How hard was that? Context effects on judgments of effort

by

Michelle Roshan Marie Ashburner

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2024

© Michelle Roshan Marie Ashburner 2024

Examining Committee Membership

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

External Examiner	Dr. Maggie Toplak Professor, Department of Psychology York University, Ontario, Canada
Supervisor	Dr. Evan F. Risko Professor, Department of Psychology University of Waterloo, Ontario, Canada
Internal Member	Dr. Derek Koehler Professor, Department of Psychology University of Waterloo, Ontario, Canada
Internal-external Member	Dr. Elizabeth Irving Professor, School of Optometry University of Waterloo, Ontario, Canada
Other Member	Dr. James Danckert Professor, Department of Psychology University of Waterloo, Ontario, Canada

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

Michelle Ashburner was the sole author for Chapters 1 and 4, written under the supervision of Dr. Evan Risko. Chapters 1 and 4 were not written for publication elsewhere.

This thesis includes two manuscripts written for publication, one comprising Chapter 2 and one comprising Chapter 3. Research presented in these chapters was conducted at the University of Waterloo by Michelle Ashburner under the supervision of Dr. Evan Risko. As the lead author of these chapters, my responsibilities included contribution to the conceptualization and design of studies, data collection and analysis, writing, and submission of the associated manuscripts. Dr. Evan Risko provided guidance during the research process and provided feedback on draft manuscripts.

Chapter 2 citation:

Ashburner, M., & Risko, E. F. (2022). On the influence of evaluation context on judgments of effort. *Journal of Experimental Psychology: Human Perception and Performance, 48*(8), 790-811.

Chapter 3 citation:

Ashburner, M. & Risko, E. F. (2024). Judgments of effort and associated cues are influenced by stimulus context. [Manuscript submitted for publication].

Note. Chapter 2 does not reflect the exact published work but has been edited for presentation in this thesis.

Abstract

How do individuals make judgments of effort? Despite cognitive effort being a central construct in scholarship, as well as an influential concept in day-to-day life, we have a limited understanding of how individuals determine the effort associated with cognitive acts. Recent work has demonstrated that judgments of effort can be influenced by the context in which they are made (i.e., the *judgment context*). I employed a reading task and stimulus set that has produced a reliable dissociation between judgments of effort and cognitive demand to further investigate contextual influences on effort judgments. Specifically, I manipulated the context (i.e., the *evaluation context*) in which individuals read and judged the stimuli; collected individuals' reasons for their effort judgments; and measured objective demand (i.e., reading times, error counts). In Experiments 1-4, I determined that this context manipulation did not seem to eliminate the dissociation between judgments of effort and objective demand; however, I revealed that evaluation context has a robust effect on judgments of effort. Furthermore, individuals' reasons varied just as markedly across evaluation contexts. In Experiments 5-7, I extended this work by manipulating the context in which individuals read the stimuli (i.e., the *stimulus context*) while holding the judgment context constant. Individuals' reasons for judgment suggested that the cues used to make effort judgments are influenced by the stimulus context, with both judgments and reasons exhibiting notable changes across stimulus contexts. Implications of these results, including how they guide our understanding of the effort judgment process, are discussed.

Acknowledgements

I first thank my research supervisor Dr. Evan Risko for his dedication to my academic training. Many thanks also to my Examining Committee: Dr. Maggie Toplak, Dr. Derek Koehler, Dr. Elizabeth Irving, and Dr. James Danckert.

I have accumulated a lot of wonderful interactions with folks in the Department of Psychology over the years and I thank the Students, Faculty, and Staff who have provided such unwavering support. I begrudgingly thank the IT People.

I am grateful for the unconditional support from my friends, both on and off campus. My life has been made immeasurably better because of you!

I close by thanking my husband Nizar, with whom food tastes better, winter feels warmer, and nature looks brighter. To me, he has been a model of what it means to be loved without bound. I aim to be the same to him.

Dedication

This work is dedicated to the memory of William Ronald Childs.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xii

Chapter 1 Introduction 1

1.1 Motivation 1

1.2 Perspectives on Subjective Effort 1

1.3 Current Investigation 4

Chapter 2 On the influence of evaluation context on judgments of effort 7

2.1 Introduction 7

2.1.1 Joint vs. Separate Evaluation 7

2.1.2 Joint vs. Separate Evaluation of Effort 11

2.2 Experiments 1a and 1b 13

2.2.1 Method 14

2.2.2 Results 17

2.2.3 Discussion 22

2.3 Experiments 2a and 2b 24

2.3.1 Method 24

2.3.2 Results 27
2.3.3 Discussion 31
2.4 Experiment 3 33
2.4.1 Method 33
2.4.2 Results 34
2.4.3 Discussion 38
2.5 Experiment 4 40
2.5.1 Method 42
2.5.2 Results 45
2.5.3 Discussion 57
2.6 General Discussion 58
2.7 Conclusion 63
Chapter 3 Judgments of effort and associated cues are influenced by stimulus context 64
3.1 Introduction 64
3.2 Experiment 5 67
3.2.1 Method 68
3.2.2 Results 70
3.2.3 Discussion 78
3.3 Experiment 6 80
3.3.1 Method 80
3.3.2 Results 80
3.3.3 Discussion 86
3.4 Experiment 7 90
3.4.1 Method 90

3.4.2 Results 90
3.4.3 Discussion 96
3.5 General Discussion 97
3.6 Conclusion 100
Chapter 4 General Discussion 102
4.1 Summary of Main Results 102
4.2 How do Individuals Make Judgments of Effort? 103
4.3 Empirical Questions Addressed 105
4.3.1 Evaluation Context and Measures of Objective Demand 105
4.3.2 Reasons for Judgment and Cue Use Across Contexts 107
4.3.3 Explicit Cue Reports do not Fully Explain Judgments 108
4.4 Limitations and Further Considerations 110
4.4.1 Employing Explicit Cue Reports 110
4.4.2 Other Contextual Influences on Cue Use in Judgments of Effort 111
4.4.3 Relative Influence of Stimulus Context vs. Judgment Context 112
4.4.4 Construction of Individuals' Effort Judgments 112
4.4.5 Cues as a Causal Mechanism for Changes in Judgments 113
4.5 Conclusion 114
References 116

List of Figures

Figure 1. *Examples of stimuli used in Experiments 1a and 1b* 15
Figure 2. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1a* 20
Figure 3. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1b* 21
Figure 4. *Examples of stimuli used in Experiments 2a and 2b* 26
Figure 5. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2a* 28
Figure 6. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2b* 30
Figure 7. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the experience condition* 36
Figure 8. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the no experience condition* 38
Figure 9. *Mean reading times (in seconds) as a function of stimulus type grouped by evaluation mode* 48
Figure 10. *Mean effort judgments as a function of stimulus type grouped by evaluation mode* 49
Figure 11. *Percent of total responses in favour of each feature for the open response data, split by evaluation condition* 53
Figure 12. *Percent of total responses in favour of each feature for the selected response data, split by evaluation condition* 53
Figure 13. *Percent of responses in favour of each of the four main features for open response and selected response data* 56
Figure 14. *Effort judgments as a function of stimulus type and stimulus context in Experiment 5* 72
Figure 15. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 5* . 77
Figure 16. *Effort judgments as a function of stimulus type and stimulus context in Experiment 6* 82
Figure 17. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 6* . 86
Figure 18. *Effort judgments as a function of stimulus type and stimulus context in Experiment 7* 92
Figure 19. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 7* . 94

List of Tables

Table 1. *Mean effort judgments by stimulus type and evaluation mode for Experiments 1-3* 18
Table 2. *Mean response times, error counts, and effort judgments by stimulus type and evaluation mode for Experiment 4* 47
Table 3. *Percent of total responses in favour of each feature by evaluation context* 52
Table 4. *Proportions of participants endorsing a structure-based reason within stimulus type and stimulus context* 74
Table 5. *Effort judgments and cell sizes by stimulus context, stimulus type, and reason for Experiments 5-7* 76
Table 6. *Proportion of single cue use (compared to two cues) split by stimulus context for Experiments 5-7* 78
Table 7. *Means, standard deviations, and cell sizes for rotation estimates in degrees split by primary reason and stimulus context* 96

Chapter 1
Introduction

1.1 Motivation

Cognitive effort has long been a construct of interest within the psychological sciences. Information gathered from the study of cognitive effort can enable a better understanding of how people behave and make decisions. Moreover, the understanding of how individuals experience cognitive load and prioritize effort expenditures is essential to many fields outside of psychology, including education (David et al., 2024; Paas et al., 2005), human factors engineering (Couceiro et al., 2019; Sun, Yao, & Carretero, 2013; Wickens, 2024), human-computer interaction (Luong et al., 2019; Mosaly at al., 2018), and library and information sciences (McGregor, Azzopardi, & Halvey, 2021; Schwieder, 2016). Indeed, Zipf's work on the Principle of Least Effort (i.e., we tend to seek out the *path of least resistance*, all else being equal; Zipf, 1949) has had widespread interdisciplinary influence (Chang, 2016). Despite having received such dedicated attention, understanding how individuals judge the effort associated with a given task remains an open question. My dissertation focuses on deepening our understanding of this fundamental question: How do individuals judge the effort associated with a cognitive task? If effort represents one of the primary drivers of human behaviour (e.g., Zipf, 1949), then progress on this question seems central to our effort in psychological science. In this chapter, I provide a brief survey of various perspectives on subjective effort to set the stage for the following two empirical chapters that are themselves published (Chapter 2) or submitted (Chapter 3); as such, they contain introductory material in their own right.

1.2 Perspectives on Subjective Effort

An intuitive account of judgments of effort are that they simply reflect a kind of readout of one's experience of effortful processing. Theories of what make processing effortful often hold that the use of central resources, which can be defined as a finite pool of resources that enable the processing of information as well as higher-order cognition (Gopher & Sanders, 1984), drive the experience of cognitive effort. For example, Yeh and Wickens (1984) state

that “Most variables that have been found to affect subjective workload are related to the demands on working memory.” (Yeh & Wickens, 1984, p. 6). This includes variables such as the number of concurrent tasks undertaken, and the amount of information held in working memory (Yeh & Wickens, 1988). Working memory requires central resources; therefore, from this perspective, a task will feel effortful to the extent that central resources are deployed toward its completion. Thus, judgements of effort could reflect a readout of this experience of effortfulness or central resource use.

A related but distinct proposition about the experience of effort is that it reflects a signal to the value of foregone activities (Kurzban et al., 2013; Kurzban, 2016). Specifically, Kurzban has argued that an individual’s sense of effort is the output of a computation of opportunity costs associated with that task. In this framework, the human cognitive system is held to have a finite amount of central resources at its disposal. As such, there exists a constant problem one must solve due to *simultaneity* (i.e., tasks requiring similar computational mechanisms often cannot be completed simultaneously). The solution is one of prioritization, the requirement to choose a task based on a ‘benefits vs. costs’ computation. Furthermore, the costs associated with the controlled allocation of resources to complete a given task is equivalent to the costs of not engaging in the next-best task available (i.e., *opportunity costs*). Similar ideas involve cost-benefit computations for the task at hand; for example, a recent proposed modification to the Expected Value Theory (Szekely & Michael, 2021) predicts that effort is felt as aversive only when costs outweigh benefits. Presumably, a part of such a cost-based computation could include the amount of central resources required to complete the task under consideration (i.e., whether it is the current task, or the next-best alternative). Given that, in these accounts, the experience of effort is theorized to reflect the outcome of a cost-based computation, what would that mean for judgments of effort? One could speculate that relatively greater costs would correspond to relatively higher judgments of effort for the task under consideration, reflecting some manner of access to the computed costs (e.g., a sense of aversion commensurate with the computed costs). That is, one could

extend such accounts of the experience of effort and suggest that judgements of effort reflect a readout of these computations (e.g., opportunity costs).

While a direct readout view of judgements of effort seems intuitive, such an account can have trouble explaining demonstrations that judgements of effort can be influenced by aspects of the task environment (e.g., Dunn, Gaspar, & Risko, 2019; Dunn, Lutes, & Risko, 2019; Dunn & Risko, 2016; Efklides, et al., 2006; Raaijmakers et al., 2017). That is, the same task with presumably the same central resource demands can be judged as more or less effortful depending on the context in which that judgement is made. Therefore, contextual influences on judgments of effort can present a problem for accounts that suggest that individuals' judgements of effort are based on a direct readout of the experience of effort that is a driven by central resource use or related computations. In my dissertation I adopt a different perspective on judgements of effort. Specifically, I consider judgments of effort from the perspective that they are a type of metacognitive judgment (i.e., a judgment based on an awareness of one's own cognition; American Psychological Association, n.d.). According to an influential framework, metacognitive judgments are inferences based on a variety of cues (Koriat, 1997). For example, in a memorization task cues such as how long an item appears on screen or how many times an item is displayed during a study session can inform individuals' judgments of how likely they will be to recall that item. The process of making such a judgment typically involves the application of a metacognitive belief about a particular cue; for example, the belief that *the more times an item is displayed, the easier it will be to remember* may result in a judgment of relatively high likelihood of later recall for an item that is displayed four times compared to only once. Another type of cue within this framework can be described as the subjective experience of ease associated with the processing of information (i.e., "fluency" or "ease of processing"; Alter & Oppenheimer, 2009). This kind of cue is thought to be immediately available to an individual during online processing and can operate as a cue informing judgments. In the example above, an item presented for the fourth time may feel easier to process than an item presented for the first time, and this feeling of comparatively greater ease of processing could influence a

judgement of the likelihood of later recall (regardless of whether this feeling is consciously available to the judge; see Metcalfe, 2000). Judgments of effort can be conceptualized similarly, in that a variety of cues can contribute to their construction, leaving open the possibility that changes in context can result in different judgments for the same task.

1.3 Current Investigation

This investigation comprises seven experiments (N = 4,502) wherein judgments of effort are elicited for a relatively simple reading task. In Chapter 2, participants judged the effort associated with reading 25 unrelated high-frequency words arranged in a lattice at four distinct orientations (i.e., *stimulus types*). Critically, some participants were assigned to a *joint evaluation condition*, wherein they experienced and judged all four stimulus types; and others were assigned to a *separate evaluation condition*, wherein they experienced only one stimulus type and then judged only that type. Furthermore, Experiments 1a, 1b, 2a, 2b, and 3 investigated a known dissociation between objective effort measures and self-reported effort for this reading task. Specifically, previous research using a judgement context akin to joint evaluation demonstrated that judgments of effort for the stimulus types followed a pattern that did not completely track with demand measures (e.g., reading time, error commission; Ashburner & Risko, 2021; Dunn & Risko, 2016). Thus, these studies addressed (a) whether evaluation context (i.e., joint vs. separate) has an influence on judgments of effort; and (b) whether a separate evaluation context brings judgments of effort more in line with objective demand measures. Within these experiments, prospective (i.e., in anticipation of completing the task at a later time) and retrospective (i.e., reflecting on the completed task) judgments were collected to examine whether the results differed across elicitation timings. Furthermore, prospective judgments following experience reading the stimuli as well as without such experience were collected to separate the effect of experiencing the reading task firsthand from elicitation timing.

Experiment 4 involved the collection and analysis of written responses to a probe for participants' reason(s) for their effort judgments (i.e., *explicit cue reports).* I sought to address several questions within this study to ascertain whether the reasons individuals

provided differed across evaluation contexts. Specifically, (a) Did more responses primarily address the stimulus type manipulation in the joint evaluation context, as compared to the separate evaluation context? (b) Did the reasons in the separate evaluation context exhibit more variation than those in the joint evaluation context? (c) Did the proportion of reasons change across evaluation contexts? Chapter 2 also addresses effort judgments as viewed through the lens of General Evaluability Theory (GET; Hsee & Zhang, 2010), where I discuss why this particular manipulation might lead to a shift in the cues employed across evaluation contexts, following up on Dunn, Koehler, and Risko (2017).

In Chapter 3, I follow up on the Chapter 2 findings that effort judgments were heavily influenced by the evaluation context in which they were made; and that individuals' explicit cue reports indicate a corresponding shift in cues used across evaluation contexts. Specifically, Experiments 5 and 6 address a conflation between the paradigm's *stimulus context* (i.e., the phase of the study when participants experience the display type(s)) and *judgment context* (i.e., the phase of the study when participants judge the display type(s)). Specifically, I address whether the differences in effort judgments observed in Chapter 2, and the concomitant shift in cues, are due to experiencing multiple (i.e., *mixed* stimulus context) vs. a single (i.e., *pure* stimulus context) stimulus type in the reading phase or due to the opportunity to compare (or not) stimulus types in the judgment phase. Related to questions about elicitation timing in Chapter 2, Experiment 5 employs a post-experience prospective elicitation and Experiment 6 employs a retrospective elicitation. Participants' reasons for judgment were collected in a similar way to the method detailed in Chapter 2. Focusing on cues related to the stimulus rotation, I examined the influence of stimulus context on the likelihood of endorsing a rotation-based reason. Within these two studies, I also examined the extent that participants' explicit cue reports captured the pattern of their judgments of effort. Specifically, I addressed (a) whether individuals who endorsed a reason related to the rotation of the stimuli (i.e., a *structure-based* reason) demonstrated the expected pattern of judgments (i.e., upright < rotated); and (b) whether individuals who did *not* endorse a structure-based reason judged the stimulus types as about equally effortful. Furthermore, as

participants often indicated more than one type of reason in their explicit cue reports, I investigated whether the proportion of single cue use (vs. two cues) changed across stimulus context.

Experiment 7 aimed to replicate the previous findings in Chapter 3 as well as investigate an unexpected finding from the previous two studies, which may point to individuals using the same cue in conceptually different ways across stimulus contexts. I addressed one interpretation of this account by testing whether experiencing both an upright and rotated stimulus in the mixed stimulus context made the rotated stimulus appear 'more' rotated, compared to just seeing rotated stimuli in the pure stimulus context. Finally, I further examined a possible stimulus context influence on single vs. multiple cue use by having participants select multiple reasons from a list and rank their selections. Altogether, Chapter 3 provides further support for the idea that individuals make effort judgments by appealing to available cues, while revealing a limitation of explicit cue reports in addressing questions about cue use in effort judgments.

In the reports that follow, links to the project pages and preregistrations on an open registry (i.e., Open Science Framework) for all experiments are provided, as well as data, analysis code, and the coding scheme for coding written responses. Deviations from the preregistered analysis plans are almost always additions and are explicitly reported.

Chapter 2
On the influence of evaluation context on judgments of effort

2.1 Introduction

The anticipation or experience of cognitive effort is thought to affect how individuals make decisions (Bitgood, Dukes, & Abbey, 2006; Dunn, Inzlicht, & Risko, 2019; Garbarino & Edell, 1997; Gray et al., 2006; Inzlicht, Shenhav, & Olivola, 2018; Kurzban et al., 2013; Potts, Pastel, & Rosenbaum, 2018). In recent years, many researchers have focused on finding the determinants of individuals' judgments of effort (Dunn, Koehler, & Risko, 2017; Song & Schwarz, 2008; Thompson et al., 2013; Undorf & Erdfelder, 2011, 2013, 2015; Westbrook, Kester, & Braver, 2013; Yeh & Wickens, 1988). According to one framework, judgments of effort can be viewed as a kind of cue-based inference similar to other metacognitive judgments (Dunn & Risko, 2016; Dunn, Gaspar, & Risko, 2019; Efklides, et al., 2006; Koriat, 1997; Raijmakers et al., 2017) and judgments in general. An interesting prediction derived from this idea is that judgments of effort should be sensitive to the context in which they are made (e.g., Birnbaum, 1982, 1999; Dunn et al. 2017; Hsee & Tang, 2007; Morewedge et al., 2009; Parducci, 1965, 1968). For example, Ashburner and Risko (2021) provided evidence that judgments of effort for a given task changed as a function of when individuals were asked to provide their judgment (i.e., immediately after performing a single trial of a task vs. after performing multiple trials of the same task). Provided the central role effort is often argued to play in our day-to-day lives – we are "cognitive misers" (De Neys, Rossi, & O Houde, 2013; Liu & Goodhue, 2012; Stanovich, 2018) following a "principle of least effort" (Waters, 1937; Zipf, 1949) – understanding how judgments of effort for a given cognitive act vary with the judgment context will provide important insights into the perception of effort.

2.1.1 Joint vs. Separate Evaluation

An important aspect of the judgment context is whether the judge is provided a relevant reference point against which options can be compared. In general, in what can be referred to

as a *joint evaluation context*, individuals are presented with more than one item and they are asked to evaluate the items simultaneously. Conversely, in a *separate evaluation context*, individuals are presented with a single item and asked to evaluate it in isolation (Bazerman et al., 1999; Hsee, 2000). The distinction between evaluation contexts mirrors the comparison between within-participant (i.e., joint evaluation) and between-participant (i.e., separate evaluation) value judgments (Birnbaum, 1982, 1999). The pronounced impact of evaluation context on decisions has been clearly illustrated across a number of diverse contexts (Bazerman, Loewenstein, & White, 1992; Bazerman et al., 1994; Bazerman et al., 2011; Birnbaum, 1982, 1999; Blount & Bazerman, 1996; Bohnet, van Geen, & Bazerman, 2016; Buiten & Keren, 2009; Christopoulos et al., 2011; Davidai & Shafir, 2018; Gonzalez-Vallejo & Moran, 2001; Gould & Kramer, 2009; Hsee, 1998; Hsee et al., 2003; Hsee et al., 1999; Hsee et al., 2013; Hsee & Zhang, 2004, 2010; Irwin et al., 1993; Kahneman & Ritov, 1994; Nowlis & Simonson, 1997; Ritov & Baron, 2011; Sevdalis & Harvey, 2006; Tanner, 2008; Willemsen & Keren, 2004). For example, manipulations of evaluation context on willingness to pay or choice paradigms can often lead to drastic differences in the magnitude and/or ordering of individuals' preferences, including preference reversals.

Hsee (1996) investigated willingness to pay judgments for two used music dictionaries across evaluation contexts. Two attributes were assigned to each dictionary: the number of entries and the condition of the book. In the joint evaluation context, participants were asked to indicate the amount they were willing to pay for each of two dictionaries, where Dictionary A had 10,000 entries and looked like new; and Dictionary B had 20,000 entries and looked like new except for a torn cover. In the separate evaluation context, participants were presented with either Dictionary A or Dictionary B and asked to indicate the amount they were willing to pay for the dictionary shown. In the joint evaluation context, participants reported higher willingness to pay for Dictionary B than for Dictionary A; but in the separate evaluation context, participants were willing to pay significantly more for Dictionary A. Thus, across evaluation contexts, individuals' preferences reversed (Bazerman et al., 1992; Bazerman et al., 1999; Hsee, 2000, Hsee et al., 1999; Hsee & Zhang, 2010). The

use of separate evaluation can also, in some circumstances, lead to rather nonsensical results. For example, Birnbaum (1999) demonstrated that, in a between-participant design (i.e., separate evaluation), individuals rated the number 9 as greater than the number 221 in terms of magnitude.

Hsee and colleagues (Hsee, 1996, 2000; Hsee & Zhang, 2010) have explained differences in judgments (i.e., value sensitivity) across evaluation modes via the General Evaluability Theory. According to this theory, value sensitivity is determined by the *evaluability* of the attribute. For a given individual, a relatively evaluable attribute would correspond to a high degree of value sensitivity (i.e., changes to the attribute would result in detectable changes to its value judgment); while a relatively inevaluable attribute would correspond to a low degree of value sensitivity (i.e., changes to the attribute would not result in detectable changes to its value judgment). Within the General Evaluability Theory, the type and amount of information possessed by an individual about a to-be-judged attribute determines whether an attribute is relatively evaluable or inevaluable. This information can be partitioned into three factors: knowledge, nature, and mode. Knowledge refers to information about a given attribute such as its average value and distributional properties, which are acquired by individuals through experience. Nature refers to whether an inherent reference system is possessed with which to evaluate a given attribute (e.g., temperature is an innately evaluable attribute due to humans' physiological ability to sense which temperatures are comfortable; Hsee & Zhang, 2010). Mode refers to whether an individual is making their value judgment in a joint – vs. separate – evaluation context. The joint evaluation mode – by virtue of enabling comparison across items – provides a judgment context that, all else being equal, increases the evaluability (and thus, the value sensitivity) of the attribute being evaluated. By contrast, the separate evaluation mode is a judgment context that does not allow for one item's value to serve as a reference point for the evaluation of another item. This lack of reference information can decrease the evaluability of an item's attribute, as compared to the joint evaluation mode, when an attribute is relatively inevaluable on the knowledge and/or nature dimension. For example, individuals without much knowledge of

diamonds will likely display a greater sensitivity for the size of a diamond in the joint evaluation mode than in the separate evaluation mode (Hsee & Zhang, 2010).

With respect to the preference reversal discussed above, a torn book cover is categorically different from an intact cover; most individuals need not have both items in front of them to determine that the latter is preferable. Conversely, the relation between the number of entries in a music dictionary and value may not be obvious to the average individual. That is, the condition of a book cover is a relatively evaluable attribute as compared to the number of entries – leading the former to have a greater impact on judgments in the separate evaluation mode. In the joint evaluation mode, however, the difference in the number of entries becomes more salient and, given its putative relevance to the value of a dictionary relative to the condition of the cover, ends up having a greater impact on judgments.

Birnbaum (1999) provided a similar explanation for the observation that 9 was judged as greater than 221 in a separate evaluation context. He suggested that the different stimuli (i.e., 9, 221) invoked different contexts and that these different contexts yielded different judgments when individuals reported how large the number was on a 10-point scale. In particular, the smaller digit was thought to bring to mind other single-digit numbers, of which 9 is relatively large; whereas 221 might bring to mind other three-digit numbers, of which 221 is relatively small. If one were to assume that the judgment functions – the mappings from the subjective value of the stimulus to the judgment scale – were commensurate across conditions, Birnbaum (1999)'s observation would lead to the conclusion that people believe 9 is greater than 221. In this case it is clear that the judgment functions employed by individuals in separate evaluation were context dependent.

Returning to the notion that judgments of effort reflect a kind of cue-based inference, these perspectives on the influence of evaluation mode highlight the importance of considering the potential influence of evaluation context on how individuals draw inferences from the cues available. That is, comparisons across evaluation context might serve to better

reveal the cues employed in effort judgments and how those cues are used to make an effort judgement.

2.1.2 Joint vs. Separate Evaluation of Effort

As noted above, given the important role effort putatively plays in decisions, the idea that individuals' judgments of the effortfulness of a given action depends on the judgment context is important to understand. For example, if the critical role of effort in decisions derives from the need to conserve energy (Boksem & Tops, 2008; Croxson et al., 2009; Kurzban et al., 2013; Richter, Gendolla, & Wright, 2016), it would seem counterproductive if our judgments of effort for equivalent actions were to vary considerably as a function of the judgment context.

Theoretically, one might expect effort to be relatively evaluable given our extensive experience (i.e., knowledge) and the potential that an inherent reference system (i.e., nature) exists (e.g., that which produces feelings of fluency; Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005; Thompson et al., 2013; Undorf & Erdfelder, 2011, 2013, 2015). Initial work examining this general idea, by Dunn et al. (2017), has demonstrated that the evaluability of effort (as determined by comparisons of similar actions across evaluation contexts) can depend on the task. Participants were randomly assigned to an evaluation mode (i.e., separate or joint) and were asked for their prospective judgments of effort for three tasks: a motor task, where the judged attribute was the amount of weight to be lifted; a memory task, where the judged attribute was the number of elements to be remembered; and a perceptual task, where the judged attribute was the angle of rotation of a word. Across three experiments, Dunn et al. (2017) provided evidence that weight in the motor task and set size in the to-be-remembered set were relatively evaluable attributes. That is, there was little difference across evaluation contexts in effort judgments for these attributes. However, individuals' judgments of effort were less affected by angle of rotation for the perceptual task in the separate evaluation context, as compared to judgments in the joint evaluation context. Dunn et al. (2017) interpreted this as evidence that effort in terms of stimulus rotation was relatively inevaluable.

The influence of different types of stimulus rotation on judgments of effort has also been used in recent work to examine the relation between effort judgments, demand selection, and performance (Ashburner & Risko, 2021; Dunn et al., 2019; Dunn, Lutes, & Risko, 2016; Dunn & Risko, 2016). Specifically, Dunn and Risko (2016) had participants complete multiple trials of a reading task for which there were four stimulus types, each containing 25 words: upright words in an upright frame (UW-UF); upright words in a rotated frame (UW-RF); rotated words in an upright frame (RW-UF); and rotated words in a rotated frame (RW-RF; Figure 1). After the reading task, participants provided judgments of effort for each stimulus type on a 7-point scale. Dunn and colleagues' results revealed that reading times across the stimulus types generally tracked with judgments of effort: Participants were fastest when reading the UW-UF display, slower when reading the UW-RF display, and slower still when reading the RW-UF and RW-RF displays, with the latter two being equivalent. However, a dissociation was observed: Namely, the RW-RF stimulus type was consistently judged as more effortful. Furthermore, Dunn et al. (2016) had participants complete a demand selection task with these displays (all 25 words in the displays were "WORD"), wherein individuals were presented with two of the four stimulus types side-by-side on a screen (e.g., UW-UF and UW-RF; or RW-UF and RW-RF) and were asked to choose the display they would prefer to read. Participants also completed the reading task, during which reading times, errors made, and blink rates were collected. Dunn et al. (2016) demonstrated that all of the objective demand measures were equivalent across the RW-UF and RW-RF stimulus types in the reading task (there were in fact no differences in errors made across the stimulus types); however, the RW-UF stimulus type was preferred over the RW-RF stimulus type in the demand selection task. These dissociations provide a unique opportunity to investigate the contributions to judgments of effort that are seemingly independent of measures of cognitive demand.

In attempting to explain the observed dissociations, Dunn and Risko (2016) suggested that the RW-RF stimulus type was judged to be more effortful to read because it has two forms of rotation, whereas the RW-UF stimulus type only has one (Figure 1). Such an

account suggests that the opportunity to make comparisons across the different types of rotation might be necessary for the observed dissociation. That is, in Dunn and Risko (2016; see also Dunn et al., 2016; Risko et al., 2014) participants were exposed to all of the different types of stimulus rotation (i.e., UW-UF, UW-RF, RW-UF, RW-RF). Indeed, the type of rotation was the only attribute of the task that systematically varied across trials. This judgment context might have contributed to the outsized impact of stimulus rotation – relative to other putative metrics of cognitive demand – on judgments of effort. From this perspective, the joint evaluation context draws attention to an attribute of the stimuli that might be misleading in terms of its mapping to objective measures of cognitive demand. This raises the interesting possibility that removing the ability of individuals to compare across stimulus types – using a pure separate evaluation design – would lead to an elimination of the observed dissociation. In particular, separate evaluation might eliminate the perceived difference in effort between reading the RW-UF and RW-RF displays, possibly by bringing to the fore aspects of the stimulus that lead judgments of effort to better align with objective measures of demand. Here we provide a direct test of this general prediction by comparing judgments of effort across the four stimulus types used by Dunn and Risko (2016) in both joint and separate evaluation contexts. We examine this general question across four experiments wherein we also manipulate other aspects of the judgement context. In addition, in Experiment 4 we examine individuals' reasons for their effort judgments across the joint and separate evaluation contexts. Taken together, the present investigation provides novel insight into the sensitivity of judgments of effort to judgement context and through doing so a deeper understanding of how individuals make effort judgments.

2.2 Experiments 1a and 1b

To establish whether judgments of effort for the stimulus types from Dunn et al. (2016; Figure 1) differed according to evaluation mode, participants were randomly assigned to a joint evaluation mode or a separate evaluation mode. In the former, participants viewed all four stimulus types, while in the latter, only one stimulus type was randomly chosen for presentation. Across conditions, the stimulus types comprised 25 instances of the word

"WORD", thus only differing by their orientations. Participants provided prospective judgments of effort for the task of reading the display(s) as quickly and as accurately as possible while keeping their head upright. We conducted Experiment 1b to replicate the results of Experiment 1a with a larger sample size as well as to preregister the analysis examining the expected interaction between stimulus type and evaluation mode, which was exploratory in Experiment 1a.

2.2.1 Method

2.2.1.1 Participants

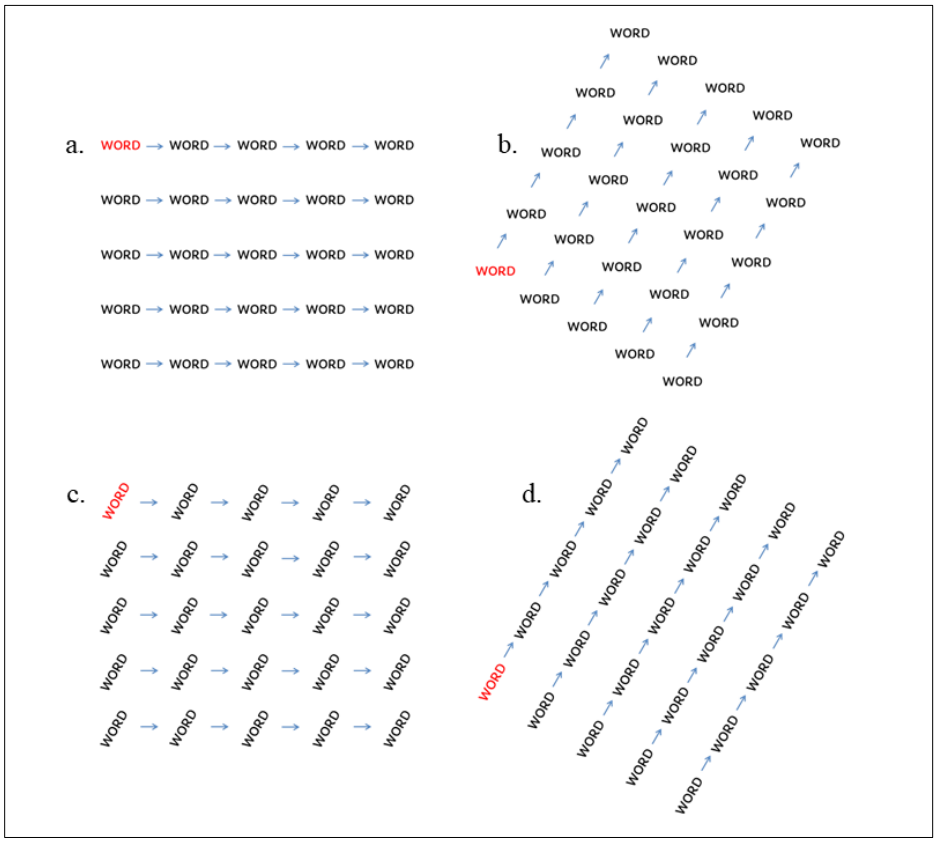
Participants located within the US or Canada were recruited from Amazon Mechanical Turk. Specifically, workers with at least 100 HITs of experience and at least a 95% approval rating, were eligible for participation. A sample size of 320 was determined using previous research (Ashburner & Risko, 2021); however, to facilitate complete counterbalancing in the within-participants conditions, and in anticipation of some participants providing incorrect responses to the attention check, we aimed to recruit 500 and 720 participants in Experiment 1a and 1b, respectively.

2.2.1.2 Stimuli and Design

A one-factor (stimulus type) within-participants design was employed to measure effort ratings in the joint evaluation mode; and a one-factor (stimulus type) between-participants design was employed to measure effort ratings in the separate evaluation mode. There were four stimulus types: displays with upright words and an upright frame (UW-UF), displays with upright words and a rotated frame (UW-RF), displays with rotated words and an upright frame (RW-UF), and displays with rotated words and a rotated frame (RW-RF). All rotations were at 60° and could be counterclockwise or clockwise. All arrays comprised 5 rows of 5 words each, where every word was "WORD" typed in 18-point black Courier New font on a white background. The starting "WORD" in each display was coloured red, and blue single-headed arrows marked the direction in which one would read. A 7-point effort scale (*1 = Not at all Effortful*; *7 = Very Effortful*) was provided below the stimulus. For the joint evaluation

mode, all four stimulus types were presented on the same display – with each stimulus occupying one quadrant of the display – and labelled a. through d. Four effort scales corresponding to each of the four stimulus types (i.e., labelled a. through d.) were provided below the display. The position of each stimulus (i.e., upper left, upper right, lower left, lower right) was counterbalanced, as was direction of rotation. Thus, there were 24 counterbalances for each direction of rotation, resulting in 48 counterbalances in total for the joint evaluation mode. Figure 1 is a visualization of the stimuli in the separate and joint evaluation modes.

Figure 1. *Examples of stimuli used in Experiments 1a and 1b*



Note. a. UW-UF; b. UW-RF; c. RW-UF; d. RW-RF, rotated 60° clockwise. For the joint evaluation condition, participants were shown all four displays simultaneously, as above, and provided four corresponding prospective effort judgments. For the separate evaluation condition, participants were shown exactly one of these displays and provided a single effort judgment.

2.2.1.3 Procedure

Interested participants were provided a link with which to access a survey that was hosted by Qualtrics. For ease of data coding and analysis, six different survey links were distributed: one link for each of the four displays in the separate evaluation mode; one for all counterclockwise rotated displays in the joint evaluation mode (as per Figure 1); and one for all clockwise rotated displays in the joint evaluation mode. Randomization into one of two evaluations modes, and subsequent randomization into one of four conditions in the separate evaluation mode (UW-UF, UW-RF, RW-UF, RW-RF) was therefore achieved via the distribution of separate surveys. After online consent was provided, participants were presented with an instruction page, followed by the effort judgment task, then an attention check question, after which the study concluded.

In the separate evaluation condition, the instructions stated that participants would be shown a display of words, and that their task was to imagine reading actual words in that display aloud as quickly and as accurately as possible while keeping their head upright. Participants were explicitly instructed to refrain from reading the words as shown, and instead to only imagine the described task. Upon advancing to the next screen, instructions remained at the top of the page, followed by an image of one of the four stimulus types, and a 7-point effort scale (*1 = Not at all Effortful*; *7 = Very Effortful*) on which to indicate the effort they would associate with reading actual words in the display shown.

The instructions in the joint evaluation condition were identical to those in the separate evaluation condition except that it was indicated that multiple displays would be shown at the same time, and corresponding effort scales would be provided. The task in the joint evaluation condition was also identical to that in the separate evaluation condition – the placement of instructions, display image, and effort scales was the same – except for there being four displays, below which were four effort scales. In both the joint and separate evaluation conditions, the attention check question asked participants to read ten statements about the task they just completed and select the one false statement. Following this, participants were provided written feedback on the study and a link to receive payment.

2.2.2 Results

2.2.2.1 Experiment 1a

A total of 588 participants responded to one of the six surveys. Of these, 206 were excluded from analysis. While attempts were made to prevent the same participant from completing more than one survey, a search through the metadata showed that there were four participants who provided data in two different surveys. We were able to find and remove three of these four participants' data. Out of the remaining 582 participants, 200 provided an incorrect response to the attention check question, four of which also did not finish the survey. Each analysis was conducted with data from participants (total N = 382 minus required exclusions for the Erlebacher's analysis as discussed below) who passed the attention check; and another set of data, the *full dataset*, that includes the participants who failed the attention check (total N = 578 minus exclusions). Overall, the results were qualitatively similar across datasets and all reported results that follow are based on the dataset excluding participants who failed the attention check. Any differences will be noted in the Supplementary Materials section. This study was preregistered; the preregistrations, data, analysis scripts; and registration summary are available on the Open Science Framework: https://osf.io/s478n/. Table 1 displays descriptive statistics for Experiments 1 through 3. All analyses were conducted using the open-source statistical analysis software R, version 4.0.2.

Table 1. *Mean effort judgments by stimulus type and evaluation mode for Experiments 1-3*

	Experiment 1a	*Experiment 1b*	*Experiment 2a*	*Experiment 2b*	*Experiment 3*
Joint Evaluation					
UW-UF	1.54 (1.33)	1.87 (1.40)	1.90 (1.66)	1.70 (1.64)	1.61 (1.23)
UW-RF	3.09 (1.70)	3.69 (1.75)	3.41 (1.86)	3.34 (1.68)	3.35 (1.71)
RW-UF	4.10 (1.69)	4.01 (1.70)	3.84 (1.83)	3.61 (1.79)	3.69 (1.63)
RW-RF	4.54 (1.68)	4.70 (1.69)	4.81 (1.60)	4.41 (1.68)	4.49 (1.69)
Separate Evaluation					
UW-UF	3.37 (1.84)	3.09 (1.60)	3.67 (1.75)	4.00 (1.90)	2.88 (1.63)
UW-RF	3.84 (1.63)	3.74 (1.58)	4.35 (1.80)	4.01 (1.69)	3.39 (1.63)
RW-UF	3.59 (1.75)	3.70 (1.74)	3.83 (1.89)	3.90 (1.81)	3.12 (1.75)
RW-RF	4.07 (1.71)	3.43 (1.67)	3.79 (1.70)	3.87 (1.69)	3.43 (1.63)

Note. Standard deviations in parentheses. Data comprise observations from participants who provided the correct response to the attention check question.

2.2.2.1.1 Erlebacher's Analysis

To test for an interaction between stimulus type and evaluation mode (i.e., the within-participants vs. the between-participants design), an Erlebacher's analysis was conducted using code developed by Merritt, Cook, and Wang (2014). This analysis was developed for situations where the effect of a condition in a between-participants design is to be compared with the same effect in a within-participants design. As such, the *F*-statistic for the interaction between condition and design is not biased, resulting in a lower likelihood of Type II error (Erlebacher, 1977). As this analysis required all cells to have equal size, further exclusion of data was necessary. Specifically, as the lowest cell size across evaluation modes was 68, this was the largest cell size that could be achieved when using the Erlebacher's analysis.[1] In all experiments, the same analyses as reported below were conducted on a dataset including observations from participants who failed the attention check, the *full*

[1] Additional data from the other conditions were removed beginning with the latest survey start time as recorded by Qualtrics and working chronologically backwards until all cells contained 68 observations.

dataset. Any qualitative differences between the results from the *full dataset* and the main results are discussed inline.

A 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) Erlebacher's ANOVA was conducted with data from 340 participants (n = 68 observations per cell). This analysis revealed a main effect of stimulus type, $F(3, 268) = 35.08$, $p < .001$, $\eta^2 = .13$; and a main effect of evaluation mode, $F(1, 268) = 5.43$, $p = .021$, $\eta^2 = .01$, such that the mean judgment of effort in the joint evaluation mode ($M = 3.32$, $SD = 1.97$) was significantly lower than the mean judgment of effort in the separate evaluation mode ($M = 3.72$, $SD = 1.74$). Finally, there was a significant interaction between stimulus type and evaluation mode, $F(3, 268) = 17.62$, $p < .001$, $\eta^2 = .07$. Analyses to further examine this interaction were not preregistered; as such, they should be considered exploratory and are reported below.

To investigate the stimulus type by evaluation mode interaction, we conducted two separate ANOVAs with stimulus type as the factor and judgment of effort as the outcome variable: A single factor repeated measures ANOVA and a single factor between-participants ANOVA, for the joint and separate evaluation mode data, respectively. In each analysis, there were 68 observations per condition.

2.2.2.1.2 Judgments of Effort in the Joint Evaluation Mode

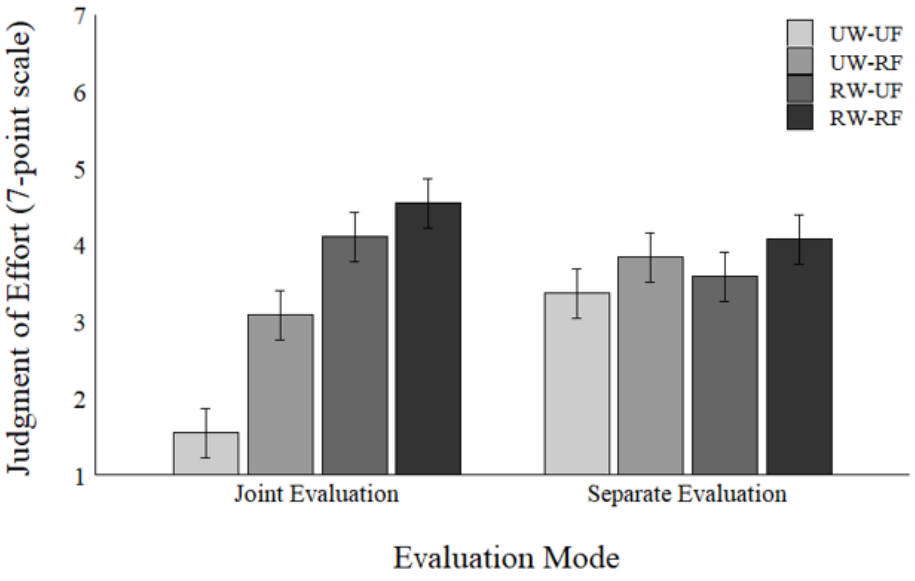
A single factor repeated measures ANOVA revealed a significant effect of stimulus type, $F(2.41, 161.74) = 66.38$, $p < .001$, $\eta_g^2 = .34$. Paired comparisons indicated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(67) = 8.12$, $p < .001$, $d = 1.00$[2]; which was judged as less effortful than the RW-UF stimulus type, $t(67) = 4.02$, $p < .001$, $d = 0.60$; which was judged as less effortful than the RW-RF stimulus type, $t(67) = 2.38$, $p = .020$, $d = 0.26$.

[2] Cohen's *d* computed using the effsize::cohen.d function.

2.2.2.1.3 Judgments of Effort in the Separate Evaluation Mode

A single factor between-participants ANOVA revealed no significant effect of stimulus type, $F(3, 268) = 2.12$, $p = .098$, $\eta_g^2 = .02$. Figure 2 is a visualization of the data. When the full dataset was analyzed ($n = 108$ per cell), there was a significant effect of stimulus type, $F(3, 428) = 7.07$, $p < .001$, $\eta_g^2 = .05$; however, the pattern across stimulus types was the same as that for the main analyses. Specifically, the UW-RF and RW-RF stimulus types were judged to be more effortful than the UW-UF and RW-UF stimulus types.

Figure 2. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1a*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs (Loftus & Masson, 1994) for separate and joint evaluation data, respectively.

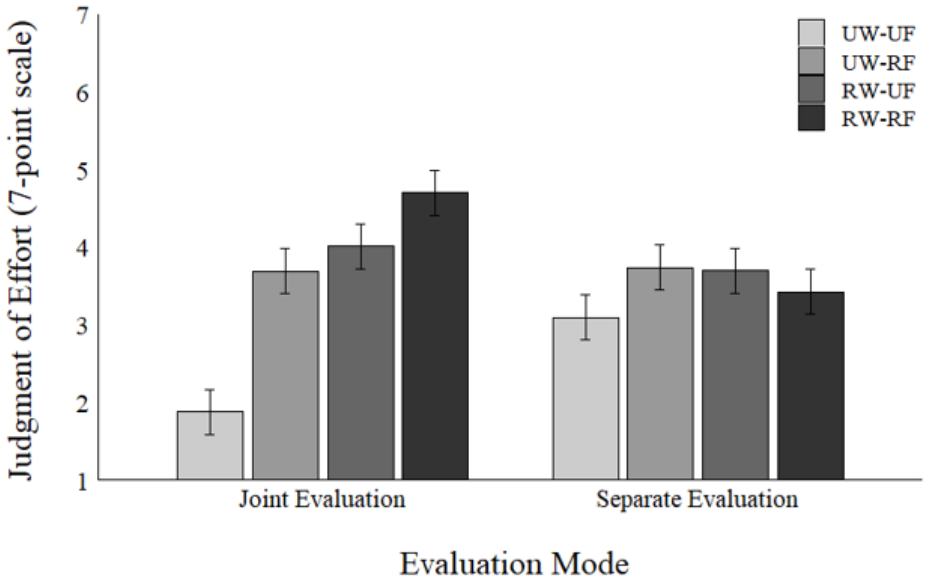
2.2.2.2 Experiment 1b

A total of 794 participants responded to one of the six surveys. Of these, 298 were excluded from analysis due to providing an incorrect response to the attention check question.

2.2.2.2.1 Erlebacher's Analysis

A 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) Erlebacher's ANOVA was conducted with data from 435 participants (n = 87 observations per cell). This analysis revealed a significant effect of stimulus type, $F(3, 344) = 38.30$, $p < .001$, $\eta^2 = .12$. There was no main effect of evaluation mode, $F(1, 344) < 1$, such that the mean judgment of effort in the Joint Evaluation condition ($M = 3.57$, $SD = 1.94$) was not significantly different than the mean judgment of effort in the separate evaluation mode ($M = 3.49$, $SD = 1.67$). When the full dataset was analyzed, judgments were on average significantly higher in the separate evaluation mode. Critically, there was a significant interaction between stimulus type and evaluation mode, $F(3, 344) = 20.05$, $p < .001$, $\eta^2 = .06$. To examine this interaction, two separate ANOVAs were computed as described above. Figure 3 is a visualization of these data.

Figure 3. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1b*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.2.2.2.2 Judgments of Effort in the Joint Evaluation Mode

A single factor repeated measures ANOVA, with 87 observations per condition, revealed a significant effect of stimulus type, $F(3, 258) = 67.14$, $p < .001$, $\eta_g^2 = .29$. Paired comparisons indicated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(86) = 9.15$, $p < .001$, $d = 1.14$; but there was no significant difference in effort judgment between the UW-RF ($M_{\text{UW-RF}} = 3.69$, $SD_{\text{UW-RF}} = 1.75$) and the RW-UF ($M_{\text{RW-UF}} = 4.01$, $SD_{\text{RW-UF}} = 1.70$) stimulus type, $t(86) = 1.50$, $p = .136$, $d = 0.19$. Finally, the RW-RF stimulus type was judged as less effortful than the RW-RF stimulus type, $t(86) = 3.57$, $p = .001$, $d = 0.41$.

2.2.2.2.3 Judgments of Effort in the Separate Evaluation Mode

A single factor between-participants ANOVA, with 87 observations per condition, revealed a significant effect of stimulus type, $F(3, 344) = 2.85$, $p = .037$, $\eta_g^2 = .02$. Pairwise comparisons revealed that the UW-UF stimulus type was judged as less effortful than the UW-RF and RW-UF stimulus types, $t(171.96) = 2.67$, $p = .008$, $d = 0.40$ and $t(170.88) = 2.40$, $p = .017$, $d = 0.36$, respectively. No other differences were significant; however, when the full dataset was analyzed, there was also a significant difference in effort judgment between the UW-UF ($M_{\text{UW-UF}} = 3.44$, $SD_{\text{UW-UF}} = 1.72$) and the RW-RF ($M_{\text{RW-RF}} = 3.86$, $SD_{\text{RW-RF}} = 1.76$) stimulus types.

2.2.3 Discussion

The results of Experiments 1a and 1b demonstrated that the pattern of prospective judgments of effort of rotated stimuli depends greatly on evaluation context. This was supported by a significant interaction between stimulus type and evaluation mode in both Experiments. Follow-up analyses revealed the nature of this interaction. Specifically, in the joint evaluation mode, judgments of effort follow a clear pattern wherein the UW-UF stimulus type is judged as least effortful, followed by UW-RF, RW-UF, and then RW-RF (although in Experiment 1b the increase in effort from the UW-RF to the RW-UF stimulus type was not significant). Importantly, this graded pattern has been observed repeatedly in previous research where

participants experienced each of the four stimulus types (Ashburner & Risko, 2021; Dunn & Risko, 2016). In contrast, in the separate evaluation mode, judgments of effort do not follow this pattern. Specifically, effort judgments made in the separate evaluation mode appear to be completely insensitive to the three different types of stimulus rotation and much less sensitive to the putative difference in effortfulness across the upright and rotated stimuli.

It is vital to note at this juncture that our prediction regarding effort judgments in the separate evaluation mode was that the pattern across the stimulus types would demonstrate a stronger relation with objective demand measures. Consequentially, as stated in the Introduction, the difference in effort judgments across the RW-UF and RW-RF stimulus types would be expected to vanish in the separate evaluation mode. While this was certainly the case, the results of Experiments 1a and 1b suggest a much more profound effect of evaluation mode than we expected. Interestingly, the pattern of effort judgments in the separate evaluation mode also does not coincide with measures of objective demand for these stimuli (with the exception of errors made, which was typically unaffected by stimulus type in previous experiments). In previous research (Ashburner & Risko, 2021; Dunn et al., 2016; Dunn & Risko, 2016), in terms of reading time, participants were fastest when reading the UW-UF display, slower when reading the UW-RF display, and slower still when reading the RW-UF and RW-RF displays, with the latter two being equivalent. A similar pattern was revealed when using an indirect measure of demand (i.e., blink rate) and selections in a demand selection task (e.g., “Select the display that would be easiest to read aloud.”).

One potential explanation for the presently observed dissociation between these objective measures of demand and judgments of effort in the separate evaluation mode is that individuals were making prospective judgments. That is, participants made judgments predicting how effortful it would be to read a display but did not actually engage in the reading task. This removes the experience of reading, which could provide additional information from which participants can draw when making a judgment of effort. Indeed, experience plays a central role in the General Evaluability Theory in increasing evaluability (Hsee & Zhang, 2010). In addition, the common experience of ease/difficulty (e.g., fluency;

Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005; Thompson et al., 2013; Undorf & Erdfelder, 2011, 2013, 2015) might be seen as drawing individuals towards a more similar judgment context across evaluation modes (Birnbaum, 1999). To examine if the lack of such an experience was responsible for the lack of sensitivity to the different types of stimulus rotation in the separate evaluation mode, Experiments 2a and 2b involved the elicitation of a retrospective judgment of effort after participants read the display(s).

2.3 Experiments 2a and 2b

In Experiments 1a and 1b, prospective judgments of effort were requested. In Experiments 2a and 2b we used a similar paradigm, but had participants silently read actual words before making a judgment of effort associated with their reading experience. If the lack of sensitivity to the different types of stimulus rotation in the separate evaluation mode of Experiments 1a and 1b was due to the lack of experience reading the displays, then we should see a pattern more closely resembling what we find using objective measures of demand in the separate evaluation mode. The same results are expected in the joint evaluation mode.

2.3.1 Method

2.3.1.1 Participants

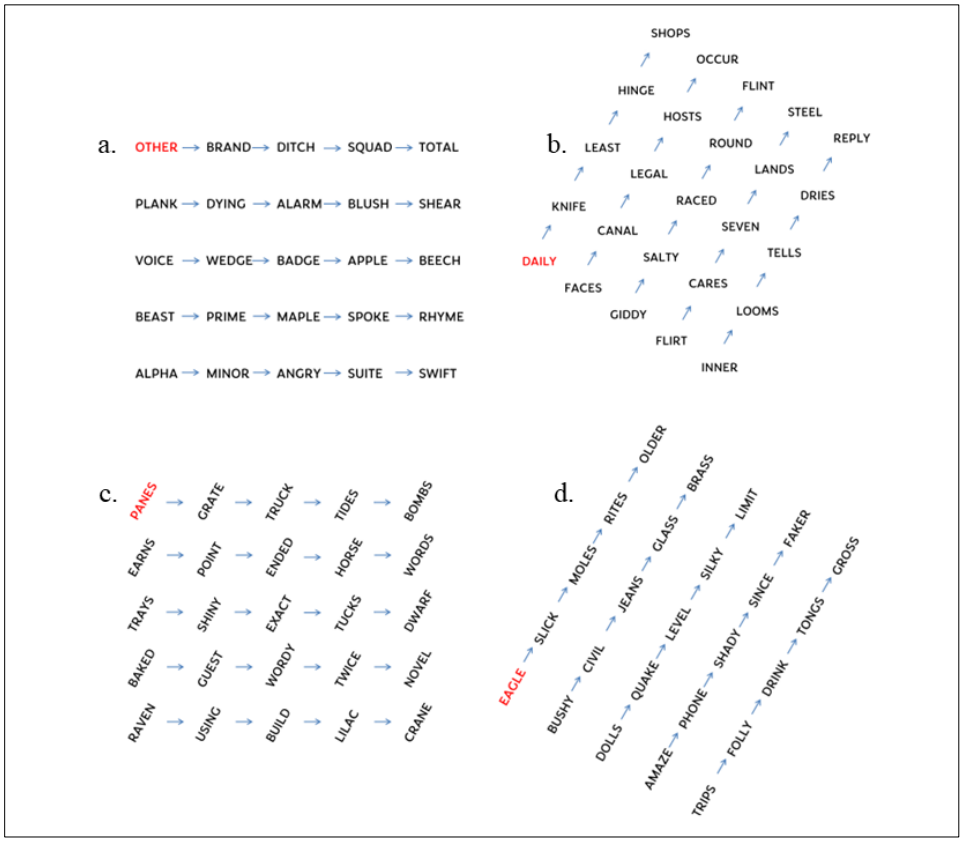
As in Experiment 1b, we aimed to recruit 720 participants from Amazon Mechanical Turk.

2.3.1.2 Stimuli and Design

The stimuli were similar to those used in Experiments 1a and 1b; however, the 25 "WORD" items were replaced with 25 different five letter words. Four such stimuli were chosen at random from the set of stimuli used in Ashburner and Risko (2021) so that participants in the joint evaluation mode would read each stimulus exactly once. In the joint evaluation mode, the four stimuli were randomly assigned to a place in the presentation order (i.e., first, second, third, or fourth display presented for reading) for each of the direction counterbalances (i.e., counterclockwise or clockwise rotation). Figure 4 is a visualization of

one of the counterclockwise counterbalances in the joint evaluation mode. In the separate evaluation mode, participants were randomly assigned to one of two direction counterbalances as well as to one of four stimulus counterbalances in addition to one of the four stimulus types. Due to the large proportion of participants excluded via the attention check question in Experiments 1a and 1b, coupled with the general consistency of results reported above with those obtained using the full dataset, changes were made to the exposition of the instructions and the attention check question to reduce the number of exclusions. Specifically, headings were added to the top of the instruction page prompting participants to read them carefully and that they would be asked about the instructions at the end of the study; and the heading on the attention check question page was revised to be in red lettering.

Figure 4. *Examples of stimuli used in Experiments 2a and 2b*



Note. a. UW-UF "Other"; b. UW-RF "Daily"; c. RW-UF "Panes"; d. RW-RF "Eagle", rotated 60° clockwise. For the joint evaluation condition, participants were shown all four displays in series. For the example here, the presentation order was: a, b, c, d. An effort scale was provided below each display such that participants provided four retrospective effort judgments in total. For the separate evaluation condition, participants were shown exactly one of these displays and provided a single effort judgment.

2.3.1.3 Procedure

The procedure was very similar to that of Experiments 1a and 1b, with the following differences: In the separate evaluation mode, the instructions stated that participants would be shown a display of words, and that their task was to silently read each word in the display (i.e., to visually read each word in the order indicated without having to vocalize) as quickly and as accurately as possible while keeping their head upright. The lack of requirement to read aloud was chosen to ensure that individuals who were not in an environment where they

could vocalize (e.g., at a public library) were still able to participate. The instructions in the joint evaluation mode were identical to those in the separate evaluation mode except that each of the four displays was presented in series, one for each page, so that participants would be able to read each display without scrolling. The judgments were then provided simultaneously as in Experiments 1a and 1b. Finally, in Experiment 2b, we included a probing question immediately after judgments of effort were provided. An exploratory analysis was conducted on these data as part of the development of a coding scheme. Specifically, participants were presented with a free form response field in which to tell the experimenter about their reason(s) for selecting the effort level or levels they selected.

2.3.2 Results

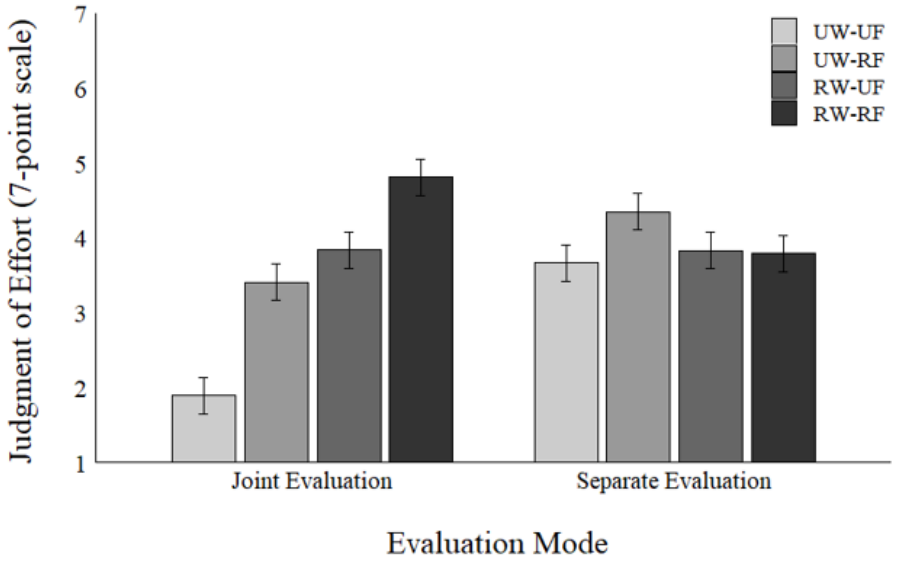
2.3.2.1 Experiment 2a

A total of 736 participants responded to one of the six surveys. Of these, 157 were excluded. All but one of these participants provided an incorrect response to the attention check question, and one participant did not finish the survey. As above, we report the results of analysis with the data from participants who passed the attention check (N = 579).

2.3.2.1.1 Erlebacher's Analysis

Data used for the 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) Erlebacher's ANOVA contained data from 525 participants (105 observations per cell). This analysis revealed a significant effect of stimulus type, $F(3, 416) = 36.77$, $p < .001$, $\eta^2 = .08$; as well as a significant effect of evaluation mode, $F(1, 416) = 7.08$, $p = .012$, $\eta^2 = .01$, such that judgments of effort were on average higher in the separate evaluation mode ($M = 3.91$, $SD = 1.80$) than in the joint evaluation mode ($M = 3.49$, $SD = 2.03$). There was a significant interaction between stimulus type and evaluation mode, $F(3, 416) = 31.89$, $p < .001$, $\eta^2 = .07$. To examine this interaction, further analyses were conducted. Figure 5 is a visualization of these data.

Figure 5. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2a*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.3.2.1.2 Judgments of Effort in the Joint Evaluation Mode

A single factor repeated measures ANOVA, with 105 observations per condition, revealed a significant effect of stimulus type, $F(2.70, 280.80) = 96.65$, $p < .001$, $\eta_g^2 = .27$. Paired comparisons indicated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(104) = 9.07$, $p < .001$, $d = 0.85$; which was judged as less effortful than the RW-UF stimulus type, $t(104) = 2.23$, $p = .028$, $d = 0.23$; which was judged as less effortful than the RW-RF stimulus type, $t(104) = 6.54$, $p < .001$, $d = 0.56$.

2.3.2.1.3 Judgments of Effort in the Separate Evaluation Mode

A single factor between-participants ANOVA, with 105 observations per condition, revealed a significant effect of stimulus type, $F(3, 416) = 3.03$, $p = .029$, $\eta_g^2 = .02$. Pairwise comparisons revealed that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(207.83) = 2.80$, $p = .006$, $d = 0.39$; moreover, the UW-RF stimulus

type was judged as more effortful than both the RW-UF and RW-RF stimulus types, $t(207.49) = 2.06$, $p = .041$, $d = 0.28$ and $t(207.31) = 2.33$, $p = .021$, $d = 0.32$, respectively. No other differences were significant.

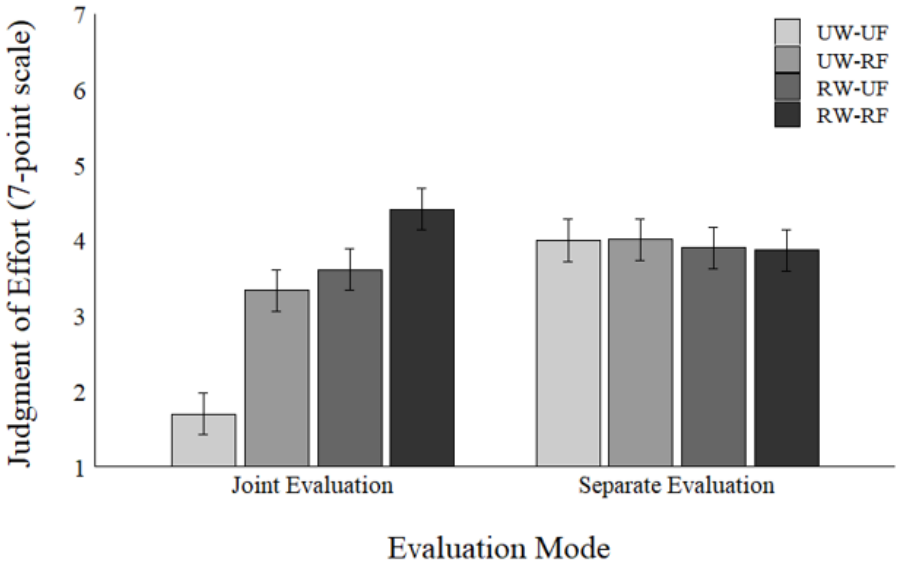
2.3.2.2 Experiment 2b

A total of 743 participants responded to one of the six surveys. Of these, 245 were excluded. All but one of these participants provided an incorrect response to the attention check question, and one participant did not finish the survey.

2.3.2.2.1 Erlebacher's Analysis

Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) ANOVA contained data from 415 participants (83 observations per cell). This analysis revealed a significant effect of stimulus type, $F(3, 328) = 20.53$, $p < .001$, $\eta^2 = .06$; as well as a significant effect of evaluation mode, $F(1, 328) = 15.67$, $p < .001$, $\eta^2 = .03$, such that judgments of effort were on average higher in the separate evaluation mode ($M = 3.95$, $SD = 1.77$) than in the joint evaluation mode ($M = 3.27$, $SD = 1.96$). There was a significant interaction between stimulus type and evaluation mode, $F(3, 328) = 24.82$, $p < .001$, $\eta^2 = .07$. To examine this interaction, further analyses were conducted. Figure 6 is a visualization of these data.

Figure 6. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2b*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.3.2.2.2 Judgments of Effort in the Joint Evaluation Mode

A single factor repeated measures ANOVA, with 83 observations per condition, revealed a significant effect of stimulus type, $F(2.77, 227.06) = 66.43$, $p < .001$, $\eta_g^2 = .25$. Paired comparisons indicated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(82) = 9.22$, $p < .001$, $d = 0.99$; however, the UW-RF and RW-UF stimulus types did not differ in judgments of effort, $t(82) = 1.33$, $p = .187$, $d = 0.16$. The RW-UF stimulus type was judged as less effortful than the RW-RF stimulus type, $t(82) = 4.17$, $p < .001$, $d = 0.46$.

2.3.2.2.3 Judgments of Effort in the Separate Evaluation Mode

A single factor between-participants ANOVA, with 105 observations per condition, indicated no difference across the stimulus types, $F(3, 328) < 1$, $\eta_g^2 < .01$. When the full dataset was analyzed, a significant effect was revealed, $F(3, 584) = 3.19$, $p = .023$, $\eta_g^2 = .02$. Pairwise

comparisons indicated that – consistent with Experiment 2a – the UW-RF stimulus type was judged as more effortful than the RW-RF stimulus type, $t(291.64) = 2.88$, $p = .004$, $d = 0.34$. No other differences were significant.

2.3.2.3 Exploratory Cross-Experiment Analysis

Due to the subtle differences in judgment patterns in the separate evaluation context across experiments, an analysis of data from Experiments 1a, 1b, 2a, and 2b was conducted. Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (elicitation timing: prospective, retrospective) between-subjects ANOVA contained data from 1,522 participants. This analysis revealed significant main effects of stimulus type, $F(3, 1502) = 5.83$, $p = .001$, $\eta_g^2 = .01$, and elicitation timing, $F(1, 1502) = 11.11$, $p = .001$, $\eta_g^2 = .01$. The interaction was also significant, $F(3, 1502) = 2.76$, $p = .041$, $\eta_g^2 = .01$. Simple effects tests within each elicitation timing were computed, revealing a significant effect of stimulus type for the prospective elicitation, $F(3, 677) = 5.83$, $p = .001$, $\eta_g^2 = .03$, and for the retrospective elicitation, $F(3, 825) = 2.66$, $p = .047$, $\eta_g^2 = .01$.

Pairwise comparisons for prospective judgments (i.e., Experiments 1a and 1b) demonstrated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(343.67) = 3.62$, $p < .001$, $d = 0.40$; the RW-UF stimulus type, $t(330.26) = 2.93$, $p = .004$, $d = 0.32$; and the RW-RF stimulus type, $t(348.83) = 3.47$, $p = .001$, $d = 0.37$. No other differences were significant, ts <1.

Pairwise comparisons for retrospective judgments (i.e., Experiments 2a and 2b) demonstrated that the UW-RF stimulus type was judged as more effortful than the UW-UF stimulus type, $t(402.41) = 2.32$, $p = .021$, $d = 0.23$; the RW-UF stimulus type, $t(385.00) = 2.20$, $p = .028$, $d = 0.22$; and the RW-RF stimulus type, $t(393.91) = 2.52$, $p = .012$, $d = 0.25$. No other differences were significant, ts < 1.

2.3.3 Discussion

When contrasting Experiments 1a and 1b, where individuals made prospective judgments, to Experiments 2a and 2b, where individuals made retrospective judgments, it seems clear the

same general pattern emerges. That is, when moving from joint to separate evaluation, the distinctions between stimulus types seem to largely disappear. At a finer grain, there appeared to be some subtle differences in the separate evaluation mode across the prospective and retrospective experiments. Specifically, the difference between the upright stimulus and the rotated stimuli for prospective judgments appeared absent for retrospective judgments.

As in Experiments 1a and 1b, evaluation mode greatly influenced judgments of effort, as supported by a significant interaction between stimulus type and evaluation mode. Follow-up analyses revealed the nature of this interaction. Specifically, in the joint evaluation mode, judgments of effort follow a clear pattern wherein the UW-UF stimulus type is judged as least effortful, followed by UW-RF, RW-UF, and then RW-RF (although in Experiment 2b the increase in effort from the UW-RF to the RW-UF stimulus type was not significant). Judgments of effort in the separate evaluation mode did not follow this pattern, as there was no difference in effort judgments across the stimulus types in the separate evaluation mode save for the UW-RF stimulus type, which was judged as more effortful than the other three in Experiment 2a. Note that this pattern largely failed to replicate in Experiment 2b.

Taken together, retrospective effort judgments are influenced by evaluation mode in much the same way as prospective effort judgments. The prediction that experience reading the displays would bring effort judgments more in line with what is typically found with measures of objective demand was clearly not supported. While the goal of the latter experiments was to compare judgments of effort made with no direct experience to those made after having experienced reading the displays, we also altered whether the judged act of reading was in the future or in the past. That is, individuals in Experiments 1a and 1b made a prospective effort judgment, whereas in Experiments 2a and 2b the judgment was directed at their recent experience. This confound need not be the case. Individuals could be provided experience reading the displays, followed by a request to make a prospective judgment of effort. In Experiment 3, we conduct just such an experiment.

2.4 Experiment 3

In Experiment 3, participants were first provided with the same experience reading a display as in Experiments 2a and 2b, after which they were asked to provide a *prospective* judgment of effort corresponding to reading another display of the same stimulus type, similar to Experiments 1a and 1b. In addition, we included the prospective (i.e., prospective judgment with no experience) conditions from Experiments 1a and 1b to allow a direct comparison across these two judgment types.

2.4.1 Method

2.4.1.1 Participants, Stimuli, Design, and Procedure

Participants were recruited from Amazon Mechanical Turk. In an attempt to minimize loss of data via incorrect responses to the attention check, workers with at least 1000 HITs experience and at least a 99% approval rating were eligible to participate. We aimed to recruit approximately 100 participants per cell (a total of 5(100) x 2 = 1000 total), to analyze data from 90 participants per cell (i.e., 900 total). Two relatively easy fraud detection questions were included at the very beginning of the survey, and if either of these questions was not answered correctly, access to the survey was not provided.

In the *experience condition*, participants silently read the same displays as in Experiments 2a and 2b (i.e., comprising 25 random five letter words), then made prospective judgments (i.e., for silently reading another display of the same stimulus type) while viewing the displays used in Experiments 1a and 1b (i.e., "WORD" occupying each of the 25 positions). Participants assigned to the joint evaluation mode read four different displays, one for each stimulus type, counterbalanced for order of presentation and display type (Figure 4). After reading these four displays, participants were presented with a page containing all four stimulus types, with each stimulus occupying one quadrant of the display and labelled a. through d. Four effort scales corresponding to each of the four stimulus types (i.e., labelled a. through d.) were provided below the display. In the *no experience condition*, the stimuli and

procedure were identical to that of Experiments 1a and 1b, except that it was not specified that participants imagine reading “aloud”.

Minor changes were made to the exposition of the instructions, namely, instructions in all conditions were formatted as a bulleted list – rather than prose as in previous experiments – to increase readability. Moreover, an additional request not to memorize the words was added to instructions in the experience condition. This was done in response to some participants reporting they were attempting to memorize the words as they completed the reading task in Experiment 2b in anticipation of a test.

As with Experiment 2b, immediately after judgments of effort were reported, participants were presented with a free form response field in which to tell the experimenter about their reason(s) for selecting the effort level they selected. An exploratory analysis was conducted on these data as part of the development of a coding scheme. Finally, a change was made to the attention check question, to avoid possible misinterpretation. Due to the addition of the experience condition, 12 surveys were employed to collect data for this experiment (compared to the six surveys employed in previous experiments).

2.4.2 Results

A total of 1017 participants responded to one of the four surveys. Of these, 97 were excluded. All but two of these participants provided an incorrect response to the attention check question, and two participants did not finish the survey. The results of analysis with the data from participants who passed the attention check (N = 920) are reported below.

2.4.2.1 Erlebacher’s Analysis

To test the interaction between stimulus type and evaluation mode, Erlebacher’s analyses were conducted, first on the data from the experience condition, and second on the data from the no experience condition. To support this preregistered analysis, a 3-way Erlebacher’s ANOVA was computed using code based on Erlebacher (1978) Case 1 partitioning of variance. This code was written by the authors and included in the analysis script within the OSF project pages. Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF)

x 2 (experience condition: no experience, experience) x 2 (evaluation mode: joint, separate) ANOVA contained data from 850 participants (85 observations per cell). This analysis revealed a marginally significant three-way interaction, $F(3, 672) = 2.57$, $p = .053$, $\eta^2 = .01$. Though only marginal, we felt it prudent to explore this interaction to shed some light on the distinction between the experience and no experience conditions. The simple effects tests for this purpose are below.

2.4.2.2 Erlebacher's Analysis – Experience Condition

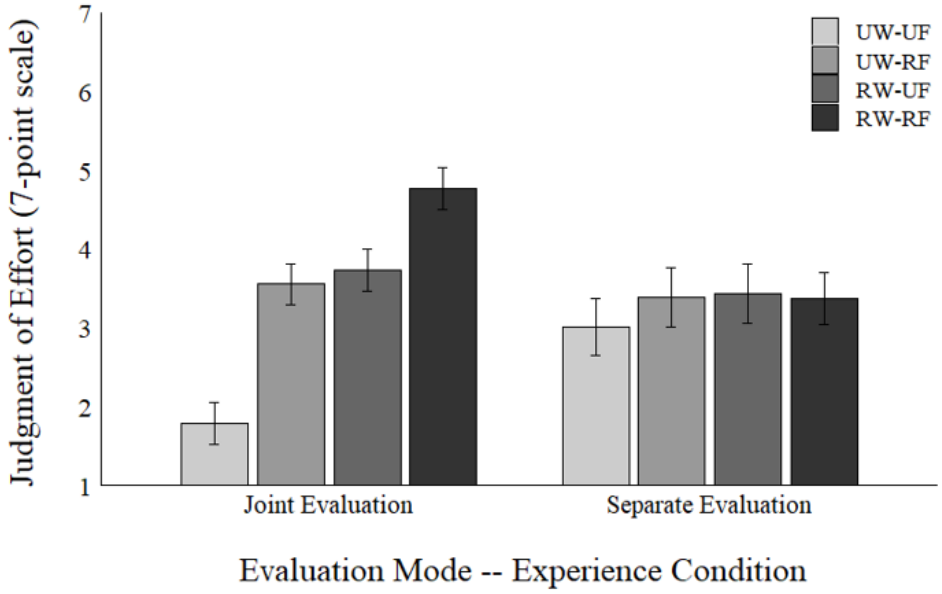
This 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) Erlebacher's ANOVA revealed a significant main effect of stimulus type, $F(3, 336) = 38.79$, $p < .001$, $\eta^2 = .11$; as well as a significant interaction, $F(3, 336) = 22.45$, $p < .001$, $\eta^2 = .07$. The main effect of design was not significant, $F(1, 336) < 1$.

2.4.2.2.1 Simple Effects Analysis – Experience Condition

A single factor within-participants ANOVA with stimulus type as the factor revealed a significant effect of stimulus type on judgments of effort, $F(3, 252) = 84.28$, $p < .001$, $\eta_g^2 = .30$. By contrast, a single-factor between-participants ANOVA with stimulus type as the factor revealed a non-significant effect of stimulus type on judgments of effort, $F(3, 336) = 1.16$, $p = .324$, $\eta_g^2 = .01$. Figure 7 is a visualization of these data.

Paired comparisons conducted on the joint evaluation mode data indicate that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(84) = 9.33$, $p < .001$, $d = 1.07$; which did not differ in effort judgment with the RW-UF stimulus type, $t(84) < 1$, $d = 0.10$. The RW-UF stimulus type was judged as less effortful than the RW-RF stimulus type, $t(84) = 5.86$, $p < .001$, $d = 0.63$.

Figure 7. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the experience condition*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.4.2.3 Erlebacher's Analysis – No Experience Condition

Mirroring the analysis with the experience data, the 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) Erlebacher's ANOVA on the no experience data revealed a significant main effect of stimulus type, $F(3, 336) = 48.62$, $p < .001$, $\eta^2 = .15$; as well as a significant interaction, $F(3, 336) = 20.56$, $p < .001$, $\eta^2 = .06$. The main effect of design was not significant, $F(1, 336) < 1$. The analyses below investigate the simple effect of stimulus type in each evaluation mode for the experience and the no experience data separately.

2.4.2.3.1 Simple Effects Analysis – No Experience Condition

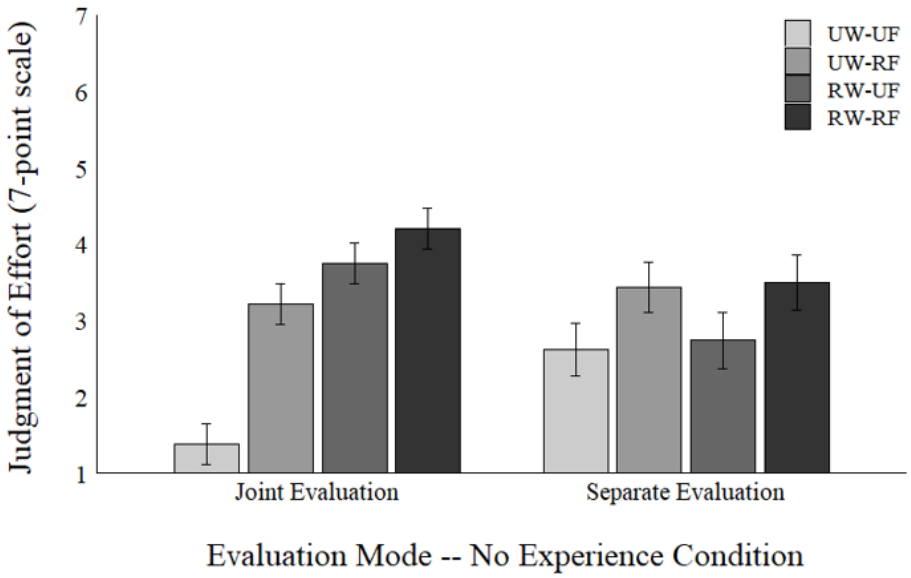
A single factor within-participants ANOVA with stimulus type as the factor revealed a significant effect of stimulus type on judgments of effort, $F(3, 252) = 84.42$, $p < .001$, $\eta_g^2 = .35$. A single-factor between-participants ANOVA with stimulus type as the factor also

revealed a significant effect of stimulus type, $F(3, 336) = 6.42$, $p < .001$, $\eta_g^2 = .05$. Figure 8 is a visualization of these data.

Paired comparisons conducted on the joint evaluation mode data indicate that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(84) = 10.51$, $p < .001$, $d = 1.35$; which was judged as less effortful than the RW-UF stimulus type, $t(84) = 2.46$, $p = .016$, $d = 0.34$; which was judged as less effortful than the RW-RF stimulus type, $t(84) = 2.51$, $p = .014$, $d = 0.45$.

Pairwise comparisons on the separate evaluation mode data indicate that the UW-UF stimulus type was judged as less effortful than both the UW-RF and the RW-RF stimulus types, $t(167.84) = 3.32$, $p = .001$, $d = 0.51$ and $t(167.51) = 3.41$, $p = .001$, $d = 0.52$, respectively. Moreover, the RW-UF stimulus type was also judged as less effortful than both the UW-RF and the RW-RF stimulus types, $t(166.62) = 2.75$, $p = .007$ and $t(167.99) = 2.86$, $p = .005$, $d = 0.44$, respectively. Neither the UW-UF and RW-UF, nor the UW-RF and RW-RF stimulus types were judged as differentially effortful, ts < 1.

Figure 8. *Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the no experience condition*



Note. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.4.3 Discussion

The results from Experiment 3 demonstrate that, regardless of whether experience precedes the prospective judgment of effort, evaluation context drastically influences the judgment pattern. Specifically, there was a large difference between the UW-UF stimulus and the rotated stimuli in the joint evaluation mode that was either lessened or absent in the separate evaluation mode. The results of Experiment 3 also suggest that having experience with the task before making a prospective judgment in the separate evaluation mode seems to attenuate any subtle differences between the upright and rotated stimuli. In other words, the retrospective judgments of effort elicited in Experiments 2a and 2b have a very similar pattern to the prospective judgments of effort following experience in Experiment 3. The results obtained from this experiment, designed to decouple experience and judgment timing (i.e., prospective vs. retrospective), seems to support the idea that the lack of difference

between the upright and disoriented stimuli in the separate evaluation context in Experiments 2a and 2b (as compared to Experiments 1a and 1b) is related to the experience itself. In general, this is opposite to the effect of experience we had predicted.

Turning to the subtle differences across stimulus types with prospective judgments in the separate evaluation mode, it seems that eliciting prospective judgments – with or without prior experience – may preserve some cue(s) that distinguishes the stimulus types. Of these subtle differences, the most replicable appears to be the relatively low judgment of the UW-UF stimulus type as compared to the three rotated stimuli. One possible explanation for this may be that, without any experience reading the UW-UF stimulus, individuals may find salient the comparison between it and typical upright text, resulting in a prospective judgment of relatively low effort (as compared to the analogous retrospective judgment). The experience of reading a UW-UF stimulus as quickly and as accurately as possible, however, may cue individuals to the various distinctions between the UW-UF stimulus and reading typical upright text (e.g., reading random words quickly and accurately is experienced differently, perhaps with less ease, than normal reading).

Furthermore, it was observed that, when no experience was provided, participants in the RW-UF condition reported a lower effort judgment than those in the UW-RF and RW-RF conditions. Notably, this was observed only when the displays were rotated counterclockwise. No such difference was observed when displays were rotated clockwise, suggesting that the relatively low effort judgment for the RW-UF stimulus type in the no experience condition may be driven by a sensitivity to the direction of rotation. However, an analysis conducted with the full dataset demonstrated the same pattern as that pictured in Figure 8 (i.e., no experience, separate evaluation condition), yet the effect of direction of rotation was absent, leading us to conclude that direction of rotation may not reliably inform participants' effort judgments when evaluated separately. Also worth noting is that the UW-RF stimulus is judged as relatively more effortful in Experiment 2a, an effect that came through in the cross-experimental analysis for retrospective judgments. This was also the case in the no experience condition, excepting the RW-RF stimulus type, in Experiment 3. It may be that the UW-RF stimulus is arranged in a manner that causes some difficulty

determining reading direction from the starting word, which can influence judgments in the separate evaluation mode, but seems to do so sporadically.

All that said, future research may be warranted to further investigate these subtle differences between prospective and retrospective judgments of effort in separate evaluation mode. In the present investigation, we continue to focus on the robust effect of evaluation mode on judgments of effort by introducing measures of objective demand and attempting to probe deeper into why effort judgments are changing so profoundly across evaluation contexts.

2.5 Experiment 4

Experiments 1 through 3 demonstrate that judgments of effort are strongly influenced by the evaluation context, largely independent of whether judgments are collected prospectively, prospectively after experiencing the task, or retrospectively. The form of the dissociation between demand measures and judgments of effort across stimulus types reported in previous research (Ashburner & Risko, 2021; Dunn & Risko, 2016) therefore seems to depend largely on the evaluation context in which judgments are made. That is, neither judgments of effort collected in a joint evaluation context nor judgments of effort collected in a separate evaluation context match the pattern revealed in demand measures (i.e., reading time, error count) and neither do they match each other. That said, one potentially important consideration is that, in previous investigations, demand measures have been collected in a *mixed* stimulus context (i.e., wherein participants experience all stimulus types). Thus, it remains possible that, if demand measures were collected in a *pure* stimulus context (i.e., wherein participants experience a single stimulus type), the pattern would reflect that found in Experiments 1-3, when judgments of effort are collected in a separate evaluation context (wherein participants evaluate a single stimulus type). For example, the effect of stimulus type on reading time might be attenuated in a pure stimulus context due to the lack of unpredictability, as compared to when stimuli are mixed (e.g., Ilan & Miller, 1994). In Experiment 4 we examine this possibility by collecting reading time and error data in addition to effort judgments from participants in either a mixed or pure stimulus design.

In Experiment 4 we also start to examine one potential explanation for the pronounced effect of evaluation context on judgments of effort. Research reviewed in the Introduction (e.g., Hsee & Zhang, 2004, 2010) often explained dissociations between joint and separate evaluation contexts by recourse to the general idea that the dimensions contributing to the judgment (e.g., the torn book cover vs. the number of words) in question shifted across evaluation modes. Thus, from the perspective that judgments of effort (see Dunn & Risko, 2016; Dunn et al., 2019) reflect an inferential process based on available cues, one means of understanding the results presented here is to consider the reading task as containing a number of cues on which a participant might base their effort judgment, and the evaluation context influencing which cues contribute (or, how the contribution of the cues are weighted). Under this interpretation, the evaluation context affected judgments of effort because the cue or collection of cues that were most salient to individuals changed across evaluation mode. One means of exploring this idea is to examine the reasons individuals provide for their effort judgments. Thus, in addition to the collection of judgments of effort and demand data, participants in Experiment 4 were also asked to provide the reason for their effort judgment(s) using two methods of elicitation, a free-form response field (i.e., the same one from Experiments 2b and 3), as well as a multiple-choice question based on the coding scheme developed.

As noted above, the effect of evaluation condition on judgments of effort might be due to the evaluation context making salient certain features of the stimuli on which to base one's judgments. This idea leads to the general prediction that participants' reasons for selecting a particular effort judgment should vary across evaluation conditions. For example, judgments in the joint evaluation condition may have a greater likelihood of being the output of a process that is grounded in the structure of the displays, compared to the separate evaluation condition. A potential corollary of this prediction is that, overall, participants should endorse fewer reasons in the joint evaluation condition, compared to the separate evaluation. This could reflect a kind of focusing effect, whereby participants in the joint evaluation condition have a high degree of consensus on which features inform their judgments. Whereas, in the separate evaluation context, this would be less likely to be the

case. Lastly, collecting participants' reasons for judgments also provides an opportunity to learn more about why judgments in the separate evaluation context appear uninfluenced by stimulus type. For example, if individuals in the separate evaluation condition tend to report reasons for which the stimulus types do not differ (e.g., if they were to base their effort judgement on qualities of the words in the display which do not vary across the stimulus types), then this would provide one potential explanation for the observations in Experiments 1-3. In a similar vein, if individuals across the stimulus types tend to select different reasons, this might provide another means of explaining the lack of a systematic effect of stimulus type in the single evaluation condition.

We can therefore ask a number of questions of the reasons individuals provide: (1) Is there a difference in the reasons provided across evaluation conditions (e.g., do more responses primarily address the stimulus type manipulation in the joint evaluation condition)? (2) Do the reasons in the separate evaluation condition exhibit more variation than in the joint evaluation condition? (3) Is there a difference in the proportion of reasons provided across the stimulus types for participants in the separate evaluation condition (e.g., do more responses primarily address the nature of the words in the UW-UF condition than the RW-RF condition)? Answers to these questions will provide a novel perspective on how evaluation context might influence judgments of effort.

2.5.1 Method

2.5.1.1 Participants

Participants located within the US, Canada, or the UK were recruited from the Prolific platform. Specifically, workers with at least 100 studies completed and at least a 99% approval rating, were eligible for participation. We aimed to analyze data from 730 participants to detect a small pairwise effect (i.e., Cohen's $d = 0.33$) across the stimulus types in the separate evaluation mode.

2.5.1.2 Apparatus

Two survey programming tools were employed: Qualtrics (as in previous Experiments) and Phonic AI, a service that allows secure collection of audio data from online participants. Code developed by Phonic was embedded into the Qualtrics survey, enabling the participants' voice while reading to be captured and stored.

2.5.1.3 Stimuli and Design

Sixteen 25-word displays were used for test trials, with four additional displays reserved for practice trials. Included in this stimulus set were the four displays used in Experiments 2a and 2b, with the other displays similarly constructed. The design was similar to that of Experiment 3, with individuals gaining experience with the reading task before making a prospective judgment. Therefore, the displays composed of "WORD" from Experiments 1a, 1b, and 3 were employed as a reference on the page when judgments were elicited. The main differences here are that participants performed multiple trials of the reading task as they were recorded reading out loud (note that participants were instructed to read silently to themselves in our previous experiments, as it was thought that compliance with this instruction would be more widespread compared to instructing individuals to read out loud when not being watched by an experimenter).

2.5.1.4 Procedure

After providing correct responses to the pre-survey questions and providing consent, participants were presented with an instructional video that detailed how to audio record themselves while reading the displays. All other instructions were presented as per Experiment 3 (i.e., read as quickly as possible; do not memorize the words; etc.). A short quiz was then taken to ensure that the main instructions had been understood, wherein any incorrect response was followed by a statement of the correct one. Following this, four practice trials were presented[3] in one of four pre-determined orders. On a given trial, a participant would start on a *pause screen* (i.e., a screen on which they were not being

[3] The data from these trials were not analyzed.

recorded, containing reminders such as “Remember to read as quickly as possible.”, and occasionally a note about how far along they were in the experiment at that point). When the participant was ready, they clicked a button to advance to the display screen and the audio capture started automatically. Participants were instructed to complete the trial, then scroll down until they could click the ‘stop recording’ button. To submit their recording and leave the page, participants had to click the ‘submit’ button, resulting in them being presented with another pause screen. The experimental trials comprised 16 displays, one per page. For participants in the separate evaluation mode, all 16 displays were of the same stimulus type; for participants in the joint evaluation mode, there were 4 of each type. The direction of rotation (i.e., counterclockwise vs. clockwise) was manipulated between-participants, to maintain consistency with the previous Experiments. As with the practice trials, participants were randomly assigned to 4 possible orders of the 16 displays. Ideally, a newly generated randomization for each participant would have been employed; however, due to the limitations of the programming software, it was not possible to track the order of the displays for coding and analysis purposes.

Following the 16 experimental trials, the procedure followed that of Experiment 3, wherein a display comprising the word “WORD” was provided in the same orientation of those the participant just read (four displays for participants in the joint evaluation mode), and a prospective judgment of effort was elicited (four judgments for participants in the joint evaluation mode). Participants then wrote about the reason(s) for their judgment(s) in the free-form response field. Following this, participants were asked to select the statement that best represented their reason from a list of nine, including “Something specific about the characteristics of the words?” and “Something specific about the way the words were organized on the screen?” These statements directly corresponded to the coding scheme we applied to the free-form responses in this Experiment. The remainder of the procedure was as per Experiment 3.

2.5.2 Results

A total of 1230 participants' responses were collected via the Prolific platform. One hundred and fifty-two participants' data were removed due to either: the participant attempted or completed the study more than once (i.e., duplicate Prolific IDs were detected in the dataset); the participant reported a technical problem that necessitated a restart of the survey; the response was flagged as having timed-out (i.e., it took longer than 63 minutes for submission of the approximately 20 minute survey); a participant used a mobile phone despite instructions to only use a desktop computer; the attention check question was answered incorrectly[4]; or the response was incomplete.

Upon pre-analysis a further 197 participants' data were removed using the threshold of greater than 2 of the 16 experimental trials having been deemed unusable. This occurred because either: trial recordings were of poor enough quality that behavioural data could not be reliably extracted from the recording; the participant ended their audio recording before starting the last word in the display; the participant was interacting with another voice while recording or was otherwise distracted due to background noise; or the participant was coughing, commenting, or sipping a beverage during the trial. The remaining 881 participants' data[5] (700 and 181 in the separate and joint evaluation condition, respectively) were then subjected to outlier removals prior to the main analysis.

2.5.2.1 Outlier Detection and Removal

Reading times were coded from the onset of recording until the onset of the final word in the display. Outlier reading times were defined as having an absolute *z*-score greater than 2.5 and were detected via a within-participant search. For the joint evaluation mode data, the additional parameter of within-condition (i.e., stimulus type) was applied to the search. This

[4] As the proportion of individuals who provided an incorrect answer to the attention check question was minimal (26 participants in total), data including such responses was not analyzed.

[5] Due to experimenter error, a total of 12 respondents were able to complete two surveys within this study. That is, there are 12 observations that should have been flagged as duplicates and removed; however, by the time this error was caught, the data were completely de-identified. Due to the large sample size, the inclusion of these 12 observations is unlikely to affect the results reported here.

detection process resulted in the removal of 121 trials from the separate evaluation data and 23 trials from the joint evaluation data. The introduction of greater than two unusable trials to three participants' data resulted in their exclusion from the separate evaluation data. Thus, after reading time outlier removal, there were 697 and 181 participants' data in the separate and joint evaluation condition, respectively. Upon pre-analysis, relatively slow reading times (in seconds) provided by a participant in the separate evaluation condition was detected (M = 41.37, SD = 2.71). The results did not qualitatively differ based on whether this participant's data were included. Therefore, the results reported below include these data.

Error counts were coded as an addition of 1 to the count for any mistake made while reading the display. Specifically, repeating or skipping a word; starting the word more than once (each attempt counting as 1 error); or another type of mispronunciation each counted as an error. In the case of a premature press of the stop recording button, the cut-off final word was represented as an additional error. A cursory analysis demonstrated non-significant effects of stimulus type, evaluation mode, and the interaction on error counts (as per previous research; Ashburner & Risko, 2021; Dunn & Risko, 2016); as such, error counts were only coded for a subset (i.e., 460 and 101 participants in the separate and joint evaluation condition, respectively) of the reading time data. The same procedure as described above was employed to detect and remove error count outliers, after which there were 349 and 79 participants' data in the separate and joint evaluation condition, respectively.

2.5.2.2 Erlebacher's Analysis – Reading Times

Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) ANOVA, with reading time as the outcome variable, contained data from 805 participants (161 observations per cell). This analysis revealed a significant effect of stimulus type, $F(3, 640) = 12.57$, $p < .001$, $\eta^2 = .02$. Both the effect of evaluation mode and the interaction were not significant, F's < 1. Table 2 comprises descriptive statistics of reading times, error counts, and judgments of effort; and Figure 9 is a visualization of these data. To examine this main effect, pairwise comparisons were conducted collapsing across evaluation mode; specifically, independent t-tests were computed. As these comparisons

were not pre-registered, they should be considered exploratory and are presented here for the purpose of demonstrating that our current findings align with those from previous research.

The UW-UF stimulus type was read marginally faster than the UW-RF, $t(641.66) = 1.76$, $p = .079$, $d = 0.14$; and was read significantly faster than both the RW-UF and RW-RF stimulus types, $t(624.78) = 3.83$, $p < .001$, $d = 0.30$ and $t(640.89) = 4.25$, $p < .001$, $d = 0.34$, respectively. The UW-RF stimulus type was read significantly faster than both the RW-UF and RW-RF stimulus types, $t(620.03) = 2.26$, $p = .024$, $d = 0.18$ and $t(639.34) = 2.58$, $p = .010$, $d = 0.20$, respectively. Critically, the difference between reading times for the RW-UF and the RW-RF stimulus types were not significant, $t(632.09) < 1$, $p = .889$, $d = 0.01$.

2.5.2.3 Erlebacher's Analysis – Error Counts

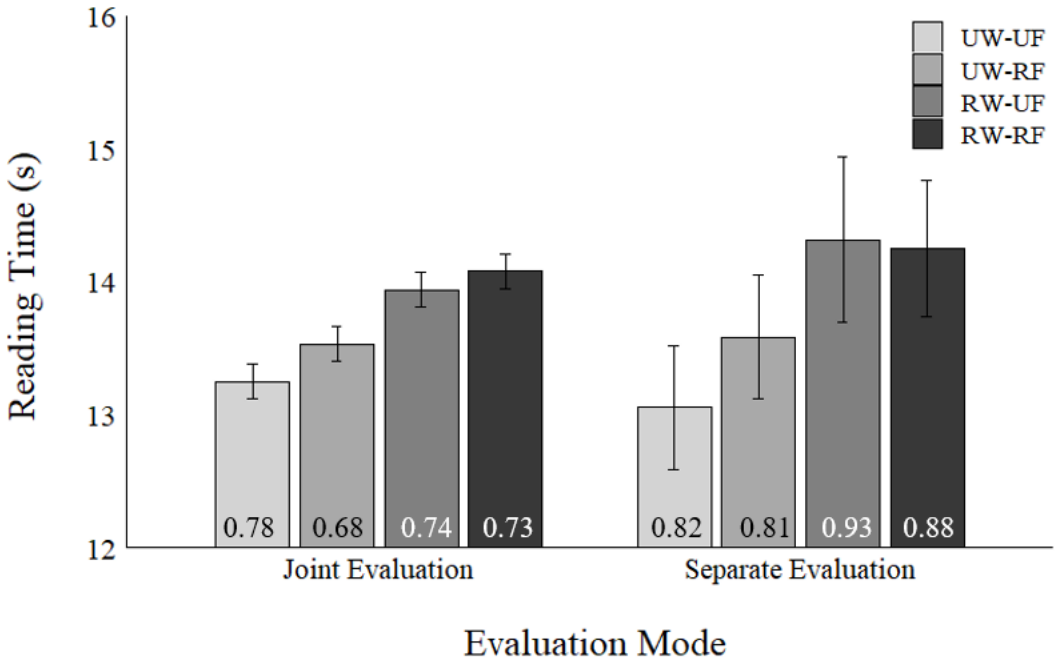
Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) ANOVA, with error count as the outcome variable, contained data from 305 participants (79 observations per cell). This analysis revealed that the effect of stimulus type, evaluation mode, and the interaction were non-significant, all Fs < 1.

Table 2. *Mean response times, error counts, and effort judgments by stimulus type and evaluation mode for Experiment 4*

	Reading Time	*Error Count*	*Effort Judgment*
Joint Evaluation			
UW-UF	13.21 (2.84)	0.78 (0.69)	1.53 (0.86)
UW-RF	13.46 (2.75)	0.68 (0.63)	3.07 (1.57)
RW-UF	13.89 (2.79)	0.74 (0.72)	3.75 (1.60)
RW-RF	14.06 (2.78)	0.73 (0.70)	4.46 (1.70)
Separate Evaluation			
UW-UF	13.02 (2.95)	0.82 (0.70)	3.04 (1.52)
UW-RF	13.59 (3.00)	0.81 (0.72)	3.12 (1.54)
RW-UF	14.31 (3.97)	0.93 (1.21)	3.04 (1.63)
RW-RF	14.20 (3.25)	0.88 (0.59)	3.42 (1.60)

Note. Standard deviations in parentheses. Response times in seconds. Error counts measure raw number of errors.

Figure 9. *Mean reading times (in seconds) as a function of stimulus type grouped by evaluation mode*

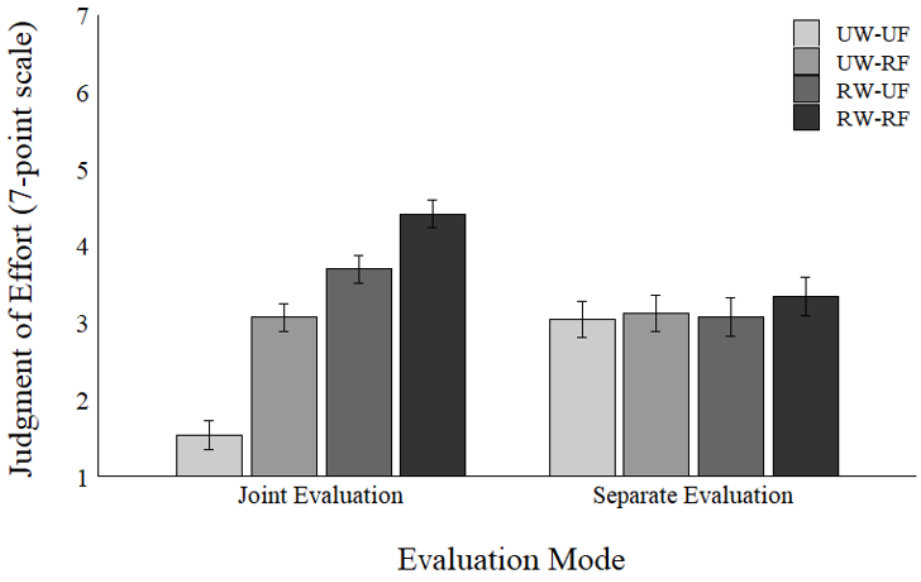


Note. Error bars represent 95% CIs (using DF_{error} from the Erlebacher's analysis) and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively. Error counts displayed at bottom of bars.

2.5.2.4 Erlebacher's Analysis – Judgments of Effort

Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) x 2 (evaluation mode: joint, separate) ANOVA, with judgment of effort as the outcome variable, contained data from the same 805 participants as above. This analysis revealed a significant effect of stimulus type, $F(3, 640) = 73.10$, $p < .001$, $\eta^2 = .11$, as well as a significant interaction, $F(3, 640) = 52.46$, $p < .001$, $\eta^2 = .08$. The effect of evaluation mode was not significant, $F(1, 640) < 1$. Follow-up tests were conducted to determine the nature of the interaction, Figure 10 is a visualization of these data.

Figure 10. *Mean effort judgments as a function of stimulus type grouped by evaluation mode*



Note. Error bars represent 95% CIs (using DF_{error} from the Erlebacher's analysis) and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

2.5.2.4.1 Judgments of Effort in the Joint Evaluation Mode

A single factor repeated measures ANOVA revealed a significant effect of stimulus type, $F(2.72, 434.83) = 174.04$, $p < .001$, $\eta_g^2 = .34$. Paired comparisons indicated that the UW-UF stimulus type was judged as less effortful than the UW-RF stimulus type, $t(160) = 12.31$, $p < .001$, $d = 1.18$; which was judged as less effortful than the RW-UF stimulus type, $t(160) = 4.49$, $p < .001$, $d = 0.42$; which was judged as less effortful than the RW-RF stimulus type, $t(160) = 6.28$, $p < .001$, $d = 0.42$.

2.5.2.4.2 Judgments of Effort in the Separate Evaluation Mode

A single factor between-participants ANOVA indicated no difference across the stimulus types, $F(3, 640) = 1.16$, $p = .325$, $\eta_g^2 = .01$.

2.5.2.5 Open Field Responses and Participant Self-Coding

A coding scheme was developed using the reasons participants provided for their effort judgments in Experiments 2b and 3. Specifically, participants' responses were first designated as specific to the task (e.g., "*the words were simple*") or not (e.g., "*it just felt easy*"). Those responses coded as specific were then assigned features based on whether their response referenced (a) the structure of the displays (i.e., how the words were arranged on the screen; key terms included *left-to-right*; *diagonal*; *tilted*), (b) the words themselves, (i.e., physical features such as font and colour, and non-physical features such as word frequency or randomness), (c) the number of words in the display, (d) keeping their head upright while reading, (e) performance requirements (i.e., having to read quickly and/or accurately), (f) being able to read coming into the study (e.g., "*I am fluent in English, so it wasn't too hard*"), or (g) a reason that did not fit into the other categories or was confusing to the coder. Two independent coders provided these feature assignments to each response. Coders could choose more than one category but listed a primary category that served as the best feature associated with each response. The analysis reported below involves only the primary assigned categories. Furthermore, during the coding process, both coders reported some responses that clearly indicated the participant misunderstood the judgment portion of the task. In this case, a code was assigned to the response to indicate this, which will be discussed further below.

In addition to coding the participants' responses, we asked them to "self-code" their response by choosing just one of the developed features discussed above. In addition to these seven features, participants were also presented with the option of "Nothing specific about the task," as well as "Something specific about reading out loud.[6]" In sum, we collected coded data provided by two independent human coders alongside multiple-choice answers

[6] This option was accidentally left in as an option in the self-coding question for Experiment 4. It was originally intended for comparison across the previous experiments, wherein participants were instructed to silently read, with Experiment 4 wherein participants had to read out loud. As this option was chosen with very low frequency, it seems sensible to conclude the responses would be distributed across the remaining frequencies similarly should this option had been removed.

provided by participants themselves, using the dataset comprising of 878 participants (697 and 181 in the separate and joint evaluation condition, respectively) before the removal of data for the Erlebacher's analyses.

Responses indicating a misunderstanding of the judgment task were removed from both coders' datasets. There were 9.0% and 10.3% such responses in the first and second coder's data, respectfully. Coder agreement was computed omitting such responses in either coder's data. There was an 89.7% agreement on task specific vs. not task-specific coding and, from those responses coded as task specific by both coders, a 77.4% agreement on the primary feature selected. Results obtained from one of the two coders' data (N = 799) are reported below, and no qualitative differences were observed across coders' data. To allow comparison across the *open response* (i.e., coded from participants' written responses) data and the participants' *selected response* (i.e., their selection from the multiple-choice question) data, as well as to ensure high quality of data to be analyzed, those responses that were flagged as a misunderstanding of the judgment task by both coders were also removed from the selected response data[7].

2.5.2.5.1 Features Across Evaluation Condition

The frequency of responses coded as specific to the task (as compared to those that were not) was significantly higher in the joint evaluation mode, as compared to the separate evaluation mode in the open response data, $\chi^2(1, n = 799) = 15.32$, $p < .001$. Interestingly, this was not the case for the selected response data as, in both the joint and the separate evaluation mode, about 95% of participants chose a specific reason, $\chi^2(1, n = 805) = 0.32$, $p = .574$. Table 3 comprises frequencies of each feature discussed here, split by evaluation condition; and Figures 11 and 12 are visualizations of these data.

[7] The results did not qualitatively differ when including these responses in the self response data.

Table 3. *Percent of total responses in favour of each feature by evaluation context*

	Joint		*Separate*	
	Open	*Selected*	*Open*	*Selected*
Specific	92.2	96.1	79.2	94.7
Feature				
Structure	77.1	54.3	25.9	21.8
Words	2.4	4.0	30.8	12.8
Head Upright	11.4	24.3	9.2	8.4
Performance	3.6	9.8	14.9	26.5
Reading	-	4.0	-	18.9

Note. All other features comprised less than 10% of the total counts, so per the preregistration these features were not included in the analysis.

The frequency of responses coded as relating to the structure of the display (as compared to all other task-specific responses) was significantly higher in the joint evaluation mode in the open response data, $\chi^2(1, n = 656) = 134.59$, $p < .001$. Consistent with this, significantly more participants in the joint evaluation mode chose this reason in the selected response data, as compared to those in the separate evaluation mode, $\chi^2(1, n = 765) = 67.09$, $p < .001$. Turning to responses categorized as relating to the words themselves, a significantly lower frequency was detected in the joint evaluation mode in the open response data, $\chi^2(1, n = 656) = 53.88$, $p < .001$, a pattern that held in the selected response data, $\chi^2(1, n = 765) = 9.81$, $p = .002$.

Figure 11. *Percent of total responses in favour of each feature for the open response data, split by evaluation condition*

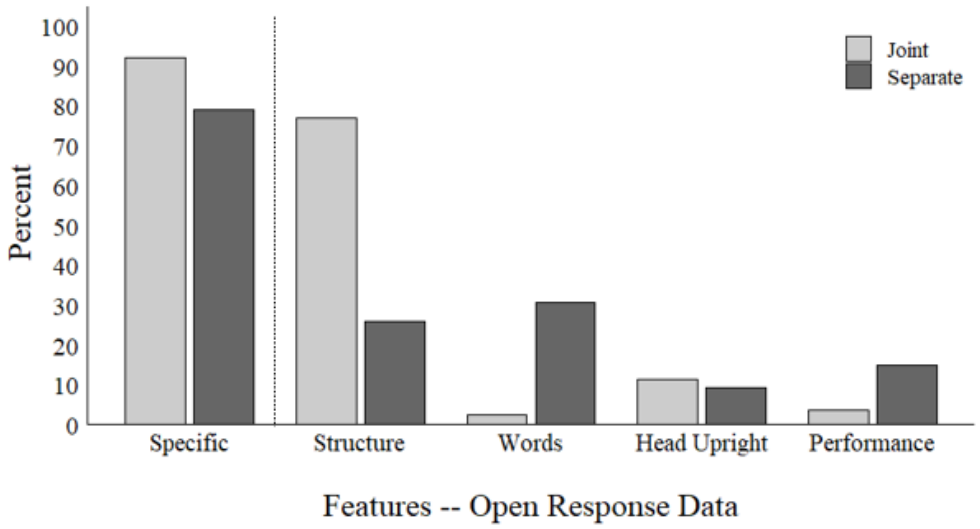
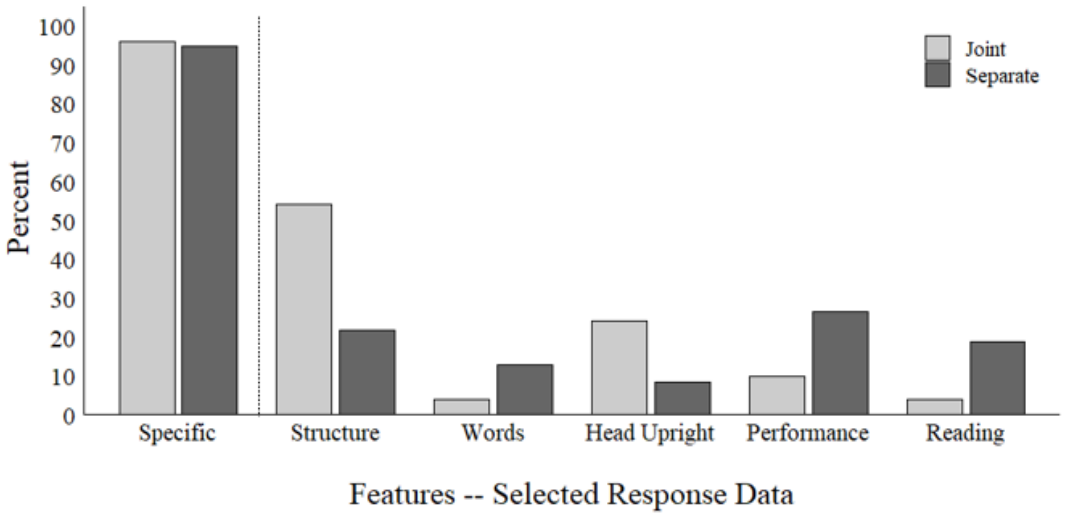


Figure 12. *Percent of total responses in favour of each feature for the selected response data, split by evaluation condition*



The frequency of responses coded as primarily relating to the requirement for participants to keep their head upright did not significantly differ across evaluation modes in the open response data, $\chi^2(1, n = 656) = 0.49$, $p = .485$. Notably, this was not the case for the

selected response data, for whom the frequency was significantly higher in the joint evaluation mode, $\chi^2(1, n = 765) = 30.24$, $p < .001$. Another feature that was employed relatively often in the open response data (and selected by over 10% of participants in the selected response data) was *performance requirements*. The frequency of responses coded as primarily relating to performance during the task was significantly lower in the joint evaluation mode in the open response data, $\chi^2(1, n = 656) = 13.86$, $p < .001$, and participants' choices in the selected response data were consistent with this determination, $\chi^2(1, n = 765) = 20.29$, $p < .001$. Finally, over 10% of participants selected *(f) being able to read coming into the study*. Since this was not the case in the open response data, a comparison was computed only for the selected response data, which revealed that significantly less participants in the joint evaluation mode attributed their judgment to their beliefs about reading as an acquired skill, $\chi^2(1, n = 765) = 21.43$, $p < .001$.

2.5.2.5.2 Diversity of Features Across Evaluation Conditions

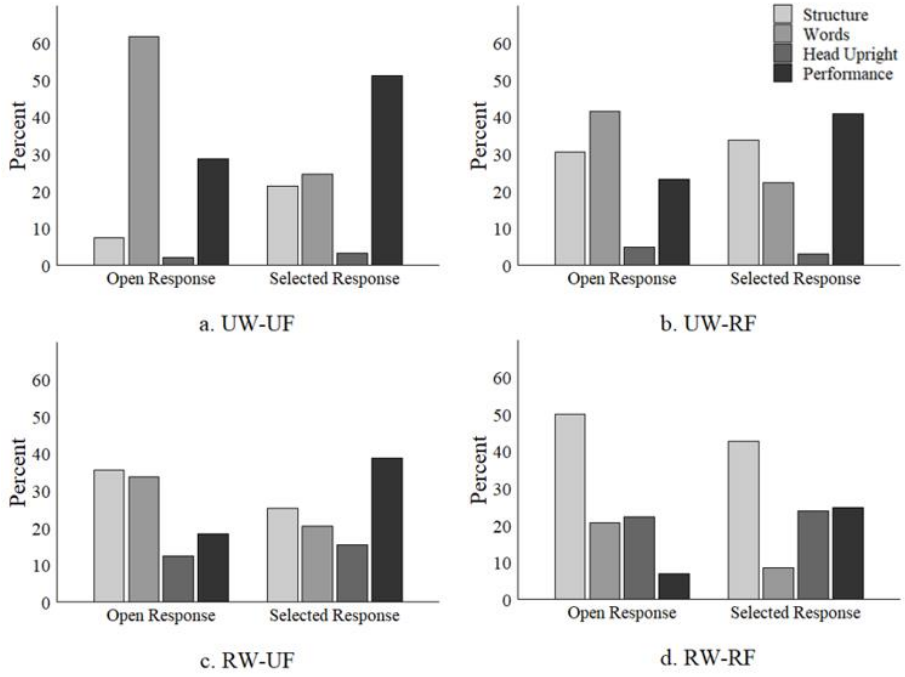
One way to determine whether the separate evaluation data exhibit more variability than the joint evaluation data is via the use of the Gini-Simpson (GS) Index[8] (Caso & Angeles, 1988), a measure of diversity or "unalikeability" across categories. The GS Index can take values from 0 to 1, where 0 represents no diversity (i.e., all responses were coded as a single feature) and as the GS Index approaches 1, this represents increasing diversity in the sample. The joint evaluation data, not including the responses coded as *(g) a reason that did not fit into the other categories or was confusing to the coder*, was found to have a GS Index of .34 ($m = 6$) and .61 ($m = 7$) for open response and selected response data, respectively. In the separate evaluation condition, corresponding GS Indices were higher: .76 ($m = 6$) and .80 ($m = 7$) for open response and selected response data, respectively. Thus, there is some evidence that in the separate evaluation mode, reasons for individuals' judgments are distributed with more variability than those in the joint evaluation condition.

[8] $GS = 1 - (p_1^2 + p_2^2 + \ldots + p_m^2)$, where m is the number of distinct categories and p_i is the proportion of responses in the i^{th} category.

2.5.2.5.3 Features Across Stimulus Types in Separate Evaluation

To determine whether the coded responses were distributed differentially across stimulus types in the separate evaluation context, comparisons were computed using a subset of the separate evaluation data comprising individuals who endorsed either the *structure*, *words*, *head upright*, or *performance requirements* feature (i.e., the four main features). Figure 13 is a visualization of these data split by stimulus type. An omnibus chi-squared test of independence revealed a significant difference in the distribution of reasons across stimulus type in the open response data and the selected response data, $\chi^2(9, n = 396) = 91.05$, $p < .001$ and $\chi^2(9, n = 412) = 55.44$, $p < .001$, respectively. Four follow-up chi-squared goodness of fit tests, one for each stimulus type, were computed to determine whether the proportion of features varied significantly. The frequency of responses in the UW-UF condition varied significantly across the four main features in the open response data, $\chi^2(3, n = 94) = 82.43$, $p < .001$. The same was demonstrated for the selected response data, $\chi^2(3, n = 94) = 43.96$, $p < .001$. As seen in Figure 13a, the UW-UF stimulus type is associated with relatively more words-based and performance-based reasons, compared to the very low proportion of structure and head upright reasons.

Figure 13. *Percent of responses in favour of each of the four main features for open response and selected response data*



Note. a) proportion of individuals in UW-UF; b) proportion of individuals in UW-RF; c) proportion of individuals in RW-UF; d) proportion of individuals in RW-RF.

The frequency of responses in the UW-RF condition differed significantly across the four main reasons in the open response data and the selected response data, $\chi^2(3, n = 82) = 23.27$, $p < .001$ and $\chi^2(3, n = 98) = 31.88$, $p < .001$. As visualized in Figure 13b, it seems that the head upright reason was least cited as the primary reason for individuals' judgments in the UW-RF condition. The main difference between the open response and selected response data in this condition is that the words were cited most frequently in the open response data, while the performance requirements were the most cited reason in the selected response data. Turning to the RW-UF condition, a significant difference across the four main reasons was detected in the open response data and in the selected response data, $\chi^2(3, n = 104) = 16.15$, $p = .001$ and $\chi^2(3, n = 103) = 12.46$, $p = .006$, respectively. In the open response data, this difference seems due to the relatively more frequent citing of the structure and words

reasons; whereas in the selected response data, the performance requirements reason was the most frequently endorsed. Finally, the frequency of responses in the RW-RF significantly varied across the four main reasons in the open response data and in the selected response data, $\chi^2(3, n = 116) = 45.38$, $p < .001$ and $\chi^2(3, n = 117) = 27.44$, $p < .001$, respectively. Figure 13d demonstrates this difference, which seems driven by the comparatively high proportions of structure-based reasons and head upright requirement reasons for the RW-RF stimulus type.

Finally, a comparison was conducted for the selected response data regarding the reading as an acquired skill reason across the four stimulus types. This comparison was computed using the dataset including all reasons except: *Something specific about the task, and the option is not listed here*. No difference was detected across stimulus types, $\chi^2(3, n = 592) = 2.79$, $p = .426$; and this did not change when removing the UW-UF data from analysis, $\chi^2(2, n = 443) = 0.08$, $p = .961$.

2.5.3 Discussion

In Experiment 4, we were able to directly compare the pattern of demand measures across stimulus types in a pure stimulus condition to a mixed stimulus condition similar to that employed in previous research. The pattern of demand measures across the stimulus types did not differ greatly across stimulus condition. Specifically, no interaction between reading times and evaluation condition was detected; and neither stimulus type, evaluation condition, nor the interaction had a significant effect on error counts, mirroring reports from previous studies (Ashburner & Risko, 2021; Dunn & Risko, 2016). Further, it was demonstrated that the pattern of effort judgments matched that from the previous experiments such that judgments in the joint evaluation condition mode followed the graded pattern wherein UW-UF < UW-RF < RW-UF < RW-RF; and judgments in the separate evaluation mode followed the typical flat pattern. Taken together, this set of results allows us to put aside the idea that the pattern of objective demand could differ substantially across evaluation conditions.

The analysis of participants' open and selected responses provided a new perspective on the nature of the influence of evaluation context on judgments of effort. Returning to the three main lines of inquiry discussed above, the results demonstrated that there was a significant difference in individuals' reasons for their judgments of effort across evaluation context (i.e., a higher proportion of individuals in joint evaluation endorsed stimulus type as their reason); and the reasons given in the separate evaluation condition were collectively more diverse – or unalike – than those in the joint evaluation condition. These two results provide preliminary support for the notion that the pronounced effect of evaluation mode on judgments of effort here is because evaluation mode shifts the basis of individuals' judgments. From the perspective of cue-utilization theories of metacognitive judgments (Koriat, 1997), this might be understood as a shift in the cues used to infer the effortfulness of a mental activity. Finally, within the separate evaluation context the reasons provided varied considerably across the different stimulus types. We examine the implications of these observations further in the General Discussion.

2.6 General Discussion

Across four experiments, we demonstrated that judgments of effort can be markedly influenced by evaluation context. Effort judgments in the joint evaluation condition consistently followed the pattern demonstrated by past research (e.g., Ashburner & Risko, 2021; Dunn et al., 2016; Dunn & Risko, 2016), namely, the RW-RF stimulus type was judged as more effortful than the RW-UF stimulus type, which was judged as more effortful than the UW-RF stimulus type (though this difference was small and not significant in some instances), which was judged as more effortful than the UW-UF stimulus type. In contrast, effort judgments in the separate evaluation condition showed little in the way of discernable ordering with respect to the stimulus types. This pattern was consistent across three different temporal contexts: (1) a prospective effort judgment without experience reading the display(s) (Experiments 1a, 1b, and 3); (2) a retrospective judgment after having read the display(s) (Experiments 2a and 2b); and (3) a prospective effort judgment after having read the display(s) (Experiments 3 and 4). Notably, the stark contrast of prospective effort

judgments across evaluation contexts in Experiment 3, wherein individuals read one display silently before judgment, persisted when individuals gave prospective effort judgments after having read multiple displays aloud (as in Experiment 4). Thus, despite the several different types of judgment elicitation, involving various amounts and types of task experience, the rather drastic effect of evaluation context (i.e., joint vs. single evaluation) was consistently observed.

We began this investigation with the prediction that eliciting judgments of effort within a separate evaluation context would lead to an elimination of the previously observed dissociations between judgments of effort and proxies for objective demand (e.g., reading time). That is, in separate evaluation, judgments of effort would more closely match measures of objective demand than previous work examining judgments of effort in joint evaluation. The idea underlying this prediction was that it was the joint evaluation context that was leading participants to exaggerate the effort requirements associated with the disorientated displays. While the separate evaluation context clearly changed the pattern of effort judgments, the resulting pattern in fact also diverges – arguably more so – from patterns of objective demand than those made in the joint evaluation context. Thus, separate evaluation did not bring judgments of effort in line with proxies for objective demand. This leaves open the question of the basis for the original dissociation between objective demand measures and judgments of effort in the joint evaluation context and introduces a novel dissociation between the former and judgments of effort in the separate evaluation context.

While the latter statement is clearly true when we consider reading time (and blink rate; Dunn et al., 2016), which has been the focus of previous work, it is interesting to note that the relatively unchanging judgments of effort across the stimulus types in the separate evaluation condition does align with the relative number of errors made across stimulus types. That is, there is typically no effect of display type on reading errors. If the number of errors committed provides a better proxy for objective demand than the other measures employed (i.e., reading time, blink rates), then one might consider separate evaluation to have, in fact, brought judgments of effort in line with objective demand. One challenge with

this perspective is that the task parameters typically employed are unlikely to render error counts as a particularly sensitive measure of demand. For example, individuals had an unlimited amount of time to read the displays. Therefore, a stronger test of this idea would require, for example, imposing a time limit on reading.

Stepping away from the relation between objective demand and judgments of effort, the most salient result in the present investigation was the pronounced effect of evaluation mode on judgments of effort. The stimulus types here have a clear and consistent effect on effort judgments in a joint evaluation context but little in the way of apparent systematic effect in a separate evaluation context. Critically, the pattern was largely the same whether individuals were making judgments of effort prospectively (having read the displays or not) or retrospectively. According to the General Evaluability Theory, a feature of evaluable attributes is that they are evaluated similarly in the joint and separate evaluation contexts. From this perspective, the variations in stimulus disorientation used here are clearly not evaluable in the context of effort judgments. Moving beyond this classification, we sought further insight into the processes underlying the dissociation across evaluation modes. Based on previous work, we proposed, within the context of a cue-utilization account of effort judgments, that the evaluation mode shifts the cues individuals use to make their judgments. The reason data from Experiment 4 are generally consistent with this notion. That is, in addition to a stark shift in effort judgments across evaluation contexts, there was an equally stark shift in the reasons individuals provided for their effort judgments. In the joint evaluation context, most individuals stated reasons based on disorientation-relevant features (e.g., structure of the display, head orientation) which was not the case in separate evaluation. The latter condition was also characterized by a greater diversity of reasons. While the shift in reasons across evaluation conditions was clear, individuals did tend to select from the same "universe" of reasons across conditions (i.e., the four most frequently cited reasons were the same across evaluation contexts). Therefore, this shift represents a change in the pattern of use across these reasons.

Why then might the stimulus type used here be inevaluable, or similarly, why might stimulus type tend to lead to shifts in the cues employed across evaluation contexts? According to the General Evaluability Theory, dissociations as a function of evaluation mode suggest that effort-based judgments of the stimuli used here are low on what is referred to as the knowledge (i.e., acquired expertise or experience) and nature (i.e., presence of an innate reference system) dimensions. One way to think about being low on these dimensions in the present context is that, across individuals, it will be unlikely that a common attribute/cue (or set of attributes/cues) will be used to generate effort judgments – both across the evaluation contexts and across the stimulus types within the separate evaluation condition. Absent a common cue/attribute, consistency across evaluation mode will likely be difficult to achieve. For example, if one group is basing their effort judgment on the extent of disorientation, which varies across stimulus types, and another is basing their effort judgment on the familiarity of the words, which does not vary across stimulus types, then in the aggregate, there is unlikely to be agreement in the pattern of effort judgments across stimulus types. The reasons individuals provided for their effort judgment(s) provide some support for this idea. The pattern of reasons provided in the separate evaluation context was clearly different from those provided in the joint evaluation context. In addition, the reasons provided differed across stimulus type within the separate evaluation condition.

Thus, one means of understanding why there is a dissociation here between joint and separate evaluation contexts is that the stimulus type manipulation does not spontaneously lead individuals to select the same attributes/cues when making judgments of effort across evaluation contexts. Consistent with this idea is the observation that, in a situation wherein we might expect individuals to be more likely to spontaneously select the same attributes/cues across evaluation contexts, there is consistency in judgments of effort across joint and separate evaluation contexts. As noted previously, Dunn et al. (2017) manipulated set size for effort judgments in a memory task, as well as the evaluation context, and demonstrated that judgments of effort increased with increased set size in both the joint and separate evaluation contexts. When eliciting a judgement about remembering a given number

of items (Dunn et al., 2017), one can reasonably expect that individuals might base their effort judgment on the number of items irrespective of the evaluation context because of their knowledge of the importance that the number of items plays with respect to their short term memory capacity. Thus, this shared knowledge might underlie the consistency found across evaluation modes (Dunn et al., 2017).

While shifts in the cues or attributes of the cognitive activity being judged provide one potential explanation of the influence of evaluation context on effort judgments, there are likely to be others as well. Another interesting explanatory path that requires further examination is how individuals, across different evaluation contexts, generate and map internal representations of effort onto the scale provided. Understanding how this process can be influenced by the evaluation context will likely provide important insights into dissociations across evaluation contexts. As suggested by the General Evaluability Theory, this process might be considered to be particularly difficult or likely to lead to noisy response patterns in separate evaluation contexts where individuals are left to some extent unmoored (or to their own devices) in making this mapping. The lack of any effect of stimulus type here could be taken to support this notion. That said, as noted above, it is important to keep in mind that this difficulty does not appear to be simply a feature of mapping internal representations of effort onto a scale, given clear demonstrations of consistency across evaluation modes (and in a manner that makes theoretical sense) in other situations wherein individuals are making judgments of effort. Rather, the question is why particular tasks (e.g., reading) or manipulations (e.g., stimulus disorientation) in conjunction with the need to make judgments of effort might lead to difficulties in this mapping process.

Overall, these results point to the theoretical importance of considering the activities for which individuals are making judgments of effort as being composed of multiple attributes (cues) that could, in isolation or in combination, form the basis of an effort judgement. And, furthermore, that the cues used will vary based on features of the context in which those judgments are made. The methods used here (i.e., comparison across evaluation modes; collection of participants' reasons) provide one means of further exploring these

issues. That said, other methods would be valuable given the framework developed here. For example, if differences in effort judgments here reflect differences in the reasons on which those judgments are made, then one should be able to prime specific reasons and find changes in effort judgments even when evaluation context is held constant. Interestingly, related theoretical issues are just beginning to be examined in the context of other metacognitive judgments (e.g., Broder & Undorf, 2019; Undorf & Broder, 2020, 2021).

2.7 Conclusion

The present investigation aimed to determine whether judgments of effort depend on the evaluation context in which they are elicited. They clearly do, at least when we consider subjective effort as a function of the orientation of a display. Moreover, this study provided a new perspective on individuals' judgment processes in both evaluation contexts. Given individuals are cognitive misers, the sensitivity of effort judgments to evaluation context demonstrated here should give us pause in considering how effective we might be in selecting the path of least resistance.

Chapter 3
Judgments of effort and associated cues are influenced by stimulus context

3.1 Introduction

Many everyday decisions can be influenced by individuals' perceptions of effort. Since an appraisal of effort for a given cognitive act can affect behaviour, mental effort as a construct represents an important aspect of the human experience. As such, there exists a long history of research examining effort (Ashburner & Risko, 2021; Dunn, Koehler, & Risko, 2017; Gabarino & Edell, 1997; Song & Schwarz, 2008; Thompson et al., 2013; Tyler et al., 1979; Undorf & Erdfelder, 2011, 2013, 2015; Westbrook, Kester, & Braver, 2013; Wu, Ferguson, & Inzlicht, 2023; Yeh & Wickens, 1988). One recent perspective on perceptions of effort has examined them as a type of metacognitive judgment (Dunn & Risko, 2016; Dunn, Gaspar, & Risko, 2019; Efklides, et al., 2006; Koriat, 1997; Raijmakers et al., 2017; Scheiter, Ackerman, & Hoogerheide, 2020). According to extant theoretical accounts – the cue-utilization account (Koriat, 1997) and analytic-processing theory (Mueller & Dunlosky, 2017) – metacognitive judgments are inferences based on available cues. While such accounts have heavily focused on metamemory judgements (e.g., judgements of learning), they seem readily extensible to other kinds of metacognitive judgments. Therefore, it is reasonable to posit that judgments of effort are made via the same inferential process. Important for the present work, a number of investigations have demonstrated metacognitive judgments to be sensitive to the context in which they are made (e.g., Dunn et al. 2017; Rhodes & Tauber, 2011; Schwarz, 2010; Tatz & Peynircioglu, 2020; Zawadzka, Simkiss, & Hanczakowski, 2018). In the present investigation we examine this contextual sensitivity with respect to judgements of effort.

Individuals' judgment spaces contain a number of cues that may inform their metacognitive judgments. One way to understand context shifts in metacognitive judgements is that contextual shifts alter the saliency of these different cues. One way in which a change

in the saliency of different cues can occur is through a context change involving comparisons across options or attributes. Judgments in general have been shown to be sensitive to whether a relevant reference point is provided (Bazerman et al., 1994; Birnbaum, 1999; Buiten & Keren, 2009; Christopoulos et al., 2011; Davidai & Shafir, 2018; Hsee et al., 1999; Hsee & Zhang, 2010; Kahneman & Ritov, 1994). Specifically, in a *joint evaluation context*, individuals are presented with two or more items to evaluate simultaneously, such that it is possible for the judge to compare various attributes of the options. Conversely, in a *separate evaluation context*, only one item is presented and evaluated in isolation. In Chapter 2 (i.e., Ashburner and Risko, 2022), I demonstrated that judgements of effort can vary significantly depending on whether they are made in a joint or single evaluation context. In this work, participants read several displays containing 25 high frequency words each and then judged the effort associated with this task. In the joint evaluation context, there were four different *stimulus types* to read and evaluate: The words were either all upright or all rotated by 60°, and the entire display was also either upright or rotated by the same amount. In the separate evaluation context, participants only read one stimulus type and judged the effort associated with that type by selecting a value on a 7-point scale. After their effort selection, participants were asked to provide the reasons for their judgment. These written reasons were then categorized by human coders as well as the participants themselves.

This investigation revealed that evaluation context had a notable impact on judgments of effort. In the joint evaluation context, individuals judged the completely upright stimulus type as least effortful; the rotated stimulus type containing rotated words as most effortful; and the other two stimulus types in between. In the separate evaluation condition, there seemed to be no influence of stimulus type on individuals' judgments of effort. This result, replicated many times, suggested a possible shifting of cues across evaluation contexts. To test this idea, Chapter 2 involved elicitation of individuals' reasons for judgment using an open response question and the development of a coding scheme based on these data. In a later experiment, individuals were invited to select a category that represented their primary reason for judgment from a list comprising our codes in addition to providing a written

response. Upon examination of individuals' reasons for their judgment, over half of individuals in the joint evaluation context reported a reason related to the structure of the stimuli (i.e., orientation of the words or the frame). By contrast, individuals in the separate evaluation context endorsed a variety of reasons related to aspects of the stimuli other than their orientation such as the words used, or the performance requirements of the task (i.e., reading as quickly and accurately as possible). The difference in the pattern of effort judgments, combined with the change in individuals' self-reported reasons, is consistent with the idea that the cues used to infer effortfulness are shifting across the evaluation contexts.

An important aspect of the evaluation context in the studies reported in Chapter 2 is that it comprised both a reading phase, wherein participants experienced the reading task over several trials; and a judgment phase, where individuals were asked to consider the stimulus type(s) they experienced and judge the effort associated with the reading task. Therefore, participants in the joint evaluation context experienced multiple stimulus types and were presented with multiple stimulus types to judge in the judgement phase, whereas participants in the separate evaluation context experienced a single stimulus type and were presented with a single stimulus type to judge. This leaves open the question of whether the differences in effort judgements observed (and concomitantly the shift in effort cues that were attended to) was due to (1) the experience of reading multiple stimulus types in the reading phase and/or (2) the opportunity to compare (or not) the various stimulus types in the judgment phase. Previous research suggests that (2) is sufficient to influence effort judgements (Dunn, Inzlicht, & Risko, 2019; Dunn, Lutes, & Risko, 2016). Specifically, Dunn, Lutes, and Risko (2016) presented participants with pairs of stimuli (from the four stimulus types described above) and participants were asked to choose the one they preferred to read. Participants made these demand selections numerous times and the results revealed that individuals' preferences corresponded to the effort judgment patterns reported above (i.e., the fully upright stimulus type was most preferred; the rotated stimulus type containing rotated words was least preferred, and the other two were in between). In the demand selection task, individuals did not have any experience reading the displays and simply made

choices based on the stimulus types. Therefore, the direct comparison between two stimulus types was sufficient to influence choice behaviour in a way that tracks with judgments of multiple stimulus types in Chapter 2. How the influence of a "mixed" (i.e., reading multiple stimulus types) versus a "pure" (i.e., reading a single stimulus type) *stimulus context* might influence effort judgements, however, is less clear.

One route through which the exposure to multiple stimulus types (e.g., one comprising upright text and another comprising rotated text) might influence effort judgments, is that it causes individuals to attend to different cues than when only one stimulus type is experienced. For example, cues associated with the difference between the stimuli may receive more attention in a mixed environment, as a direct comparison across stimulus types is possible. In addition, the experience of fluency might also vary as a function of a pure vs. mixed stimulus context. The experience of fluency has been claimed to be relative (Greifeneder & Unkelbach, 2013). That is, the way an experience of ease influences judgments for a given cognitive task depends on comparisons with similar experiences. An example of this is a study where the ease of processing line drawings was manipulated by varying the duration of presentation or the image clarity. Subjective reports of fluency and liking ratings for each image were collected. Liking ratings tracked with fluency ratings (i.e., were positively related) only when ease of processing was manipulated within-participants. When manipulated between-participants, liking ratings did not seem to depend on subjective experiences of ease (Forster, Gerger, & Leder, 2015). The authors concluded that fluency effects on liking depend on whether there are relative differences present among the stimuli. If it is the case that a fluency experience, or at least its impact on related judgments, depends on salient differences between the stimuli, then exposure to multiple stimulus types could influence effort judgments via altering the experience of fluency relative to being exposed to only a single stimulus type.

3.2 Experiment 5

In Experiment 5, participants were randomly assigned to a stimulus context (i.e., pure vs. mixed) for the reading phase. Those in the pure context read several displays of a single

stimulus type (i.e., upright or rotated), while those in the mixed context read displays of both types. This design decouples the influence of the stimulus context from that of the judgment context, which always involved judgment of a single stimulus type (i.e., either upright or rotated). After making their judgment(s), participants provided their rationale in an open response field and characterized their reason(s) for judgment by selecting a best-fitting category from a list. Collection of these self-reports allow an examination of the types of cues individuals report using, as well as whether those cues change across stimulus contexts.

3.2.1 Method

3.2.1.1 Participants

Participants were recruited from Prolific, were residents of the United States, Canada, or the United Kingdom, had at least 100 studies completed, and had an approval rating of at least 99%. Based on previous research, we aimed to have at least 70 observations per cell to achieve at least .80 power to detect an effect size of $\eta_g^2 = .03$.

3.2.1.2 Stimuli and Design

The study was completed via a Qualtrics survey. The stimuli were displays of 25 high-frequency five letter words (five words per line and five lines total). Words were in 18-point Courier New font, with the starting word in red and the rest in black on a white background, and blue arrows indicated the direction of reading. There were two stimulus types: upright and rotated. The upright stimulus type appeared similar to typical text, with a left-to-right reading pattern; the rotated stimulus type appeared like tilted text, such that the display and the words therein were rotated by 60-degrees. Whether the rotation was clockwise or counterclockwise was counterbalanced between-participants. This was a 2 (stimulus type: upright, rotated) x 2 (stimulus context: pure, mixed) between-participants design.

3.2.1.3 Procedure

After providing consent, participants were provided instructions that the study comprised two phases: the reading phase and the judgment phase. In the reading phase, a given trial

involved reading the display shown on screen as quickly and accurately as possible, then clicking the 'next button' to proceed to another trial. All participants experienced eight trials. In the pure stimulus context, all displays were of the same stimulus type, upright or rotated; in the mixed stimulus context, there were four upright and four rotated displays (pseudo-randomly interleaved). Prior to the judgment phase, participants were only told that they would provide a judgment on a cognitive task to be described. Once all of the displays had been read, participants were instructed to provide a single prospective judgment of effort, relating to the task of reading another display in the future. Participants were randomly assigned to judge either the upright or the rotated stimulus type and were shown a template of that type (wherein all words in the display were 'WORD'). The judgment was via selection of a point on a 7-point scale (*1 = Will require little effort, 7 = Will require a great deal of effort*) similar to that used by Koriat, Nussinson, and Ackerman (2014).

Immediately after selecting their judgment, participants were asked to provide their reasoning for their selection in a free form response field. Following this, participants were asked to reflect on the reason they had just provided and select, from 7 categories, the option that best represented their reason (e.g., "Something specific about the number of words to read?"; "Something specific about the way the words were organized on the screen?"; etc.). One of these options was "Nothing specific about the task," allowing a comparison between task-specific vs. non-specific responses. Afterward, participants were asked whether they moved their head at all during the reading phase of the study, asked to complete an attention check question, and asked what type of device they used to complete the study and if scrolling or resizing their screen while reading was necessary. Following this, the study concluded.

3.2.2 Results

A total of 345 participants provided consent and participated in the study. Altogether, 42 participants' data were excluded from analysis[9]. Specifically, data provided by three participants were excluded because participants took an excessive amount of time to complete the study (i.e., over 37 minutes); data provided by eight participants were excluded due to them not completing the study; and data provided by one participant were excluded because they spent an exceptionally long time (i.e., over five minutes) on one display. Furthermore, due to a programming issue, some participants were able to attempt or complete the study more than one time, resulting in 23 observations of data being excluded from analysis; and seven participants selected an incorrect response to the attention check question. Data provided by the remaining 303 participants (with a mean age of 35.36 and 211 female, 89 male, and 3 preferring not to answer) was included in the analysis. This study was preregistered; the preregistration, data, and analysis script are available on the Open Science Framework: https://osf.io/xeywr/. All analyses were conducted using the open-source statistical analysis software R, version 4.3.2.

3.2.2.1 Subsets of Data

Of the remaining 303 participants, 27 noted they had to scroll or resize their screen while reading a display. An analysis conducted excluding these data was consistent with the results reported below, thus these participants' data were retained for all analyses. Thirty-four participants had an average reading time of less than six seconds, suggesting that these individuals did not actually read each word, but were still exposed to the stimulus type. An analysis excluding these observations was computed and the results were consistent with those from the main analysis. Additionally, 42 individuals may have misunderstood the judgment task based on their open responses (e.g., comments related to memorizing the words in the display, or finding a specific word in the display). As above, an analysis excluding these observations was computed and the results were consistent with those from

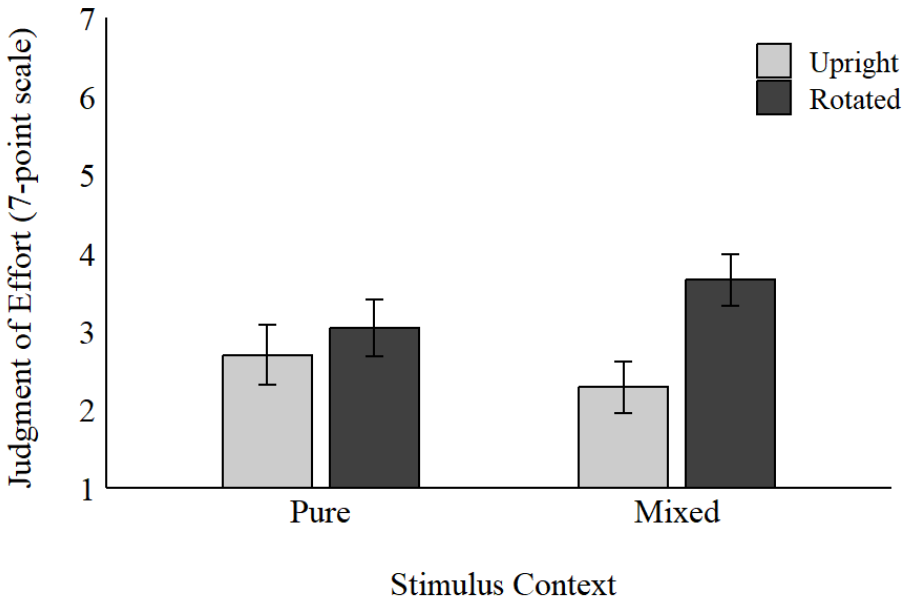
[9] Of the 33 excluded observations that included a judgment, there were 12, 14, 3, and 4 in the pure/upright, pure/rotated, mixed/upright, and mixed/rotated conditions, respectively.

the main analysis. Therefore, the results reported below comprises 303 observations unless otherwise stated.

3.2.2.2 Judgments of Effort

There were 75, 78, 72, and 78 observations in the upright pure, upright mixed, rotated pure, and rotated mixed conditions, respectively. A 2 (stimulus type: upright, rotated) x 2 (stimulus context: pure, mixed) between-subjects ANOVA with effort judgments as the output variable was computed. This analysis revealed a significant interaction, $F(1, 299) = 8.18$, $p = .003$, $\eta^2_g = .03$, such that the difference between effort judgments for the upright and rotated stimulus types was larger in the mixed stimulus context ($M_{\text{rotated-upright}} = 1.37$), compared to the pure stimulus context ($M_{\text{rotated-upright}} = 0.35$). With respect to main effects, the stimulus context was not significant, $F < 1$; and the rotated stimulus type was judged as more effortful than the upright stimulus type, $F(1, 299) = 23.11$, $p < .001$, $\eta^2_g = .07$. To better understand the interaction, simple effects comparisons were computed across stimulus context and stimulus type separately. These tests revealed that there was no difference in effort judgments between the pure and mixed context for the upright stimulus, $t(146.22) = 1.60$, $p = .111$, $d = 0.26$; however, judgments in the mixed context were significantly higher than in the pure context for the rotated stimulus, $t(145.09) = 2.44$, $p = .016$, $d = 0.40$. Moreover, while there was no difference in effort judgments between the upright and the rotated stimulus type in the pure context, $t(144.88) = 1.29$, $p = .199$, $d = 0.21$, the rotated stimulus was judged as significantly more effortful than the upright stimulus in the mixed context, $t(153.98) = 5.80$, $p < .001$, $d = 0.93$. In short, a significant rotation effect was observed in the mixed context but not in the pure context. Figure 14 is a visualization of these data.

Figure 14. *Effort judgments as a function of stimulus type and stimulus context in Experiment 5*



Note. Error bars represent 95% CIs

3.2.2.3 Participants' Explicit Cue Reports

The written reasons participants provided (i.e., the *open text responses*) for their effort judgment were coded in two stages. The first stage was a determination of whether the response discussed something specific to the reading task (e.g., *I had to move my head due to the awkward position of the words*), or instead was a more general statement about their ideas of effortfulness (e.g., *I found it in between very easy and very difficult*). The second stage was the coding of the responses that were specific to the task using 6 categories[10] (i.e., the structure of the display; the words comprising the display; the performance requirements of the task; one's ability to read in English; the number of words in the display; and a category representing something specific to the task but not mentioned). Only one reason could be selected, but if there were multiple reasons within the written response, multiple codes were

[10] These categories were part of a coding scheme that was created using open response data from previous experiments (Ashburner & Risko, 2022).

used to characterize that response and the coder indicated the ordering of reasons by importance (i.e., primary, secondary, etc.). As mentioned in the Procedure, participants also made a selection from a list of these categories to represent what they thought was their primary reason for their judgment (i.e., *selected responses*). Included in this list was a category representing a non-task-specific reason. The two coders agreed on 90% of the specific/not specific code assignments[11]. Of the responses that were specific to the task, coders agreed on 64% of the primary specific reason assigned and this increased to 84% when considering whether the coders had at least one specific reason in common between their primary and secondary reasons. A third coder resolved all differences between the original two coders, and the results involving the open responses are based on the resolved codes. All analyses involving reasons for judgment used data that excluded observations associated with a likely misunderstanding of the judgment task ($n = 42$). The results were consistent across the selected responses and the open text responses; therefore, we report the analyses conducted using the selected responses.

First, to determine the relation of stimulus type, stimulus context, and their interaction on whether a specific reason was selected, logistic regression was used. More than 95% of responses involved a specific reason for judgment. This analysis demonstrated no significant effects, $|z| < 1$.

Moving to the reasons specific to the task, we focus on whether the primary reason supplied by participants was related to the structure of the display (i.e., a *structure-based reason*). This is because a structure-based reason is arguably the only reason that is relevant with respect to the stimulus type manipulation. Across stimulus contexts, we anticipate a shift in reasons to vs. away from the structure of the stimuli, whereas we do not have specific predictions for those reasons that do not change across stimulus types.

[11] Upon review of the open text responses, we found that one coder's input was missing. The written response for this observation was similar to: *One should always invest as much effort as possible*. This response was coded as 'not specific to the task' after the fact by the authors.

A logistic regression analysis with the subset of data comprising only specific reasons was computed to determine the relation of stimulus type, stimulus context, and their interaction on the likelihood of selecting "Something specific about the way the words were organized on the screen?", as the best reason for judgment. The interaction term was not significant, B = -0.82, SE = 0.62, p = .190, so a model with only the stimulus type and stimulus context predictors is reported here. The main effect of stimulus type was significant, B = 1.61, SE = 0.29, p < .001, such that when the stimulus is rotated, the odds of selecting a structure-based reason (versus selecting a reason unrelated to the structure of the stimuli) increase about 5 times from when the stimulus is upright. The main effect of stimulus context was also significant, B = 1.34, SE = 0.29, p < .001, such that, when in the mixed stimulus context, the odds of selecting a structure-based reason increased about 3.8 times relative to the pure stimulus context. Table 4 comprises a summary of relevant proportions for all experiments.

Table 4. *Proportions of participants endorsing a structure-based reason within stimulus type and stimulus context*

	Experiment 5	*Experiment 6*	*Experiment 7*
Upright Stimulus			
Pure	.11	.37	.27
Mixed	.44	.58	.48
Rotated Stimulus			
Pure	.51	.65	.67
Mixed	.74	.80	.78

Note. Observations from participants who likely misunderstood the task have been excluded.

3.2.2.4 Exploratory Analyses

3.2.2.4.1 Structure-based vs. Not Structure-based

A remaining question of interest, given the difference in proportion of cue type (i.e., reason) selected across stimulus contexts, is whether individuals who selected a structure-based reason demonstrate the pattern of judgments we would expect from individuals inferring their judgments from the orientation of the stimulus (i.e., the rotated display would be more effortful to read than the upright display). In the same vein, we would expect individuals who did not select a reason based on the structure of the stimulus type to judge the upright and rotated displays as about equally effortful based on the idea that the other reasons for judgment involve characteristics of the stimuli that do not change across stimulus type. This analysis provides insight into the extent to which participant's explicit cue reports capture the pattern of participant's judgements of effort.

When the data are partitioned according to whether individuals selected a structure-based reason, we contend with an unbalanced design together with the fact that this imbalance is confounded with condition. That is, the proportion of individuals selecting a structure-based reason shifts across stimulus context. Therefore, we opted to provide the means along with cell sizes in Table 5 alongside the exploratory analyses below.

Table 5. *Effort judgments and cell sizes by stimulus context, stimulus type, and reason for Experiments 5-7*

	Experiment 5			*Experiment 6*			*Experiment 7*		
	M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Pure Stimulus Context									
Upright, Structure	2.50	1.64	6	2.37	1.25	30	2.57	0.98	21
Rotated, Structure	3.52	1.26	31	3.75	1.68	59	4.09	1.66	55
Upright, Not Structure	3.06	1.73	48	2.88	1.69	52	2.89	1.52	56
Rotated, Not Structure	2.73	1.72	30	3.03	1.53	32	3.04	1.19	27
Mixed Stimulus Context									
Upright, Structure	2.00	1.14	27	1.86	1.13	50	1.85	1.00	47
Rotated, Structure	4.07	1.32	55	4.07	1.52	67	4.11	1.36	75
Upright, Not Structure	2.79	1.53	34	2.50	1.70	36	2.40	1.48	50
Rotated, Not Structure	2.79	1.47	19	3.71	1.45	17	4.05	1.36	21

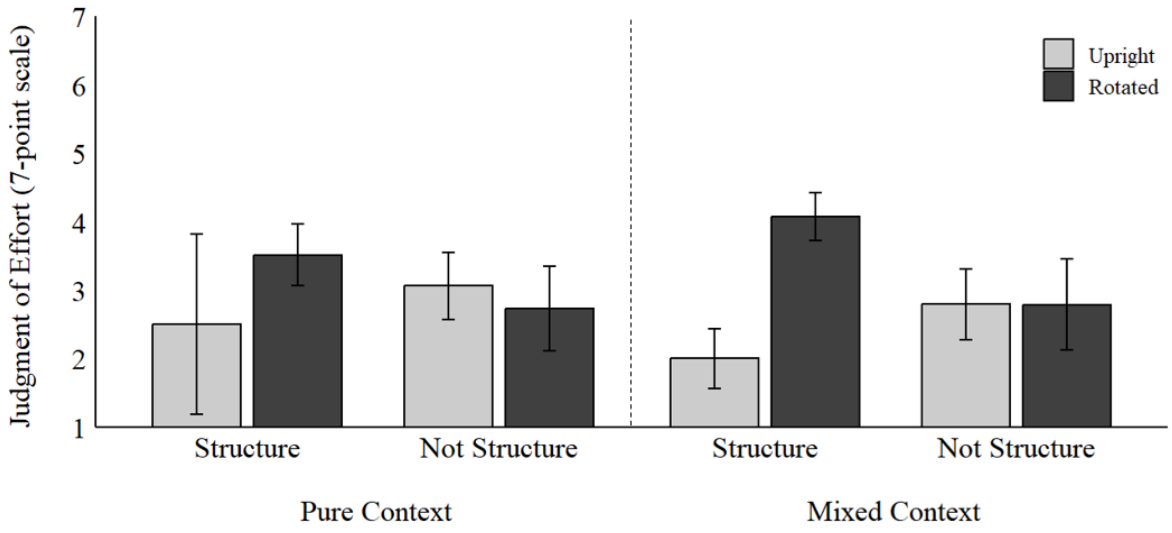
Note. Observations from participants who likely misunderstood the task have been excluded.

There appeared to be a difference in the judgment pattern across stimulus contexts, as visualized in Figure 15. A 2 (stimulus type) x 2 (stimulus context) x 2 (reason: structure-based vs. not structure-based) ANOVA demonstrated the expected main effect of stimulus type, $F(1, 242) = 8.84$, $p = .003$, $\eta^2_g = .04$, while the main effects of stimulus context and reason were not significant, Fs < 1. The interaction of stimulus type by reason was significant, $F(1, 242) = 13.64$, $p < .001$, $\eta^2_g = .05$; and neither the stimulus context by stimulus type, nor the stimulus context by reason, interactions were significant, $F(1, 242) = 2.22$, $p = .137$, $\eta^2_g = .01$ and $F < 1$, respectively. The three-way interaction was also not significant, $F < 1$.

Focusing on the stimulus type by reason interaction, pairwise comparisons indicated a significant difference between the rotated and upright stimulus ($M_{\text{rotated-upright}} = 1.78$) when a structure-based reason was selected, $t(61.69) = 6.92$, $p < .001$, $d = 1.38$. This difference

($M_{rotated\text{-}upright}$ = -0.20) did not reach significance when a non-structure-based reason was selected, $t(102.75) < 1$.

Figure 15. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 5*



Note. Error bars represent 95% CI. These data do not include observations associated with a non-specific reason or a likely misunderstanding of the task.

Focusing on the individuals who endorsed a structure-based reason, visual inspection of Figure 15 suggests that the rotation effect may be larger in the mixed context, compared to the pure context. Given the unbalanced design and the confounding of structure-based reasons with stimulus context, we thought it prudent to conduct a 2 (stimulus type) x 2 (stimulus context) between-subjects ANOVA within the structure-based reason. This analysis revealed a main effect of stimulus type, $F(1, 115) = 22.85$, $p < .001$, $\eta^2_g = .17$; neither the main effect of stimulus context nor the interaction were significant, $F < 1$ and $F(1, 115) = 2.67$, $p = .105$, $\eta^2_g = .02$, respectively. Similar results were obtained when analyzing the open text responses.

3.2.2.4.2 Number of Cues Endorsed

As mentioned above, open text responses that were specific to the task could contain more than one reason for judgment. This was not the case for selected responses (i.e., participants

could only select one response). The vast majority of responses were represented by one or two categories, with only a handful of responses associated with greater than two different categories. A chi-squared test of independence was computed to determine whether there was an effect of stimulus context on the proportion of observations associated with one vs. two categories (Table 6 comprises the values and proportions for all Experiments).

Table 6. *Proportion of single cue use (compared to two cues) split by stimulus context for Experiments 5-7*

	Experiment 5	*Experiment 6*	*Experiment 7 Selected*	*Experiment 7 Coded*
Pure Stimulus Context	.74	.59	.75	.58
Mixed Stimulus Context	.63	.71	.83	.69

Note. Observations from participants who likely misunderstood the task have been excluded.

The proportion of individuals choosing a single cue was .74 in the pure context, and .63 in the mixed context. The proportion of single- to double-category responses did not differ significantly across the stimulus contexts, $\chi^2(1, N = 233) = 2.74$, $p = .098$. There were cases wherein a participant noted two distinct reasons that are associated with the same category of our coding scheme. For example, here is a paraphrased response that was coded as two structure-based reasons: *The words were in a typical reading direction and the angle is not too difficult*. With this in mind, the same analysis was conducted with the interpretation that the unique number of reasons represented the number of cues used, and the results were consistent with those reported above.

3.2.3 Discussion

The stimulus context had an effect on individuals' judgments of effort, such that exposure to both the upright and rotated stimulus types in the mixed context resulted in a higher judgment for the rotated stimulus, as compared to the pure context. Moreover, while not significant, judgments of the upright stimulus were on average lower in the mixed context, compared to the pure context. Finally, we observed a rotation effect (i.e., the rotated stimulus

was judged as significantly more effortful than the upright stimulus) in the mixed stimulus context, but this effect did not reach significance in the pure stimulus context. Taken together, this indicates a larger effect of stimulus rotation in the mixed stimulus context. Thus, even when the judgement context is held constant (i.e., participants judge exactly one stimulus type regardless of assigned stimulus context), a change in the stimulus context preceding that judgement can have a substantial effect on effort judgements. Furthermore, if we are to interpret individuals' reported reason(s) for judgment as cues employed, this shift in effort judgements was mirrored by a shift in the cues individuals reported relying on across stimulus contexts. Specifically, the likelihood of individuals selecting the category for structure-based reasons to represent their reason for judgment was significantly higher when in the mixed (as compared to pure) stimulus context. This finding provides an explanation for the interaction between stimulus type and stimulus context, as more individuals attending to the rotation of the rotated stimulus in the mixed context is arguably associated with an increased perception of effort for that stimulus type.

Notably, in an exploratory analysis, individuals who endorse a structure-based reason judged the rotated stimulus as significantly more effortful than the upright stimulus; whereas individuals who endorsed a non-structure-based reason judged the stimuli as about equally effortful. This result provides some evidence that individuals are making their judgments based on the reasons(s) they report. That is, individuals that reported using a structure-based reason were influenced by the structure of the stimulus and those who did not report using a structure-based reason were not. Finally, with respect to the number of cues used (i.e., single vs. two) in participants' written responses about their judgments, individuals did not differ significantly depending on stimulus context, though there was a trend such that a greater proportion of individuals used a single cue in the pure context, as compared to the mixed context.

The way in which a metacognitive judgment is elicited can influence the judgment itself (e.g., metamemory judgments elicited prospectively vs. retrospectively; Luna & Albuquerque, 2022). To ensure that the results of Experiment 5 can be generalized beyond

anticipated/prospective effort judgments, we sought to replicate them eliciting a retrospective judgment of effort in Experiment 6. Furthermore, we aimed to recruit a larger sample to buffer the effect of any necessary exclusions.

3.3 Experiment 6

3.3.1 Method

3.3.1.1 Participants, Stimuli, Design, and Procedure

We aimed for a cell size of at least $n = 80$ ($N = 320$) after exclusions but were able to achieve a cell size of at least $n = 90$. A post hoc analysis using the theoretically-derived effect size reported in Experiment 5 (i.e., $\eta^2_g = .03$) and a cell size of 90 demonstrated that we had .88 power to detect the 2x2 between-subjects interaction between stimulus type and context reported below. The sample was drawn from the same participant pool as in Experiment 5 and the stimuli, design, and procedure were identical to that of Experiment 5 with one exception. Participants were asked to judge the effort that was required of them to read the display that resembled the template provided (i.e., either the upright or the rotated display template as in Experiment 5). The effort judgments were made on a 7-point scale with labels reflecting a retrospective judgment (*1 = Required little effort, 7 = Required a great deal of effort*).

3.3.2 Results

A total of 408 participants provided consent and participated in the study[12]. Altogether, 25 participants' data were excluded from analysis. Specifically, data provided by four participants were excluded due to them not completing the study; and data provided by four participants were excluded because they reported an issue with the study that may have influenced data quality. Furthermore, due to a programming issue, some participants were

[12] Of the 20 excluded observations that included a judgment, there were 6, 2, 7, and 5 in the pure/upright, pure/rotated, mixed/upright, and mixed/rotated conditions, respectively.

able to attempt or complete the study more than one time, resulting in 10 observations of data being excluded from analysis; and seven participants selected an incorrect response to the attention check question. Data provided by the remaining 383 participants (with a mean age of 33.91 and 256 female, 126 male, and 1 intersex) was included in the analysis. This study was preregistered; the preregistration, data, and analysis script are available on the Open Science Framework: https://osf.io/rd832/.

3.3.2.1 Subsets of Data

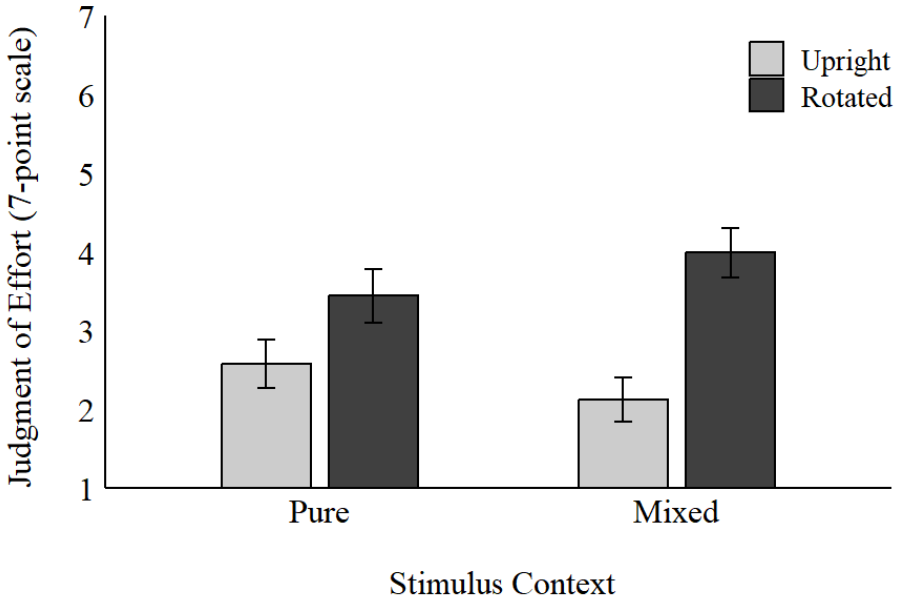
Of the remaining 383 participants, 28 noted they had to scroll or resize their screen while reading a display; thirty-one participants had an average reading time of less than six seconds; and 20 participants may have misunderstood the judgment task based on their open responses. An analysis computed excluding each of these subsets was consistent with the results from the main analysis. Therefore, the results reported below comprises 382 observations.

3.3.2.2 Judgments of Effort

There were 94, 95, 100, and 94 observations in the upright pure, upright mixed, rotated pure, and rotated mixed conditions, respectively. A 2 (stimulus type: upright, rotated) x 2 (stimulus context: pure, mixed) between-subjects ANOVA with effort judgments as the output variable was computed. This analysis revealed a significant interaction, $F(1, 379) = 10.01$, $p = .002$, $\eta^2_g = .03$, such that the difference between effort judgments for the upright and rotated stimulus types was larger in the mixed stimulus context ($M_{\text{rotated-upright}} = 1.87$), compared to the pure stimulus context ($M_{\text{rotated-upright}} = 0.87$). With respect to main effects, the stimulus context was not significant, $F < 1$; and the rotated stimulus type was judged as more effortful than the upright stimulus type, $F(1, 379) = 73.91$, $p < .001$, $\eta^2_g = .16$. To better understand the interaction, simple effects comparisons were computed, which demonstrated that effort judgments in the mixed context were significantly lower than in the pure context for the upright stimulus, $t(184.40) = 2.15$, $p = .033$, $d = 0.31$; additionally, judgments in the mixed context were significantly higher than in the pure context for the rotated stimulus,

$t(191.54) = 2.33$, $p = .021$, $d = 0.33$[13]. Moreover, there was a significant effect of rotation (i.e., the rotated stimulus type was judged as more effortful than the upright) for both the pure context, $t(191.45) = 3.68$, $p < .001$, $d = 0.53$; and the mixed context, $t(184.20) = 8.77$, $p < .001$, $d = 1.28$, as visualized in Figure 16.

Figure 16. *Effort judgments as a function of stimulus type and stimulus context in Experiment 6*



Note. Error bars represent 95% CI

3.3.2.3 Participants' Reason for Judgment

To determine the relation of stimulus type, stimulus context, and their interaction on whether a specific reason was selected, logistic regression was conducted. For this analysis,

[13] This comparison did not achieve significance when analyzing the subsets of data associated with screen resizing, low average reading time, and likely misunderstanding of the task.

20 observations were excluded, corresponding to individuals who likely misunderstood the task. About 95% of responses involved a specific reason for judgment. This analysis demonstrated no significant effects, all $|z| < 1$.

Focusing now on the subset of observations corresponding to a specific reason for judgment, a logistic regression analysis with this subset was computed to determine the influence of stimulus type and stimulus context on the likelihood of selecting a structure-based reason as the best reason for judgment (as in Experiment 5, the interaction term was not significant, $|z| < 1$, and was removed from the model). The main effect of stimulus type was significant, $B = 1.11$, $SE = 0.23$, $p < .001$, such that when the stimulus is rotated, the odds of selecting a structure-based reason (versus selecting another reason) increase about 3 times from when the stimulus is upright. The main effect of stimulus context was also significant, $B = 0.83$, $SE = 0.23$, $p < .001$, such that, when in the mixed stimulus context, the odds of selecting a structure-based reason (versus selecting another reason) increase about 2.3 times from when in the pure stimulus context.

3.3.2.4 Exploratory Analyses

3.3.2.4.1 Structure-based vs. Not Structure Based

As in Experiment 5, the data were partitioned into judgments associated with structure-based reasons and judgments associated with another, non-structure-based reason. With respect to the open responses, coders agreed on 98% of the specific/not specific code assignments. Of the responses that were specific to the task, coders agreed on 64% of the primary specific reason assigned and this increased to 87% when considering whether the coders had at least one specific reason in common between their primary and secondary reasons. Contrary to Experiment 5, there were some discrepancies between participants' selected reasons and the open text responses, which are reported below. A 3-way ANOVA testing the effects of stimulus type, stimulus context, and selected reason (i.e., structure-based vs. not structure-based) demonstrated the expected main effect of stimulus type, $F(1, 335) = 47.54$, $p < .001$, $\eta^2_g = .12$, while the main effects of stimulus context and reason were

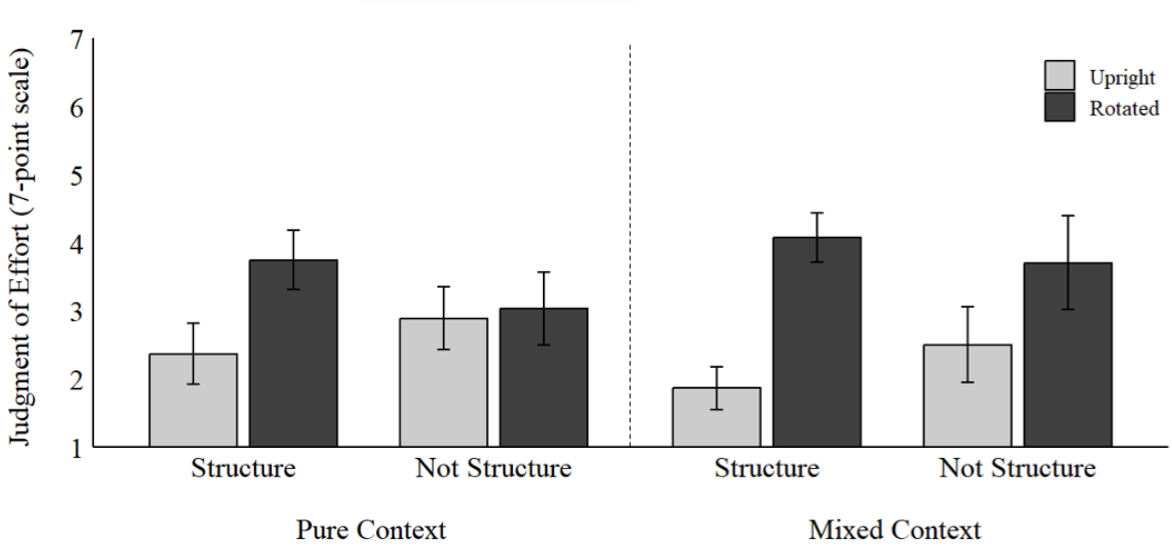
not significant, *F*s < 1. There was a significant interaction between stimulus type and stimulus context, $F(1, 335) = 6.98$, $p = .009$, $\eta^2_g = .02$; a significant interaction between stimulus type and reason, $F(1, 335) = 9.76$, $p = .002$, $\eta^2_g = .03$; and a non-significant interaction between stimulus context and reason, $F < 1$. Finally, the three-way interaction was not significant, $F < 1$.

Focusing on the stimulus type by reason interaction, pairwise comparisons revealed a significant difference between the rotated and upright stimulus ($M_{\text{rotated-upright}} = 1.87$) when a structure-based reason was selected, $t(198.80) = 9.60$, $p < .001$, $d = 1.29$. This difference ($M_{\text{rotated-upright}} = 0.54$) was not significant when a reason unrelated to structure was selected, $t(108.38) = 1.90$, $p = .060$, $d = 0.33$. When analyzing the open text responses, this difference ($M_{\text{rotated-upright}} = 0.61$) was significant, $t(172.02) = 2.68$, $p = .008$, $d = 0.39$.

Additionally, visual inspection of Figure 17 suggests that the rotation effect may be larger in the mixed context, compared to the pure context, both for individuals who endorsed a structure-based reason and individuals who endorsed non-structure-based reasons. As with Experiment 5, the unbalanced design caused partly by the higher proportion of structure-based reasons in the mixed stimulus context led us to examine this apparent difference in rotation effect by conducting separate 2 (stimulus type) x 2 (stimulus context) between-subjects ANOVAs within each level of reason. Analysis of the data associated with endorsement of a structure-based reason revealed a significant main effect of stimulus type, $F(1, 202) = 72.41$, $p < .001$, $\eta^2_g = .26$ (i.e., the rotated stimulus type was judged as more effortful); and a significant interaction, $F(1, 202) = 3.91$, $p = .049$, $\eta^2_g = .02$. The effect of stimulus context was not significant, $F < 1$. Analysis of the data associated with endorsement of a non-structure-based reason demonstrated a significant effect of stimulus type, $F(1, 133) = 5.03$, $p = .027$, $\eta^2_g = .04$ (in the expected direction); and did not yield a significant interaction, $F(1, 133) = 3.09$, $p = .081$, $\eta^2_g = .02$. The effect of stimulus context was not significant, $F < 1$. As mentioned above, there was a discrepancy between the selected reasons and the open text responses. Specifically, when these 2x2 ANOVAs were computed for the open text responses, analysis of the data associated with those endorsing a structure-based

reason did not yield a significant interaction, $F(1, 146) < 1$. (In the pure context, $M_{\text{rotated-upright}} = 2.14$ and in the mixed context, $M_{\text{rotated-upright}} = 2.26$). Furthermore, using the data associated with those endorsing a non-structure-based reason revealed a significant interaction, $F(1, 200) = 6.28$, $p = .013$, $\eta^2_g = .02$. Main effects for these analyses were qualitatively consistent with those reported for the selected reasons. To summarize, for the open text responses associated with a structure-based reason coded as the primary reason, the rotation effect was not significantly larger in the mixed context, though the analysis of the selected responses demonstrated a significant increase of the rotation effect in the mixed context. Moreover, for the open text responses associated with endorsement of only non-structure-based reasons, the rotation effect was significantly larger in the mixed context, while the analysis of selected responses only demonstrated a marginal effect. This discrepancy prompted an analysis of the open text responses partitioned by whether a structure-based reason was mentioned regardless of importance (e.g., if a written response indicated that the performance requirements of the task was the primary reason and a structure-based reason was also provided); and an analysis of the open text responses taking into account any suggestions for most important cue provided by the coder who resolved any disagreements (there were 31 such observations). These additional analyses supported the results reported above for the open text responses.

Figure 17. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 6*



Note. Error bars represent 95% CI. These data do not include observations associated with a non-specific reason or a likely misunderstanding of the task.

3.3.2.4.2 Number of Cues Endorsed

A chi-squared test of independence was computed to determine whether there was an effect of stimulus context on the proportion of observations associated with one vs. two categories. The proportion of individuals reporting a single cue was .59 in the pure context and .71 in the mixed context, which represented a significant difference, $\chi^2(1, N = 344) = 4.83$, $p = .028$. The same analysis was conducted with the interpretation that the unique number of reasons represented the number of cues used, and the results were consistent with those reported above.

3.3.3 Discussion

Experiment 6 produced the predicted interaction between stimulus type and context. Assignment to the mixed stimulus context resulted in a relatively higher effort judgment for the rotated stimulus, as well as a relatively lower judgment for the upright stimulus, as compared to judgments in the pure stimulus context (this latter simple effect did not reach

significance in Experiment 5). Furthermore, this shift in effort judgments across stimulus contexts tracked with the commensurate shift of reported cue usage. Specifically, the likelihood of individuals selecting the "structure" category to represent their reason for judgment was significantly higher when in the mixed (as compared to pure) stimulus context.

The exploratory analysis of judgments based on the reason provided as well as stimulus type and stimulus context revealed that individuals who report using a structure-based reason to make their judgment did seem to do so. That is, these individuals are making their judgments based on their reported cues. Interestingly, the rotation effect for those in the mixed context who selected the structure category was significantly larger than those in the pure context who selected the structure category; however, this result did not hold when analyzing the open text responses. Moreover, for those endorsing a non-structure-based cue, the rotation effect visualized in Figure 5 was not significant but did achieve significance when analyzing the open text responses. In fact, a consistent significant rotation effect was revealed for those who did *not* (primarily) report relying on a structure-based cue in the open text responses. Despite the statistical differences between participants' selected responses and the open text responses, the data demonstrate a similar pattern. Specifically, the rotation effect appears more pronounced in the mixed context, both for individuals who endorse a structure-based cue and individuals who do not.

It is difficult to explain a difference in the magnitude of the rotation effect on effort judgement across contexts for those who do not endorse a structure-based cue under the supposition that participants are making effort judgments based *only* on the cues they report. That is, if participants report using the same cue and that is a veridical report of the information used, then, without additional ad-hoc assumptions, the pattern of effort judgements should be comparable. Thus, if the observed difference in the rotation effect across contexts for those who do not endorse a structure-based reason represents a true effect, it would pose an interesting challenge to the sufficiency of explicit cue report as a window into cue usage. There exist a number of non-mutually exclusive ideas worth considering in light of such a finding (i.e., participants report using the same cue across contexts, but their

behaviour, presumably a product of the usage of that cue, varies across those contexts). One possible explanation here is that the mixed context is generating a type of experiential cue that is not readily available upon reflection. That is, individuals may be processing a fluency experience and the signal provided by that experience differs across stimulus contexts. As discussed in the Introduction, fluency effects can differentially influence judgments depending on whether there are relative differences present among the stimuli. It seems sensible that a direct comparison of two stimulus types such as that offered by the mixed context could therefore change the way in which a fluency signal translates to an individuals' judgment of effort. If the mixed context is providing individuals comparatively stronger experiential cues, due to the experience of reading both an upright and rotated stimulus multiple times, it might be that a kind of sheer subjective experience (Koriat, 2007) could unconsciously influence judgments. As such, this account suggests an influence on effort judgements that escapes explicit report.

A separate idea is that across contexts the same reported cue (i.e., structure-based) is being used, but used in a conceptually different manner. For example, an individual in the pure context might appeal to a structure-based cue with a relatively lower weighting than an individual in the mixed context. That is, the structure-based cue could be the secondary cue reported by both individuals, but in the mixed context, it carries more importance relative to other reported cues. Another possible difference in how the structure cue could be used is illustrated by considering judgments for the upright stimulus type across contexts. Specifically, an individual in the pure context appealing to the left-to-right reading direction might select the structure category; and an individual in the mixed context appealing to the upright (vs. rotated) nature of the text might select the structure category. In any case, this account suggests that, while the explicit cue report might capture the actual cue(s) used, in order to capture judgement behaviour a complete account would have to grapple with variations in how cues are mentally conceptualized. In Experiment 7 we test one potential variant of this kind of idea. Specifically, we examined whether participants in the mixed context perceived the rotation of the stimulus as greater than in the pure context. That is, does

experiencing both an upright and rotated stimulus make the rotated stimulus appear more rotated? If so, then this would provide a potential explanation for the greater effect of rotation observed in the mixed context for individuals primarily endorsing a structure-based reason. On this account, participants may be in fact relying on the same cue (i.e., structure-based), but the mixed context is altering what might be considered the input to the process of converting attribute values on that cue into effort judgements. For example, if in some rough manner degrees of disorientation are mapped to increasing effort when one relies on a structure-based cue, then perceiving a stimulus as more rotated would lead to a higher judgement of effort, all else held equal. To investigate this, Experiment 7 includes a question asking participants to estimate the degree of disorientation of the rotated stimuli they read.

An additional finding worth noting is the difference in terms of number of cues (one vs. two) used across Experiments. In Experiment 5, there was no difference in proportion of single-cue vs. two-cue reasons across context; whereas in Experiment 6, the number of individuals endorsing a single category was greater than those endorsing two categories in the mixed context, as compared to the pure context. One limitation of Experiments 1 and 2 with respect to examining shifts in the number of cues used is that participants did not have the opportunity to select more than one category regardless of whether their written responses suggested using more than one. That is, when selecting from the list of categories, participants had to select the most important category to represent their open responses. In Experiment 7, we provide a method by which participants may select more than one cue category after providing their written responses and indicate the order of importance of each category they select. This better resembles the process by which the open text responses are coded. Finally, Experiment 7 involves preregistering previously exploratory analyses.

3.4 Experiment 7

3.4.1 Method

3.4.1.1 Participants, Stimuli, Design, and Procedure

We aimed for a cell size of at least $n = 100$ ($N = 400$) after exclusions, which was achieved. A post hoc analysis using this cell size and the theoretically-derived effect size of $\eta^2_g = .03$ demonstrated that we had .94 power to detect the 2x2 between-subjects interaction between stimulus type and context reported below. The same participant pool used for Experiments 1 and 2 was used for Experiment 7. As in Experiment 6, participants were asked to provide a retrospective judgment of effort. Additionally, to allow for more accurate comparison between the human coding of written responses and the self-selected categories, participants were invited to choose as many categories as they wanted to represent their written reason and were asked to rank their selections from most to least important. After the ranking was completed, participants who encountered a rotated stimulus were asked to estimate the number of degrees by which they thought it was rotated by entering a number in the provided field. Participants who did not encounter a rotated stimulus were instructed to place a "0" in the field.

3.4.2 Results

A total of 465 participants provided consent and participated in the study. Altogether, 19 participants' data were excluded[14] from analysis. Specifically, four participants took a disproportionately long time to read one display (e.g., over 3 minutes); one participant took 40 minutes to complete the study; 13 participants selected an incorrect response to the attention check question; and one observation was accidentally deleted due to experimenter error. Data provided by the remaining 446 participants (with a mean age of 39.06 and 263 female, 181 male, and 2 preferring not to answer) was included in the analysis. This study was preregistered; the preregistration, data, and analysis script are available on the Open

[14] There were 7, 3, 3, and 6 observations in the pure/upright, pure/rotated, mixed/upright, and mixed/rotated conditions, respectively.

Science Framework: https://osf.io/8c3z6/. All analyses were conducted using R, version 4.3.2.

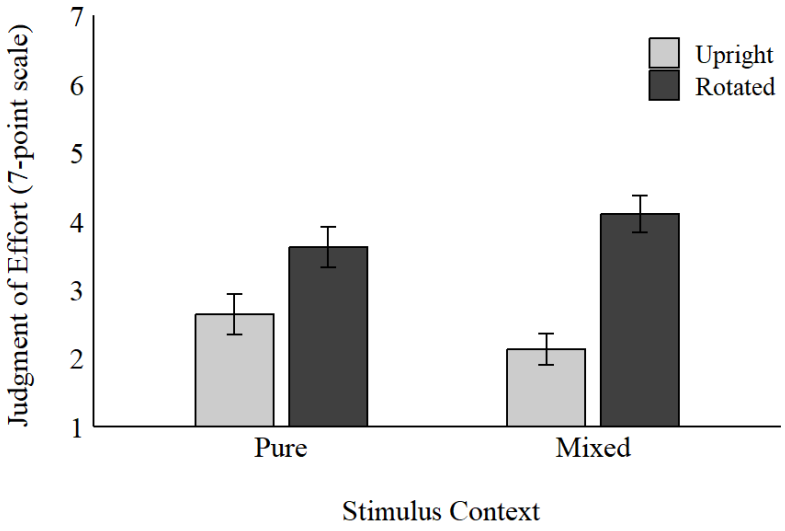
3.4.2.1 Subsets of Data

Of the remaining 446 participants, 34 noted that they had to scroll or resize their screen while reading a display; 30 participants had an average reading time of less than six seconds; and 22 participants likely misunderstood the task based on their open responses. Analyses computed excluding each of these subsets revealed results consistent with the results from the main analysis (N = 446), which are reported below.

3.4.2.2 Judgments of Effort

There were 105, 115, 117, and 109 observations in the upright pure, upright mixed, rotated pure, and rotated mixed conditions, respectively. A 2 (stimulus type: upright, rotated) x 2 (stimulus context: pure, mixed) between-subjects ANOVA with effort judgements as the output variable was computed. A significant interaction was revealed, $F(1, 442) = 12.99$, $p < .001$, $\eta^2_g = .03$, such that the difference between the upright and rotated stimulus type was larger in the mixed stimulus context ($M_{\text{rotated-upright}} = 1.98$), compared to the pure stimulus context ($M_{\text{rotated-upright}} = 0.98$). With respect to main effects, the stimulus context was not significant, $F < 1$; and the rotated stimulus type was judged as more effortful than the upright stimulus type, $F(1, 442) = 113.14$, $p < .001$, $\eta^2_g = .20$. To better understand the interaction, simple effects comparisons were computed, which demonstrated that effort judgments in the mixed context were significantly lower than in the pure context for the upright stimulus, $t(201.1) = 2.72$, $p = .007$, $d = 0.37$; and judgments in the mixed context were significantly higher than in the pure context for the rotated stimulus, $t(222.43) = 2.39$, $p = .018$, $d = 0.32$. Moreover, there was a significant rotation effect in both the pure context, $t(219.68) = 4.58$, $p < .001$, $d = 0.61$, and in the mixed context, $t(215.7) = 11.10$, $p < .001$, $d = 1.49$. Figure 18 is a visualization of these data.

Figure 18. *Effort judgments as a function of stimulus type and stimulus context in Experiment 7*



Note. Error bars represent 95% CI

3.4.2.3 Participants' Reason for Judgment

To determine whether stimulus type, stimulus context, and their interaction influenced the selection of a specific reason, logistic regression was employed. For this analysis, we removed 22 observations corresponding to individuals who likely misunderstood the task, 14 where individuals did not rank their selections, 32 who had ties for a most important reason, and 5 who selected "Nothing specific about the task," as well as one or more task-specific categories. About 95% of participants' selections involved a specific reason for judgment. Neither the main effect of stimulus type, $|z| < 1$; the main effect of stimulus context, $|z| < 1$; or the interaction, $B = 1.18$, $SE = 0.97$, $p = .223$, were significant.

With respect to the subset of observations corresponding to a specific reason for judgment, logistic regression was employed to determine the effect of stimulus type and stimulus context on the likelihood of selecting a structure-based reason as the most important reason for judgment (recall that participants were able to select multiple reasons in this experiment)[15]. After removal of the interaction term ($|z| < 1$), the main effect of stimulus type was significant, $B = 1.50$, $SE = 0.24$, $p < .001$, such that when the stimulus is rotated, the odds of selecting a structure-based reason increase about 4.5 times from when the stimulus is upright. The main effect of stimulus context was also significant, $B = 0.75$, $SE = 0.24$, $p = .001$, such that, when in the mixed stimulus context, the odds of selecting a structure-based reason increases about 2 times from when in the pure stimulus context. Results based on the open text responses[16] were consistent with the above reports.

3.4.2.4 Structure-based vs. Not Structure-based

The data were partitioned into judgments made by participants who selected the structure category to represent their most important or only reason, and judgments for which the structure category was a secondary selection or not at all selected. The pattern of judgments appeared to differ across stimulus contexts, as visualized in Figure 19. A 3-way ANOVA testing the effects of stimulus type, stimulus context, and reason (i.e., whether a structure-based reason was ranked as most important) demonstrated the expected main effect of stimulus type, $F(1, 344) = 72.45$, $p < .001$, $\eta^2_g = .17$, while the main effects of stimulus context and reason were not significant, Fs < 1. There was a significant interaction between stimulus type and stimulus context, $F(1, 344) = 11.73$, $p = .001$, $\eta^2_g = .03$; and a significant interaction between stimulus type and reason, $F(1, 344) = 9.20$, $p = .003$, $\eta^2_g = .03$; and the stimulus context by reason interaction was not significant, $F(1, 344) = 3.49$, $p = .062$, $\eta^2_g =$

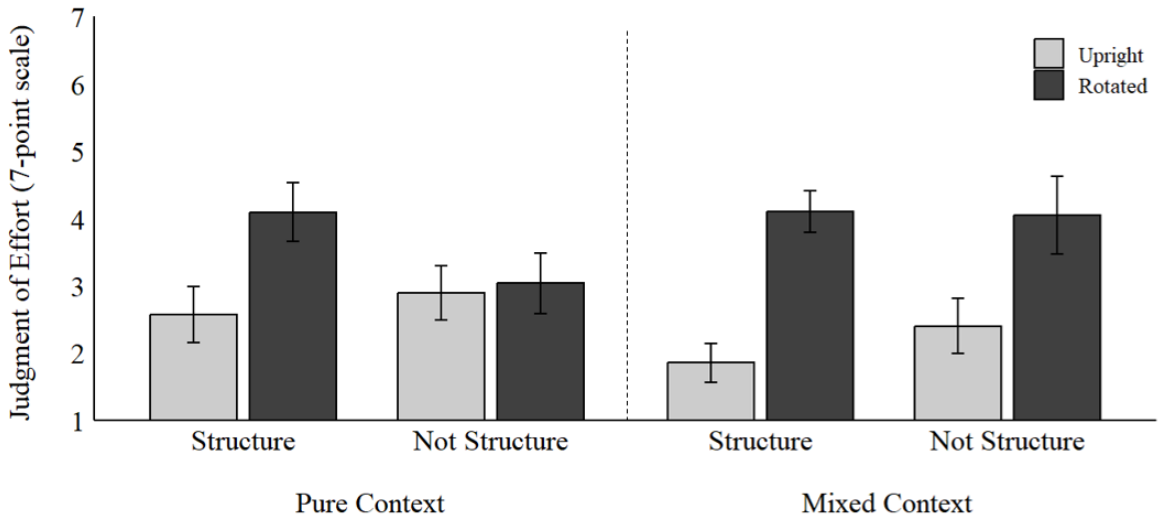
[15] Regression models with a modified criterion variable (structure was selected, regardless of importance vs. structure was not selected) were consistent with the reported results.

[16] Coders agreed on 94% of the specific/not specific code assignments. Of the responses that were specific to the task, coders agreed on 80% of the primary specific reason assigned and this increased to 93% when considering whether the coders had at least one specific reason in common between their primary and secondary reasons.

.01. Finally, the three-way interaction was not significant, $F(1, 344) = 1.38$, $p = .242$, $\eta^2_g < .01$.

Focusing on the stimulus type by reason interaction, pairwise comparisons revealed a significant difference between the rotated and the upright stimulus ($M_{rotated\text{-}upright} = 2.03$) when the structure category was ranked first, $t(179.97) = 11.16$, $p < .001$, $d = 1.50$. This difference ($M_{rotated\text{-}upright} = 0.83$) remained significant when the structure category was either not ranked first or not selected, $t(101.14) = 3.35$, $p = .001$, $d = 0.56$.

Figure 19. *Judgments of effort as a function of stimulus type, context, and reason for Experiment 7*



Note. Error bars represent 95% CI. These data do not include observations associated with a non-specific reason or a likely misunderstanding of the task.

Additionally, visual inspection of Figure 19 suggests that the rotation effect may be larger in the mixed context, compared to the pure context, both for individuals who primarily endorsed a structure-based reason and individuals who primarily endorsed a non-structure-based reason. Despite the observation of a non-significant three-way interaction here and in previous experiments, it seemed sensible to conduct separate 2 (stimulus type) x 2 (stimulus context) between-subjects ANOVAs at each level of reason, as in previous experiments. When analyzing data associated with primarily endorsing a structure-based reason, there was a significant main effect of stimulus type, $F(1, 194) = 78.49$, $p < .001$, $\eta^2_g = .29$ (i.e., the

rotated stimulus type was judged as more effortful); a non-significant main effect of stimulus context, $F(1, 194) = 2.73$, p = .010, $\eta^2_g = .01$; and the interaction was not significant, $F(1, 194) = 2.98$, $p = .086$, $\eta^2_g = .02$. When analyzing data associated with not primarily endorsing a structure-based reason, there was a significant effect of stimulus type, $F(1, 150) = 12.71$, $p < .001$, $\eta^2_g = .08$ (in the expected direction); a non-significant effect of stimulus context, $F(1, 150) = 1.06$, $p = .305$, $\eta^2_g = .02$; and a non-significant interaction, $F(1, 133) = 3.09$, $p = .081$, $\eta^2_g = .02$. Due to participants' ability to select multiple categories and rank their selections, it was also possible to interpret the reason variable as distinguishing between whether the structure category was selected (regardless of importance). Under this interpretation, an analysis of the data associated with endorsing a structure-based reason achieved significance, $F(1, 217) = 4.97$, $p = .027$, $\eta^2_g = .02$. Main effects for this analysis were qualitatively consistent with those reported above. Finally, when the data associated with not endorsing a structure-based reason was analyzed, the results were consistent with those reported above.

3.4.2.5 Number of Cues Endorsed

A chi-squared test of independence was computed to determine whether there was an effect of stimulus context on the proportion of observations associated with one vs. two categories. The proportion of individuals who selected one, vs. two, categories was .75 in the pure context and .83 in the mixed context. An analysis of participants' selections demonstrated the effect of context did not reach significance, $\chi^2(1, N = 351) = 3.19$, $p = .074$. Similarly, the proportion of open text responses corresponding to one vs. two cues was .58 in the pure context and .69 in the mixed context, $\chi^2(1, N = 324) = 3.51$, $p = .061$. When interpreting the data such that multiple cues could exist per category, the proportion of single cue responses was .61 and .75 in the pure and mixed context, respectively, which represented a significant difference, $\chi^2(1, N = 345) = 7.15$, $p = .007$.

3.4.2.6 Estimate of Rotation

This analysis included the same data as the previous sections ($N = 352$); however, 77 participants did not experience a rotated stimulus, so their data were excluded. Moreover, 21

participants who did encounter a rotated display entered degree estimates of over 90, suggesting a misunderstanding of this task and these data were also excluded. A 2 (stimulus context) x 2 (reason: "structure" as primary category vs. "structure" as secondary category or not selected) between-subjects ANOVA with rotation estimate as the outcome variable revealed no effects of context or reason on participants' estimates, all $F(1, 250) < 1$. Regardless of condition, individuals underestimated the rotated stimuli to be at about 45 degrees from upright (as summarized in Table 4). When employing the modified reason variable (i.e., the structure category was selected, regardless of importance vs. the structure category was not selected), the results were consistent with the above.

Table 7. *Means, standard deviations, and cell sizes for rotation estimates in degrees split by primary reason and stimulus context*

	M	*SD*	*n*
Structure-based			
Pure	46.80	15.61	50
Mixed	46.02	10.52	115
Non-structure-based			
Pure	47.29	15.46	24
Mixed	45.18	9.36	65

Note. Observations from participants who likely misunderstood the task have been excluded. Estimates are in degrees.

3.4.3 Discussion

The expected interaction between stimulus context and stimulus type was again observed, as was the greater likelihood of individuals selecting a structure-based reason for judgment in the mixed stimulus context. Consistent with Experiment 6, individuals who did not endorse a structure-based reason still judged the rotated stimulus as more effortful, but only in the mixed context. There was also a greater difference between the upright and rotated judgments amongst individuals endorsing a structure-based reason in the mixed context, though this effect sometimes did not reach significance. It seems that stimulus rotation is

influencing the judgment process in the mixed context, even for those who do not (primarily) base their judgments on a structure-based cue. As discussed above, one explanation of this observation is that the mixed context is generating a type of fluency signal, which may be unconsciously influencing judgments. Additionally, but not exclusively, it could be that individuals weight a structure-based cue differently based on context; or conceptualize the structure category differently depending on the context. We tested one instance of such an account, namely that experiencing both an upright and a rotated stimulus makes the rotated stimulus appear more rotated. Notably, average estimates of degrees rotated did not differ across conditions, which indicates that the context in which the stimuli were read did not have a differential effect on how tilted the rotated stimuli appeared. Thus, the greater difference between upright and rotated stimuli in the mixed context for those endorsing a structure-based cue is unlikely to be due to a perceptual difference informing judgments of effort.

With respect to the number of cues used to infer judgment, there was some discrepancy in terms of significant vs. marginal effects between participants' selected reasons and the open text responses. That said the overall picture appears to be that more individuals selected a single cue (compared to two cues) in the mixed context, as compared to the pure context and this is consistent with Experiment 6. One possible explanation for the observation that more individuals rely on a single cue when in the mixed context could be that the increase in the saliency of structure-based cues in the mixed context results in other cues becoming comparatively less important to participants when reflecting on their reasons for judgment. That is, a salient cue crowds out others.

3.5 General Discussion

Across three experiments, we demonstrated that judgments of effort are sensitive to the stimulus context even when individuals make their effort judgment in a separate evaluation context. Specifically, the interaction between stimulus context and stimulus type indicated a larger difference for upright vs. rotated stimuli in the mixed context, as compared to the pure context. In all experiments, a significant rotation effect (i.e., the rotated stimulus was judged

as more effortful than the upright stimulus) was observed in the mixed stimulus context. Together with Chapter 2, this suggests that effort judgements are sensitive to stimulus context even when a single judgment is made. Moreover, in Experiments 6 and 7 the rotation effect was also significant in the pure stimulus context. Additionally, individuals endorsed a structure-based cue more frequently in the mixed context, consistent with the idea that the difference in judgments across stimulus contexts is related to a shift in the cues used to make these judgments. Upon a closer examination of the relation between individuals' judgments of effort as a function of participants' reported reasons (i.e., interpreted as a structure-based vs. not structure-based cue) as well as stimulus context and stimulus type, we found that, in general, individuals who endorsed a structure-based cue judged the rotated stimulus as more effortful than the upright stimulus. This provided some reassurance that participants who endorse a structure-based cue are actually inferring effort at least partly by appealing to the structure of the stimuli. Furthermore, in Experiments 6 and 7, while the unbalanced design makes it difficult to interpret formal analyses, it seemed as if the rotation effect was changing to some degree across contexts. Specifically, the rotation effect was descriptively greater in the mixed stimulus context for those who endorsed a structure-based cue as well as for those who did *not* endorse a structure-based cue. This difference was also characterized by a larger rotation effect in the mixed condition. Lastly, an examination of the number of cues used in the pure vs. mixed stimulus context revealed some inconsistency across experiments. Specifically, in Experiment 5 there was no significant difference between contexts in terms of proportion of single vs. two cue responses. In Experiment 6, significantly more individuals in the mixed context endorsed a single cue vs. two cues, as compared to the pure context, a pattern similar to that demonstrated in Experiment 7. That said, a combined analysis across all experiments did not yield a significant effect of context, $\chi^2(1, N = 901) = 2.77$, $p = .096$. The influence of stimulus context on the number of cues used deserves further scrutiny, as does the discrepancy between results across experiments.

We began this investigation with the idea that individuals' judgments of effort may change across pure versus mixed stimulus contexts and that this would be associated with a

shift in the cues that individuals report relying on to inform their effort judgement. As discussed above, the results provide supporting evidence for this idea. This shift could be precipitated by individuals in the mixed context having the opportunity to attend to a direct comparison between their experience reading upright and rotated stimuli. This account is limited by the observation that the larger difference between upright and rotated stimuli in the mixed context for those who endorsed structure in Experiments 6 and 7 cannot be solely explained by the idea that a shift in explicitly reportable cues (e.g., structure-based cues) is the driver of individuals' effort judgments. Focusing on the individuals who primarily endorse a structure-based cue, we observed a larger rotation effect in the mixed context as compared to the pure context, an observation that was marginally significant or significant depending on how the reason variable was formed in Experiments 6 and 7. As previously discussed, it could be that some of the participants made their effort judgment, at least in part, based on their memory of a sheer fluency experience. More specifically, given the presence of highly distinguishable stimuli (i.e., upright, rotated) in the mixed context, and the understanding that fluency experiences can influence metacognitive judgments more readily when there are such salient distinctions, it is sensible to postulate that this experiential or feeling-based cue influenced judgments more so in the mixed context. In the Introduction, we discussed that fluency experiences are thought to be constructed in a relative fashion (even without an apparent comparison; Greifeneder & Unkelbach, 2013), and that the ability to experience two different stimuli in the same reading session arguably makes the difference between the stimuli more salient. In other words, in the pure context, individuals will still unconsciously appeal to a fluency signal as part of their judgment process, but in the mixed context this signal may be amplified by the direct contrast between the stimuli. Further support for this idea comes from the observation of a rotation effect in the mixed context for those who did not endorse any cue related to the structure of the stimuli. If a relatively stronger fluency signal is detected by virtue of the mixed context, then individuals may not be able to report their use of it (Koriat, 2000; Metcalfe, 2000). As such, individuals in the mixed context may have used a cue unrelated to the structure of the stimuli, but part of their effort judgment could have depended on an unconscious inference based on sheer experience

that was sensitive to the orientation of the stimulus. Importantly, across all experiments, judgments in the pure stimulus context from those who reported using a non-structure-based cue exhibited negligible rotation effects.

There is an aspect of these results that are inconsistent with previous research. Namely, the rotation effect in the pure stimulus context seems rather robust, reaching significance in Experiments 6 and 7. The effect size associated with a combined analysis across all experiments is $d = 0.48$ ($M_{upright} = 2.63$, $M_{rotated} = 3.41$) On the other hand, in Chapter 2, a significant difference between the rotated and the upright stimulus was not consistently reported in a similar pure condition. A combined analysis with data from Experiments 1-4 ($N = 1348$) demonstrated a significant rotation effect in the pure condition but with a smaller effect size, $t(1344.70) = 3.72$, $p < .001$, $d = 0.20$ ($M_{upright} = 3.24$, $M_{rotated} = 3.58$). Thus, as discussed in Chapter 2, there was some sensitivity to rotation in the pure context; however, the effect was smaller. One explanation for this inconsistency is a methodological change made in the current study, involving a change to the labels on the 7-point effort scale. As mentioned in the Methods, we used a scale with a minimum (1 = *Required little effort*) and maximum (*7 = Required a great deal of effort*) effort label. By contrast, in Chapter 2, the end labels were 1 = *Not at all Effortful* and *7 = Very Effortful*. Individuals employing the scale with the latter labelling may have felt reluctant to quantify their effort expenditure as "Not at all effortful," as several trials of reading random words may require a non-zero amount of effort. Individuals in the current study may have been comparatively more comfortable selecting the low end of the effort scale to represent a very easy task due to the less strict label. This might explain why the upright stimulus is judged as less effortful on average in the current study within the pure stimulus context.

3.6 Conclusion

The present experiments provide evidence that a stimulus context can influence effort judgments; and that, via explicit cue reports, individuals may do so because the stimulus context leads to a shift in the cues individuals use to make their effort judgements. That said, the experiments also revealed an influence of stimulus context that was not seemingly

captured by these reports. Together, these results further our understanding of cue-use in effort judgments and provide a deeper understanding of the effort construct.

Chapter 4
General Discussion

4.1 Summary of Main Results

Taken together, Chapters 2 and 3 contribute to our understanding of judgments of effort, viewed as an inferential process. Beginning with the pattern of judgments of effort across contexts, this work has demonstrated that judgments of effort can be markedly influenced by evaluation context. In particular, judgements made in a joint evaluation context showed a clear differentiation across stimulus types; however, in the separate evaluation context, the stimulus type manipulation appeared to have negligible influence on judgments of effort. This pattern of judgments did not seem more in line with the associated pattern of objective demand measures as originally hypothesized. Nonetheless, there was a clear influence of evaluation context on effort judgements. When separating the stimulus context (i.e., pure vs. mixed stimulus context) from the judgment context (i.e., joint judgment context vs. separate judgment context) in Chapter 3, it was demonstrated that the difference between the upright and rotated stimulus type (i.e., the *rotation effect*) was overall larger in the mixed stimulus context, compared to the pure stimulus context, when eliciting a single judgment. Thus, judgements of effort were again demonstrated to be sensitive to context – in this case, the stimulus context.

Individuals' reasons for judgment also differed across contexts; specifically, there were more structure-based reasons in the joint evaluation context as compared to the separate evaluation context in Chapter 2. Moreover, the reasons in the separate evaluation context exhibited more diversity than those in the joint evaluation context. Turning to stimulus context, there was a greater likelihood that individuals endorsed a structure-based reason in the mixed context, regardless of the stimulus type being judged. Taken together, these results are consistent with the idea that the contextual effects observed on judgments of effort were a product of shifts in the cues individuals were using to infer effort; specifically, evaluation context, and stimulus context, influence the cues used when making such judgments. With

respect to the likelihood of endorsing a reason specific to the task, I consistently observed that over 90% of individuals endorsed a specific reason; and there was no statistical difference in this proportion across contexts. Lastly, in Chapter 3 I also examined the effect of stimulus context as a function of the number of cues endorsed (i.e., one vs. two cues), demonstrating conflicting results across experiments.

Turning to the different elicitation timings (i.e., prospective vs. retrospective), these did not seem to greatly influence the pattern of judgments much beyond the context manipulations. That said, in Chapter 2, retrospective judgments may have exhibited a slightly flatter pattern in the separate evaluation context than did prospective judgments. Moreover, when prospective judgments were also elicited after experience reading, they seemed to produce a pattern that was a bit flatter compared to when there was no experience.

4.2 How do Individuals Make Judgments of Effort?

How does this dissertation inform our understanding of how people make judgments of effort? Judgments of effort are likely made in a similar way to other judgments, which involve a weighting of information (i.e., cues) depending on, among other things, how salient that information is to the judge (Hamilton et al., 1978). In the context of research on probabilistic inference, Wallsten and colleagues (Wallsten, 1976; Wallsten & Barton, 1982) theorize that relevant dimensions of the stimuli are analyzed in series, from most to least salient. Moreover, once dimension values have been determined, they additively contribute to the judge's opinion. While I do not have stake in whether cues are processed strictly in series, or whether the total effect of dimension values combine additively, the idea of multiple cues being processed into dimensions strongly reflects my idea of how judgments of effort are constructed and seems most consistent with the data present in my dissertation.

There are a number of cues that could be available to an individual when judging the effortfulness associated with processing a given stimulus. In a similar vein to Wallsten and colleagues, I propose that from this set of available cues, only a subset of relatively salient cues are selected to contribute to an individual's judgment. While cues themselves are often

described qualitatively (e.g., stimulus colour, stimulus size), each of the selected cues needs to be conceptualized as a dimension, a representation of that cue with an ordering that is often pictured as a number line (e.g., Wallsten & Barton, 1982). For example, if judging the stimuli employed in the present work, and supposing the number of words to read was a relatively salient cue, the dimension associated with this cue might be conceptualized as an ordered list of integers. Moreover, the value associated with a cue must be located on its associated dimension. In the example with the cue of number of words to read, the value would be 25. Once the value of a given cue has been established, it can be mapped onto an associated internal effort dimension, constructed with prior beliefs about effort (e.g., "doing more is harder than doing less"). In the case of the number of words, an internal effort dimension for a task in a 20-minute study might have a left endpoint of 1 word (as the least effortful to read) and a right endpoint of 500 words (as the most effortful to read), leaving 25 words nearer to lowest effort.

At this point, relevant cues have been processed into internal effort values; the judge will also assign a weighting to each. A reasonable assumption, bearing resemblance to Hamilton et al. (1978), is that this weighting assignment represents the relative importance of each of the cues being used. For example, the most important cue of the judge's 'salient cue subset' would have the largest weighting and the least important cue, the smallest. That said, I believe that collecting effort judgments and individuals' weighting of their reasons will provide more insight into how weightings are determined; this is briefly discussed in the next section. Cues and weightings would then be combined, perhaps as a linear combination, resulting in an internal value of subjective effort. Given the measuring tool made available, individuals map this value onto the scale to express their judgment of effort. A simple model of how this occurs, which I currently favour, involves the computed effort value being converted to a percentage, where 0% corresponds to the individuals' low end of the effort dimension and 100% corresponds to its high end. This percentage is then converted to the associated value on the effort scale provided. For example, if an individual determined the effort associated with reading a display was at 50% on their internal effort dimension, it

would correspond to a 4 on the 7-point effort scale. Importantly, effort cues do not necessarily lend themselves to easily be conceptualized as number lines. For example, the dimension associated with the average word frequency of the words in the displays would likely vary more so in its construction across individuals, as compared to the dimension associated with the number of words used. As such, assumptions must be made to formalize this verbal model, which themselves can be tested by building a computational model of the cue inference process described here and comparing simulated data to experimental data. Thus, building a computational model of effort judgments as determined by cue use is a sensible next step in cementing our understanding of this process.

Returning to the summary of findings, judgments were demonstrated to change across contexts, with reported cues shifting in a commensurate manner. More structure-based cues were endorsed in the joint evaluation context and in the mixed stimulus context, which can be interpreted via this model as more individuals selecting structure-based cues as relatively salient when assigned to these conditions. Moreover, judgments in these contexts reflect the relatively high [low] effort associated with reading the disorientated [upright] stimuli, which might result from the determination of a relatively higher value of structure-based cues and/or a relatively greater weighting for those cues.

4.3 Empirical Questions Addressed

4.3.1 Evaluation Context and Measures of Objective Demand

In Chapter 2, I sought to determine whether a separate evaluation context would bring judgments of effort more in line with objective measures. While overall this did not seem to be the case, there could be an argument for judgments being brought more in sync with errors made while reading, as these did not significantly differ across the four stimulus types. Previous work supports the idea that individuals perceive likelihood of making errors as a cue for effortfulness. Specifically, participants were given a choice between a task that takes a lot of time but has a low likelihood of error commission, and a task taking less time but with a higher likelihood of error commission (Dunn, Inzlicht, & Risko, 2019). The authors

demonstrated that individuals more so associated a higher likelihood of error commission, as compared with longer time on task, with greater effort. To further investigate this idea, a reading environment that generally elicits more variation in errors made than those employed in Chapter 2 (e.g., a time limit on reading or use of non-words or words of lower frequency) could shed some light on whether a separate evaluation context brings effort judgments more in line with objective measures. Importantly, individuals' perception/memory of errors made could presumably become a relatively salient cue within such a proposed paradigm. For example, if objective error counts were observed to significantly differ across the UW-UF and UW-RF stimulus types, individuals in the separate evaluation context may notice this difference and use likelihood of error commission as a predominant cue to inform judgment.

As discussed in Chapter 2, previous work (Dunn et al., 2017) demonstrated that the anticipated effort associated with reading rotated vs. upright words seemed relatively inevaluable, compared with other types of effort judgments such as memorizing a given number of items. That is, effort judgments increased with an increase in the number of to-be-remembered items regardless of evaluation context; while the degree of rotation of text only influenced effort judgments in the joint evaluation context. Chapter 2 provides supporting evidence for the idea that this relative inevaluability of rotation could be attributed to the shifting of cues made across evaluation contexts. That is, in addition to judgments, individuals' cues differed starkly across the evaluation contexts; and a greater diversity of cues were endorsed in the separate evaluation context. One reason why the patterns of judgments for the memory task from Dunn et al. (2017) were similar across evaluation contexts could be that a salient cue brought to mind during judgment would be list size, with a relatively common belief that the larger the list, the more effortful it will be to remember every item. Such a pre-experimental belief would presumably be robust to context changes, such that individuals used the same cue (i.e., list size) leading to the same pattern across evaluation contexts in their judgments of effort. One might ask why no similar pre-experimental belief about rotated text seemed to influence effort judgments in the current work. While rotated text is likely judged as more effortful to read than upright text, reading

short, high-frequency words is such a practiced skill that, even with rotation, it may still seem somewhat easy. This would mean that anticipated effort associated with reading the disoriented text might not differ much on the 7-point scale compared to the upright text.

4.3.2 Reasons for Judgment and Cue Use Across Contexts

In a model outlining the process of how effort judgments are made, like the one described above, a crucial component is the predominant cue/cues employed by the judge. This thesis provides one way of examining these cues, namely via the collection of individuals' explicit cue reports. Starting with Chapter 2, I revealed that there was a difference in reasons provided across evaluation contexts. Specifically, more responses addressed structure-related aspects of the stimuli in the joint evaluation context, as compared to the separate evaluation context. Moreover, there was more variation in reasons from individuals in the separate evaluation context, with the proportion of reasons changing over context such that more individuals in the separate evaluation context endorsed reasons pertaining to the words used or the performance requirements of the task, compared to the joint evaluation context. Since cues such as the characteristics of the words and the requirement to read as quickly as possible did not vary across stimulus types, use of such cues would likely combine to form the aggregate pattern of effort in the separate evaluation context. Importantly, even within the separate evaluation context, individuals' proportion of reasons shifted with stimulus type. For example, of those judging the UW-UF stimulus type, more endorsed performance requirements than structure-based reasons; whereas for those judging the RW-RF stimulus type this pattern reversed, with more endorsing structure-based reasons than performance requirements. This suggests that, though individuals may be less likely to spontaneously select the same cue in a separate evaluation context, the reason data still demonstrate a sensitivity to the stimulus type manipulation.

In a similar vein, one question addressed by Chapter 3 was to what extent participants' explicit cue reports capture the pattern of their judgments of effort, thereby testing whether individuals are likely reporting the cues they actually used to make their judgments. For this analysis, the data were partitioned by category of reason endorsed,

namely primarily structure-based vs. primarily not structure-based. I demonstrated that the pattern of judgments seems consistent with the response category. Specifically, judgments by individuals who endorsed a structure-based reason (or reasons) demonstrated a rotation effect in both the pure and the mixed stimulus context. Moreover, judgments by individuals who endorsed a non-structure reason (or reasons) did not demonstrate much of a difference across stimulus types in the pure context. Similar patterns were observed when partitioning the data in different ways; for example, selections from the list of categories vs. human coding of explicit cue reports, or coded responses that included a structure-based reason vs. those that did not. Therefore, it appears that the explicit cue reports capture the general expected pattern of judgments of effort for the stimuli, indicating that individuals are accessing the cues they used to infer their judgments.

In Experiment 7, I tested one instantiation of the idea that reported cue use is the same across stimulus contexts but effortfulness is inferred in distinct ways. Specifically, I addressed the question of whether experiencing both an upright and rotated stimulus makes the rotated stimulus appear more rotated, compared to experiencing the rotated stimulus in isolation. Individuals judged the rotated stimulus as almost equally rotated across stimulus contexts, and there was no influence of whether a structure-related cue was predominantly endorsed. Based on this result and the aforementioned pattern of judgments when controlling for cue used, further investigation into different ways individuals use a given cue across contexts is warranted.

4.3.3 Explicit Cue Reports do not Fully Explain Judgments

An interesting observation with respect to individuals' explicit cue reports was that, most noticeably in Experiments 6 and 7, the rotation effect appeared larger in the mixed stimulus context, compared to the pure context, for individuals who endorsed a structure-based reason. The account of effort judgments developed here is that the context shifts the cues used to make judgments of effort; however, this account does not predict that the mixed context can increase the rotation effect from that in the pure context for those endorsing the same cue. This may point to a limitation of the process of collecting explicit cue reports.

Interestingly, the rotation effect also appeared larger in the mixed context for individuals who did *not* endorse a structure-based reason. It was as if, in the mixed context, these individuals were sensitive to a difference between the stimuli but reported use of cues that did not significantly differ across the stimuli. For example, reporting that the upright stimulus was less effortful to process than the rotated one because of the number of words (but the number of words was the same). As discussed in Chapter 3, fluency (i.e., ease of processing) signals could influence judgments to some extent regardless of an individual's stimulus context condition. However, the direct comparison of upright and rotated stimuli experiences offered by the mixed context would likely enhance signals associated with the difficulty of recognizing rotated words when upright displays are experienced at the same time. If fluency is indeed a more prominent cue in the mixed context, then this might explain the larger effect of rotation in that context. Relatedly, I conducted an exploratory analysis on a subset of data from Chapter 3 representing only those individuals who spent less than an average of six seconds on each trial. Most of these participants spent less than four seconds on average reading each stimulus, so they very likely did not read each word carefully and instead skimmed or simply viewed the display. When examining these data separately, the same judgment pattern emerged as described in Chapter 3, namely the rotation effect was significantly greater in the mixed stimulus context. This raises the question of whether the observed difference in the rotation effect across contexts, controlling for reported cue use, would replicate if individuals were not able to experience the processing of the stimuli at all (i.e., not even read one or two words while scrolling to advance the page, as may have been the case for the excessively speedy participants). One possible design that could shed light on this involves the use of descriptions of the stimuli for participants to read, followed by a judgment elicitation. For example, instructions could include that one imagine they read a page of 25 short, easy words separated by arrows and arranged in a 5 by 5 lattice configuration. The rotation of each word could be verbally described as well as depicted in an image to ensure the format was understood without any words actually being processed. In such a paradigm, controlling for whether a structure-based cue was used, if the rotation effect was observed to be the same across stimulus contexts, this would suggest that the

increase in the rotation effect in the mixed stimulus context was due to a differential appeal to experiential (i.e., fluency) cues across contexts. On the other hand, if the pattern of judgments resembled those reported in Chapter 3, one would conclude that cues based on processing experience were not the main driver of the effect of the mixed context.

Aside from the possibility that individuals' explicit cue reports partly reflect an (unconscious) appeal to fluency cues, the observed pattern across stimulus contexts might be partly due to individuals making inferences based on inaccurate information. For example, one participant indicated that they thought the words in the rotated stimulus were more difficult, as compared to those in the upright stimulus, but selected the structure category because they attended to the rotation of the stimulus. If someone found multiple dimensions for which the rotated stimulus was more effortful than the upright stimulus, then it could result in relatively higher judgments of the rotated stimulus in the mixed context. The same kind of logic could apply to judgments of the upright stimulus in the mixed context; however, it is important to note that not many of the responses seemed based on inaccurate perceptions.

4.4 Limitations and Further Considerations

4.4.1 Employing Explicit Cue Reports

This work has revealed some limitations of employing explicit cue reports. There is the possibility that misunderstandings occurred either within the process of coding individuals' written responses into categories; or individuals' interpretation of the categories they selected from the list. To speak to the latter issue, the results using the coded responses and the selected responses were generally consistent, such that noise generated by such communication issues did not seem to affect the results. The above discussion on differences in the nature of structure-based responses across stimulus context brings to mind the idea that the categories are also relatively broad, leaving room for more interpretations within each category. In the future, an investigation could involve the human and/or artificial categorization of responses into finer-grained categories. For example, the structure-based

category, "…the way the words were organized on the screen," could be partitioned into multiple structure-related concepts, such as "the direction of reading," "the need to tilt one's head," "the angle of the words," "finding each new line in the display," and so on. It could be that across these more targeted categories, the pattern of judgments across stimulus contexts controlling for cues used may appear different, shedding light on the underlying driver(s) of this pattern. Altogether, experiments designed to address issues related to limitations of the current methodology of determining cue use, such as those suggested above, would be a solid next step in learning more about how cue use influences judgments of effort.

Another question regarding cue use across stimulus contexts is whether the number of reported cues differed. As discussed in Chapter 3, the results were inconsistent across experiments. If an effect is emerging from the current investigation, it seems to be that the mixed stimulus context increases the likelihood of individuals endorsing a single cue, compared to two distinct cues. This could be in part due to the mixed context accentuating the structure-based manipulation, such that structure-based cues could be computationally weighted more in individuals' judgment process, compared to the pure context. A natural extension of Experiment 7 could not only involve individuals providing a ranking for their multiple category selections, but also providing a type of numerical weighting for each. All in all, the question of whether the number of cues used is influenced by context changes remains unresolved at this time.

4.4.2 Other Contextual Influences on Cue Use in Judgments of Effort

In Chapter 3, the pattern of judgments in the pure context differed from that in the separate evaluation context in Chapter 2, despite the paradigms being very similar. I posited that this was because the left scale endpoint had changed in a qualitative manner, inviting more individuals to choose it in Chapter 3, as compared to Chapter 2. This difference in judgments across scale types illustrates other aspects of contextual influences on judgments of effort, such as the elicitation format, which includes wording and scale labelling. One example of the elicitation format influencing judgments of effort is asking individuals to provide their level of *required* effort (i.e., effort related to the demands of the task), versus asking for their

invested effort. Recent work has demonstrated measurable differences between these two types of judgments (Van Kessel, Ashburner, & Risko, 2024). Specifically, required effort reports were negatively associated with judgments of learning on a post-lecture comprehension test when controlling for invested effort; however, invested effort was positively associated with judgments of learning, controlling for required effort. Parallel to the current investigation, recent work has demonstrated judgments of effort to be sensitive to other kinds of contextual manipulations (Desjardins et al., 2024; Embrey et al., 2024).

4.4.3 Relative Influence of Stimulus Context vs. Judgment Context

In Chapter 3, I demonstrated that the differences in judgment patterns, as well as the reasons endorsed, held when individuals judged a single stimulus type but experienced reading the stimuli in a pure vs. mixed list format. Therefore, the aspect of the joint evaluation context that invited comparison of the stimulus types during the judgment phase of the study was not the only influential element to the observed judgment pattern. Additionally, evidence from previous work supports the idea that judgments of effort made in a joint judgment context without any experience in a stimulus context closely resemble those made in the joint evaluation context from Chapter 2 (Dunn et al., 2016). Based on this past work and the account developed here, one could predict that judgments made in a joint judgment context would exhibit a greater rotation effect than those made in a separate judgment context, whatever the stimulus context; and that judgments made in a fully joint evaluation context (i.e., mixed stimulus and joint judgment) would exhibit a greater rotation effect compared to judgments made about experience in a mixed stimulus context and in a separate judgment context. Testing this prediction would shed some light on the relative contributions of cues brought to mind during the reading experience on individuals' judgments of effort; and as such would be a valuable future endeavour.

4.4.4 Construction of Individuals' Effort Judgments

The work herein has inspired some considerations in addition to addressing questions about the way in which effort judgments are constructed. Firstly, what is the process by which

individuals map their internal representations of effort onto a provided scale? This was discussed with respect to the flat pattern of judgments in the separate evaluation context, with the proposition that perhaps the act of selecting the effort level on the 7-point scale was influenced by evaluation context. Perhaps the separate evaluation context could lead to difficulties in this process, as discussed in Chapter 2; however, a broader question is whether this mapping process changes substantially across certain contexts for judgments of effort. The account developed here, namely that changes in context can shift the cues used to make judgments of effort, does not directly address the way in which an effort value is determined and mapped onto a scale after prominent cues are attended to. One type of investigation that could shed some light on this process involves selecting a cue that, like an effort scale, has measurable endpoints about which most people would share knowledge. The counterclockwise rotation of a word, for example, can be at 0°, 180°, or any value in between. If individuals judge several different stimuli at different rotations, then their judgments may provide a pattern that indicates how rotation value is translated into a level of effortfulness. For example, individuals could select values that correspond to a linear increase in effort with rotation value, or a relatively non-linear pattern, either of which would provide insight into how the scale is used.

4.4.5 Cues as a Causal Mechanism for Changes in Judgments

A prediction derived from the account developed here is that one should be able to prime the use of specific cues and find associated changes in judgment patterns despite the context being held constant. Such priming may take the form of work done by Mueller and Dunlosky (2017), wherein participants read instructions that focused them on certain cues (i.e., the colour of words in a memory task). For example, in a potential extension to Chapter 3, some participants could be assigned to instructions that centered around the effort required to read rotated text; while others could be assigned to instructions about how longer words or more complex words take more effort to process. Collection of participants' reasons for their judgments in such experiments would function as a manipulation check; while patterns of

judgment are predicted to more so reflect the idea that the rotated stimulus is harder to read than the upright stimulus in the priming condition that centered around rotation.

A related question is whether one can say that individuals' cue selections mediate the relation between context and effort judgments. The central results involve interactions between a context variable and a stimulus type variable, which presents some challenges when building a mediation model that has one independent variable. That said, an analysis with more observations, incorporating the interaction effect would be valuable.

Finally, in the Introduction, I argue that one important reason to understand the drivers of judgments of effort is because individuals' judgments are thought to influence their behaviour. How contextual changes influence behaviour was not directly examined in the current investigation, as the main line of inquiry was about the process underlying judgments of effort. That said, future work could include a choice component in addition to collecting judgments via a scale, allowing a comparison between rating and choice data. Comparing choice data to judgments would be a start towards establishing to what degree individuals' judgments influence their behaviour.

4.5 Conclusion

This dissertation began with the idea that cognitive effort is a ubiquitous construct, both within areas of scholarship and for the general public during day-to-day activities. Moreover, in many cases, judgments of effort can affect subsequent behaviour, such as whether to complete a task or leave it unfinished. As such, work examining the underlying factors that can drive changes in judgments of effort for the same task is an important undertaking. The current investigation addresses several questions central to understanding context effects on effort judgments and introduces methodology to collect data representing the cues individuals use to make their judgments. As discussed in the Introduction, accounts that suggest judgments of effort are based on a direct readout of an experience of effort associated with central resource use can be problematic when considering how context-sensitive judgments of effort can be. The cue-inference account of judgments of effort detailed here,

viewed through the lens of metacognitive judgments, offers an alternative to such 'direct readout' accounts. In summary, this work has furthered our understanding of effort, and contributed to the literature on how cues are utilized when making judgments of effort.

References

American Psychological Association. (n.d.). Metacognition. In *APA dictionary of psychology*. Retrieved July 15, 2024, from dictionary.apa.org/metacognition

Unkelbach, C., & Greifeneder, R. (Eds.). (2012). The Experience of Thinking: How the Fluency of Mental Processes Influences Cognition and Behaviour (1st ed.). Psychology Press. https://doi.org/10.4324/9780203078938

Ashburner, M. & Risko, E. F. (2021). Judgments of effort as a function of post-trial vs. post-task elicitation. *Quarterly Journal of Experimental Psychology, 74*(6), 991-1006.

Ashburner, M., & Risko, E. F. (2022). On the influence of evaluation context on judgments of effort. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 790-811.

Bazerman, M. H., Gino, F., Shu, L. L., & Tsay, C. J. (2011). Joint evaluation as a real-world tool for managing emotional assessments of morality. *Emotion Review, 3*(3), 290-292.

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*(2), 220-240.

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior & Organization, 39*(1), 41-58.

Bazerman, M. H., Schroth, H. A., Shah, P. P., Diekmann, K. A., & Tenbrunsel, A. E. (1994). The inconsistent role of comparison others and procedural justice in reactions to hypothetical job descriptions: Implications for job acceptance decisions. *Organizational Behavior and Human Decision Processes*, *60*(3), 326-352.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55-68.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychological measurement*. Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 401-185.

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., & Abbey, L. (2006). Interest and effort as predictors of reading: A test of the general value principle. *Current Trends in Audience Research*, *19*(2), 2-6.

Blount, S., & Bazerman, M. H. (1996). The inconsistent evaluation of absolute versus comparative payoffs in labor supply and bargaining. *Journal of Economic Behavior & Organization*, *30*(2), 227-240.

Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225-1234.

Boksem, M. A., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain research reviews*, *59*(1), 125-139.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta psychologica*, *197*, 153-165.

Caso, C., & Angeles gil, M. (1988). The Gini-Simpson index of diversity: estimation in the stratified sampling. *Communications in Statistics-Theory and Methods*, *17*(9), 2981-2995.

Chang, Y. W. (2016). Influence of the principle of least effort across disciplines. *Scientometrics*, *106*(3), 1117-1133.

Christopoulos, G., Kokkinaki, F., Harvey, N., & Sevdalis, N. (2011). Paying for no reason? (Mis-) perceptions of product attributes in separate vs. joint product evaluation. *Journal of Economic Psychology*, *32*(5), 857-864.

Couceiro, R., Duarte, G., Durães, J., Castelhano, J., Duarte, C., Teixeira, C., ... & Madeira, H. (2019, May). Biofeedback augmented software engineering: monitoring of programmers' mental effort. In *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)* (pp. 37-40). IEEE.

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E., & Rushworth, M. F. (2009). Effort-based cost–benefit valuation and the human brain. *Journal of Neuroscience*, *29*(14), 4531-4541.

David, L., Biwer, F., Baars, M., Wijnia, L., Paas, F., & De Bruin, A. (2024). The relation between perceived mental effort, monitoring judgments, and learning outcomes: A meta-analysis. *Educational Psychology Review*, *36*:66.

Davidai, S., & Shafir, E. (2018). Are ‘nudges’ getting a fair shot? Joint versus separate evaluation. *Behavioural Public Policy*, 1-19.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*(2), 269-273.

Desjardins, S., Tang, R., Yip, S., Roy, M., & Otto, A. R. (2024). Context effects in cognitive effort evaluation. *Psychonomic Bulletin & Review*, 1-10.

Dunn, T. L., Gaspar, C., & Risko, E. F. (2019). Cue awareness in avoiding effortful control. *Neuropsychologia, 123*, 77-91.

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological Research, 83*(5), 1033-1056.

Dunn, T. L., Koehler, D. J., & Risko, E. F. (2017). Evaluating Effort: Influences of evaluation mode on judgments of task-specific efforts. *Journal of Behavioural Decision Making, 30*(4), 869-888.

Dunn, T. L., Lutes, D. J., & 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., 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.

Embrey, J. R., Mason, A., & Newell, B. R. (2024). Too hard, too easy, or just right? The effects of context on effort and boredom aversion. *Psychonomic Bulletin & Review*, 1-10.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within-and between-subjects manipulation of the independent variable. *Psychological Bulletin, 84*(2), 212-219.

Erlebacher, A. (1978). The analysis of multifactor experiments designed to contrast the within-and between-subjects manipulation of the independent variables. *Behavior Research Methods & Instrumentation, 10*(6), 833-840.

Forster, M., Gerger, G., & Leder, H. (2015). Everything's relative? Relative differences in processing fluency and the effects on liking. *Plos One*, *10*(8), e0135944.

Garbarino, E. C., & Edell, J. A. (1997). Cognitive effort, affect, and choice. *Journal of consumer research*, *24*(2), 147-158.

González-Vallejo, C., & Moran, E. (2001). The evaluability hypothesis revisited: Joint and separate evaluation preference reversal as a function of attribute importance. *Organizational Behavior and Human Decision Processes*, *86*(2), 216-233.

Gopher, D., & Sanders, A. F. (1984). S-Oh-R: Oh stages! Oh resources!. In Prinz, W., Sanders, A. F. (eds) *Cognition and motor processes* (pp. 231-253). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gould, S. J., & Kramer, T. (2009). "What's it Worth to Me?" Three interpretive studies of the relative roles of task-oriented and reflexive processes in separate versus joint value construction. *Journal of economic psychology*, *30*(6), 840-858.

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 behaviour. *Psychological Review, 113*(3), 461-482.

Greifeneder, R., Unkelbach, C. (2013). Experiencing thinking. In R. Greifeneder & C Unkelbach (Eds.), *The Experience of Thinking: How the fluency of mental processes influences cognition and behaviour*. Taylor & Francis Group.

Hamilton, D. L., Fallot, R. D., & Hautaluoma, J. (1978). Information salience and order effects in impression formation. *Personality and Social Psychology Bulletin*, *4*(1), 44-47.

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. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*(2), 107-121.

Hsee, C. K. (2000). Attribute evaluability and its implications for joint-separate evaluation reversals and beyond. In D. Kahneman & A. Tversky (Ed.), *Choices, Values, and Frames* (pp. 543-563). Cambridge University Press.

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., & Tang, J. N. (2007). Sun and water: On a modulus-based measurement of happiness. *Emotion*, *7*(1), 213-218.

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., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making*, *16*(4), 257-272.

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.

Ilan, A. B., & Miller, J. (1994). A violation of pure insertion: mental rotation and choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(3), 520-536.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The Effort Paradox: Effort is both costly and Valued. *Trends in Cognitive Science, 22*(4), 337-349.

Irwin, J. R., Slovic, P., Lichtenstein, S., & McClelland, G. H. (1993). Preference reversals and the measurement of environmental values. *Journal of Risk and Uncertainty*, *6*(1), 5-18.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty*, *9*(1), 5-37.

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. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and cognition*, *9*(2), 149-171.

Koriat, A. (2007). Metacognition and Consciousness. In P. D. Zelazo, M. Moscovitch, E. Thompson (Eds.), *The Cambridge handbook of consciousness*. Cambridge University Press.

Koriat, A. & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478-492.

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*(6), 661-679.

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology, 7*, 67-70.

Liu, B. Q., & Goodhue, D. L. (2012). Two worlds of trust for potential e-commerce users: Humans as cognitive misers. *Information Systems Research*, *23*(4), 1246-1262.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490.

Luna, K., & Albuquerque, P. B. (2022). Do beliefs about font size affect retrospective metamemory judgments in addition to prospective judgments?. *Experimental Psychology*, 69(3), 172-184.

Luong, T., Martin, N., Argelaguet, F., & Lécuyer, A. (2019, March). Studying the mental effort in virtual versus real environments. In *2019 IEEE Conference on virtual reality and 3D user interfaces (VR)* (pp. 809-816). IEEE.

McGregor, M., Azzopardi, L., & Halvey, M. (2021, March). Untangling cost, effort, and load in information seeking and retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval* (pp. 151-161).

Merritt, P.S., Cook, G.I., & Wang, M. (2014). *Erlebacher's method for contrasting the within and between-subjects manipulation of the independent variable using R and SPSS*. Unpublished manuscript.

Metcalfe, J. (2000). Feelings and judgments of knowing: is there a special noetic state?. *Consciousness and cognition*, *9*(2), 178-186.

Morewedge, C. K., Kassam, K. S., Hsee, C. K., & Caruso, E. M. (2009). Duration sensitivity depends on stimulus familiarity. *Journal of Experimental Psychology: General, 138*(2), 177-186.

Mosaly, P. R., Mazur, L. M., Yu, F., Guo, H., Derek, M., Laidlaw, D. H., ... & Mostafa, J. (2018). Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). *International Journal of Human–Computer Interaction, 34*(5), 467-475.

Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language*, *93*, 245-258.

Nowlis, S. M., & Simonson, I. (1997). Attribute–task compatibility as a determinant of consumer preference reversals. *Journal of marketing research, 34*(2), 205-218.

Paas, F., Tuovinen, J. E., Van Merrienboer, J. J., & Aubteen Darabi, A. (2005). A

motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational technology research and development, 53,* 25-34.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American, 219*(6), 84-93.

Potts, C. A., Pastel, S., & Rosenbaum, D. A. (2018). How are cognitive and physical difficulty compared?. *Attention, Perception, & Psychophysics*, *80*(2), 500-511.

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction, 51,* 36-46.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological bulletin*, *137*(1), 131-148.

Richter, M., Gendolla, G. H., & Wright, R. A. (2016). Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. In Elliot, A. J. *Advances in motivation science* (Vol. 3, pp. 149-186). Elsevier.

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.

Ritov, I., & Baron, J. (2011). Joint presentation reduces the effect of emotion on evaluation of public actions. *Cognition and Emotion*, *25*(4), 657-675.

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased?. *Educational Psychology Review*, *32*(4), 1003-1027.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (eds.), The mind in context (pp. 105 -125). New York: Guilford

Schwieder, D. (2016). Low-effort information searching: The heuristic information-seeking toolkit. *Behavioral & Social Sciences Librarian*, *35*(4), 171-187.

Sevdalis, N., Harvey, N., & Yip, M. (2006). Regret triggers inaction inertia–but which regret and how?. *British Journal of Social Psychology*, *45*(4), 839-853.

Song, H., & Schwarz, N. (2008). If It's Hard to Read, It's Hard to Do: Processing Fluency

Affects Effort Prediction and Motivation. *Psychological Science*, *19*(10), 986–988.

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444.

Sun, G., Yao, S., & Carretero, J. A. (2013, September). An Investigation of the Relation between the Complexity of problem Structure and Mental Effort. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 260-264). Sage CA: Los Angeles, CA: SAGE Publications.

Székely, M., & Michael, J. (2021). The sense of effort: A cost-benefit theory of the phenomenology of mental effort. *Review of Philosophy and Psychology*, *12*(4), 889-904.

Tanner, C. (2008). Context Effects in Environmental Judgments: Assimilation and Contrast Effects in Separate and Joint Evaluation Modes 1. *Journal of Applied Social Psychology*, *38*(11), 2759-2786.

Tatz, J. R., & Peynircioğlu, Z. F. (2020). Judgments of learning in context: Backgrounds can both reduce and produce metamemory illusions. *Memory & Cognition*, *48*, 581-595.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237-251.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607-617.

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629-642.

Undorf, M., & Bröder, A. (2021). Metamemory for pictures of naturalistic scenes: Assessment of accuracy and cue utilization. *Memory & cognition*, *49*(7), 1405-1422.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264-1269.

Undorf, M., & Erdfelder, E. (2013). Separation of Encoding Fluency and Item Difficulty Effects on Judgments of Learning. *Quarterly Journal of Experimental Psychology*, *66*(10), 2060-2072.

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A

closer look at the contribution of processing fluency. *Memory & Cognition*, *43*(4), 647-658.

Van Buiten, M., & Keren, G. (2009). Speaker–listener incompatibility: Joint and separate processing in risky choice framing. *Organizational Behavior and Human Decision Processes*, *108*(1), 106-115.

Van Kessel, K., Ashburner, M., & Risko, E. F. (2024). Dissociations between data-driven and goal-driven effort reports: Performance, metacognition, and affect. *Quarterly Journal of Experimental Psychology*, *77*(5), 983-993.

Wallsten, T. S. (1976). Using conjoint-measurement models to investigate a theory about probabilistic information processing. *Journal of Mathematical Psychology*, *14*(2), 144-185.

Wallsten, T. S., & Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(5), 361.

Waters, R. H. (1937). The principle of least effort in learning. *The Journal of General Psychology*, *16*(1), 3-20.

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), 1-23. e68210.

Wickens, C. D. (2024, September). The Multiple Resource Theory and Model. Some Misconceptions in Data Interpretations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (p. 10711813241260740). Sage CA: Los Angeles, CA: SAGE Publications.

Willemsen, M. C., & Keren, G. (2004). The role of negative features in joint and separate evaluation. *Journal of Behavioral Decision Making*, *17*(4), 313-329.

Wu, R., Ferguson, A. M., & Inzlicht, M. (2023). Do humans prefer cognitive effort over doing nothing?. *Journal of Experimental Psychology: General*, *152*(4), 1069-1079.

Yeh, Y. H., & Wickens, C. D. (1984). *The dissociation of subjective measures of mental workload and performance* (No. NASA-CR-176609).

Yeh, Y. Y. & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *The Journal of the Human Factors and Ergonomics Society*, *30*(1), 111-120.

Zawadzka, K., Simkiss, N., & Hanczakowski, M. (2018). Remind me of the context: Memory and metacognition at restudy. *Journal of Memory and Language*, *101*, 1-17.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.