

On the Influence of Evaluation Context on Judgments of Effort

Michelle Ashburner and Evan F. Risko

University of Waterloo

© American Psychology Association, 2022. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: [10.1037/xhp0001026](https://doi.org/10.1037/xhp0001026)

Author Note

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), an Early Researcher Award from the Province of Ontario, and funding from the Canada Research Chairs program to E.F.R. Correspondence concerning this article should be addressed to Michelle Ashburner, 200 University Ave W, Waterloo, ON, N2L 3G1. E-mail: mrmashbu@uwaterloo.ca

Word count: 15 500

Abstract

Cognitive effort is a central construct in our lives, yet our understanding of the processes underlying our judgments of effort are limited. Recent work has suggested that our judgments of effort are sensitive to the context in which they are made (i.e., the judgment context). Using a cognitive task and stimulus set that has produced a reliable dissociation between judgments of effort and cognitive demand (as measured by performance and other indirect measures of demand), we examined whether evaluation context might be able to eliminate this dissociation (i.e., bring judgments of effort more in line with measures of cognitive demand). To address this question, we conducted four experiments manipulating a number of aspects of the judgment context including, principally, a manipulation of joint vs. separate evaluation; whether the judgment was prospective, or retrospective; and whether prospective judgments were made with or without having experienced the cognitive task. Additionally, we collected objective demand measures and examined participants' self-reported reasons for their judgments of effort across the joint and separate evaluation contexts. Results demonstrated that evaluation context has a marked effect on judgments of effort; however, no judgment context appeared to yield a pattern more similar to what is found using measures of cognitive demand. Moreover, the reasons individuals cited for their judgments varied across evaluation contexts. Implications of the present work for our understanding of judgments of effort are discussed.

Keywords: Effort; Cognitive Effort; Judgment Context; Evaluation Mode; Evaluation Context;

Metacognitive Judgments

Public Significance Statement: This study highlights the influence of context on judgments we make about our effort expenditures.

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.

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 (GET). According to GET, 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 GET, 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.

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, Lutes, & Risko, 2016; Dunn & Risko, 2016, 2019). 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; see 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 (see 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.

Experiments 1a and 1b

To establish whether judgments of effort for the stimulus types from Dunn et al. (2016; see 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.

Method

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.

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. See Figure 1 for a visualization of the stimuli in the separate and joint evaluation modes.

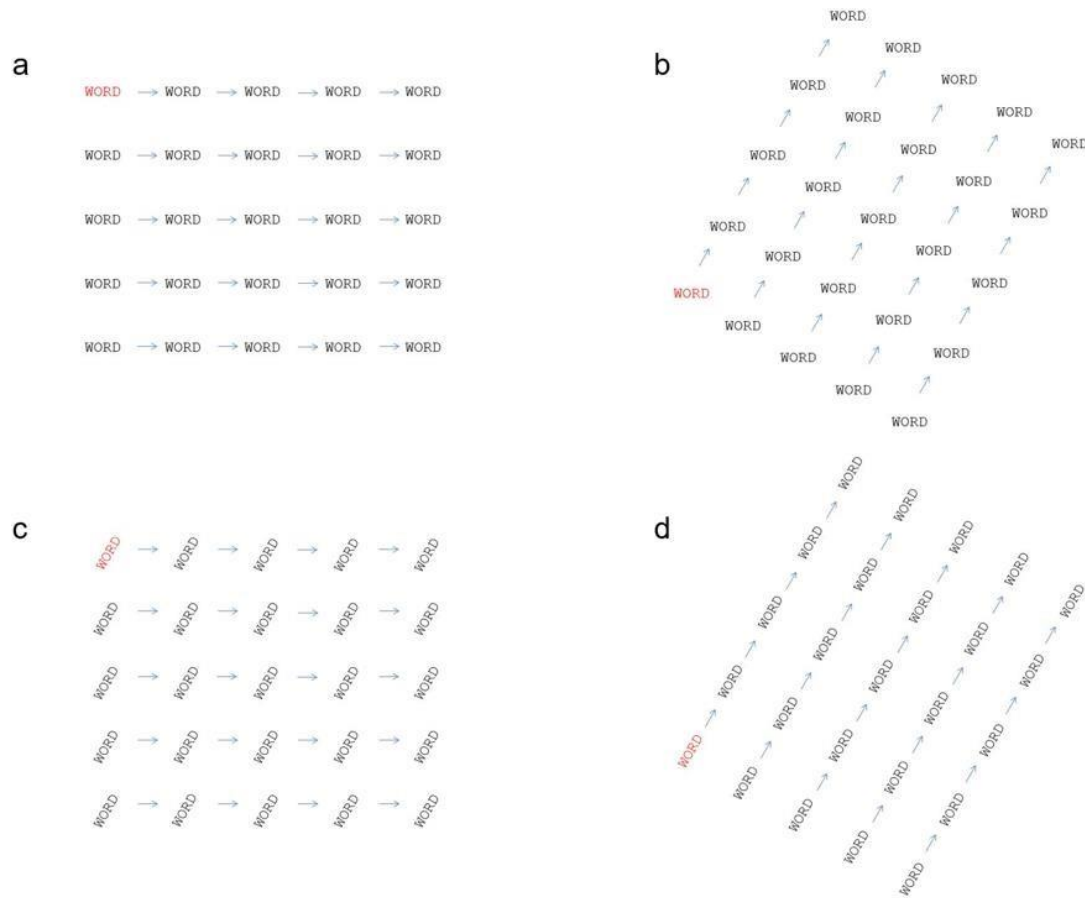


Figure 1. Example of stimuli used in Experiments 1 and 2: a. UW-UF; b. UW-RF; c. RW-UF; d. RW-RF, clockwise counterbalance. For the JE condition, participants were shown all four displays as above, and provided four corresponding prospective effort judgments. For the SE condition, participants were shown exactly one of these displays and provided a single effort judgment.

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 (see Figure 1 for one such counterbalance); 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 SE 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 JE condition were identical to those in the SE 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 JE condition was also identical to that in the SE 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 JE and SE condition, the attention check question asked participants to read ten statements about the task they just completed and select the one false statement (see the Supplementary Materials section). Following this, participants were provided written feedback on the study and a link to receive payment.

Results

Experiment 1a

Five hundred and eighty-eight 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*

	<i>Experiment 1a</i>	<i>Experiment 1b</i>	<i>Experiment 2a</i>	<i>Experiment 2b</i>	<i>Experiment 3</i>
<i>Joint Evaluation</i>					
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)
<i>Separate Evaluation</i>					
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.

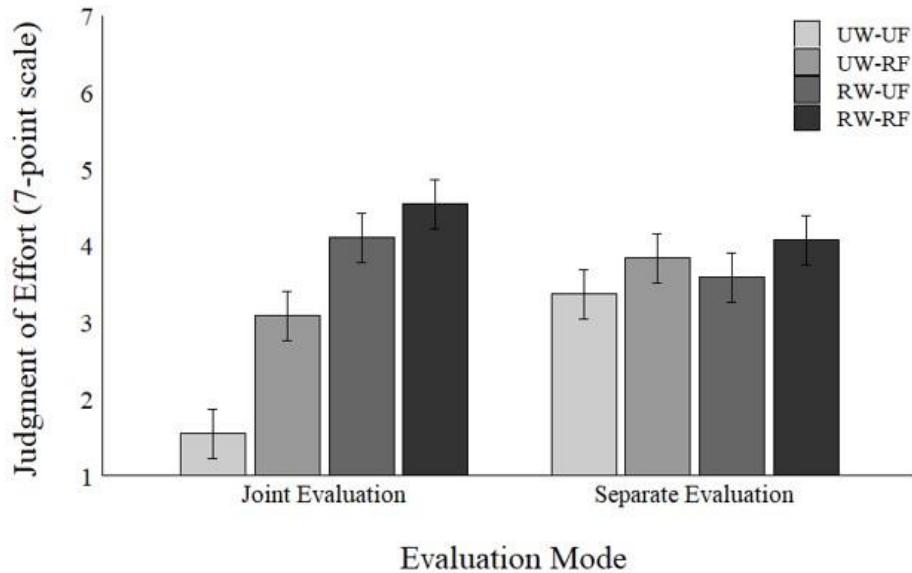


Figure 2. Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1a. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

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 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.

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.

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.

Judgments of Effort in the Joint Evaluation Mode. A single factor repeated measures ANOVA revealed a significant effect of stimulus type, $F(2.41, 161.74) = 66.38, p < .001, \eta^2 = .34$. 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, $t(67) = 2.38, p = .020, d = 0.26$.

Judgments of Effort in the Separate Evaluation Mode. A single factor betweenparticipants ANOVA revealed no significant effect of stimulus type, $F(3, 268) = 2.12, p = .098, \eta^2 = .02$. See Figure 2 for a visualization of the data. When the full dataset was analyzed ($n = 108$ per cell), there was a significant effect of stimulus type, $F(3, 428) = 7.07, p < .001, \eta^2 = .05$; 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.

² Cohen's d computed using the `effsize::cohen.d` function.

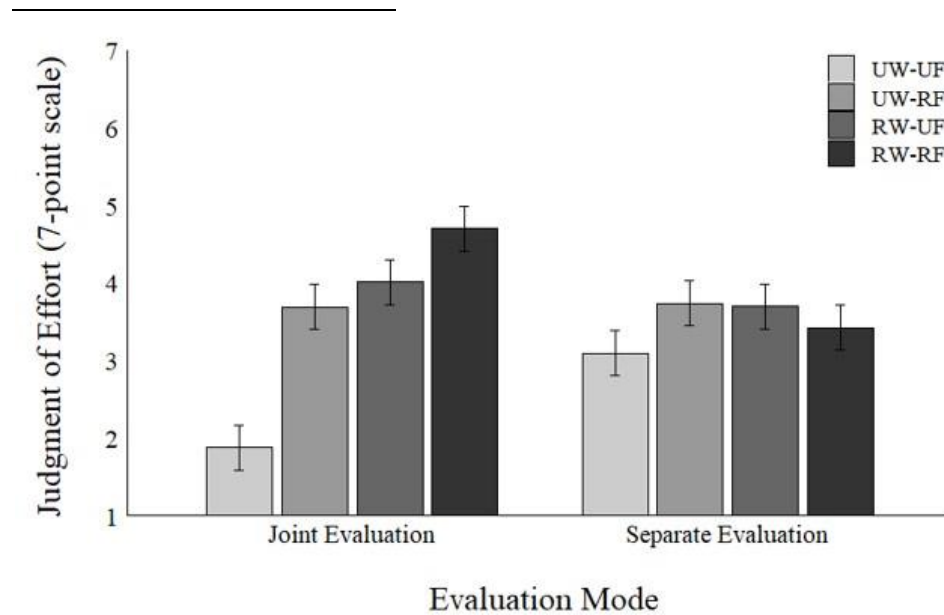
Experiment 1b

Figure 3. Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 1b. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

Seven hundred and ninety-four 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.

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 ($M_{JE} = 3.57, SD_{JE} = 1.94$) was not significantly different than the mean judgment of effort in the separate evaluation mode ($M_{SE} = 3.49, SD_{SE} = 1.67$). 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.

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 ($M_{UW-RF} = 3.69, SD_{UW-RF} = 1.75$) and the RW-UF ($M_{RW-UF} = 4.01, SD_{RW-UF} = 1.70$) 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, $t(86) = 3.57, p = .001, d = 0.41$.

Judgments of Effort in the Separate Evaluation Mode. A single factor betweenparticipants 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 UWUF 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.

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 UWRF 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 GET 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).

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.

Method

Participants

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

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). See Figure 4 for 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. Therefore, one participant in the separate evaluation condition might read a clockwise-rotated RW-UF display starting with the word “EAGLE”, as in Figure 4a. 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.

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, a probing question was included immediately after judgments of effort were reported in Experiment 2b only. 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.³

³ An exploratory analysis was conducted on these data as part of the development of a coding scheme.

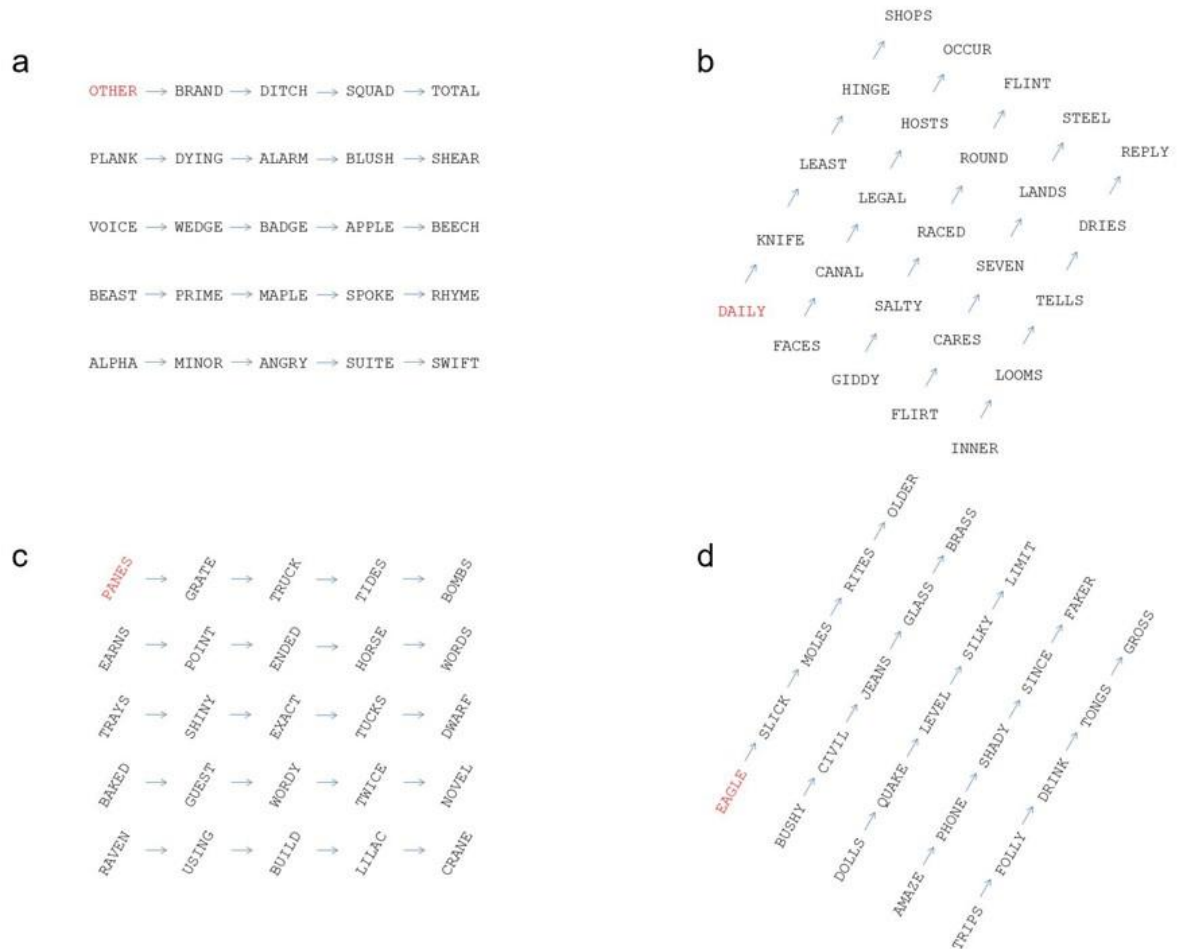


Figure 4. Example of stimuli used in Experiments 3 and 4: a. UW-UF “Other”; b. UW-RF “Daily”; c. RW-UF “Paness”; d. RW-RF “Eagle”, clockwise counterbalance. For the JE condition, participants were shown all four displays in series. For the example above, the order was: a, b, c, d. An effort scale was provided below each display such that participants provided four effort judgments in total. For the SE condition, participants were shown exactly one display and provided a single effort judgment.

Results

Experiment 2a

Seven hundred and thirty-six 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 ($N = 579$).

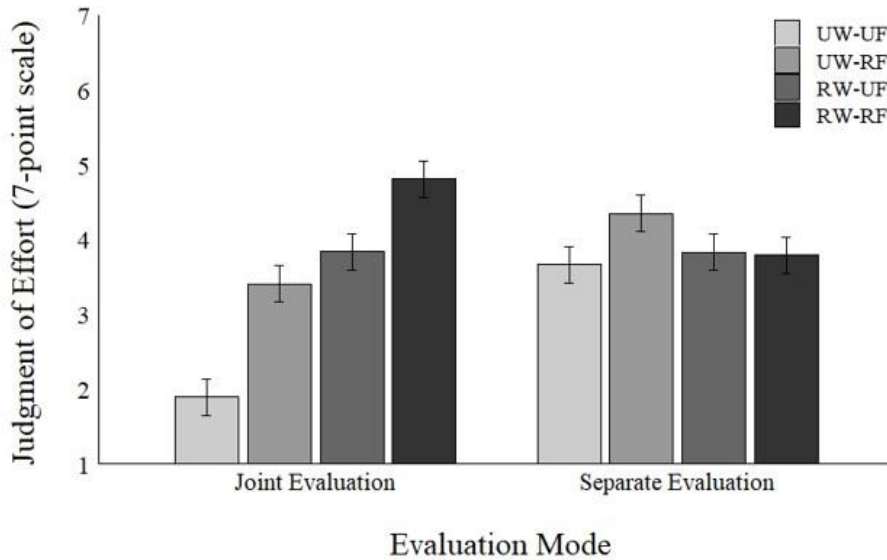


Figure 5. Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2a. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

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.

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, $F(2.70, 280.80) = 96.65, p < .001, \eta^2_g = .27$. 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, $t(104) = 6.54, p < .001, d = 0.56$.

Judgments of Effort in the Separate Evaluation Mode. A single factor betweenparticipants ANOVA, with 105 observations per condition, revealed a significant effect of stimulus type, $F(3, 416) = 3.03, p = .029, \eta^2_g = .02$. Pairwise comparisons revealed that the UWUF 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.

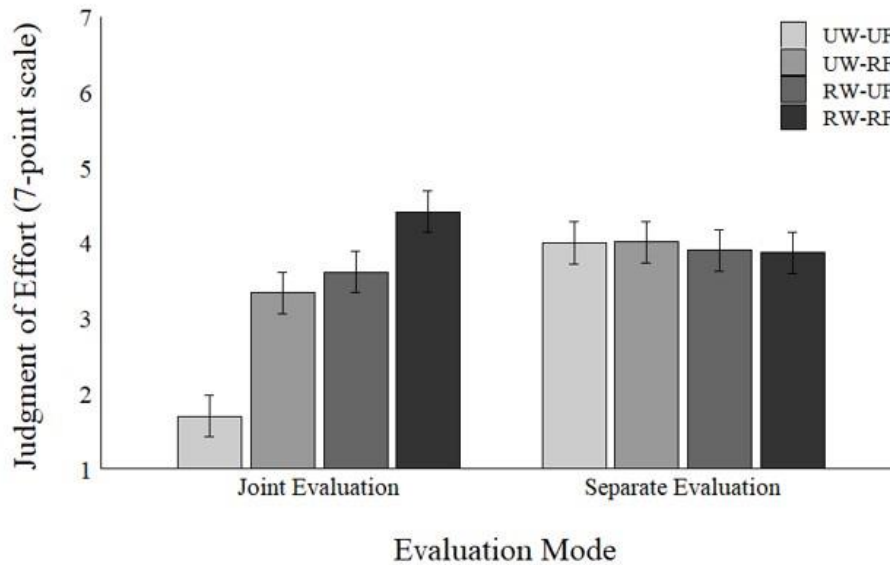
Experiment 2b

Figure 6. Mean judgments of effort as a function of stimulus type grouped by evaluation mode for Experiment 2b. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

Seven hundred and forty-three 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.

Erlebacher's Analysis. Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) $\times$ 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.

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, $F(2.77, 227.06) = 66.43, p < .001, \eta^2 = .25$. 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, $t(82) = 4.17, p < .001, d = 0.46$.

Judgments of Effort in the Separate Evaluation Mode. A single factor betweenparticipants ANOVA, with 105 observations per condition, indicated no difference across the stimulus types, $F(3, 328) < 1, \eta^2 < .01$. When the full dataset was analyzed, a significant effect was revealed, $F(3, 584) = 3.19, p = .023, \eta^2 = .02$. 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.

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.

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.

Method

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 $5(100) \times 2 = 1000$ 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 (see 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.⁴ 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).

⁴ An exploratory analysis was conducted on these data as part of the development of a coding scheme.

Results

One thousand and seventeen 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.

Erlebacher's Analysis – Omnibus

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.

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, $F(1, 336) < 1$.

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, $F(3, 336) = 1.16$, $p = .324$, $\eta_g^2 = .01$. See Figure 7 for a visualization of these data.

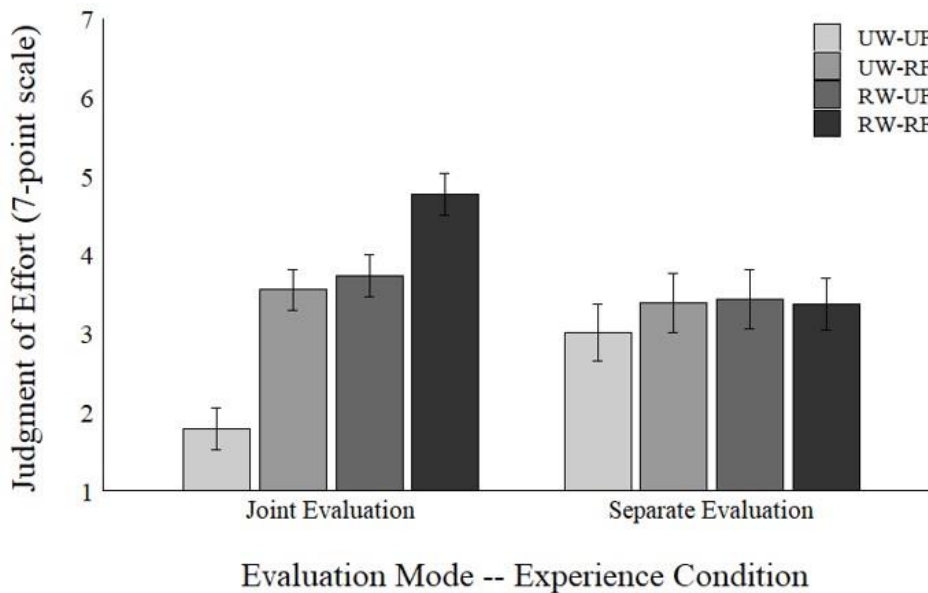


Figure 7. Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the Experience Condition. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

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, $t(84)$

$= 5.86, p < .001, d = 0.63$.

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.

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, $F(3, 336) = 6.42, p < .001, \eta_g^2 = .05$. See Figure 8 for a visualization of these data.

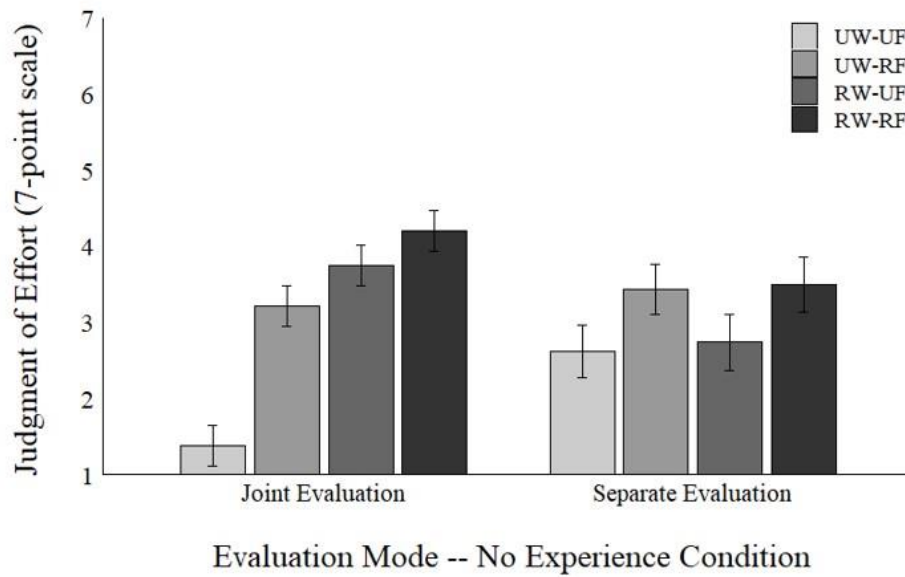


Figure 8. Mean judgments of effort as a function of stimulus type grouped by evaluation mode in the No Experience Condition. Error bars represent 95% CIs and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

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, $ts < 1$.

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), serves to support the idea that the comparative flatness of the separate evaluation judgement pattern 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, 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, which involved retrospective judgments. 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.

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 (wherein participants experience all stimulus types). Thus, it remains possible that, if demand measures were collected in a pure stimulus context (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 judgement (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.

Method

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.

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.

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. In addition, the display composed of "WORD" from Experiments 1a, 1b, and 3 was employed. The design was similar to that of Experiment 3, with individuals gaining experience with the reading task before making a prospective judgment. 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).

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 presented⁵ 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 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 of the word “WORD” was provided in the same orientation of

⁵ The data from these trials were not analyzed.

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 freeform 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.

Results

One thousand two hundred and thirty 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 incorrectly⁶; 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

⁶ 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.

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 (700 and 181 in the separate and joint evaluation condition, respectively) were then subjected to outlier removals prior to the main analysis.

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 a z-score size 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 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 provided by a participant in the separate evaluation condition were detected ($M = 41\,372$, $SD = 2708$). 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 of the reading time data. After outliers were detected and removed using the same procedure as described above, there were 349 and 79 participants' data in the separate and joint evaluation condition, respectively.

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 . See Table 2 for descriptive statistics of reading times, error counts, and judgments of effort; and Figure 9 for 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, $t(632.09) < 1, p = .889, d = 0.01$.

Table 2

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

	<i>Reading Time</i>	<i>Error Count</i>	<i>Effort Judgment</i>
<i>Joint Evaluation</i>			
UW-UF	13,210 (2,835)	0.78 (0.69)	1.53 (0.86)
UW-RF	13,462 (2,749)	0.68 (0.63)	3.07 (1.57)
RW-UF	13,893 (2,791)	0.74 (0.72)	3.75 (1.60)
RW-RF	14,059 (2,780)	0.73 (0.70)	4.46 (1.70)
<i>Separate Evaluation</i>			
UW-UF	13,017 (2,945)	0.82 (0.70)	3.04 (1.52)
UW-RF	13,585 (2,998)	0.81 (0.72)	3.12 (1.54)
RW-UF	14,313 (3,970)	0.93 (1.21)	3.04 (1.63)
RW-RF	14,197 (3,254)	0.88 (0.59)	3.42 (1.60)

Note: Standard deviations in parentheses. Response times in milliseconds.

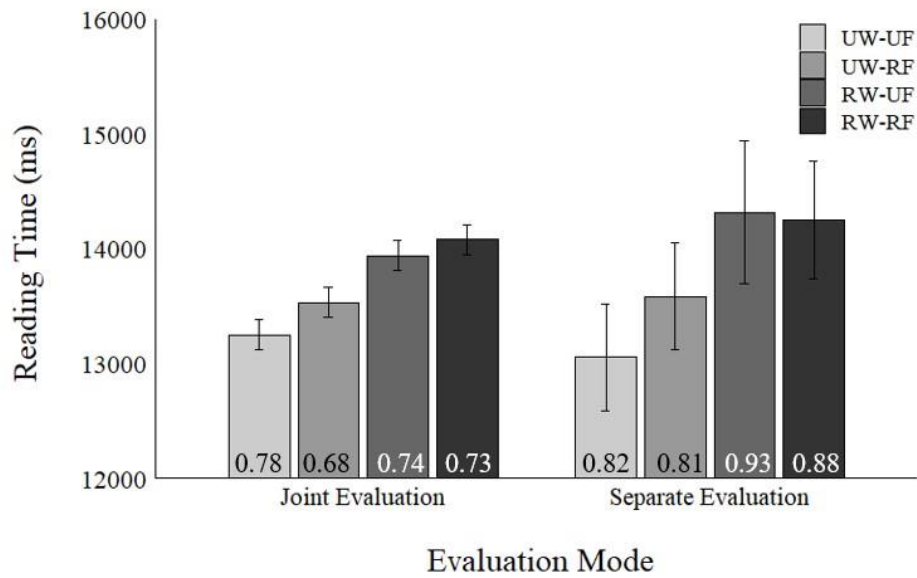


Figure 9. Mean reading times (in milliseconds) as a function of stimulus type grouped by evaluation mode. Error bars represent 95% CIs (using DF_{error} from the Erlebacher's analysis) and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively. Error counts displayed at bottom of bars.

Erlebacher's Analysis – Judgments of Effort

Data used for this 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) $\times$ 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, see Figure 10 for a visualization of these data.

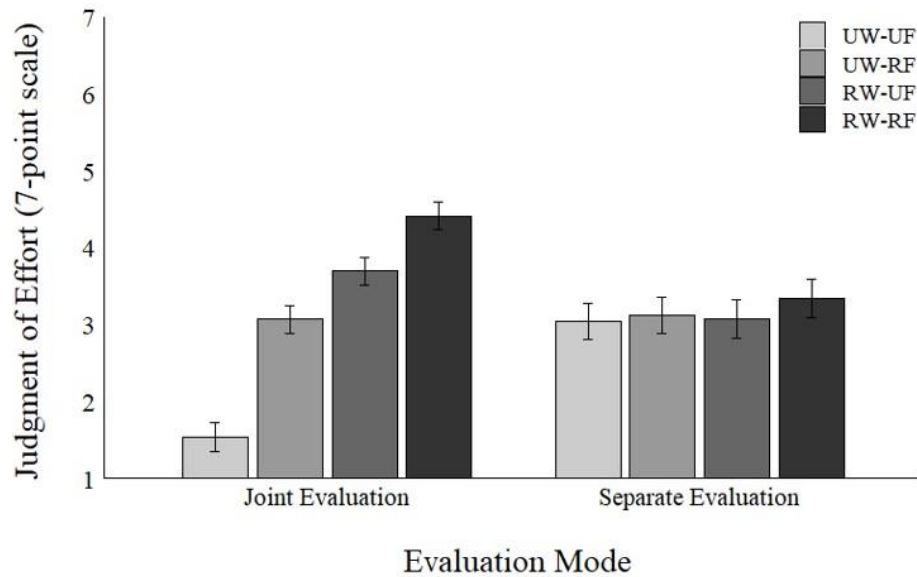


Figure 10. Mean effort judgments as a function of stimulus type grouped by evaluation mode. Error bars represent 95% CIs (using DF_{error} from the Erlebacher's analysis) and Masson-Loftus 95% CIs for separate and joint evaluation data, respectively.

Judgments of Effort in the Joint Evaluation Mode. A single factor repeated measures

ANOVA revealed a significant effect of stimulus type, $F(2.72, 434.83) = 174.04, p < .001, \eta_p^2 = .34$. 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, $t(160) = 6.28, p < .001, d = 0.42$.

Judgments of Effort in the Separate Evaluation Mode. A single factor between-participants

ANOVA indicated no difference across the stimulus types, $F(3, 640) = 1.16, p = .325, \eta_p^2 = .01$.

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 $F_s < 1$. See Figure 11 below.

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.”⁷ In sum, we collected coded data provided by two independent human coders alongside multiple-choice answers 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 data⁸.

⁷ This option was accidentally left 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.

⁸ The results did not qualitatively differ when including these responses in the self response data.

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, $\chi^2(1, n = 799) = 15.32, p < .001$, respectively. 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, $\chi^2(1, n = 805) = 0.32, p = .574$. See Table 3 for frequencies of each feature discussed here, split by evaluation condition; and Figure 11a and Figure 11b for a visualization of these data.

Table 3

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

	<i>Joint</i>		<i>Separate</i>	
	<i>Open</i>	<i>Selected</i>	<i>Open</i>	<i>Selected</i>
Specific	92.2	96.1	79.2	94.7
<i>Feature</i>				
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 were not included in the analysis.

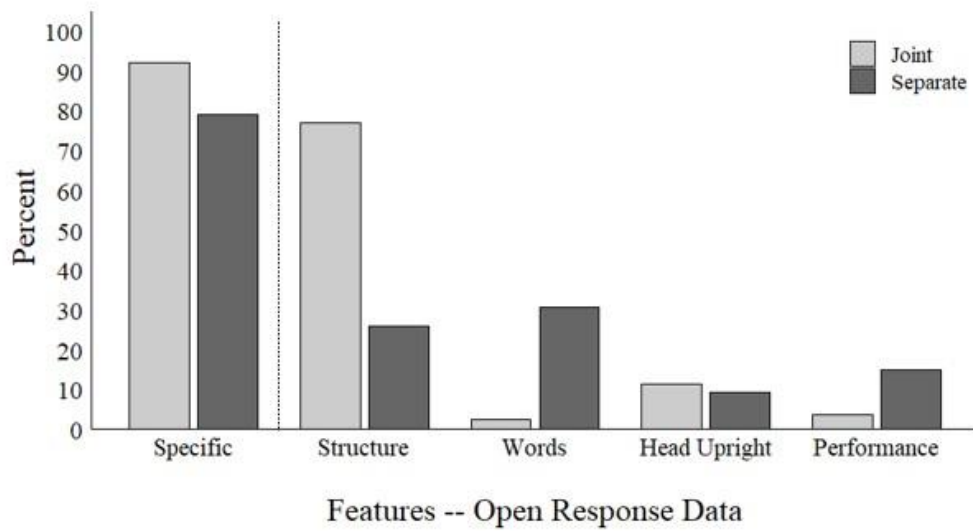


Figure 11a. Percent of total responses in favor of each feature for the open response data, split by evaluation condition.

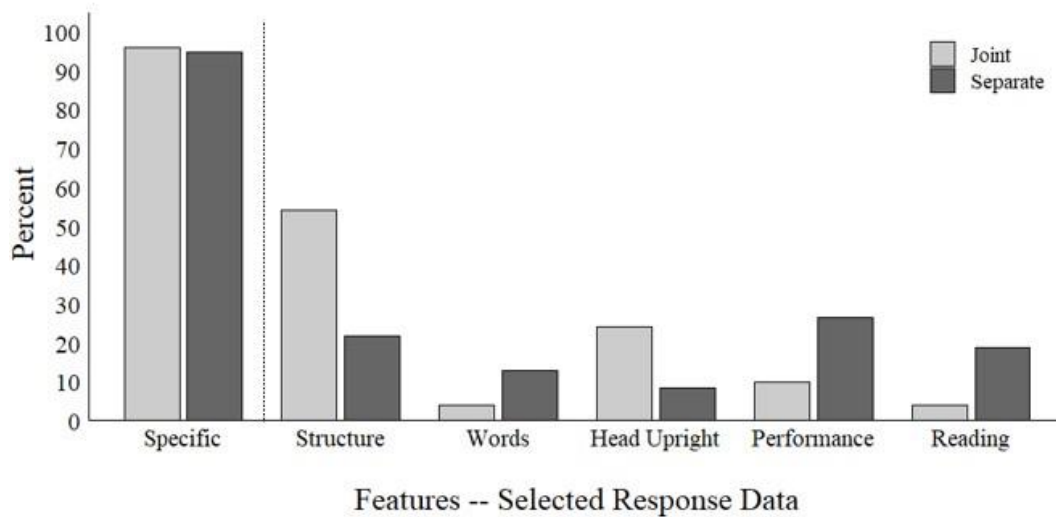


Figure 11b. Percent of total responses in favor of each feature for the selected response data, split by evaluation condition.

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, $\chi^2(1, n = 765) = 9.81, p = .002$.

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, $\chi^2(1, n = 765) = 21.43, p < .001$.

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) Index⁹ (Caso & Angeles, 1988), a measure of diversity or

“unlikeability” 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. Notably, participants in the joint evaluation condition provided a sample (i.e., the selected response data) that appears more diverse than the open response data. 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.

Features Across Stimulus Type 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). See Figure 12 for 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

⁹ $GS = 1 - (p_1^2 + p_2^2 + \dots + p_m^2)$, where m is the number of distinct categories and p_i is the proportion of responses in the i^{th} category.

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 12a, 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.

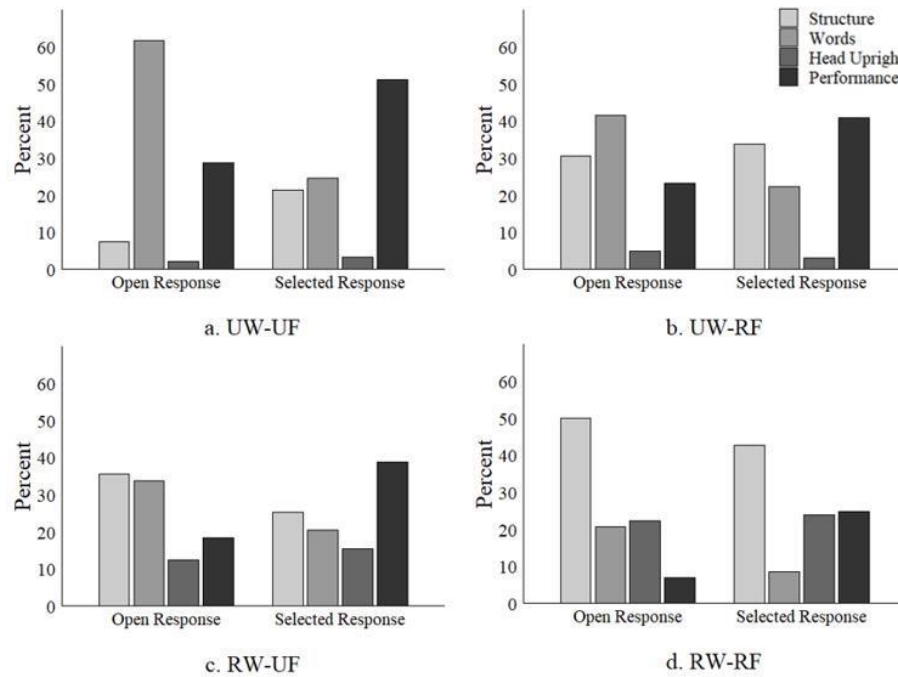


Figure 12. Percent of responses in favor of each of the four main features for Open Response and Selected Response data. Top left: proportion of individuals in UW-UF; top right: proportion of individuals in UW-RF; bottom left: proportion of individuals in RW-UF; bottom right: proportion of individuals 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, $\chi^2(3, n = 82) = 23.27, p < .001$.

.001 and $\chi^2(3, n = 98) = 31.88, p < .001$. As visualized in Figure 12b, 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 12d 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, $\chi^2(2, n = 443) = 0.08, p = .961$.

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 unlike – 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 cueutilization 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.

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 GET, 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 GET, 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 GET, 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; 2021a).

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.

References

- M. Ashburner & E.F. Risko (2021). Judgments of effort as a function of post-trial vs. post- task elicitation. *Quarterly Journal of Experimental Psychology*, 74(6), 991-1006. doi: 10.1177/17470218211005759
- 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. doi: 10.1037/0096-3445.127.1.55
- 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. doi: 10.2307/2393222
- 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. doi: 10.1016/S0167-2681(99)00025-6
- 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. doi: 10.1006/obhd.1994.1088
- 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-participants design. *Psychological Methods*, 4(3), 243-249. doi: 10.1037/1082-989X.4.3.243
- 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. doi: 10.1016/S0167-2681(96)00891-8
- Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5), 1225-1234. doi: 10.1287/mnsc.2015.2186

- Boksem, M. A., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain research reviews*, 59(1), 125-139. doi: 10.1016/j.brainresrev.2008.07.001
- 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.
- 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. doi: 10.1016/j.joep.2011.05.003
- 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. doi: 10.1523/JNEUROSCI.4515-08.2009
- Davidai, S., & Shafir, E. (2018). Are 'nudges' getting a fair shot? Joint versus separate evaluation. *Behavioural Public Policy*, 1-19. doi: 10.1017/bpp.2018.9
- 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. doi: 10.3758/s13423-013-0384-5
- Dunn, T. L., Gaspar, C., & Risko, E. F. (2019). Cue awareness in avoiding effortful control. *Neuropsychologia*, 123, 77-91. doi: 10.1016/j.neuropsychologia.2018.05.011
- 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. doi: 10.1007/s00426-017-0943-x
- 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. doi: 10.1002/bdm.2018

- 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. doi = 10.1037/xhp0000236
- Dunn, T. L., & Risko, E. F. (2016). Toward a Metacognitive Account of Cognitive Offloading. *Cognitive Science*, 40(5), 1080-1127. doi: 10.1111/cogs.12273
- 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(1), 33-49.
- 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. doi: 10.1037/0033-2909.84.2.212
- Garbarino, E. C., & Edell, J. A. (1997). Cognitive effort, affect, and choice. *Journal of consumer research*, 24(2), 147-158. doi: 10.1086/209500
- 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. doi:10.1006/obhd.2001.2953
- 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. doi: 10.1016/j.joep.2009.08.003
- 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.

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. (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. (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. doi: 10.1109/TVCG.2018.2872577

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. doi: 10.1037/0033-2909.125.5.576

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., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making*, 16(4), 257-272. doi: 10.1002/bdm.445

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. doi: 10.1037/0022-3514.86.5.680

Hsee, C. K., & Zhang, J. (2010). General Evaluability Theory. *Perspectives on Psychological*

Science, 5(4), 343-355.

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. doi: 10.1086/669484

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.

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. doi: 10.1016/j.tics.2018.01.007

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. doi: 10.1007/BF01065347

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. doi: 10.1007/BF01073401

Kogut, T., & Ritov, I. (2005). The “identified victim” effect: An identified group, or just a single individual?. *Journal of Behavioral Decision Making*, 18(3), 157-167. doi: 10.1002/bdm.492

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. doi: 10.1016/j.jml.2005.01.001

- 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. doi: 10.1017/S0140525X12003196
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within- subject designs. *Psychonomic Bulletin & Review*, 1(4), 476-490. doi: 10.3758/BF03210951
- 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.
- McBride, R. L. (1982). Range bias in sensory evaluation. *International Journal of Food Science & Technology*, 17(3), 405-410. doi: 10.1111/j.1365-2621.1982.tb00195.x
- 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.
- 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.
- Nowlis, S. M., & Simonson, I. (1997). Attribute–task compatibility as a determinant of consumer preference reversals. *Journal of marketing research*, 34(2), 205-218. doi: 10.1177/002224379703400202
- Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational behavior and human decision processes*, 109(2), 134-141.
- Parducci, A. (1964). Sequential effects in judgment. *Psychological Bulletin*, 61(3), 163-167. doi: 10.1037/h0048411

- 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. doi: 10.3758/s13414-017-1434-2
- 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. doi: 10.1016/j.learninstruc.2016.12.002
- 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 *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. doi: 10.1111/cogs.12087
- Ritov, I., & Baron, J. (2011). Joint presentation reduces the effect of emotion on evaluation of public actions. *Cognition and Emotion*, 25(4), 657-675. doi: 10.1080/02699931.2010.512512
- Schkade, D. A. (1995). Issues in the valuation of environmental resources: A perspective from the psychology of decision making. *Journal of Contemporary Water Research and Education*, 96(1), 3.
- Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. *Well-being: The foundations of hedonic psychology*, 7, 61-84.
- 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. doi:

10.1016/j.joep.2005.07.001

Sher, S., & McKenzie, C. R. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General*, 143(3), 1127.

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. doi:

10.1111/j.1467-9280.2008.02189.x

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

10.1080/13546783.2018.1459314

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. doi: 10.1111/j.1559-1816.2008.00413.x

Thompson, V. A., Prowse Turner, J. A., 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. doi: 10.1016/j.cognition.2012.90.012

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. doi: 10.1037/a0023719

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. doi: 10.1080/17470218.2013.777751

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. doi:

10.3758/s1342

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. doi: 10.1016/j.obhdp.2008.03.002

Waters, R. H. (1937). The principle of least effort in learning. *The Journal of General*

Psychology, 16(1), 3-20. doi: 10.1080/00221309.1937.9917938

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. doi: 10.1371/journal.pone.0068210

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. doi:

10.1002/bdm.476

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.

doi: 10.1177/001872088803000110

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

Wesley.