raftery, 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. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150,* 18-36.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds), *Heuristics of intuitive judgment: Extensions and applications* (pp. 49-81). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society, 47*(2), 263-292.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582-591.

Kelly, C. L., Sünram-Lea, S. I., & Crawford, T. J. (2015). The role of motivation, glucose and self-control in the antisaccade task. *PloS One*, *10*(3), e0122218.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 95,* 329-343.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*(4), 513-549.

Kool, W., & Botvinick, M. M. (2013). The intrinsic cost of cognitive control. *Behavioral and Brain Sciences, 36*(6), 697-698.

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.

Kolers, P. A. (1975). Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory, 1*(6), 689-701.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370.

Koriat, A. (2007). Metacognition and consciousness. In P.D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge, England: Cambridge University Press.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 34-53.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36-69.

Koriat, A., & Norman, J. (1984). What is rotated in mental rotation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 421-434.

Koriat, A., & Norman, J. (1985). Reading rotated words. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(4), 490-508.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing subjective and objective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147-162.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573-603.

Kurniawan, I., Guitart-Masip, M., & Dolan, R. (2011). Dopamine and effort-based decision making. *Frontiers in Neuroscience*, *5*, 47-56.

Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology*, *8*(2), 244-259.

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 of 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.

Lange, F., & Eggert, F. (2014). Sweet delusion. Glucose drinks fail to counteract ego depletion. *Appetite*, *75*, 54-63.

Lawrence, M. A. (2015). ez: Easy analysis and visualization of factorial experiments. R package version 4.3. Retrieved from http://CRAN.R project.org/package=ez

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course.* New York: Cambridge University Press.

Lichtenstein, S., & Slovic, P. (Eds.) (2006). *The construction of preference.* New York: Cambridge University Press.

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction of judgments of learning for massed and spaced rehearsal. *Metacognition and Learning, 7*(3), 175-195.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., Von Bastian, C. C., Carruth, N. P., & Miyake, A. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A pre-registered study. *PloS One*, *11*(2), e0147770.

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.

Maglio, P. P., Wenger, M. J., & Copeland, A. M. (2008). Evidence of the role of self-priming in epistemic action: Expertise and the effective use of memory. *Acta Psychologica, 127*(1), 72-88.

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.

Mangan, B. (2001). Sensation's ghost: The non-sensory "fringe" of consciousness. *Psyche, 7*(18). Retrieved from psyche.cs.monash.edu.au/v7/psyche-7-18-mangan.html

Marsh, B., Schuck-Paim, C., & Kacelnik, A. (2004). Energetic state during learning affects foraging choices in starlings. *Behavioral Ecology*, *15*(3), 396-399.

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.

Martin, T., & Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science*, *29*(4), 587-625.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57*(3), 203-220.

Maupertuis, de P. L.M. (1750). Essai cosmologie (Essay on cosmology). Amsterdam.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47-60.

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.

Medalla, M., & Barbas, H. (2010). Anterior cingulate synapses in prefrontal areas 10 and 46 suggest differential influence in cognitive control. *The Journal of Neuroscience, 30*(48), 16068-16081.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 851-861.

Michaelian, K. (2012). Metacognition and endorsement. *Mind & Language, 27*(3), 284-307.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167-202.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

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.

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

Morgan, P. L., Patrick, J., Waldron, S. M., King, S. L. & Patrick, T. (2009). Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. *Journal of Experimental Psychology: Applied, 15*(4), 291-306.

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.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10,* 257-260.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. Bower (Ed.), *The psychology of learning and memory* (pp. 125-173). San Diego, CA: Academic Press.

Newell, A. & Simon, H. A., (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nieuwenhuis, R., Grotenhuis, M. T., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models. *R Journal, 4*(2), 38-47.

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.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

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.

Nowlis, S. M., & Simonson, I. (1997). Attribute task compatibility as a determinant of consumer preference-reversals. *Journal of Marketing Research, 34,* 205-218.

Otto, T., Zijlstra, F. R., & Goebel, R. (2014). Neural correlates of mental effort evaluation involvement of structures related to self-awareness. *Social, Cognitive, and Affective Neuroscience*, *9*(3), 307-315.

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. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534-552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* New York: Cambridge University Press.

Poole, D. C., Ward, S. A., Gardner, G. W., & Whipp, B. J. (1988). Metabolic and respiratory profile of the upper limit for prolonged exercise in man. *Ergonomics*, *31*(9), 1265-1279.

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.

Raichle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, *29*, 449-476.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219-235.

Reber, R., Fazendeiro, T. A., & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness. *Psyche, 8*(10), 175-188.

Reber, R., & Schwarz, N. (2002). The hot fringes of consciousness: Perceptual fluency and affect. *Consciousness & Emotion, 2*(2), 223-231.

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45-48.

Reber, R., Wurtz, P., & Zimmermann, T. D. (2004). Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition, 13*(1), 47-60.

Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology, 11*(2), 374-385.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615-625.

Risko, E. F., & Dunn, T. L. (2015). Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and Cognition, 36,* 61-74.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676-688.

Risko, E. F., Medimorec, S., Chisholm, J., & Kingstone, A. (2014). Rotating with rotated text: A natural behavior approach to investigating cognitive offloading. *Cognitive Science, 38*(3), 537-564.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction times and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82,* 416-425.

Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings.* Cambridge, MA: MIT Press.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomics Bulletin & Review, 21*(2), 301-308.

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics, 35,* 991-1009.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schweitzer, N. J., Baker, D. A., & Risko, E. F. (2013). Fooled by the brain: Re-examining the influence of neuroimages. *Cognition*, *129*(3), 501-511.

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite change performance. *Psychological Science,25*(12), 2199-2208.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feeling of knowing. *Psychonomic Bulletin and Review, 1,* 357-375.

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 ofcognitive 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.

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *The American Economic Review, 73*(4), 596-605.

Solomon, R. L. (1948). The influence of work on behavior. *Psychological Bulletin, 45,* 1-40.

Son, L. K., & Metcalfe, L. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204-221.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33*(6)*,* 1116-1129.

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 15-38). Cambridge, England: Cambridge University Press.

Stanovich, K. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of human fatigue. *Human Factors, 36,* 285-297.

Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink. *Psychophysiology, 21*(1), 22-33.

Suarez, R. K. (1996). Upper limits to mass-specific metabolic rates. *Annual Review of Physiology*, *58*(1), 583-605.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects.*Psychonomic Bulletin & Review*, *2*(1), 55-82.

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.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in actions: Consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1412-1423.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371-384.

Vadillo, M. A., Gold, N., & Osman, M. (2016). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychological Science*. Advanced online publication. doi: 10.1177/0956797616654911

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section, 47*(3)*,* 631-650.

Veltman, J. A., & Gaillard, A. W. L. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*(5), 656-669.

Vernon, D., & Usher, M. (2003). Dynamics of metacognitive judgments: Pre- and post retrieval mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(3)*,* 339-346.

Vokey, J. R., Baker, J. G., Hayman, G., & Jacoby, L. L. (1986). Perceptual identification of visually degraded stimuli. *Behavior Research Methods, Instruments, & Computers, 18*(1), 1-9.

Walsh, M. M., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology*, *58*(3), 416-440.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience, 15*(2), 395-415.

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.

Whittlesea, B. W. A. (1997). Production, evaluation and preservation of experiences: Constructive processing in remembering and performance tasks. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 211-264). San Diego, CA: Academic Press.

Whittlesea, B. W. A. (2003). On the construction of behavior and subjective experience: The production and evaluation of performance. In C. J. Marsolek & J. S. Bowers (Eds.), *Rethinking implicit memory* (pp. 239-260). London, England: Oxford University Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision, 8*(5):7, 1-10.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology, 12*(1), 3-18.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625-636.

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.

Wood, S.N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika, 100*(1), 221-228.

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.

Appendices

Appendix A

Self-report Questionnaire Deployed in Experiment 4

Q1. What was it like performing the task in part 1 (1-2 sentences)?

Q2. How did you choose between the decks in the last (3^{rd}) block (1-2 sentences)?

Q3. Did you develop a least-effortful preference for one of the decks in the last (3^{rd}) block (circle one)?

YES NO

Q4. Was there any difference between the decks in part 1 (circle one)?

YES NO

Q5. How confident did you feel that there was/was not a difference between the decks (from 0% - 100%)?

Q6. For some participants, one of the two decks had the tendency to switch between colors more often while the other deck tended to repeat the same color. Did it seem like this was the case for you (circle one)?

YES NO

Note: Individuals completed all seven questions upon completion of both the rotation pair and switching pair (i.e., part 1 and part 2).

Appendix B

Self-report Questionnaire Deployed in Experiments 5 and 6

Q1. What was it like performing the task in part 1 (1-2 sentences)?

Q2. Was there any difference between the decks in part 1 (circle one)?

YES NO

Q3. How confident did you feel that there was/was not a difference between the decks (from 0% - 100%)?

Q4. If you answered YES to question 2 above, what do you believe the difference(s) was between the decks (1 sentence)?

Q5. How did you choose between the decks in the last (3^{rd}) block in part 1(1 sentence)?

Q6. Did you develop a least-effortful preference for one of the decks in the last (3^{rd}) block (circle one)?

YES NO

Q7. If you answered YES to question 6 above, what was the attribute(s) that you used to decide that one deck as less effortful than the other?

Note: Individuals completed all seven questions upon completion of both the rotation pair and switching pair (i.e., part 1 and part 2) in Experiment 2. The seven questions were only completed once upon completion of the DST in Experiment 3.