Original Article

QJEP

Quarterly Journal of Experimental Psychology
2021, Vol. 74(6) 991–1006

DOI: 10.1177/17470218211005759
qjep.sagepub.com
SAGE

Judgements of effort as a function of post-trial versus post-task elicitation

Michelle Ashburner and Evan F Risko

Abstract

Cognitive effort is a central construct in our lives, yet our understanding of the processes underlying our perception of effort is limited. Performance is typically used as one way to assess effort in cognitive tasks (e.g., tasks that take longer are generally thought to be more effortful); however, Dunn and Risko reported a recent case where such "objective" measures of effort were dissociated from judgements of effort (i.e., subjective effort). This dissociation occurred when participants either made their judgements of effort after the task (i.e., reading stimuli composed of rotated words) or without ever performing the task. This leaves open the possibility that if participants made their judgements of effort more proximal to the actual experience of performing the task (e.g., right after a given trial) that these judgements might better correspond to putatively "objective" measures of effort. To address this question, we conducted two experiments replicating Dunn and Risko with additional probes for post-trial judgements of effort (i.e., a judgement of effort made right after each trial). Results provided some support for the notion that judgements of effort more closely follow reading times when made post-trial as opposed to post-task. Implications of the present work for our understanding of judgements of effort are discussed.

Keywords

Cognitive effort; judgement of effort; metacognitive judgement; post-trial elicitation; post-task elicitation

Received: 3 June 2020; revised: 3 June 2020; accepted: 22 July 2020

Introduction

While there exists a long tradition of research investigating subjective workload (Eggemeier & Stadler, 1984; Jex, 1988; Moray, 1982; Yeh & Wickens, 1988), the psychological basis of cognitive effort has received increased attention from researchers over the last 5 years (Dunn et al., 2017; Dunn, Gaspar, & Risko, 2019; Dunn, Inzlicht, & Risko, 2019; Inzlicht et al., 2018; Kool & Botvinick, 2018; Kurzban, 2016; Potts et al., 2018; Yildirim et al., 2019). An important dimension of this line of research is attempting to address *how* individuals judge the effort anticipated or experienced on a given task or trial of a given task (Dunn et al., 2017; Foo et al., 2009; Gweon et al., 2017; Marshall, 2002; Song & Schwarz, 2008; Westbrook et al., 2013). That is, when someone is asked how effortful they found a task to be (or will be in the case of a prospective judgement), what factors determine their judgement of effort (i.e., subjective effort)? In the present investigation, we examine whether the type of judgement modulates the information brought to bear on those judgements.

One way to think about judgements of effort, at least in the context of the types of cognitive tasks discussed herein, is as a type of metacognitive judgement (a judgement about one's cognitive processes; Dunn, Gaspar, & Risko, 2019; Dunn & Risko, 2016; Koriat et al., 2014; Raaijmakers et al., 2017; Schmeck et al., 2015; van Gog et al., 2012). Metacognitive judgements, according to an influential framework, are viewed as inferential in nature (Koriat, 1993). Stated broadly, individuals rely on available information to infer, for example, the likelihood that they will recall some information in the future (a judgement of learning [JOL]) or, as is the focus here, the effortfulness of a given task (a judgement of effort). These judgements might be more experiential, for example, based on the experience of fluency, or more belief-based, for example, based on the belief that Task A is more inherently effortful than is Task B because of some characteristic of that task (e.g., involves more elements to process). In the context of

University of Waterloo, Waterloo, Ontario, Canada

Corresponding author:
Michelle Ashburner, University of Waterloo, 200 University Ave. W., Waterloo, Ontario N2L 3G1, Canada.
Email: mrmashbu@uwaterloo.ca

judgements of learning, Mueller et al. (2013) provided evidence in a paired-associate recall task that judgements were more related to participants' beliefs (i.e., belief-based) regarding relatedness than to processing fluency (i.e., experiential). Conversely, Undorf and Erdfelder (2015) demonstrated, using a similar paradigm, contributions of both fluency and beliefs to judgements of learning. Undorf and Ackerman (2017) also illustrated that experiences of fluency may be selectively utilised in the formation of judgements of learning for a recall task; that is, study time was only related to JOLs when the latter were on the higher end. This effort to understand the contributions of different types of information to metacognitive judgements extends beyond judgements of learning (e.g., feeling of rightness, judgements of solvability; Ackerman & Beller, 2017; Ackerman & Thompson, 2017; for a review, see Ackerman, 2019). Taken together, these studies underline the importance, within each type of metacognitive judgement, to understand what information is relied on to make these judgements, and the conditions under which this information might change.

Research investigating metacognitive judgements has demonstrated that they can be influenced by their placement relative to critical cognitive events. For example, prospective judgements of confidence (i.e., how well one thinks they will perform on an upcoming trial) are less related to task performance than are retrospective judgements (i.e., how well one thinks they performed on the previous trial; Boldt & Gilbert, 2019; Fleming et al., 2016; Gilbert, 2015; Siedlecka et al., 2016). The benefit for retrospective judgements likely derives from the act of task performance, which itself may inform confidence judgements (e.g., fluency; Fleming et al., 2016). In a similar vein, Nelson and Dunlosky (1991) demonstrated that participants' judgements of learning on a paired-associate recall task were more strongly correlated with performance when made after a delay than when made immediately after learning an item. Nelson and Dunlosky (1991; see also Dunlosky & Nelson, 1997) suggested that delay impacted JOL accuracy because, immediately after study, individuals access both short-term and long-term memory to inform their judgements, where the contribution of the former is misleading. When the item is no longer in short-term memory (i.e., after a delay), a more accurate judgement can be made based on the item's current retrievability from long-term memory (Nelson & Dunlosky, 1991; see also Metcalfe & Finn, 2008; Scheck et al., 2004; Weaver & Kelemen, 1997). Consistent with this general idea, Koriat and Ma'ayan (2005) demonstrated that JOLs for paired associates elicited via a pre-JOL recall test immediately after study were more strongly associated with information available at encoding (e.g., pre-JOL encoding fluency), whereas JOLs obtained after a delay between study and the pre-JOL recall test were more strongly associated with information available at recall (e.g., pre-JOL retrieval fluency). Moreover, Koriat et al. (2006) demonstrated that the strength of the correlation between JOLs and study time was significantly greater for immediate, as compared with delayed, JOLs.

More relevant to the present work, van Gog et al. (2012) and Schmeck et al. (2015) investigated differences in retrospectively reported mental effort on a problem-solving task using post-trial judgements and post-task judgements. Specifically, van Gog et al. (2012) measured perceived effort immediately following each of six problem-solving exercises, and then obtained participants' overall perceived effort after completion of all exercises in the block. These two judgement types differ in terms of both their temporal proximity to the cognitive event (i.e., the post-trial judgements are made closer in time to the cognitive event in question) and their scope (i.e., the post-trial judgement refers to a specific cognitive event whereas the judgement made at the end of the task refers to a group of cognitive events). van Gog et al. (2012) demonstrated that perceived mental effort was higher when provided post-task, as compared with the average of the post-trial judgements. Importantly, this result did not depend on whether the post-trial and post-task judgements were provided between-subjects or within-subjects. Schmeck et al. (2015) also examined these two types of judgements using the same paradigm, with a focus on the extent to which measures of subjective cognitive load predicted performance. With respect to post-trial versus post-task judgements, Schmeck et al. (2015) replicated the results of van Gog et al. (2012), demonstrating that post-task effort ratings were significantly higher than the average of the post-trial ratings. Schmeck et al. (2015) suggested that post-task judgements may be higher than the average of the post-trial judgements due to the former being perceived as a single judgement of one long, multicomponent task. Taken together, the research outlined above demonstrates that manipulating the type of metacognitive judgements (i.e., post-trial vs. post-task)—including judgements of effort—can affect the judgements themselves. Here, we examine the influence of judgement type (i.e., post-trial, post-task) on the relation between reading time and judgements of effort.

Several recent studies (Baars et al., 2017; Dunn et al., 2016; Dunn, Gaspar, & Risko, 2019; Dunn, Inzlicht, & Risko, 2019; Dunn & Risko, 2016; Korbach et al., 2017; Potts et al., 2018; Schmeck et al., 2015; van Gog et al., 2012) have examined the relation between various sources of information and judgements of effort (e.g., time, errors, intrinsic properties of the stimuli). In particular, Dunn and colleagues focused on the extent to which time informs individuals' judgements of effort, because researchers often use the time to complete a task as an index of fluency or ease of processing (Benjamin et al., 1998; Koriat & Ma'ayan, 2005; Thompson et al., 2013; Undorf & Erdfelder, 2011, 2013, 2015). In addition, time costs appear to be a central factor in making decisions about resource allocation (i.e., time as the currency in making effort-based decisions; Gray et al., 2006). Dunn and Risko (2016) had participants complete multiple trials of a reading task for which there

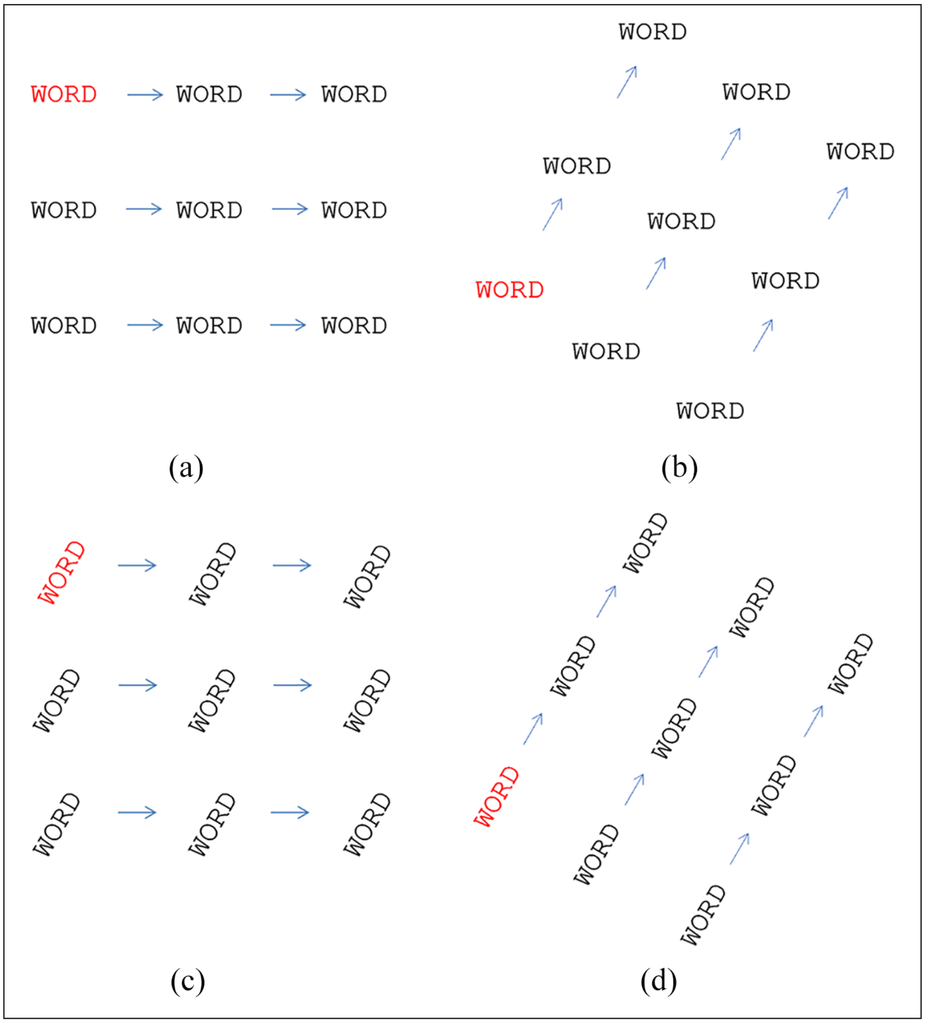

Figure 1. Examples of each stimulus type: (a) UW-UF, (b) UW-RF, (c) RW-UF, and (d) RW-RF. All rotations above are 60° counterclockwise. For illustrative purposes, each stimulus display above contains nine words; however, the displays used by Dunn and Risko (2016) and here contained 25 words.

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). After the reading task, participants provided judgements 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 judgements of effort; however, a consistent dissociation was observed: namely, reading times for the RW-RF stimulus type and the RW-UF stimulus type were equivalent, yet the RW-RF stimulus type was consistently judged as more effortful (e.g., Dunn & Risko, 2016). One possible explanation for this dissociation, put forward by Dunn and Risko (2016), is that individuals, rather than relying on time (or an underlying feeling of fluency related to time), inferred effortfulness based on stimulus orientation and their beliefs about their perceptual/cognitive systems (e.g., processing a disoriented stimulus is hard). For example, the RW-RF stimulus type may be believed 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).

The dissociation between reading time and judgements of effort observed by Dunn and Risko (2016) provides a unique opportunity to investigate how judgement type (i.e., post-trial vs. post-task) might influence the use of a potentially important experience related to effort expenditure (i.e., time, fluency). As described above, Dunn and Risko (2016) used post-task judgements of effort such that participants read numerous displays of each type then completed a judgement of effort using a generic instance of each type. Thus, as noted above, participants were making judgements of effort separated from the experience of reading each display. This raises the interesting possibility that the dissociation between reading time and judgements of effort might reflect a greater reliance on beliefs, as opposed to experiences, due to the judgement being elicited post-task. We provide a direct test of this idea here.

In the present study, we used the same stimulus types and task as Dunn and Risko (2016) and added probes for post-trial judgements of effort in addition to a similar series of post-task judgements of effort.

If the dissociation between judgements of effort and reading times reported by Dunn and Risko (2016) is predominately due to the judgement having been elicited post-task, then this dissociation should be reduced or eliminated using post-trial judgements of effort. Alternatively, if this dissociation is not due to the judgement having been elicited post-task, similar results could be observed for post-trial judgements as for post-task judgements. That said, it is also possible that this dissociation is observed for both post-trial and post-task judgements, yet the basis for those judgements might differ. Finally, the present design affords the opportunity to examine whether the observation by van Gog et al. (2012) and Schmeck et al. (2015), that on average judgements of effort are higher when provided post-task, generalises to a completely different context.

Experiment 1

Method

Participants. Thirty-two University of Waterloo undergraduate students participated in this experiment in exchange for course credit. This sample size was chosen based on previous research (Dunn & Risko, 2016).

Apparatus. The presentation of stimuli and recording of participants' responses were handled by E-Prime 3.0 software. Participants viewed all stimuli and instructions on a widescreen 22-inch LG monitor while seated at a desk and used a QWERTY keyboard for response entry. A Logitech web camera fixed on top of the monitor was used to capture audio and video of each session; the former was used to measure reading time and error count, and the latter for coding head movement.

Stimuli and design. A one-factor (stimulus type) within-subjects design was employed. Each slide consisted of a stimulus composed of 25 words: five rows of five words each. Words were typed in 18-point black Courier New font on a white background. For the practice and test trials, all words were five letters in length and contained one or two syllables.[1] Stimuli had average SubtlexUS (Brysbaert & New, 2009) word frequencies ranging from 9.03 to 336.57. For the post-test slides, all words were "WORD." There were four stimulus types: stimuli with UW-UF, stimuli with UW-RF, stimuli with RW-UF, and stimuli with RW-RF. Each stimulus type was presented eight times during test; for each of the disoriented stimulus types (i.e., UW-RF, RW-UF, and RW-RF), four of the eight were rotated counterclockwise by 60° and four were rotated clockwise by 60°.

The starting word on each slide was coloured red and blue single-headed arrows marked the direction in which participants were instructed to read. See Figure 1 for an illustrative example of stimulus types containing three rows of three words each. A pre-ordered list of trials was made for each participant consisting of 32 randomly selected and ordered stimuli (i.e., 25-word displays). There were 33 stimuli in total from which to draw and we intended to randomly choose a subset of size 32 from the set of available stimuli. Therefore, each participant's list was formed by selecting 32 of these 33 displays. A list of five additional stimuli that did not vary across participants comprised the practice trials. Finally, seven slides of post-test stimuli were composed entirely of the word "WORD," one slide per stimulus type and, for disoriented stimuli types, one slide per direction of rotation.

Procedure. Participants entered the testing room and were seated facing the centre of the monitor at about eye level, and the keyboard was on the desk directly in front of them. After providing consent, video and audio recording began. Instructions were displayed on the screen as a research assistant (who knew about the nature of the experiment) read them and answered any questions. Participants were instructed to read each word on the slide aloud as quickly and as accurately as possible, while keeping their head upright and limiting movement. Participants were not instructed to correct errors or to ignore them; only to proceed in the way they felt would best follow the given instructions. Moreover, errors were coded offline using audio capture, so as not to distract participants. Once finished, participants were instructed to press the spacebar to advance the slide to a 7-point effort scale (1=*not at all effortful*, 4=*somewhat effortful*, 7=*very effortful*) where they were asked to consider the stimulus on the previous slide and, using the keyboard, make a judgement of effort regarding their experience of reading. Upon entering their judgement, the next trial began. The five practice stimuli were first presented, in the same order for every participant: UW-UF, RW-UF, UW-RF, RW-UF, and RW-RF. After the practice trials were complete, the research assistant left the room. These practice trials were followed by 32 test trials and corresponding judgements of effort. There was no feedback of any kind (i.e., regarding reading time or error count) provided to participants. Following this were post-task judgements of effort, wherein participants were asked to view a generic stimulus type (i.e., each word was "WORD") of each orientation and direction, for a total of seven slides (one UW-UF display and two of each disoriented display), presented in random order to each participant. Participants were provided written instructions to view the displays, but not to read them, and provide an overall judgement of effort for each stimulus type on a 7-point scale (1=*not at all effortful*, 7=*very effortful*) using the keyboard. The experiment then concluded and participants were debriefed.

Table 1. Experiment 1 mean reading time (ms), mean error count, and mean judgements of effort.

Dependent variable	Stimulus type			
	UW-UF	UW-RF	RW-UF	RW-RF
Reading time	15,876 (3,201)	16,244 (3,291)	17,211 (3,668)	17,413 (3,992)
Error count	1.34 (0.89)	1.17 (0.82)	1.29 (0.86)	1.40 (1.00)
Post-trial judgements	2.16 (0.68)	2.62 (0.81)	3.15 (1.00)	3.58 (1.10)
Post-task judgements	1.47 (0.80)	2.33 (1.16)	2.78 (1.17)	3.52 (1.42)

Standard deviations in parentheses. Post-trial and post-task judgements of effort are on 7-point scales. UW-UF: upright words in an upright frame; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.

Results

Three participants were replaced. One participant experienced difficulty pronouncing many of the words; another read each word at a pace of over 2 s, and another did not wish to be video recorded. An operational error was detected at the time of coding due to instructions having been given incorrectly. Particularly, participants were given information about the nature of the experiment that could have biased their judgements. This error resulted in the need to remove these data and recruit 12 additional participants. Due to a programming error, lists for two participants were not equally composed of each stimulus type; these data were retained. The final sample had 32 participants. Any trials during which an obvious head movement was made to facilitate reading were removed from analyses. Thirty trials in all were removed for this reason: 0.8% (of total) UW-UF, 0.8% UW-RF, 2.0% RW-UF, and 8.2% RW-RF.[2] In an additional 7.3% of trials, participants prematurely pressed the spacebar before reading the last word on the slide; these trials were also removed. Finally, one trial was removed because the participant paused for an extended time (2,422 ms) after a mispronunciation. After these exclusions, a within-subject, within stimulus type search for outliers at the trial level found no reading times, error count scores, or immediate judgements of effort with $|z| > 3$. Therefore, no observations were trimmed from these data. In total, 10.2% of observations were removed.[3] Provided the uneven exclusions for head rotations (see above), we conducted a second set of analyses excluding participants who had a disproportionate rate of head rotation for the RW-RF stimulus type, to have an approximately equal distribution. We excluded three participants to form a *head-tilt control subset* and the resulting proportion of trials in these data for each stimulus type was 3.9% UW-UF, 1.3% UW-RF, 0.0% RW-UF, and 0.4% RW-RF. Overall, the results were similar. When an important deviation from the reported results (i.e., the complete sample) was found, we note it in the appropriate section. All analyses were run using the open-source statistical analysis software R, Version 3.4.4. The code and data are available on the Open Science Framework project webpage: https://osf.io/tgx85/. See Table 1 for mean reading time, error count, post-trial judgements of effort, and post-task judgements of effort by stimulus type.

Reading times. Reading times for each trial were collected by the E-Prime software, measured in milliseconds from stimulus onset to when the spacebar was pressed. A one-way repeated-measures analysis of variance (ANOVA) with stimulus type as the factor, corrected for sphericity violations, revealed a significant effect of stimulus type on reading time, $F(2.11, 65.52) = 16.92$, $p < .001$, $\eta_g^2 = .03$. Pairwise *t*-tests were conducted across stimulus types. Compared with UW-UF trials, individuals were no slower on UW-RF trials, $t(31) = 1.74$, $p = .092$, $d = 0.11$;[4] however, individuals were slower on RW-UF trials, $t(31) = 5.19$, $p < .001$, $d = 0.39$, and on RW-RF trials, $t(31) = 4.87$, $p < .001$, $d = 0.42$. Compared with UW-RF trials, individuals were slower on RW-UF trials, $t(31) = 4.26$, $p < .001$, $d = 0.28$, and RW-RF trials, $t(31) = 3.82$, $p = .001$, $d = 0.32$. Individuals were not significantly slower on RW-RF trials than on RW-UF trials, $t(31) = 1.07$, $p = .292$, $d = 0.05$, $BF_{01} = 3.13$. Provided the importance of the contrast between the RW-UF and the RW-RF stimulus types, the Bayes Factors are presented exclusively for this last contrast. A Bayes Factor in support of the null hypothesis (i.e., BF_{01}) is a value indicating how much more likely the given results are to occur if the null hypothesis (i.e., the effect size is zero) is true, compared with if the alternative hypothesis (i.e., the effect size is non-zero) is true (Jarosz & Wiley, 2014). In the case above, it is 3.13 times more likely that the null hypothesis is true, compared with the alternative hypothesis.[5] When conducting this analysis on the head-tilt control subset, the results were qualitatively similar except that individuals were significantly slower on UW-RF trials than on UW-UF trials, $t(28) = 2.53$, $p = .017$, $d = 0.16$. Moreover, while individuals were not significantly slower on RW-RF trials than on RW-UF trials, the Bayes Factor in support of the null for this contrast was 1.85.

Error count. The error count for each trial was coded as the number of errors made while reading. An error was added when a sound or syllable in a word was repeated more than once, when a word was repeated more than once, when a word was missed, or other serious mispronunciations, including pluralising a singular word or reading the singular version of a word presented in plural form. Pauses within words were not counted as errors, nor were

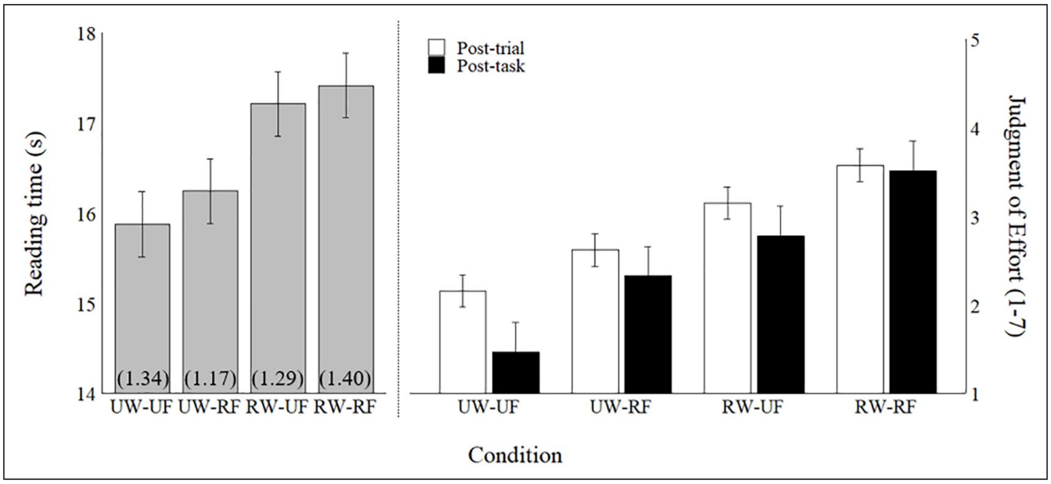

Figure 2. Reading times by stimulus type (left panel) and post-trial versus post-task judgements of effort by stimulus type (right panel) for Experiment 1. Average error count per stimulus type reported in parentheses (left panel). Error bars are Masson–Loftus 95% CI (Loftus & Masson, 1994).

utterances of "um" or similar filler words. A one-way repeated-measures ANOVA with stimulus type as the factor revealed that there was no effect of stimulus type on error count, $F(2.29, 71.12)=1.95$, $p=.128$, $\eta_g^2=.01$.

Effort—post-trial judgement. Post-trial judgements of effort were assessed through the report of participants' perceived effort immediately after each trial. Each post-trial judgement of effort was between 1 (*not at all effortful*) and 7 (*very effortful*). A one-way repeated-measures ANOVA with stimulus type as the factor revealed a significant effect of stimulus type on post-trial judgements of effort, $F(3, 93)=44.23$, $p<.001$, $\eta_g^2=.26$. Pairwise *t*-tests were conducted across stimulus types. Compared with UW-UF trials, individuals reported more effort on UW-RF trials, $t(31)=3.85$, $p=.001$, $d=0.62$; on RW-UF trials, $t(31)=6.63$, $p<.001$, $d=1.16$; and on RW-RF trials, $t(31)=9.77$, $p<.001$, $d=1.55$. Compared with UW-RF trials, individuals reported more effort on RW-UF trials, $t(31)=4.09$, $p<.001$, $d=0.58$, and on RW-RF trials, $t(31)=7.37$, $p<.001$, $d=0.99$. Critically, individuals reported significantly more effort on RW-RF trials than on RW-UF trials, $t(31)=3.86$, $p=.001$, $d=0.40$, $BF_{10}=56.21$. A Bayes Factor in support of the alternative hypothesis (i.e., BF_{10}) is the reciprocal of BF_{01}. Therefore, in the case above, it is 56.21 times more likely that the alternative hypothesis is true, compared with the null hypothesis.

Effort—post-task judgement. A one-way repeated-measures ANOVA with stimulus type as the factor revealed a significant effect of stimulus type on post-task judgements of effort, $F(3, 93)=25.87$, $p<.001$, $\eta_g^2=.30$. Pairwise *t*-tests were conducted across stimulus types. Compared with the UW-UF condition, individuals reported greater effort for the UW-RF condition, $t(31)=3.70$, $p=.001$, $d=0.86$; for the RW-UF condition, $t(31)=6.24$, $p<.001$, $d=1.31$; and for the RW-RF condition, $t(31)=7.05$, $p<.001$, $d=1.78$. Compared with the UW-RF condition, individuals reported more effort for the RW-UF condition, $t(31)=2.05$, $p=.049$, $d=0.39$, and the RW-RF condition, $t(31)=5.24$, $p<.001$, $d=0.92$. Finally, individuals reported significantly more effort for the RW-RF condition than for the RW-UF condition, $t(31)=3.06$, $p=.004$, $d=0.57$, $BF_{10}=8.74$.

Exploratory analysis

Stimulus type by judgement type interaction. A two-way repeated-measures ANOVA with stimulus type and judgement type (i.e., post-trial or post-task judgement of effort) as factors revealed a significant effect of judgement type, $F(1, 31)=6.10$, $p=.019$, $\eta_g^2=.03$, such that post-task judgements were on average lower than post-trial judgements. Moreover, there was a significant interaction between stimulus type and judgement type, $F(3, 93)=2.72$, $p=.049$, $\eta_g^2=.01$. This interaction reflected a large difference in reported effort for the UW-UF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.69$), which reduced in magnitude (and was not statistically significant) for the UW-RF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.29$) and the RW-UF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.37$), and was absent for the RW-RF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.06$; see Figure 2). To further explore the nature of the interaction between stimulus type and judgement type, we conducted three 2 (stimulus type) × 2 (judgement type) ANOVAs, comparing UW-UF with UW-RF, UW-RF with RW-UF, and RW-UF with RW-RF. This revealed a significant interaction between stimulus type and judgement type when comparing UW-UF with UW-RF stimulus types, $F(1, 31)=4.35$, $p=.045$, $\eta_g^2=.01$. As seen in Figure 2, the effect of judgement type was more pronounced on judgements of effort for the UW-UF stimulus type than for the UW-RF stimulus type. There was no significant interaction between stimulus type and judgement type when comparing UW-RF with RW-UF stimulus types, $F<1$, or when com-

Table 2. Multilevel regression models predicting effort judgements for Experiment 1.

Predictor variable	Post-trial effort		Post-task effort	
	B	*SE*	*B*	*SE*
Intercept	0.62	0.52	2.26***	0.62
Reading time	0.09**	0.03	–0.06	0.04
Error count	0.11	0.1	0.08	0.14
Stimulus type				
UW-RF	0.45***	0.13	0.89***	0.24
RW-UF	0.88***	0.13	1.39***	0.24
RW-RF	1.28***	0.13	2.13***	0.25

SE: standard error; UW-UF: upright words in an upright frame; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.
Each participant contributes four observations, one per stimulus type. The stimulus type factor is treatment coded, with UW-UF as the reference group. Reading time and error count were included in the model in their raw forms.
$^{**}p<.01$; $^{***}p<.001$.

paring RW-UF with RW-RF stimulus types, $F(1, 31)=1.47$, $p=.234$, $\eta_g^2 < .01$. When the analysis was conducted on the head-tilt control subset, the interaction between stimulus type and judgement type when comparing UW-UF with UW-RF stimulus types did not reach significance, $F(1, 28)=2.59, p=.119$, $\eta_g^2=.01$.

Multilevel regression analysis. The notion that individuals' post-trial judgements are more so informed by reading times could be further examined using a regression approach wherein reading time is viewed as a predictor of effort judgements. To this end, we built two multilevel regression models, one with post-trial judgements as the dependent variable and the other with post-task judgements as the dependent variable. Both models included reading time, error count, and stimulus type as predictor variables. The data used for each of these two analyses were aggregated such that there were four observations for each subject. This was done to allow a comparison across the effort types, as there were at most two measurements of post-task effort per subject per stimulus type. Each model included random intercepts for each subject. These analyses demonstrated a significant effect of reading time on post-trial effort, $B=0.09$, standard error $(SE)=0.03$, $p=.009$, while reading time did not significantly predict post-task effort, $B=–0.06$, $SE=0.04$, $p=.137$. Consistent with the results reported above, the stimulus type manipulation was also significantly related to post-trial, as well as post-task, judgements of effort. See Table 2 for the full results.

In addition, a third multilevel regression model, with post-trial judgements of effort as the dependent variable, was computed. This model included trial number—an integer ranging from 1 to 32—in addition to the other predictor variables mentioned above. This model included random slopes and intercepts for each participant, and random intercepts for each display (i.e., including a random effect of item). As per Table 3, reading time, error count, and trial were all significant predictors of post-trial effort. That is, for trials wherein participants read more slowly, or committed more errors while reading, trial-by-trial effort judgements were higher. Interestingly, as participants completed more trials, judgements of effort significantly increased.[6]

Table 3. Multilevel regression model predicting post-trial effort judgements.

Predictor variable	Experiment 1		Experiment 2		Combined	
	B	*SE*	*B*	*SE*	*B*	*SE*
Intercept	−1.86***	0.30	1.34***	0.31	–2.30***	0.27
Reading time	0.23***	0.02	0.27***	0.02	0.25***	0.01
Error count	0.08**	0.03	0.11***	0.03	0.10***	0.02
Trial	0.02***	0.00	0.01***	0.00	0.02***	0.00
Experiment	–	–	–	–	1.29***	0.22
Stimulus type						
UW-RF	0.40***	0.10	0.56***	0.07	0.50***	0.07
RW-UF	0.71***	0.13	0.78***	0.08	0.76***	0.08
RW-RF	1.04***	0.14	0.96***	0.08	0.99***	0.10

SE: standard error; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.
$^{**}p < .01$; $^{***}p < .001$.

Discussion

The results of Experiment 1, with respect to reading times, were consistent with previous research (Dunn & Risko, 2016). Specifically, participants were fastest when reading UW-UF and UW-RF stimulus types and slowest when reading the RW-UF and RW-RF stimulus types. There was no difference in reading times between the UW-UF and UW-RF stimulus types. Critically, there was no difference in reading times between the RW-UF and RW-RF stimulus types. In a similar vein, the post-task judgements of effort followed the same pattern observed by Dunn and Risko (2016), namely, UW-UF < UW-RF < RW-UF < RW-RF. Thus, the dissociation between reading times (i.e., no difference between RW-UF and RW-RF) and judgements of effort (i.e., a significant difference between RW-UF and RW-RF) was again observed. With respect to the former dissociation, the pattern of post-trial judgements of effort followed a similar pattern to that of the post-task judgements. Specifically, the RW-RF stimulus type was judged as significantly more effortful than the RW-UF stimulus type. Thus, the difference in judgements of effort between the RW-UF and RW-RF stimulus types reported by Dunn and Risko (2016) does not appear to be solely due to the separation between participants' reading experience and their judgements of effort. That is, even when the judgement of effort follows immediately after reading, there was still a marked dissociation between judgements of effort and reading time. While the pattern across the RW-UF and

RW-RF stimulus types was similar across post-trial and post-task judgements, there was clearly an impact of judgement type on judgements of effort. Namely, when making judgements post-task, relative to post-trial, judgements of effort were much lower for the UW-UF stimulus type, slightly lower for the UW-RF and RW-UF stimulus types, and the RW-RF stimulus type was unaffected. This interaction might reflect post-trial judgements better aligning with reading time. For example, the reduced difference between the UW-UF and UW-RF stimulus types in post-trial judgements relative to post-task judgements could be construed as closer to the modest difference in reading times between those stimulus types. Furthermore, the multilevel regression models provide evidence that reading time was related to post-trial judgements; however, this was not the case for post-task judgements. Finally, post-task judgements were overall lower than were post-trial judgements. Thus, Experiment 1 provides some modest support for the idea that post-trial judgements of effort more closely follow time. We next replicate and extend Experiment 1 to further examine these effects.

Experiment 2

We decided to replicate Experiment 1 using a larger sample and alter our instructions to encourage participants to increase their reading speed (i.e., not favour accuracy over speed when reading). Moreover, a replication allowed us to address the issue of participants pressing the spacebar before they had read the 25th word, enabling us to measure reading times from stimulus onset to onset of the vocalisation of the 25th word (as opposed to when the spacebar was pressed).

Method

Participants. Forty-eight University of Waterloo undergraduate students participated in this experiment in exchange for course credit. As we wanted to replicate the results of Experiment 1 with a larger sample, we increased our previous sample size by 50%.

Apparatus. The presentation of stimuli and audio/video recording of participants' responses were identical to those in Experiment 1.

Stimuli and design. As in Experiment 1, a one-factor (stimulus type) within-subjects design was employed. We found that 15 words caused regular occurrences of errors due to difficulty with pronunciation in Experiment 1. These words were replaced with words of similar frequency thought to elicit fewer errors across participants. (e.g., "COCOA" replaced with "CAMEL"). In contrast to Experiment 1, wherein 33 stimuli were available to populate the randomly generated lists, only 32 stimuli[7] were used in total (the stimulus removed was that associated with the most errors). Stimuli had average SubtlexUS (Brysbaert & New, 2009) word frequencies ranging from 9.03 to 336.57. Eight unique lists were created using the 32 stimuli such that across each of these lists, each of the 32 stimuli was presented in each orientation on exactly two trials. For the disoriented stimulus types (UW-RF, RW-UF, and RW-RF), one trial was rotated counterclockwise and the other was rotated clockwise. Furthermore, the number of practice trials was reduced from five to four (with each stimulus type presented once), and their order was randomised across participants. The stimuli were otherwise identical to those used in Experiment 1.

Procedure. Participants entered the testing room and were seated facing the centre of the monitor at about eye level, and the keyboard was on the desk directly in front of them. In a change from Experiment 1 designed to decrease occurrences of head rotation, a researcher sat to the right and behind the participant throughout the experiment. Instructions were displayed on the screen as the researcher read them and answered any questions. Participants were told that, upon reading all the words on a slide, the researcher would press a key to advance them to the 7-point effort scale (1 = *not at all effortful*, 4 = *somewhat effortful*, 7 = *very effortful*). To ensure that participants were not favouring accuracy over speed when reading, reminder slides were presented after trial 6, 12, 18, and 24, with instructions to read as quickly as possible and to keep their head upright. The main trial procedure was otherwise identical to that for Experiment 1. The generic stimulus types for the post-task judgements were presented from a randomised list of 16 slides composed of two of each stimulus type and direction (the UW-UF generic stimulus type was presented a total of four times to keep the frequency of each stimulus type constant). The experiment then concluded and participants were debriefed.

Results

Two participants were excluded. One participant experienced difficulty pronouncing many of the words; and another had difficulty keeping their head upright for seven of the eight RW-RF trials and four of the eight RW-UF trials. Their data were removed from the analyses and were replaced by additional participants. Due to a technical error, two participants were not video recorded, but audio recordings were captured, and three participants wished to participate in the experiment without being video recorded. While absence of video was a criterion for exclusion in Experiment 1, the researcher's presence during test for the current experiment allowed for sufficient monitoring of head movement so that video was a helpful addition but not required. Therefore, the data from these five participants were retained. The final sample had 48 participants. As with Experiment 1, any trials during which the participant moved their head were removed from analyses. Forty-one

Table 4. Experiment 2 mean reading time (ms), mean error count, and mean judgements of effort.

Dependent variable	Stimulus type			
	UW-UF	UW-RF	RW-UF	RW-RF
Reading time (ms)	13,729 (2,240)	13,955 (2,152)	14,858 (2,597)	14,889 (2,625)
Error count	1.10 (0.73)	1.14 (0.72)	1.33 (0.87)	1.21 (0.81)
Post-trial judgements	2.79 (0.97)	3.40 (0.83)	3.90 (0.98)	4.08 (0.98)
Post-task judgements	1.61 (0.84)	2.93 (0.98)	3.40 (1.42)	4.05 (1.30)

Standard deviations in parentheses. Post-trial and post-task judgements of effort are on 7-point scales. UW-UF: upright words in an upright frame; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.

trials in all were removed for this reason; 0.5% (of total) UW-UF, 0.5% UW-RF, 1.8% RW-UF, and 7.8% RW-RF. Finally, eight trials were removed due to procedural irregularities.[8] After these exclusions, a within-subject, within stimulus type search for outliers at the trial level found no reading times, error count scores, or post-trial judgements of effort with $|z| > 3$. Therefore, no observations were trimmed from these data. In total, 3.2% of observations were removed. As in Experiment 1, provided the uneven exclusion for head rotations (see above), we conducted a second set of analyses excluding participants who had a disproportionate rate of head rotation for the RW-RF stimulus type to have an approximately equal distribution. We excluded eight participants to form a *head-tilt control subset*, and the resulting proportion of trials in these data for each stimulus type was 0.0% UW-UF, 0.4% UW-RF, 1.1% RW-UF, and 2.5% RW-RF. Overall, the results were similar. When an important deviation from the reported results (i.e., the complete sample) was found, we note it in the appropriate section. All analyses were run using the open-source statistical analysis software R, Version 3.4.4. The code and data are available on the Open Science Framework (https://osf.io/zn4mr/). See Table 4 for mean reading times, error counts, post-trial, and post-task judgements of effort by stimulus type.

Reading times. Reading times for each trial were coded by a researcher using the Audacity audio file editing software, measured in milliseconds from stimulus onset to onset of reading the 25th word. A one-way repeated-measures ANOVA with stimulus type as the factor, corrected for sphericity violations, revealed a significant effect of stimulus type on reading time, $F(2.28, 107.12) = 33.86, p < .001$, $\eta_g^2 = .05$. Pairwise *t*-tests were conducted across stimulus types. Compared with UW-UF trials, individuals were no slower on UW-RF trials, $t(47) = 1.91$, $p = .062$, $d = 0.10$; however, individuals were slower on RW-UF trials, $t(47) = 6.62$, $p < .001$, $d = 0.47$, and on RW-RF trials, $t(47) = 7.05, p < .001, d = 0.48$. Compared with UW-RF trials, individuals were slower on RW-UF trials, $t(47) = 5.93$, $p < .001$, $d = 0.38$, and on RW-RF trials, $t(47) = 6.16$, $p < .001$, $d = 0.39$. Individuals were not significantly slower on RW-RF trials than on RW-UF trials, $t(47) = 0.28$, $p = .782$, $d = 0.01$, $BF_{01} = 6.15$. When conducting this analysis on the head-tilt control subset, the results were qualitatively similar except that individuals were significantly slower on UW-RF trials than on UW-UF trials, $t(39) = 2.40, p = .021, d = 0.15$.

Error count. The error count for each stimulus type was coded as per Experiment 1. A one-way repeated-measures ANOVA with stimulus type as the factor revealed that there was a main effect of stimulus type on error count, $F(3, 141) = 3.65$, $p = .013$, $\eta_g^2 = .01$. Pairwise *t*-tests were conducted across stimulus types. Compared with UW-UF trials, individuals made no more errors on UW-RF trials, $t(47) = 0.45$, $p = .653$, $d = 0.05$; however, more errors were made on RW-UF trials, $t(47) = 3.21, p = .002, d = 0.29$. Error counts did not differ between UW-UF and RW-RF trials, $t(47) = 1.45$, $p = .153$, $d = 0.15$. Compared with UW-RF trials, individuals made more errors on RW-UF trials, $t(47) = 2.61$, $p = .012$, $d = 0.25$. Error counts did not differ between UW-RF and RW-RF trials, $t(47) = 1.07$, $p = .289$, $d = 0.10$. Finally, no more errors were made on RW-RF trials than on RW-UF trials, $t(47) = 1.50$, $p = .140$, $d = 0.14$, $BF_{01} = 2.23$. When conducting the analysis on the head-tilt control subset, individuals made more errors on RW-RF trials than on UW-UF trials, $t(39) = 2.16, p = .037, d = 0.23$.

Effort—post-trial judgements. As in Experiment 1, post-trial judgements of effort were assessed through the report of participants' perceived effort immediately after each trial. Each post-trial judgement of effort was between 1 (*not at all effortful*) and 7 (*very effortful*). A one-way repeated-measures ANOVA with stimulus type as the factor, corrected for sphericity violations, revealed that there was a significant effect of stimulus type on post-trial judgements of effort, $F(2.51, 117.74) = 56.27, p < .001$, $\eta_g^2 = .22$. Pairwise *t*-tests were conducted across stimulus types. Compared with UW-UF trials, individuals reported more effort on UW-RF trials, $t(47) = 5.55$, $p < .001$, $d = 0.67$; on RW-UF trials, $t(47) = 9.68$, $p < .001$, $d = 1.14$; and on RW-RF trials, $t(47) = 9.42, p < .001, d = 1.32$. Compared with UW-RF trials, individuals reported more effort on RW-UF trials, $t(47) = 5.32$, $p < .001$, $d = 0.56$, and on RW-RF trials, $t(47) = 6.93$, $p < .001$, $d = 0.75$. The increase in reported effort on RW-RF trials as compared with RW-UF trials was not significant, $t(47) = 1.87, p = .067, d = 0.18, BF_{10} = 0.78$.

Effort—post-task judgement. A one-way repeated-measures ANOVA with stimulus type as the factor, corrected for sphericity violations, revealed that there was a significant effect of stimulus type on post-task judgements of effort, $F(2.29,\ 107.63)=63.81$, $p<.001$, $\eta^2_g=.38$. Pairwise *t*-tests were conducted across stimulus types. Compared with the UW-UF condition, individuals reported more effort for the UW-RF condition, $t(47)=9.47$, $p<.001$, $d=1.44$; the RW-UF condition, $t(47)=8.21$, $p<.001$, $d=1.53$; and the RW-RF condition, $t(47)=12.62$, $p<.001$, $d=2.22$. Compared with the UW-RF condition, individuals reported more effort for the RW-UF condition, $t(47)=2.32$, $p=.025$, $d=0.38$, and the RW-RF condition, $t(47)=6.09$, $p<.001$, $d=0.97$. Finally, there was a significant increase in reported effort for the RW-RF condition as compared with the RW-UF condition, $t(47)=4.36$, $p<.001$, $d=0.48$, $BF_{10}=323.22$.

Exploratory analysis. The following results are not from pre-registered analyses; however, for a more complete picture of the data, we provide them here.

Stimulus type by judgement type interaction. A two-way repeated-measures ANOVA with stimulus type and judgement type (i.e., post-trial or post-task judgements of effort) as factors revealed a significant effect of judgement type, $F(1,\ 47)=19.12$, $p<.001$, $\eta^2_g=.06$, such that post-task judgements were on average lower than post-trial judgements. Moreover, there was an interaction between stimulus type and judgement type, $F(3,\ 141)=14.43$, $p<.001$, $\eta^2_g=.04$. As in Experiment 1, this interaction appears driven by a large difference in judgements of effort for the UW-UF stimulus type, with smaller differences observed for the UW-RF and RW-UF stimulus types, and no difference for the RW-RF stimulus type. Specifically, participants provided higher post-trial judgements of effort than post-task judgements for the UW-UF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=1.18$), and to a lesser extent, the UW-RF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.47$), and the RW-UF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.51$). Finally, there was no difference between post-trial and post-task judgements of effort for the RW-RF stimulus type ($M_{\text{post-trial}}-M_{\text{post-task}}=0.03$). To further explore the nature of the interaction between stimulus type and judgement type, we conducted three 2 (stimulus type) × 2 (judgement type) ANOVAs, comparing UW-UF with UW-RF, UW-RF with RW-UF, and RW-UF with RW-RF. This revealed a significant interaction between stimulus type and judgement type when comparing UW-UF with UW-RF stimulus types, $F(1,\ 47)=22.09$, $p<.001$, $\eta^2_g=.04$. As seen in Figure 3, the effect of judgement type was more pronounced on judgements of effort for the UW-UF stimulus type than for the UW-RF stimulus type. There was no significant interaction between stimulus type and judgement type when comparing UW-RF with RW-UF stimulus types, $F<1$. Unlike in Experiment 1, there was a significant interaction when comparing RW-UF with RW-RF stimulus types, $F(1,\ 47)=9.01$, $p=.004$, $\eta^2_g=.01$. Specifically, as per Figure 3, the effect of judgement type was significantly larger for the RW-UF stimulus type than for the RW-RF stimulus type.

Multilevel regression analysis. As in Experiment 1, to allow a comparison across the effort types, we employed two multilevel regression models using aggregated data. One model featured post-trial judgements as the dependent variable and the other, post-task judgements as the dependent variable. Both models included reading time, error count, and stimulus type as predictor variables. These analyses demonstrated that reading time was not a significant predictor of either post-trial or post-test judgements of effort. However, as in Experiment 1, stimulus type was significantly related to both effort types. See Table 5 for full results.

A third multilevel regression model, with post-trial judgements of effort as the dependent variable, and trial number as a predictor, was computed. As in Experiment 1, this model was at the trial level. Due to obtaining a singular fit with random slopes at the participant level, only random intercepts were included in this model. As per Table 3, reading time, error count, and trial were all significant positive predictors of post-trial effort.

Combined analysis. Given the similarity in design across experiments, a combined analysis was conducted. To test the interaction between stimulus type and judgement type, a 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) × 2 (judgement type: post-trial, post-task) × 2 (Experiment: E1, E2) mixed measures ANOVA was computed, as well as follow-up simple effects tests where applicable. In addition, paired comparisons demonstrating an inconsistent result to those reported for Experiment 1 or 2 will be reported here. Furthermore, to assess the impact of the instructions to read more quickly in Experiment 2, a 4 (stimulus type: UW-UF, UW-RF, RW-UF, RW-RF) × 2 (Experiment: E1, E2) mixed measures ANOVA with reading time as the outcome variable was computed. As above, paired comparisons demonstrating an inconsistent result to those reported previously will be reported here.

Consistent with the results reported for Experiments 1 and 2, the 4 (stimulus type) × 2 (judgement type) × 2 (experiment) mixed measures ANOVA revealed a main effect of judgement type, $F(1,\ 78)=21.84$, $p<.001$, $\eta^2_g=.04$, such that post-trial judgements of effort ($M=3.28$, standard deviation [SD]$=1.10$) were higher than post-task judgements ($M=2.81$, $SD=1.44$). There was also a main effect of experiment, $F(1,\ 78)=12.30$, $p=.001$, $\eta^2_g=.07$, such that, on average, participants' judgements of effort were higher in Experiment 2 ($M=3.27$, $SD=1.30$) than in Experiment 1 ($M=2.70$, $SD=1.23$). Interactions between experiment and judgement type, experiment and stimulus type, and the three-way interaction were all not significant, all Fs <1.17. The interaction between stimulus type and judgement type was significant, $F(3,\ 234)=13.46$, $p<.001$, $\eta^2_g=.02$.

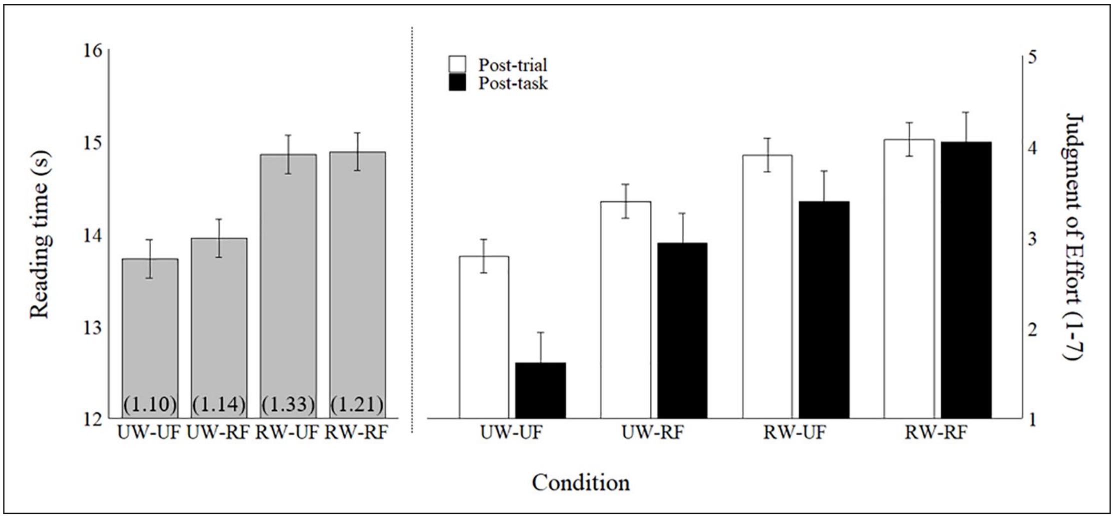

Figure 3. Reading times by stimulus type (left panel) and post-trial versus post-task judgements of effort by stimulus type (right panel) for Experiment 2. Average error count per stimulus type reported in parentheses (left panel). Error bars are Masson–Loftus 95% CI (Loftus & Masson, 1994).

Table 5. Multilevel regression model predicting effort judgements for Experiment 2.

Predictor variable	Post-trial effort		Post-task effort	
	B	*SE*	*B*	*SE*
Intercept	1.39*	0.60	0.76	0.73
Reading time	0.09	0.05	0.06	0.06
Error count	0.16	0.11	–0.02	0.16
Stimulus type				
UW-RF	0.58***	0.10	1.30***	0.18
RW-UF	0.97***	0.11	1.71***	0.19
RW-RF	1.17***	0.11	2.36***	0.19

SE: standard error; UW-UF: upright words in an upright frame; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.
Each participant contributes four observations, one per stimulus type. The stimulus type factor is treatment coded, with UW-UF as the reference group. Reading time and error count were included in the model in their raw forms.
*$p < .05$; ***$p < .001$.

As in Experiments 1 and 2, we conducted three 2 (stimulus type) × 2 (judgement type) ANOVAs, comparing UW-UF with UW-RF, UW-RF with RW-UF, and RW-UF with RW-RF stimulus types. The results of the first two ANOVAs were consistent with the results reported in Experiments 1 and 2. This analysis also revealed a significant stimulus type by judgement type interaction when comparing RW-UF with RW-RF stimulus types, $F(1, 79) = 8.69$, $p = .004$, $\eta_g^2 = .01$, such that, as in Experiment 2, the effect of judgement type was significantly larger for the RW-UF stimulus type than for the RW-RF stimulus type.

A follow-up repeated-measures ANOVA with stimulus type as the factor and post-trial judgements of effort as the outcome variable revealed an effect of stimulus type on post-trial judgements of effort, $F(2.64, 208.73) = 99.18$, $p < .001$, $\eta_g^2 = .22$. Paired comparisons demonstrated that the RW-UF stimulus type was judged as significantly less effortful than the RW-RF stimulus type, $t(79) = 3.81$, $p < .001$, $d = 0.26$, $BF_{10} = 78.40$. A follow-up repeated-measures ANOVA with stimulus types as the factor and post-task judgements of effort as the outcome variable revealed an effect of stimulus type on post-task judgements of effort, $F(2.68, 212.04) = 87.12$, $p < .001$, $\eta_g^2 = .33$. Results of paired comparisons did not deviate from the results reported in Experiment 1 or 2.

The 4 (stimulus type) × 2 (experiment) mixed measures ANOVA with reading time as the outcome variable revealed a significant main effect of stimulus type, $F(2.24, 174.33) = 47.82$, $p < .001$, $\eta_g^2 = .04$. Paired comparisons revealed that, compared with the UW-UF stimulus type, individuals were slower on the UW-RF trials, $t(79) = 2.57$, $p = .012$, $d = 0.10$. There were no other qualitative differences from the results reported for Experiment 1 or 2. In particular, the reading times across the RW-UF and RW-RF stimulus types were equivalent, even in this larger sample. Moreover, there was a significant main effect of experiment on reading time, $F(1, 78) = 13.05$, $p = .001$, $\eta_g^2 = .14$. Participants in Experiment 2 read faster ($M = 14{,}358$, $SD = 2{,}450$) than did participants in Experiment 1 ($M = 16{,}686$, $SD = 3{,}569$). The interaction between stimulus type and experiment was not significant, $F < 1.10$.

In addition to these combined analyses, three multilevel regression models were computed, two using aggregated data at the participant level and one using trial-level data (including random slopes and intercepts for each participant; and random intercepts for each item). This analysis demonstrated qualitatively similar results as those from

Table 6. Multilevel regression model predicting effort judgements for combined data.

Predictor variable	Post-trial effort		Post-task effort	
	B	*SE*	*B*	*SE*
Intercept	0.55	0.43	1.44**	0.51
Reading time	0.09**	0.03	–0.02	0.03
Error count	0.14*	0.07	0.11	0.10
Experiment	0.89***	0.20	0.44*	0.20
Stimulus type				
UW-RF	0.53***	0.08	1.14***	0.14
RW-UF	0.94***	0.08	1.60***	0.15
RW-RF	1.21***	0.09	2.29***	0.15

SE: standard error; UW-UF: upright words in an upright frame; UW-RF: upright words in a rotated frame; RW-UF: rotated words in an upright frame; RW-RF: rotated words in a rotated frame.
Each participant contributes four observations, one per stimulus type. The stimulus type factor is treatment coded, with UW-UF as the reference group. The experiment factor is treatment coded, with E1 as the reference group. Reading time and error count were included in the model in their raw forms.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

each of Experiments 1 and 2, which can be found in Table 3 (trial-level model) and Table 6 (participant-level model).

Discussion

The results of Experiment 2, with respect to reading times, were generally consistent with those of Experiment 1, as well as with previous research by Dunn and Risko (2016). Specifically, participants were fastest when reading the UW-UF and UW-RF stimulus types and slowest when reading the RW-UF and RW-RF stimulus types. As in Experiment 1, there was no difference in reading times between the UW-UF and UW-RF stimulus types (yet, again, this difference was significant when analysing the head-tilt control subset). As in Experiment 1, there was no difference in reading times between the RW-UF and RW-RF stimulus types, and as in Experiment 1, the post-task judgements of effort followed the pattern observed by Dunn and Risko (2016), that being UW-UF < UW-RF < RW-UF < RW-RF. Thus, the dissociation between reading times (i.e., no difference between RW-UF and RW-RF) and post-task judgements of effort was again observed. With respect to this dissociation, the post-trial judgements of effort followed a similar pattern to the post-task judgements; with the exception that, while the RW-RF stimulus type was judged as more effortful than the RW-UF stimulus type, this difference was not significant. Also notable was that, in Experiment 2, there was a significant interaction between stimulus type and judgement type such that the difference between the RW-UF and RW-RF stimulus types was smaller when making post-trial judgements than when making post-task judgements. The latter result is consistent with the idea that post-trial judgements of effort more closely followed reading times. Further to this conclusion, when judgements of effort were made post-task, relative to post-trial, they exhibited a pattern that seemed to more closely resemble that of reading times across conditions (see Figure 3). Specifically, as compared with post-trial judgements, post-task judgements were markedly lower for the UW-UF stimulus type, lower for the UW-RF and RW-UF stimulus types, and the RW-RF stimulus type was unaffected. This general pattern was observed in Experiment 1 though was more pronounced here and will be examined further in the "General discussion" section. Finally, as in Experiment 1, post-task judgements were overall lower than post-trial judgements.

In Experiment 2, a greater emphasis was placed on reading speed. This change in instructions appeared to have its intended effect. That is, a *post hoc* analysis with experiment as a between-subject factor revealed that, across stimulus types, individuals read significantly more quickly in Experiment 2 than in Experiment 1. Interestingly, as per the combined analysis, individuals also judged trials as significantly more effortful, whether judged post-trial or post-task. The influence of this instruction appeared to have the same effect across stimulus types, as there were no interactions between experiment and stimulus type. A feature of the stimuli in Experiment 2 was the removal of 15 relatively high error eliciting words and their replacement with words thought to evoke less errors. While this could have affected participants' performance or effort judgements, this change would arguably have had a minimal effect on the variables of interest as approximately 800 words were read. Thus, individuals appeared to invest more effort in Experiment 2, but this additional investment produced only a main effect.

General discussion

Across two experiments, we demonstrated that judgements of effort can be influenced by how the judgement is elicited. We began this investigation focused on a previously reported dissociation between reading times and judgements of effort. One potential explanation of this dissociation was that it was due to the separation of judgements of effort from the experience of reading. Overall, the present results provide moderate support for this idea. On one hand, individuals judged the RW-RF stimulus type as more effortful to read than the RW-UF stimulus type when providing post-trial, as well as post-task, judgements in Experiment 1, and in Experiment 2, this difference was not significant for post-trial judgements (but was in the same direction). When data were combined across experiments, the results were the same as in Experiment 1; namely, the RW-RF stimulus type was judged as significantly more effortful to read in the post-trial, as well as the post-task, judgement condition. From this perspective, the post-trial and post-task judgements look qualitatively similar. Importantly, however, the RW-UF versus RW-RF difference in Experiment 2 and in the combined analysis was significantly smaller for post-trial judgements than post-task. Thus, while Experiment 1 and the combined analysis provided evidence of the dissociation for post-trial judgements, it was larger for post-task judgements. This result is consistent with the hypothesis that judgements made in closer proximity to the task and with respect to a single trial more closely approximate reading times, which could be interpreted as a proxy for processing fluency (i.e., an experiential source of information). All that said, the effect of when the judgement of effort was made also appeared to have a broader influence. That is, in both experiments, there was a robust interaction between stimulus type and judgement type (i.e., post-trial vs. post-task judgements of effort), whereby the UW-UF stimulus type was judged as significantly less effortful post-task as compared with when judged post-trial, with the UW-RF and RW-UF stimulus types exhibiting this pattern but with a notably smaller magnitude, and—as noted above—no difference was observed for the RW-RF stimulus type. We examine this interaction further below along with the broader implications of the present work for our understanding of judgements of effort.

The stimulus type by judgement type interaction

As noted above, the most robust result from Experiments 1 and 2 was the interaction between stimulus type and judgement type. As articulated in the "Introduction" section, one possible explanation for such an effect starts with the notion that individuals rely more on beliefs when making post-task judgements and more on experiences when making post-trial judgements. In this regard, the largest difference across effort types was at the UW-UF stimulus type, indicating the possible involvement of a belief-based inference about upright versus disoriented text. It seems reasonable to suggest that individuals believe that it takes less effort to read an upright display than a disoriented display; thus, relying primarily on beliefs might lead to a large separation between the only upright stimulus type (i.e., UW-UF) and the three disoriented stimulus types (i.e., UW-RF, RW-UF, RW-RF). That is, the qualitative change between "not rotated" and "rotated" might weigh heavily when individuals are inferring effort based on beliefs and are relatively separated from the experience of processing the stimulus. The idea that, in certain situations, incremental differences are less salient than are categorical differences has been suggested previously (Dunn et al., 2017; Hsee & Zhang, 2010). Critically, reading times reveal relatively modest effects for certain stimulus rotations (e.g., UW-UF vs. UW-RF). If we assume that post-trial judgements are more closely tied to experiences (e.g., reading time), then this provides a plausible explanation for why the UW-UF stimulus type lies much closer to the rotated stimulus types (as is the case with reading times) when individuals are making post-trial judgements (see Figures 2 and 3).

Further evidence that the stimulus type by judgement type interaction might be due to experiences having a greater influence on post-trial judgements was present in the RW-RF versus RW-UF comparison. As described above, there was modest support for a reduction from post-task to post-trial judgements in this difference. Critically, this change brings effort judgements more in line with reading times, for which there is no difference between these stimulus types. As previously suggested (Dunn & Risko, 2016), when making judgements post-task, individuals may rely on the belief that the RW-RF stimulus type is more effortful to read than is the RW-UF stimulus type, as the former comprises rotation of the frame as well as the words (i.e., it is "more" rotated).

While aspects of the stimulus type by judgement type interaction seem compatible with post-trial judgements being more influenced by experiences than are post-task judgements, this was not universally the case. In particular, when considering the difference between the UW-RF and RW-RF stimulus types, there was no change across post-trial and post-task judgements. But, in reading time, this difference is consistently one of the largest. Indeed, reading times seem largely a product of word rotation (see Figures 2 and 3). If individuals' post-trial judgements were related more strongly to reading time, then one could reasonably expect the UW-RF versus RW-UF difference to increase in magnitude across the judgement types. Thus, there might be an alternative explanation for the interaction between stimulus type and judgement type.

With respect to the intrinsic differences between the judgement types, it is worth noting that there was a difference across judgement types with respect to the individual words that made up the stimulus displays (i.e., "WORD" for each word for post-task displays vs. unique words for the main trial displays). It is unlikely that this difference is responsible for the stimulus type by judgement type interaction. If this was the case, then we would expect to observe a main effect of judgement type (e.g., judgements might be lower overall post-task because reading "WORD" repeatedly would be easier) but no stimulus type by judgement type interaction.

As discussed in the "Introduction" section, previous studies (Schmeck et al., 2015; van Gog et al., 2012) have demonstrated that judgements of effort were higher when measured post-task, as compared with post-trial. One reason put forth for this effect was that participants may have perceived post-task judgements as a single judgement of one long task composed of several components. In the current study, the opposite effect was observed, and in fact, a post hoc analysis with the combined data confirmed that this effect remained significant after the UW-UF trials were removed. Thus, the lower judgements of effort post-task are not exclusively due to the large difference for the UW-UF stimulus type. As the post-task judgements for the current study involved providing a judgement for each stimulus type, rather than for an entire block of similar problems, one might expect a different relation between the two effort types. Future research into which tasks elicit higher, versus lower, post-task judgements of effort would help illuminate the reasons for this discrepancy.

Metacognitive framework for judgements of effort

As discussed in the "Introduction" section, a judgement of effort can be viewed as a type of metacognitive judgement. From this perspective, deciding the effortfulness of a given cognitive act involves making an inference based on available information. These judgements might be more experiential, for example, based on the experience of fluency, or more belief-based, for example, the belief that reading rotated text is effortful. While the present work was not a direct test of the metacognitive approach, the judgement type by stimulus type interaction reported in both experiments clearly supports it. That is, judgements of effort were demonstrated to be a function of the context in which the judgement was made (i.e., whether they were solicited post-trial or post-task), a result captured naturally in this framework as a shift in the sources of information relied on across judgement contexts.

The current study also speaks to the nature of metacognitive judgements more generally. That is, the results suggest that the information used to inform metacognitive judgements can change as a function of when the judgement is elicited. Specifically, a cue such as response time or fluency might have a stronger influence when the judgement is provided immediately after a trial, whereas pre-experimental beliefs (or in situ inferences) may play a larger role when the judgement more so resembles the post-task judgements here. This is consistent with the result discussed in the "Introduction" section that the relation between JOLs and self-directed study time was stronger when JOLs were elicited immediately, as compared with after a delay (Koriat et al., 2006).

An important question with respect to effort as a metacognitive judgement is whether this pattern of results would change given a significantly more effortful task. One might hypothesise that individuals' ability to monitor their effort levels trial-by-trial might be compromised by virtue of the increased cognitive load imposed by a difficult task (van Gog et al., 2012). In such a paradigm, the pattern of post-trial judgements across stimulus types may be inconsistent with the results of the current study. In any case, further investigation is required on this front. Along the same lines, as noted in the "Introduction" section, the post-trial and post-task judgements of effort differ in their temporal proximity to the task as well as in their scope. Specifically, the former is a judgement of a single trial and the latter is an overall judgement of a given stimulus type. If one were to investigate the effect of delay on judgements of effort, the same judgement type could be used with a manipulation of the temporal proximity to the task, to isolate the specific contribution of temporal proximity. In sum, future research focusing on whether the temporal proximity to other cognitive tasks plays a role in individuals' judgements of effort would be valuable.

Conclusion

The present investigation aimed to determine whether judgements of effort depend on judgement type. The critical contribution was the discovery of a stimulus type by judgement type interaction in the context of judgements of effort, which could be interpreted as a shift in the sources of information used to inform judgements of effort as a function of judgement type. Future work aiming to illuminate the underlying factors that contribute to judgements of effort, along with focusing on the effects of various contexts (e.g., single vs. joint evaluation; Dunn et al., 2017) in which judgements of effort are made, will provide a deeper understanding of decisions about our expenditures of cognitive effort.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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.

ORCID iD

Michelle Ashburner https://orcid.org/0000-0002-1839-2909

Notes

1. Some of the words were repeated across stimuli; specifically, 82 words appeared in two different stimuli, and 4 words appeared in three different stimuli.
2. This pattern of spontaneous head rotation is consistent with the results of Dunn and Risko (2016), where participants were free to rotate their heads while reading.
3. Two participants had an unequal distribution of trials. One completed nine UW-UF (upright words in an upright frame) trials and seven RW-RF (rotated words in a rotated frame) trials; another completed nine UW-RF (upright words in a rotated frame) trials and seven RW-UF (rotated words in an upright frame) trials.

4. Cohen's *d* was computed using the effsize::cohen.d function, which assumes independent samples.
5. The prior distribution used in these analyses is the default prior of 0.707, corresponding to effect sizes ranging from a Cohen's *d* of −2 to 2.
6. Similar multilevel regression models were computed with reading time, and error count, as the outcome variable. While error counts increased with the number of trials, reading times decreased. Thus, the increase in effort might reflect an attempt to read faster as trials progressed.
7. Some of the words were repeated across stimuli; specifically, 82 words appeared in two different stimuli, and 4 words appeared in three different stimuli.
8. These irregularities comprised either researcher interference in a trial to clarify instructions (e.g., reminder to keep head upright or to follow arrows indicating reading direction) or a participant coughing during a trial.

References

Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psihologijske Teme*, *28*(1), 1–20. https://doi.org/10.31820/pt.28.1.1

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgements during reasoning and memorisation. *Thinking & Reasoning*, *23*(4), 376–408. https://doi.org/10.1080.2017.1328373

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Baars, M., Wijnia, L., & Paas, F. (2017). The association between motivation, affect, and self-regulated learning when solving problems. *Frontiers in Psychology*, *8*, Article 1346. https://doi.org/10.3389/fpsyg.2017.01346

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. https://doi.org/10.1037/0096-3445.127.1.55

Boldt, A., & Gilbert, S. J. (2019). Confidence guides spontaneous cognitive offloading. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/ct52k

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the Cue for Judgments of Learning (JOL) and the Cue for Test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, *36*(1), 34–49. https://doi.org/10.1006/jmla.1996.2476

Dunn, T. L., Gaspar, C., & Risko, E. F. (2019). Cue awareness in avoiding effortful control. *Neuropsychologia*, *123*, 77–91. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/xhp0000236

Dunn, T. L., & Risko, E. F. (2016). Toward a metacognitive account of cognitive offloading. *Cognitive Science*, *40*(5), 1080–1127. https://doi.org/10.1111/cogs.12273

Eggemeier, F. T., & Stadler, M. A. (1984). Subjective workload assessment in a spatial memory task. In *Proceedings of the Human Factors Society Annual Meeting* (*Vol. 28*, No. 8, pp. 680–684). Los Angeles, CA: Sage Publications. https://doi.org/10.1177/154193128402800808

Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future: Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, *2016*(1), Article niw018. https://doi.org/10.1093/nc/niw018

Foo, M.-D., Uy, M. A., & Baron, R. A. (2009). How do feelings influence effort? An empirical study of entrepreneurs' affect and venture effort. *Journal of Applied Psychology*, *94*(4), 1086–1094. https://doi.org/10.1037/a0015599

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245–260. https://doi.org/10.1016/j.concog.2015.01.006

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.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults and preschoolers' ability to infer the difficulty of novel tasks. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 458–463). Cognitive Science Society.

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

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and Valued. *Trends in Cognitive Sciences*, *22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), Article 2. https://doi.org/10.7771/1932-6246.1167

Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. *Advances in Psychology*, *52*, 5–39. https://doi.org/10.1016/S0166-4115(08)62381-X

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*, 899–908.

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: A comparison of different objective measures. *Instructional science*, *45*(4), 515–536.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.609

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. https://doi.org/10.1016/j.jml.2005.01.001

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1624–1637. https://doi.org/10.1037/xlm0000009

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology*, *7*, 67–70. https://doi.org/10.1016/j.copsyc.2015.08.003

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants* 2002 (pp. 7–7). Scottsdale, AZ. https://doi.org/10.1109/HFPP.2002.1042860

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1084–1097. https://doi.org/10.1037/a0012580

Moray, N. (1982). Subjective mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(1), 25–40. https://doi.org/10.1177/001872088202400104

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are extremely accurate at predicting subsequent recall: The "Delayed-JOL Effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Potts, C. A., Pastel, S., & Rosenbaum, D. A. (2018). How are cognitive and physical difficulty compared? *Attention, Perception, & Psychophysics*, *80*(2), 500–511. https://doi.org/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. https://doi.org/10.1016/j.learninstruc.2016.12.002

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language*, *51*(1), 71–79. https://doi.org/10.1016/j.jml.2004.03.004

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. https://doi.org/10.1007/s11251-014-9328-3

Siedlecka, M., Paulewicz, B., & Wierzchon, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, *7*, Article 218. https://doi.org/10.3389/fpsyg.2016.00218

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. https://doi.org/10.1111/j.1467-9280.2008.02189.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. https://doi.org/10.1016/j.cognition.2012.90.012

Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review*, *24*(6), 2003–2011. https://doi.org/10.3758/s13423-017-1261-4

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/s1342

van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*(6), 833–839. https://doi.org/10.1002/acp.2883

Weaver, C. A., & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science*, *8*(4), 318–321. https://doi.org/10.1111/j.1467-9280.1997.tb00445.x

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), Article e68210. https://doi.org/10.1371/journal.pone.0068210

Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(1), 111–120. https://doi.org/10.1177/001872088803000110

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *arXiv preprint arXiv:1905.04445.*