Journal of Behavioral Decision Making, *J. Behav. Dec. Making*, **30**: 869–888 (2017)
Published online 12 April 2017 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.2018

Evaluating Effort: Influences of Evaluation Mode on Judgments of Task-specific Efforts

TIMOTHY L. DUNN,* DEREK J. KOEHLER and EVAN F. RISKO
University of Waterloo, Waterloo, ON Canada

ABSTRACT

The claim that humans adapt their actions in ways that avoid effortful processing (whether cognitive or physical) is a staple of various theories of human behavior. Although much work has been carried out focusing on the determinants of such behaviors, less attention has been given to how individuals evaluate effort. In the current set of experiments, we utilized the general evaluability theory to examine the evaluability of effort by examining subjective value functions across different evaluation modes. Individuals judged the anticipated effort of four task-specific efforts indexed by stimulus rotation, items to be remembered, weight to be lifted, and stimulus degradation across joint (i.e., judged comparatively) and single evaluation modes (i.e., judged in isolation). General evaluability theory hypothesizes that highly evaluable attributes should be consistently evaluated (i.e., demonstrate similar subjective value functions) between the two modes. Across six experiments, we demonstrate that the perceived effort associated with items to be remembered, weight to be lifted, and stimulus degradation can be considered relatively evaluable, while the effort associated with stimulus rotation may be relatively inevaluable. Results are discussed within the context of subjective evaluation, internal reference information, and strategy selection. In addition, methodological implications of evaluation modes are considered. Copyright © 2017 John Wiley & Sons, Ltd.

KEY WORDS perceived effort; evaluation mode; subjective value; reference information; judgment and decision making

The notion that individuals configure their behaviors to avoid effort is considered to be a *principle* (e.g., Clark, 2010; Zipf, 1949) or *law* (Hull, 1943) of human behavior, and consequently, has become critical to several theories spanning various sectors of psychology such as judgment and decision making (e.g., Stanovich, 2011). Recently, a specific focus on effort-based decision making, that is, avoiding an effortful task by choosing to engage in a less effortful alternative, has pushed the hypothesis to the fore as a critical behavior to understand empirically in human decision making (e.g., Botvinick & Braver, 2015; Dunn, Lutes, & Risko, 2016; Dunn & Risko, 2016; Kool, McGuire, Rosen, & Botvinick, 2010; Kool & Botvinick, 2014; Kurniawan, Guitart-Masip, & Dolan, 2011; McGuire & Botvinick, 2010; Westbrook & Braver, 2015; Westbrook, Kester, & Braver, 2013).

Although many of these endeavors have focused on the factors that drive effort avoidance (e.g., cues, Dunn et al., 2016; demands on executive control, Kool et al., 2010), less attention has been paid specifically to individuals' subjective evaluation of efforts (Bartra, McGuire, & Kable, 2013; Clithero & Rangel, 2014; Kable & Glimcher, 2009; Otto, Zijlstra, & Goebel, 2014). That is, how do we appraise the level of anticipated effort associated with some action within the decision-making process? In the current set of experiments, we examined subjective evaluations of task-specific efforts in three domains through the scope of general evaluability theory (GET: Hsee & Zhang, 2010).

General evaluability theory

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

Within GET, evaluability is dependent on three types of reference information: *mode*, *knowledge*, and *nature*. First, mode refers to what evaluation mode a person is placed in. Any evaluation takes place in either a (i) joint evaluation

*Correspondence to: Timothy L. Dunn, Department of Psychology, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada. E-mail: t2dunn@uwaterloo.ca
All data and codes are freely available via the Open Science Framework at osf.io/t4uv4.
This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to E. F. R.

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

General evaluability theory provides a theoretically motivated means of assessing the evaluability of a given attribute as it posits a straightforward relation between level of evaluability and mode: low-sensitivity attributes will produce less evaluability in the SE mode relative to the JE mode. As Hsee et al. (1999) note, difficult-to-evaluate attributes have little influence in differentiating the evaluations of values in SE, whereas in the comparative JE mode, difficult-to-evaluate attributes become easier to evaluate and hence exert a greater influence. By contrast, easier-to-evaluate attributes will have a similar impact in SE and JE. The effect of mode (SE vs. JE) should then be observable in individuals' subjective value functions (i.e., the functions that demonstrate how objective quantities map onto subjective evaluations; Kahneman & Tversky, 1979). GET proposes that subjective value functions for relatively inevaluable attributes are more linear in JE relative to SE (Figure 1) by virtue of individuals being similarly sensitive to qualitative/categorical (i.e., from some meaningful reference point) and quantitative/continuous differences (i.e., judgments not made to the reference point) across incremental values. In the SE mode, in contrast, individuals are expected to be more sensitive to qualitative differences from a reference point and show less sensitivity across quantitative values thereafter similar to a log function. Examining the relation between levels of some value, such as effort, and subjective values across evaluation modes thus provides a useful tool in determining the evaluability of a given value.

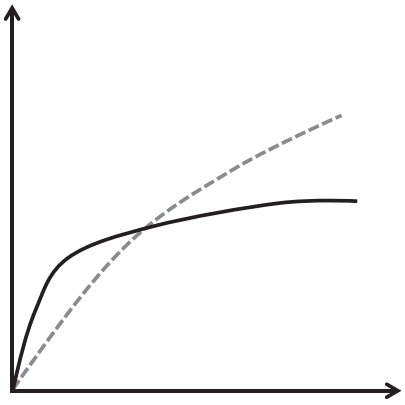

Figure 1. Subjective value functions for a relatively inevaluable attribute under single and joint evaluation modes. The subjective value function for the single evaluation (SE) mode is presented as the solid line, whereas the function for the joint evaluation (JE) mode is presented as the dashed line. General evaluability theory proposes that subjective value functions are more linear in JE across values than in SE (adapted from Hsee & Zhang, 2010)

The Evaluability of Effort

Although the evaluability of effort has yet to receive considerable attention, it would seem plausible to suggest that effort should be a highly evaluable dimension. That is, effort may be high in both knowledge, where individuals putatively have a swath of experience making effort-based decisions, and nature, where effort may be inherently evaluable. Consistent with these ideas, there have been a number of demonstrations that individuals can behave in a manner that would suggest the ability to accurately evaluate effort (e.g., Bitgood & Dukes, 2006; Gray, Sims, Fu, & Schoelles, 2006; Kool et al., 2010; Kurniawan et al., 2011; Siegler & Lemaire, 1997; Walsh & Anderson, 2009). As an example, utilizing a free-choice task where participants made decisions between simply by holding a grip device and engaging in effortful gripping (i.e., squeezing of air compressed cylinders), Kurniawan et al. (2011) demonstrated less frequent engagement in the high-effort gripping option. In a similar vein, Walsh and Anderson (2009) demonstrated that as the effort to successfully compute a solution to a multiplication problem increased, and consequently performance decreased, reliance on an external strategy of using a calculator increased. Similar findings consistent with a least effort principle have additionally been demonstrated in a variety of animal behaviors, for example, foraging (Marsh, Schuck-Paim, & Kacelnik, 2004; Stephens & Krebs, 1986).

While the evaluability of effort appears intuitive, it might be prudent to consider whether different types of effort may be more or less evaluable. One critical distinction to make may thus be between more physical and more cognitive (mental) forms of effort. With regard to physical effort, one could hypothesize that individuals possess a clear signal (e.g., energetic costs) needed to accurately evaluate the costs of expending physical effort. However, this does not appear to be the case with regard to cognitive effort. Cognitive effort is a far more challenging construct to define and study empirically (Botvinick & Braver, 2015; Dunn & Risko, 2016; Kurzban, 2016; Westbrook & Braver, 2015). Although many accounts have attempted to generalize the energetic aspect of physical effort accounts to cognitive effort accounts (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Boksem & Tops, 2008; Gailliot & Baumeister, 2007), this premise has been met with a large amount of skepticism on the basis of theoretical and empirical grounds (e.g., Botvinick & Braver, 2015; Carter & McCullough, 2013; Gibson, 2007;

DOI: 10.1002/bdm

Hockey, 2011; Inzlicht & Schmeichel, 2013, 2012; Job, Walton, Bernecker, & Dweck, 2013; Kelly, Sünram-Lea, & Crawford, 2015; Lange & Eggert, 2014; Lurquin et al., 2016; Kurzban, 2010; Kurzban, Duckworth, Kable, & Myers, 2013; Raichle & Mintun, 2006; Vadillo, Gold, & Osman, 2016; Westbrook & Braver, 2015). Furthermore, whether or not similar systems underlie evaluation of both physical and cognitive effort remains to be determined (Westbrook & Braver, 2015; cf. Boksem & Tops, 2008). Therefore, the information (e.g., reference information) available to individuals while attempting to evaluate physical and cognitive effort may indeed differ.

PRESENT INVESTIGATION

In the present investigation, we use the GET framework to examine the evaluability of task-specific efforts via manipulations of evaluation mode in the context of effort judgments across perceptual, memorial, and motor tasks. Here, we focus on evaluations of *anticipated* effort as opposed to *experienced* effort. Payne, Bettman, and Johnson (1993) make a critical distinction between the two arguing that it is the evaluations and judgments of *anticipated* (or perceived) effort that play the primary role in strategy selection and decision making, for example, avoiding engaging in a task outright, although experienced effort can indeed theoretically become anticipated effort given some hypothesized monitoring mechanism (e.g., Anzai & Simon, 1979; Koriat & Levy-Sadot, 2001; Payne et al., 1993; Son & Metcalfe, 2005; Vernon & Usher, 2003). Several recent proposals have argued that perceived effort is subjective in nature (Dunn & Risko, 2016; Dunn et al., 2016; Kool et al., 2010; Westbrook & Braver, 2015) serving as a type of "summary signal" used to select lines of action (Dunn et al., 2016). It is important to note, though, that while effort is often closely coupled with other potential determinants of behavior (e.g., task difficulty) and often covaries with similar signals (e.g., fatigue or arousal), effort can be understood as a unique cost considered in the decision-making process (for reviews on these issues, see Kurzban et al., 2013, and Westbrook & Braver, 2015).

The application of the GET framework to effort affords an examination of whether different task-specific efforts can be considered evaluable. The extent to which effort judgments may vary across JE and SE modes can provide evidence concerning the evaluability of a given type of task-specific effort. Critically, attributes that are high in evaluability do not demonstrate increased sensitivity in the JE mode because they are evaluable in the SE mode as well. Highly evaluable attributes should thus be consistently evaluated (i.e., demonstrate similar patterns of judgments) across the two modes. If a given task-specific effort does not show a susceptibility to evaluation mode, then we would expect subjective rating functions in both SE and JE to be similar (cf. Figure 1). Such a pattern, if observed consistently across different types or manipulations of effort, would provide initial evidence that effort and its determinants are evaluable. That being said, given the partial exploratory nature of Experiment 1, no specific hypothesis is offered concerning whether one specific type of effort is expected to be evaluable or not.

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

EXPERIMENT 1

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

Method

Participants

Five hundred and forty Amazon Mechanical Turk (MTurk) workers participated in the online study (Buhrmester, Kwang, & Gosling, 2011) for compensation of $US1. One hundred and eighty individuals were assigned to each of the effort dimensions, 30 in each of the six nested effort manipulation judgment groups (five in SE and one in JE; see later texts). Twenty-four percent of individuals failed at least one of three attention checks embedded in the survey (see later texts) resulting in a final N of 411 (M_{Age} = 34 years, 47% female participants, 57% reported completing a bachelor's degree or higher).

Design

A 3 (Effort Dimension: Perceptual Task, Memorial Task, Motor Task) × 5 (Effort Level: Stimulus Rotation—0°, 45°,

DOI: 10.1002/bdm

90°, 135°, 180°; Set Size—two, four, six, eight, 10 letters; Weight—5, 10, 15, 20, 25 lb) × 2 (Evaluation Mode: JE, SE) design was employed.

Stimuli

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

Procedure

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

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

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

Results

All reported analyses were conducted using R statistical software (R Development Core Team, 2014). Results are reported first for JE judgments followed by SE judgments for each effort dimension (Figures 2 and 3). For JE judgments (i.e., within-subject judgments), linear mixed models (LMMs) were constructed using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015a). All models incorporated a crossed random-effects structure including random subject slopes, and slope-by-intercept correlations[1] (Baayen, Davidson, & Bates, 2008). In addition, the *RePsychLing* package (Bates, Kliegl, Vasishth, & Baayen, 2015b) was employed to ensure random-effects structures were not over fitted (cf. Barr, Levy, Scheepers, & Tily, 2013). Significance criterion for slope terms was set as $|t| > 2$ following Baayen et al. (2008). Model assumptions were assessed using visual depictions of residuals plots using the *car* package (Fox & Weisberg, 2011). In addition, influential case analysis (i.e., Cook's

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

DOI: 10.1002/bdm

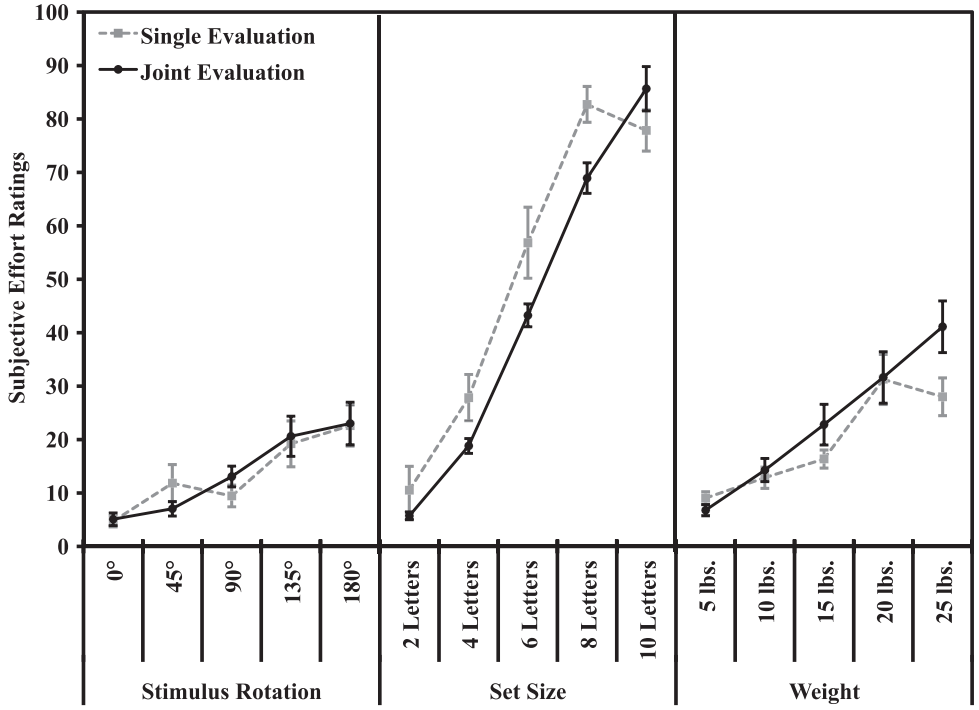

Figure 2. Single and joint evaluation subjective effort rating results in Experiment 1. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars represent ±1 standard error of mean (SEM)

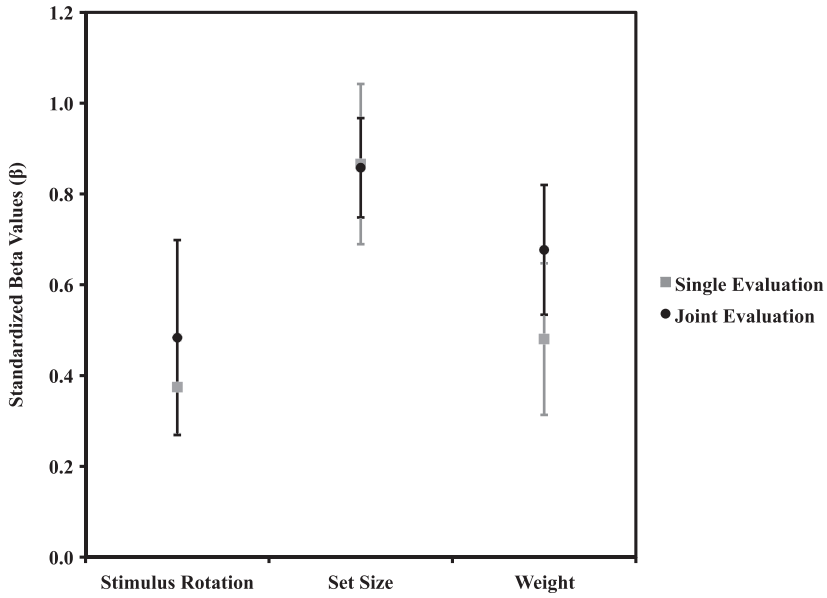

Figure 3. Standardized beta values for slopes in Experiment 1. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars represent bootstrapped 95% confidence intervals

distance) was conducted using the *influence.ME* package (Nieuwenhuis, Grotenhuis, & Pelzer, 2012). To test slope model goodness of fit, we compared the LMM containing the slope term (i.e., effort level for a particular dimension) against an intercept-only model using a log-likelihood test. SE judgments were analyzed using linear regression models (LM). Model assumptions and influential case analyses followed a similar procedure to that of LMMs. Removed cases (e.g., trials for LMMs and subjects for LMs) are reported at the start of each effort dimension section for both LMMs and LMs with all procedures following an iterative process for removal. Standardized beta values (β) and bootstrapped confidence intervals (CIs) are provided for all slope estimates. Last, all models were visually compared with loess fits to ensure linear models relative to nonlinear models were the most appropriate fit to the data post-hoc.

Stimulus rotation (perceptual)

First, approximately 2% of cases were removed for JE. LMM results demonstrated a significant positive slope associated with increased degree of stimulus rotation, b = 4.71,

DOI: 10.1002/bdm

SE = 1.07, t = 4.42, 95% CI [2.61, 6.69], β = .48, β 95% CI [.27, .69]. The slope model significantly improved model fit relative to the intercept-only model, $\chi^2(1) = 15.01$, $p < .001$. For SE, approximately 8% of cases were removed. Similar to LMM results, LM results demonstrated a significant positive slope of stimulus rotation, b = .07, *SE* = .02, t = 4.16, $p < .001$, 95% CI [.04, .1], R^2 = .14, β = .37, β 95% CI [.2, .55]. The slope model significantly improved model fit relative to the intercept-only model, $F(1, 106) = 17.29$, residual *SE* = 10.47, $p < .001$.

Set size (memorial)

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

Weight (motor)

Approximately 2% of cases were removed in JE. LMM results demonstrated a significant positive slope associated with increases in weight, b = 8.01, *SE* = .86, t = 9.29, 95% CI [6.44, 9.69], β = .68, β 95% CI [.53, .82]. The slope model significantly improved model fit relative to the intercept-only model, $\chi^2(1) = 40.21$, $p < .001$. Approximately 8% of cases were removed in SE. LM results demonstrated a significant positive slope of stimulus rotation, b = .7, *SE* = .12, t = 5.64, $p < .001$, 95% CI [.48, .98], R^2 = .23, β = .48, β 95% CI [.31, .65]. The slope model significantly improved model fit relative to the intercept-only model, $F(1, 106) = 31.78$, residual *SE* = 8.99, $p < .001$.

Discussion

Experiment 1 demonstrated several interesting findings with regard to the evaluability of task-specific efforts. First, effort judgments for the memory task produced very similar positive slopes across both the JE and SE evaluation modes. This finding provides initial evidence that effort associated with items to be remembered may be highly evaluable. Both stimulus rotation and weight to be lifted showed similar patterns of perceived effort judgments across evaluation modes, but not to the same extent as the memory task. Slopes were somewhat more positive in JE relative to SE, although both dimensions did produce significant linear functions (i.e., all effort dimensions remained relatively linear in the SE mode). That is, for all dimensions across between-subject raters, only rating a single effort level in isolation appeared to demonstrate value sensitivity in a similar fashion to individuals rating in JE where all alternatives were explicitly present. Such correspondence suggests that each task-specific effort dimension may have high levels of both knowledge and nature according to GET. Examination of the stimulus rotation effort slopes, and to some extent the weight effort slopes, though suggests that the judgments were relatively constrained towards the bottom of the rating scale. Therefore, it is possible that the overall low perceived effort for these dimensions did not allow the shape of the function to be fully demonstrated across effort manipulations. Experiment 2 looked to address this issue.

EXPERIMENT 2

Individuals in Experiment 2 performed a task similar to Experiment 1. To increase effort ratings across the evaluation scale, and provide a clearer picture of the functions associated with the stimulus rotation and weight dimensions, effort levels were increased while effort associated with the memory task was kept consistent with Experiment 1. For stimulus rotation, displays were increased from a one-word display to a nine-word 3 × 3 display. The addition of items in a rotated display is known to increase performance costs, and thus, the putative expected effort required to read the display (Risko, Medimorec, Chisholm, & Kingstone, 2014). For weight, effort levels were doubled relative to Experiment 1. Effort levels were presented in 10-lb increments starting at 10 lb and ending at 50 lb. Experiment 2 additionally affords the opportunity to replicate the main findings from Experiment 1. If all task-specific effort dimensions are evaluable, then we would expect to observe similar positive linear functions across evaluation modes for all dimensions.

Method

Participants

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

Design

A 3 (Effort Dimension: Perceptual Task, Memorial Task, Motor Task) × 5 (Effort Manipulation: Stimulus Rotation—0°, 45°, 90°, 135°, 180°; Set Size—two, four, six, eight, 10 letters; Weight—10, 20, 30, 40, 50 lb) × 2 (Evaluation Mode: JE, SE) design was employed.

Stimuli

Stimuli for the memorial task manipulations were kept the same as Experiment 1. For the perceptual task manipulations, stimulus rotation was kept constant, but set size was increased to nine words presented in a 3 × 3 display.

DOI: 10.1002/bdm

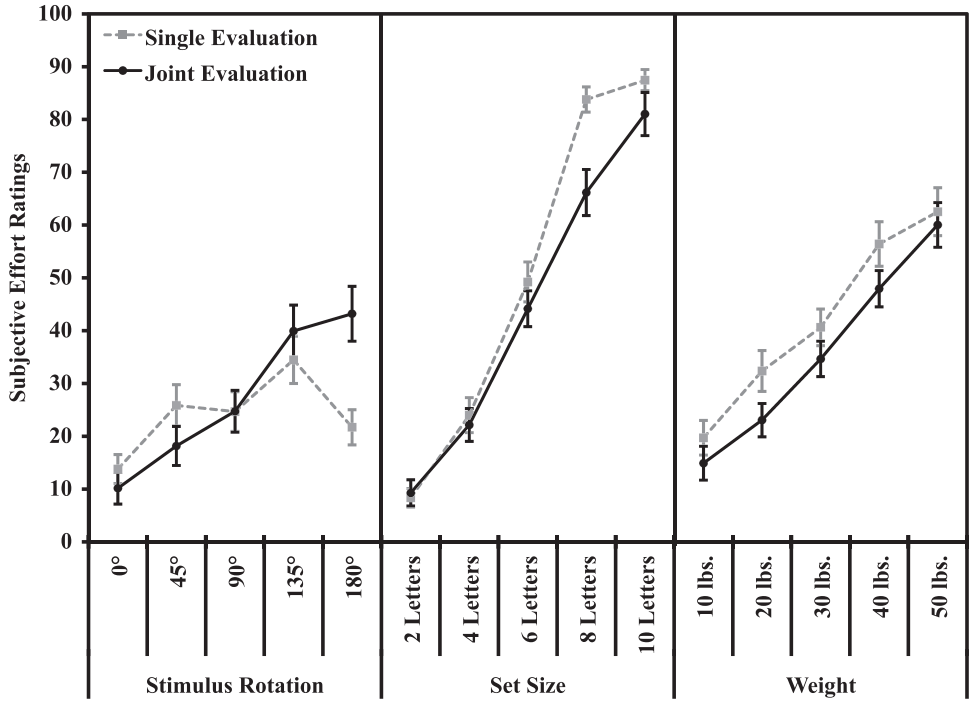

Figure 4. Single and joint evaluation subjective effort rating results in Experiment 2. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars represent ±1 standard error of mean (SEM)

Motor task manipulations were increased to 10 lb to 50 lb in 10-lb increments.

Procedure

All procedures followed Experiment 1.

Results

All data analyses and reporting procedures followed Experiment 1 (Figures 4 and 5).

Stimulus rotation

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

In addition to a demonstrated nonsignificant positive slope, visual inspection of loess fit to the data suggested that a nonlinear model would perhaps best describe the SE data. Therefore, to test a nonlinear fit, a generalized additive model (GAM) was constructed using the *mgcv* package (Wood, 2006). A cubic spline smoothing term was applied to degree of stimulus rotation utilizing three knots. Results demonstrated a significant smooth term, $edf = 1.92$, $F = 7.05$, approximate $p < .001$.[2] Furthermore, a deviance test demonstrated that the GAM produced a better fit to the SE data relative to the LM, $p < .001$, as well as a smaller Akaike information criterion (AIC; i.e., better goodness of fit; Burnham, Anderson, & Huyvaert, 2011) value relative to the LM, $AIC = 1490$ and $AIC = 1501.71$, for the GAM and LM, respectively. Thus, GAM results suggest that the nonlinear model provided a better fit to the SE data relative to the LM (Figure 6). We return to the importance of this pattern in the discussion.

Set size

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

Weight

Approximately 6% of cases were removed in JE. LMM results demonstrated a significant positive slope associated

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

DOI: 10.1002/bdm

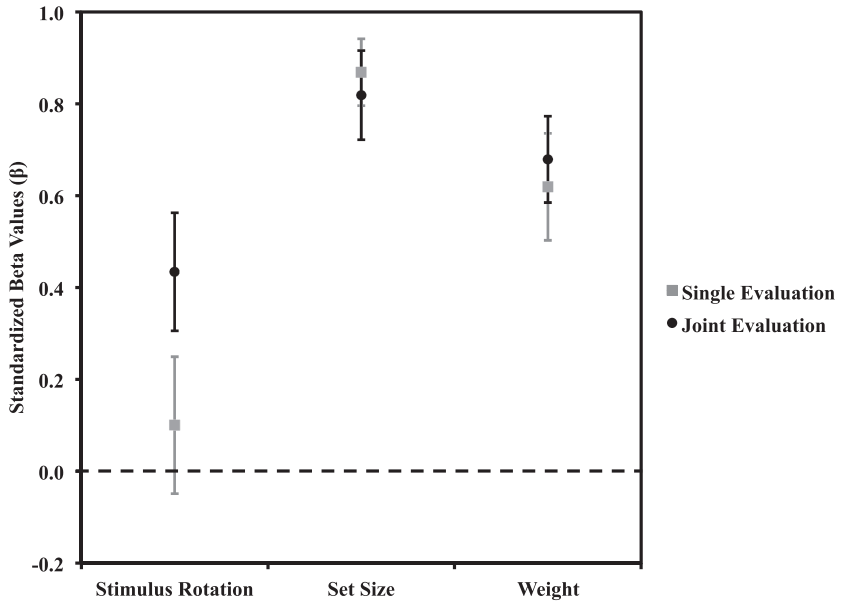

Figure 5. Standardized beta values for slopes in Experiment 2. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars represent bootstrapped 95% confidence intervals

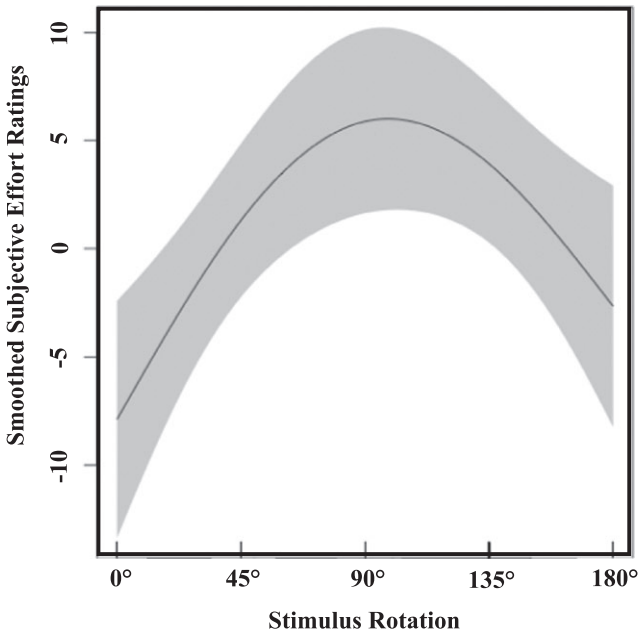

Figure 6. Generalized additive model (GAM) predicted fit for single evaluation (SE) stimulus rotation effort ratings in Experiment 2. *Note*: Shaded region represents estimated SE of the smoothed estimate

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

Discussion

Experiment 2 replicated the main finding of Experiment 1 for the memorial and motor tasks: both SE and JE produced highly similar positive linear functions of perceived effort ratings for the dimensions. Thus, Experiments 1 and 2 together suggest that the perceived effort associated with the specific memorial and motor tasks used here is evaluable. We return to potential explanations of why this may be the case for these dimensions in Experiment 5 and the General Discussion. In contrast to these tasks, stimulus rotation produced a flat slope in SE (as well as a nonlinear fit), whereas a positive linear pattern was observed in JE. As highlighted in the introduction, GET proposes that subjective value functions are more linear in JE across values relative to SE for inevaluable attributes (Figure 1) owing to individuals being similarly sensitive to quantitative and qualitative differences in JE, but only being sensitive to qualitative differences (i.e., from some meaningful reference point) with ratings asymptoting thereafter in SE. Individuals in SE seemed to be insensitive to the incremental differences in stimulus rotation past the 0–45° comparison. That is, individuals' effort judgments were sensitive to the qualitative shift from the 0° reference point, but not sensitive to the incremental differences between 45–90°, 90–135°, and 135–180°. Therefore, as opposed to Experiment 1, when the putative perceived effort was increased for stimulus rotation, the perceived effort associated with the task appeared to be relatively inevaluable.

EXPERIMENT 3

One potential explanation for the pattern of results for effort associated with stimulus rotation in Experiment 2 is that, in addition to 0° (i.e., upright), 180° stimulus rotation (i.e., upside down) may also serve as a meaningful reference point in judgments. This can be seen in the large drop in effort

DOI: 10.1002/bdm

ratings from 135° to 180° (Figure 4). Therefore, the nonlinear pattern in SE, and thus, the presumed inevaluability of the expected effort associated with stimulus rotation, may have been driven by the inclusion of two potential reference points within the effort level manipulation. Interestingly, the pattern of ratings produced in SE do somewhat follow patterns of response times (i.e., a performance index of effort) for reading simple rotated words aloud (Koriat & Norman, 1985). Thus, an alternative account is that individuals are sensitive to the effort (in terms of response times) associated with stimulus rotation in SE, but having several alternatives present in JE modulates the pattern to be more positive. To address this, Experiment 3 manipulated stimulus rotation effort at 15° increments ranging from 0° to 90°, thus excluding the potential 180° reference point. If perceived effort associated with stimulus rotation is inevaluable, then we would expect to find similar results as Experiment 2: individuals in SE should demonstrate greater sensitivity to the qualitative difference between 0° and 15°, but not demonstrate the same sensitivity to all differences thereafter. Judgments in JE should demonstrate a positive linear function across all differences. Alternatively, if individuals are sensitive to the response times associated with reading rotated words, then we would expect similar flat (or nonlinear) functions in SE and JE.

Method

Participants

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

Design

A 2 (Evaluation Mode: JE, SE) × 7 (Stimulus Rotation: 0°, 15°, 30°, 45°, 60°, 75°, 90°) design was employed.

Stimuli

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

Procedure

All procedures followed Experiments 2 and 3.

Results

All data analyses and reporting procedures follow Experiments 1 and 2.

Stimulus rotation

Approximately 6% of cases were removed in JE. LMM results revealed a significant positive slope associated with increased degree of stimulus rotation, b = .36, SE = .05, t = 7.1, 95% CI [.27, .46], β = .5, β 95% CI [.36, .64]. The slope model significantly improved model fit relative to the intercept-only model, $\chi^2(1) = 35.02$, $p < .001$. For SE, approximately 10% of cases were removed. Linear model results for SE ratings revealed a significant positive slope of stimulus rotation, b = .09, SE = .02, t = 4.26, $p < .001$, 95% CI [.05, .14], R^2 = .06, β = .24, β 95% CI [.13, .36]. Thus, both evaluation modes produced significant positive slopes; however, comparing β values across the modes reveals the slope fit in JE is more positive than the slope fit in SE (Figure 7).

Discussion

Results in Experiment 3 demonstrated that individuals in SE produced only a slightly positive linear function across the effort levels; however, the function for JE was much more positive, suggesting that individuals are not sensitive to response times associated with reading rotated words. Thus, similar to Experiment 2, there is a discrepancy between judgments made in SE relative to JE. Examination of the SE function demonstrates the large increase in ratings from the 0° reference point to the first effort level (15°) as predicted by GET. That is, individuals demonstrated sensitivity to the reference point and less sensitivity to incremental differences thereafter. Experiment 3 further lends evidence to the claim that the perceived effort associated with stimulus rotation may be less evaluable.

EXPERIMENT 4

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

DOI: 10.1002/bdm

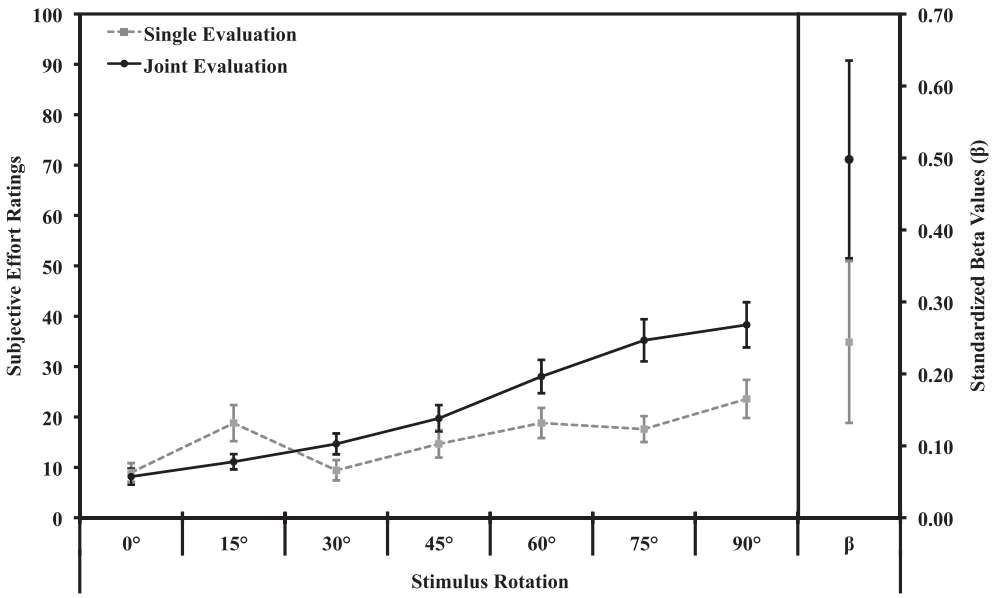

Figure 7. Single and joint evaluation subjective effort rating results (left panel) and standardized beta values for slopes (right panel) in Experiment 3. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars for the left panel represent ±1 standard error of mean (SEM). Error bars for the right panel represent bootstrapped 95% confidence intervals

with more programming experience was favored, demonstrating a preference reversal across evaluation modes confirming. The authors thus argued that GPA was more evaluable to students relative to experience. Hence, examining inconsistencies across ratings and choice represents an additional gauge of value sensitivity.

In Experiment 4, we utilized the stimulus rotation manipulation from Experiments 1–3 crossed with a set size manipulation (i.e., number of words in the display). Experiments 1–3 reveal that set size, in terms of number of items in the display, may be an evaluable attribute. Ratings in the 0° condition (i.e., upright) across the one-word (Experiment 1), two-word (Experiment 3), and nine-word (Experiment 2) set sizes demonstrate a relatively linear increase in ratings as set size increases (Figures 2, 4, and 7). Therefore, similar to the aforementioned example, a relatively evaluable attribute is pit against a relatively inevaluable one. In SE, individuals rated either a one-word display rotated 90° or a two-word display rotated only 15° on the same 0–100 scale used previously. Individuals in JE were presented with both displays and were asked to make a choice about which of the two displays would be more effortful to read aloud.

We would expect individuals to rate the two-word display rotated 15° as more effortful relative to the one-word display rotated 90° in SE. This prediction is derived from SE results from Experiments 1 and 3. Individuals assigned to SE in Experiment 1 rated the one-word display rotated 90° as less effortful to read than individuals assigned to SE in Experiment 3 rated the two-word display rotated 15° (Figures 2 and 7). If the perceived effort associated with stimulus rotation is relatively inevaluable, then stimulus rotation should exert a greater influence in JE choice relative to SE ratings. That is, individuals should choose the one-word display rotated 90° as the more effortful alternative. Such a prediction for JE choice is counterintuitive given the clear difference in objective effort (e.g., as indexed by performance) between processing one-word relative to two words (e.g., Dunn & Risko, 2016).

Method

Participants

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

Design

A 2 (Evaluation Mode: JE, SE) × 2 (Stimulus Rotation: 15°, 90°) × 2 (Set Size: one, two words) design was employed.

Stimuli

Stimuli were similar to the previous experiments.

Procedure

The procedure for SE was similar to the previous experiments. For JE, rather than providing judgments on the 0-to-100 scale, individuals were asked "Which of the two displays above do you feel would be more effortful to read aloud?" Individuals responded by selecting either "Display A." or "Display B." Owing to a program error, the position of the stimuli for JE (i.e., top or bottom position) was not counterbalanced across participants, and thus, an

DOI: 10.1002/bdm

additional 50 individuals were run through the JE condition (see previous texts) to ensure equal presentation at the positions. Results were statistically similar across the runs; therefore, the aggregated data are reported for JE. The attention check for JE asked individuals to "Please click on the Display A. button" during the instruction phase.

Results

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

First, in SE ratings, individuals judged the 15°/two-word display ($M = 10.04$, $SD = 14.03$) to be similarly effortful to read as the 90°/one-word display ($M = 12.94$, $SD = 11.99$), $t(86) = -.87$, 95% CI [−7.99, 3.1], $p = .38$. Bayesian analyses revealed a simulated mode effect size of $d = -.16$, 95% highest density intervals [−.58, .24], and positive evidence for the null, $BF_{\text{NULL}} = 3.24$. For JE choice, individuals selected the 90°/one-word display as more effortful (66%, 95% CI [56%, 75%]) relative to the 15°/two-word display, $p < .01$. Furthermore, a BF computed for the binomial data demonstrated strong evidence for the alternative (i.e., that the proportion of choices is different from chance), $BF_{\text{Alt}} = 31.95$. Thus, individuals in SE rated the two displays as similarly effortful to read, whereas in JE, individuals selected the 90°/one-word display as more effortful than the 15°/two-word display (Figure 8).

Discussion

Experiment 4 further examined the potential inevaluability of stimulus rotation by crossing stimulus rotation and a set size manipulation across a rating and choice context. In SE, individuals rated the 15°/two-word display as similarly effortful to read aloud relative to the 90°/one-word display (although the pattern was very slightly in the opposite direction as predicted). In JE, however, individuals more often choose the 90°/one-word display as the more effortful of the two alternatives. Although not a complete preference reversal, these results demonstrate an inconsistency across ratings and choice as predicted by GET if one inevaluable attribute is included as an alternative alongside an evaluable attribute. That is, stimulus rotation exerted a greater influence in JE relative to SE. This was the case even in light of clear differences in objective effort across the two options. Stimulus rotation costs for single items are relatively small (e.g., Jolicoeur, 1990; Risko et al., 2014) and would not be expected to be larger than the cost of processing an additional item (i.e., reading a one-word display relative to reading a two-word display). Therefore, results from Experiment 4 suggest, consistent with Experiments 1–3, that perceptual effort as indexed by stimulus rotation is weakly evaluable.

EXPERIMENTS 5A AND 5B

Experiments 1 through 4 have demonstrated, through the application of GET, that the perceived effort associated with stimulus rotation is relatively inevaluable in contrast to a

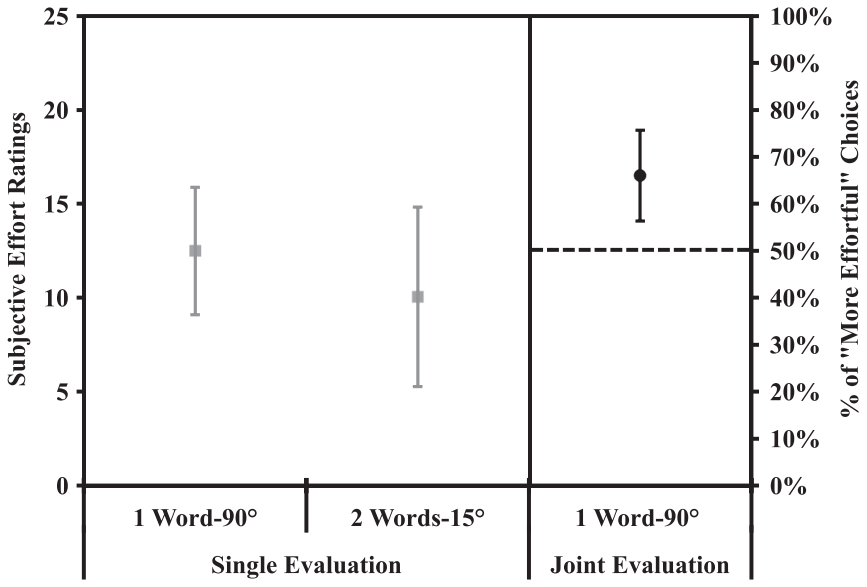

Figure 8. Single and joint evaluation results in Experiment 4. *Note*: The left panel represents between-subject SE ratings, whereas the right panel represents within-subject JE choices. Error bars in both panels represent bootstrapped 95% confidence intervals

DOI: 10.1002/bdm

memorial and motor task. One clear limitation is the constrained task-specific efforts used to this point (i.e., one task from the perceptual, motor, and memorial domains). From this, a clear question arising from Experiments 1–4 is whether the inevaluability evident with stimulus rotation is a product of perceptual effort being inherently inevaluable, or a product of stimulus rotation in and of itself being relatively inevaluable.

To examine this possibility, in Experiment 5a, individuals provided effort ratings on the basis of an additional perceptual task: identifying degraded stimuli ranging from 0% pixel removal to 88% pixel removal in intervals of 22% pixels. Although stimulus degradation and stimulus rotation can both be considered perceptual manipulations, the former arguably has associated with it a much clearer "failure point" (i.e., an upper limit in which an action cannot proceed without failure; see similarly Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). For example, such a failure point arguably exists for both set size and weight but not for the perceived effort associated with stimulus rotation (i.e., humans can process fully rotated stimuli at relatively minimal cost). Furthermore, this failure point *is* present for stimulus degradation as degradation can reach levels to the point where an item is unidentifiable. Therefore, following Experiments 1–3, if perceptual tasks are generally inevaluable, then we would expect individuals in SE to demonstrate greater sensitivity to the qualitative difference between 0% degradation and 22% degradation but not demonstrate the same sensitivity to all differences thereafter. Judgments in JE should demonstrate a relatively positive linear function across all differences. Alternatively, if evaluability is not contingent on domain-specific tasks but rather driven by an inherent failure point, then we would expect similar subjective effort functions for both the SE and JE evaluation modes.

In addition, Experiment 5b looked to extend the general results from hypothetical perceived effort as has been investigated thus far to perceived effort where individuals were aware that they would engage in the task they provide effort ratings for. Here, individuals were instructed that they would provide an effort rating (SE) or ratings (JE) for the levels of stimulus degradation and then asked to attempt to read words that were presented to them on their screen. We offer no a priori hypothesis of why the pattern of ratings would differ if awareness of engaging in the task would affect perceived effort ratings relative to not having to engage in the task. However, if awareness that engaging in a task does not affect perceived effort ratings, then the patterns of subjective effort ratings should closely match those demonstrated in Experiment E5b.

Method

Participants

For Experiments 5a and 5b, 300 MTurk workers participated in the online study for compensation of $US0.50. Fifty individuals were assigned to each of the stimulus degradation dimensions in SE (i.e., 250 total for SE), and 50 were assigned to the JE group. In Experiment 5a, approximately 8% of individuals failed at least one of three attention checks embedded in the survey resulting in a final *N* of 253 (M_{Age} = 35.5 years, 47% female participants, 61% reported completing a bachelor's degree or higher). In Experiment 5b, approximately 14% of individuals failed at least one of three attention checks embedded in the survey, resulting in a final *N* of 258 (M_{Age} = 32 years, 49% female participants, 53% reported completing a bachelor's degree or higher).

Design

A 5 (Effort Level: Stimulus Degradation—0%, 22%, 44%, 66%, 88%) × 2 (Evaluation Mode: JE, SE) design was employed for both Experiments 5a and 5b.

Stimuli

Individuals judged displays consisting of the stimulus "WORD" degraded from 0% to 88% of pixel removal in 22% increments using a diagonal grating. Individuals were asked "How effortful would it be to read this word aloud?" For Experiment 5b, five words of high-frequency nouns ($M_{WrittenFreq}$ = 326.80) were generated ("FELT", "WANT," "MIND," "DOOR," and "HELP") and counterbalanced across five lists such that each word was presented equally across the five levels of stimulus degradation.

Procedure

The procedure for JE and SE ratings in Experiment 5a followed Experiments 1–3. The procedure for Experiment 5b closely followed the aforementioned experiments. However, individuals were instructed prior to completing their effort ratings that, upon completion of the ratings, they would be asked to attempt to read a word (for SE) or words (for JE) and enter the word presented into a text box. In SE, individuals only received one word to attempt to read corresponding to the degradation level they rated. In JE, individuals received five words to attempt to read corresponding to all levels of the stimulus degradation manipulation.

Results

All data analyses and reporting procedures follow Experiments 1–3. Subjective effort rating results are first presented for Experiment 5a followed by Experiment 5b (Figures 9 and 10). Generalized additive mixed models (GAMM; see later texts) were constructed using the *mgcv* package (Wood, 2006).

Experiment 5a

Effort ratings

Joint evaluation mode. No cases were removed for JE ratings. LMM results revealed a significant positive slope associated with increased degree of stimulus degradation, b = 16.11, SE = 1.82, t = 8.87, 95% CI [12.56, 19.77],

DOI: 10.1002/bdm

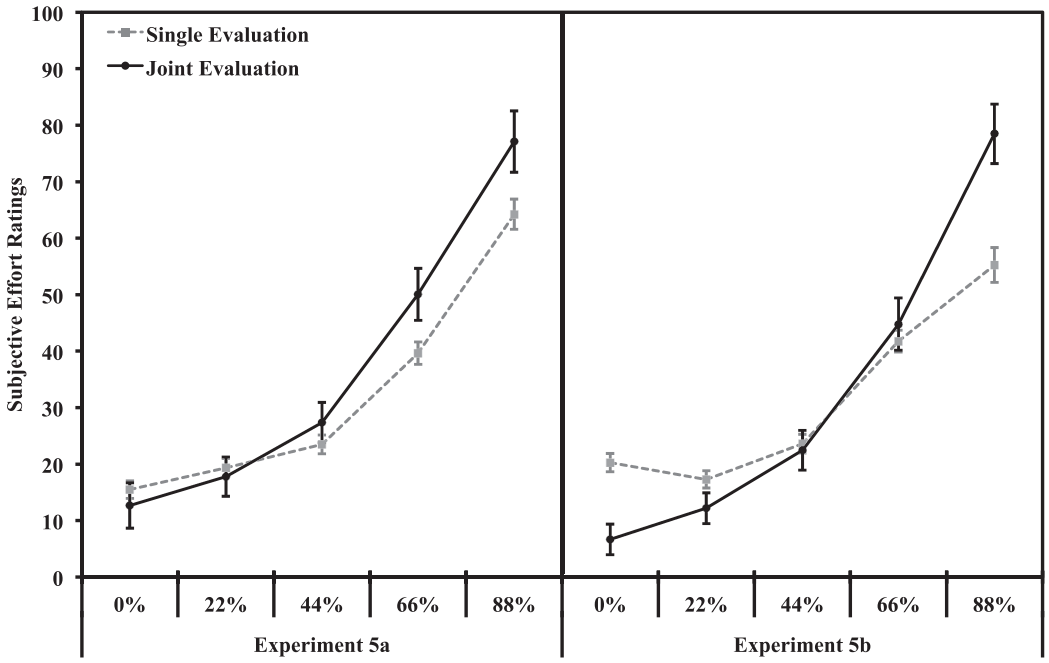

Figure 9. Single and joint evaluation subjective effort rating results for stimulus degradation in Experiments 5a (left panel) and 5b (right panel). *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars for the left panel represent ±1 standard error of mean (SEM)

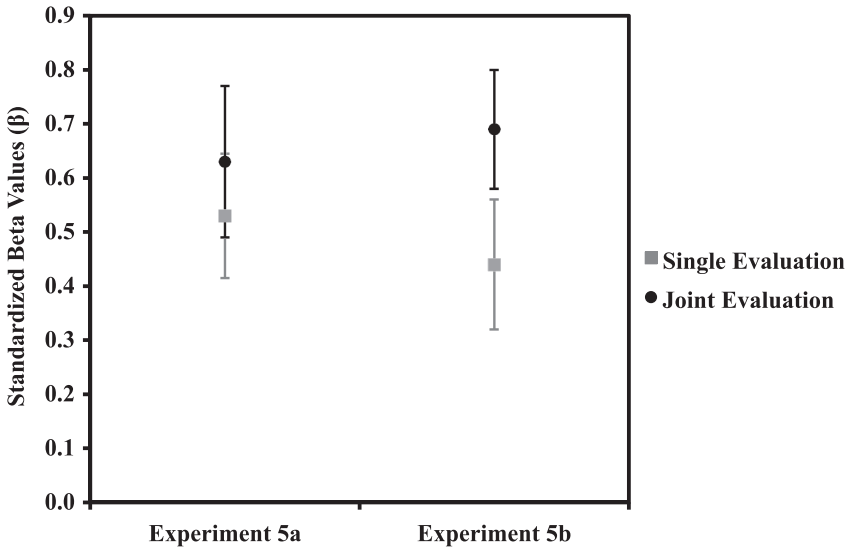

Figure 10. Standardized beta values for slopes in Experiment 5a and 5b. *Note*: Single evaluation (SE) represents between-subject ratings, whereas joint evaluation represents within-subject ratings. Error bars represent bootstrapped 95% confidence intervals

β = .63, β 95% CI [.49, .77]. The slope model significantly improved model fit relative to the intercept-only model, $\chi^2(1) = 44.98$, $p < .001$. Furthermore, visual inspection of loess fit to the data suggested a nonlinear model would best fit the JE data (i.e., an exponential fit). A GAMM demonstrated a significant smooth term, *edf* = 2.55, $F = 671.60$, approximate $p < .001$. A deviance test demonstrated that the GAMM produced a better fit to the JE data relative to the LMM, $p < .001$, as well as a smaller *AIC* value relative to LMM, *AIC* = 1795.15 and *AIC* = 1853.32, for the GAMM and LMM, respectively.

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

DOI: 10.1002/bdm

Table 1. Accuracy for word reading as a function of stimulus degradation in Experiment 5b

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

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

relative to LM, *AIC* = 1981.10 and *AIC* = 1987.76, for the GAM and LM, respectively.

Experiment 5b

Effort ratings

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

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

Reading accuracy. Given the lack of variability in accuracy across several of the stimulus degradation conditions, performance is reported here only qualitatively (Table 1). First for JE, accuracy was at ceiling (i.e., 100%) for the 0%, 22%, and 44% degradation conditions. Accuracy fell to 83% for the 66% degradation condition and to floor (0%) for the 88% degradation condition. For SE, accuracy was similarly at ceiling for the 0%, 22%, and 44% degradation conditions, and then falling to 82% accuracy for the 66% degradation condition, and finally to floor (0%) for the 88% degradation condition.

Discussion

Experiment 5a provides evidence that evaluability is not domain specific but rather may be task specific and contingent on a failure point. In contrast to the stimulus rotation manipulation used in Experiments 1–4, subjective effort ratings for the stimulus degradation manipulation used in Experiment 5a and 5b can be considered relatively evaluable. Ratings for both the SE and JE modes produced similar positive linear slopes in Experiment 5a. In addition, both functions were best fit with a nonlinear exponential function relative to the linear functions. Results for Experiment 5b, where individuals had awareness that they were required to complete the task they would rate, demonstrated a less positive slope for SE than the slope for JE. This suggests that having such awareness may affect ratings in SE but not JE, given the JE slope for E5b was extremely similar to E5a. To confirm that the flatness of the SE slope relative to the JE slope was not a result of noise, we collected additional samples for SE ratings mirroring the procedures for 5a and 5b. Results demonstrated similar slopes across the new samples with both closely matching the SE slope from E5a.[3] Thus, the difference across JE and SE slopes in E5b may indeed have been due to random variability in ratings.

Critically, these findings contrast with Experiments 2 and 3 for the divergent patterns of effort ratings of stimulus rotation where SE slopes were flat and JE slopes were positively sloped. Tasks possessing a failure point associated with some value appear to drive consistent ratings across modes given the correspondence in effort functions across modes for degradation, set size, and weight. This similarity is absent for stimulus rotation. We return to these findings in the General Discussion. Furthermore, given the similarity of functions across, it does not appear that possessing awareness of having to engage in the task alters patterns of effort ratings.

[3]For each SE rating condition, 250 participants were recruited through MTurk. In the E5a replication sample, approximately 11% of individuals failed at least one attention check resulting in a final N of 222 (M_{Age} = 32 years, 42% female participants, 53% reported completing a bachelor's degree or higher). In the E5b replication sample, approximately 12% of individuals failed at least one attention check resulting in a final N of 219 (M_{Age} = 34 years, 52% female participants, 53% reported completing a bachelor's degree or higher). Both samples produced extremely similar positive slopes, β = .51, β 95% CI [.39, .63], β = .54, β 95% CI [.42, .65], for the E5a replication and E5b replication samples, respectively, and are thus very comparable with the slope for SE in the original E5a experiment, β = .53, β 95% CI [.41, .64].

DOI: 10.1002/bdm

GENERAL DISCUSSION

General evaluability theory has been applied to various domains within economic, business, and management contexts (e.g., Bazerman, Loewenstein, & White, 1992; Hsee, 1993, 1996; Nowlis & Simonson, 1997) in an attempt to gauge the putative sensitivity of a value of interest. The current set of experiments applied the GET logic to an additional determinant of decision making: effort (see Table 2 for a review).

In Experiment 1, individuals provided judgments of expected effort in either the SE or JE mode across three task-specific efforts: stimulus rotation, set size, and weight. Results demonstrated that judgments of perceived effort for set size produced similar positive slopes across evaluation modes. The perceived effort associated with stimulus rotation and weight demonstrated similar patterns across modes, however, not to the same degree as set size, suggesting that each type of task-specific effort may be evaluable. Experiment 2 addressed a potential concern of a floor effect in Experiment 1 by increasing the putative levels of effort to be judged for the stimulus rotation and weight conditions. Results replicated the patterns found across modes for set size and weight, with both the SE and JE functions being very similar. Stimulus rotation, however, produced a positive linear function in JE and a flatter function in SE, suggesting that perceptual effort in terms of stimulus rotation may not be highly evaluable. Experiment 3 looked to further examine this notion by focusing solely on judgments of stimulus rotations at finer rotation increments relative to Experiment 2. Again, JE produced a positive linear function across increments, whereas the linear function in SE was less positive. Taken together, Experiments 2 and 3 suggested that perceptual effort in terms of stimulus rotation may not be highly evaluable.

In Experiment 4, we further examined this possibility by crossing the relatively inevaluable stimulus rotation with a manipulation of set size (i.e., number of items in the display) across both ratings and choice (i.e., "Which is more effortful?"). Within this context, GET hypothesizes that a low-evaluability value will exert greater influence in JE (choice) relative to SE (ratings). Individuals more often selected the condition associated with greater levels of stimulus rotation as the more effortful in JE, while both stimulus rotation and set size were rated as similarly effortful in SE. Experiments 5a and 5b looked to test the hypothesis that evaluability is not domain specific but rather specific to tasks that are associated with a failure point, and whether awareness of completing the task that is to be rated affects subjective effort functions. Specifically, using stimulus degradation, both SE and JE produced similar positive slopes across effort levels, as well as exponential fits to the ratings data. In the following, we discuss potential determinants of effort evaluability, highlight potential shortcomings of the current set of studies, and suggest avenues for future research.

The criticality of reference information

At a normative level, all decision making can be thought of as occurring between alternatives even when alternatives are not explicit (Hsee et al., 1999). Thus, what type of

Table 2. Summary of the current experiments

			Slope terms		
	Task-specific effort(s)	JE–SE relation	Mode	β	95% CI
Experiment 1	Stimulus rotation	Similar positive slopes	JE	.48	[.27, .69]
			SE	.37	[.20, .55]
	Set size	Similar positive slopes	JE	.86	[.75, .98]
			SE	.87	[.76, .97]
	Weight	Similar positive slopes	JE	.68	[.53, .82]
			SE	.48	[.31, .65]
Experiment 2	Stimulus rotation	Positive slope in JE, flat slope in SE	JE	.43	[.31, .56]
			SE	.10[a]	[−.05, .25]
	Set size	Similar positive slopes	JE	.82	[.72, .92]
			SE	.87	[.80, .94]
	Weight	Similar positive slopes	JE	.68	[.58, .77]
			SE	.62	[.50, .74]
Experiment 3	Stimulus rotation	More positive slope in JE relative to SE	JE	.5	[.36, .64]
			SE	.24	[.13, .36]
Experiment 4	Stimulus rotation and set size	Inconsistency across choices and ratings	—	—	—
			—	—	—
Experiment 5a	Stimulus degradation	Similar positive slopes	JE	.63[a]	[.49, .77]
			SE	.53[a]	[.41, .64]
Experiment 5b	Stimulus degradation	More positive slope in JE relative to SE	JE	.69[a]	[.59, .81]
			SE	.44[a]	[.32, .56]
Experiments 5a and 5b SE mode replications	Stimulus degradation	—	5a	.51	[.39, .63]
			5b	.54	[.42, .65]

Note: The replications of the SE mode slopes for Experiments 5a and 5b are reported in footnote 3.
JE, joint evaluation; SE, single evaluation; CI, confidence interval.
[a]Nonlinear model (generalized additive model or generalized additive mixed models) produced a better fit to the data relative to the linear model.

DOI: 10.1002/bdm

reference information is available to individuals in SE that would lead to consistent judgments, when no explicit reference information is afforded? Within GET, both knowledge (i.e., distributional information, such as the variability and average, gained through experience) and nature (i.e., whether individuals possess a stable physiological or psychological reference system) reflect internal reference information available when generating judgments in the absence of explicit reference points. In contrast, mode as utilized in the present investigation can be considered ad hoc evaluability (Hsee & Zhang, 2010). Importantly, knowledge and nature both acting as internal forms of references may provide the necessary information needed to produce high levels of value sensitivity in the absence of explicit reference information with all factors conjunctively working towards evaluability.

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

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

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

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

Subjective versus objective effort functions

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

In Experiments 1–3, there was a close relation between JE and SE judgments of perceived effort associated with set size suggesting, according to GET, that this attribute is highly evaluable. However, if we measure effort through accuracy, we know that this is not the pattern we would expect. For example, Risko and Dunn (2015) demonstrated, using similar stimuli, that accuracy was near ceiling for smaller set sizes (two and four letters) but fell radically at the medium set size (six letters) and was near floor for the larger set sizes (eight and 10 letters). As would be expected from the experiments reported here, when individuals provided subjective ratings of accuracy and effort, results demonstrated relatively linear functions for both dimensions, where perceived effort increased as the number of items increased and perceived accuracy decreased as the number of items increased. Therefore, the objective effort function (as indexed by accuracy) mimicked more of a

DOI: 10.1002/bdm

decaying logistic function rather than a linear function as produced from individuals' subjective judgments.

A type of paradox then exists between subjective and objective effort functions: on the basis of the GET logic, a specific effort may be evaluable at the subjective level but not coincide with the associated objective function as indexed by some putative measure of objective effort (see also Hsee & Zhang, 2004, for a similar issue with regard to choice-experience inconsistencies). Resolving this apparent paradox requires taking seriously the notion that effort is a subjective phenomenon not necessarily tightly tied to objective processing demands (Dunn et al., 2016; Kool & Botvinick, 2013; Westbrook & Braver, 2015). Indeed, several demonstrations exist of individuals' decisions and subjective evaluations based on effort being dissociated from objective demands as indexed by various indirect measures (e.g., Dunn & Risko, submitted; Dunn et al., 2016; Westbrook et al., 2013). From this perspective, evaluability would be defined by consistency in subjective evaluations (e.g., across modes) without any expectation that those evaluations accurately map onto processing demand per se (or indirect measures of it). Separating the concept of evaluability from "accuracy" in this manner does not diminish the importance of understanding the former. This is because it is individuals' subjective experience of effort that is proposed to drive behavior.

Evaluating "apples and oranges"

The current findings provide important insight with the regard to the relation between cognitive and physical judgments of effort. Kable and Glimcher (2009) note that subjective values allow for evaluation across options in such a way that decisions between "apples and oranges" are possible. Intriguingly, at specific effort levels in Experiments 1 and 2, individuals judged perceptual and memorial efforts as more effortful relative to physical effort and vice versa. For example, in Experiment 2, individuals' effort ratings demonstrated that attempting to hold 10 letters in memory and recalling them accurately was perceived to be more effortful than lifting 50 lb of weight from the ground, whereas lifting the same amount of weight was perceived to be more effortful than attempting to hold six letters in memory. These situations represent interesting instances where, if placed in a decision-making context, individuals would be faced with trading off one type of effort for another.

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

Methodological implications

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

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

Undoubtedly, further examples exist where some effect is present in within-subjects designs but researchers fail to observe the same effect in between-subjects designs (or vice versa), and further consideration of the issue of what evaluation mode to utilize would be of methodological

DOI: 10.1002/bdm

importance across a wide range of contexts (e.g., Birnbaum, 1999). For example, within the context of heuristics and biases, Kahneman and Tversky (1996) argue that the two designs answer explicitly different questions. A between-subjects design tests whether an individual relies on a given heuristic, whereas a within-subjects design addresses how the conflict between heuristic use and some formal rule is resolved. That is, between-subjects designs arguably offer a more realistic view of individuals' reasoning (Kahneman & Frederick, 2002; Tversky & Kahneman, 1983).

CONCLUSION

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*(1), 83–113.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

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

Carter, E. C., & McCullough, M. E. (2013). Is ego depletion too incredible? Evidence for the overestimation of the depletion effect. *Behavioral and Brain Sciences*, *36*(06), 683–684.

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

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

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

Dunn, T. L., Lutes, D. J. C., & Risko, E. F. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1372–1387.

Dunn, T. L., & Risko, E. F. (2016). Toward a metacognitive account of cognitive offloading. *Cognitive Science*, *40*(5), 1080–1127.

Dunn, T. L., & Risko, E. F. (submitted). Understanding the cognitive miser: Cue utilization in effort avoidance. Retrieved from https://www.researchgate.net/publication/303543690_Understanding_the_Cognive_Miser_Cue-utilization_in_Effort_Avoidance

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

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

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

Gilbert, S. J. (2015). Strategic offloading of delayed intentions into the external environment. *The Quarterly Journal of Experimental Psychology*, *68*(5), 971–992.

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

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

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482.

Hockey, G. R. J. (2011). A motivational control theory of cognitive fatigue. In P. L. Ackerman (Ed.), *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (, pp. 167–188). Washington, DC: American Psychological Association.

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

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

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

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

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

DOI: 10.1002/bdm

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

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

Inzlicht, M., & Schmeichel, B. J. (2013). Beyond simple utility in predicting self-control fatigue: A proximate alternative to the opportunity cost model. *Behavioral and Brain Sciences*, *36*(6), 695.

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

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

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

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

Kaas, R. E., & Raferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: Consensus and controversy. *Neuron*, *63*(6), 733–745.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.

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

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

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

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

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

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

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 131–141.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665–682.

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

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

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

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

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

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

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

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

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

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

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

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

McGuire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences*, *107*(17), 7922–7926.

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

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.11-1. Retrieved from http://CRAN.Rproject.org/package=BayesFactor

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

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

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

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

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

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

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

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

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

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

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

Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*(1), 71–92.

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

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

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

DOI: 10.1002/bdm

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

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

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

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

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

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

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

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

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

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

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

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS One*, *8*(7), e68210.

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

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

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

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

Authors' biographies:

Timothy Dunn is a PhD candidate in the University of Waterloo, Canada, Cognitive Psychology program. His research has investigated the metacognition of effort, effort-based decision making, and distributed cognition. He enjoys watching sports and laughing at his dog.

Derek Koehler is Professor of Psychology at the University of Waterloo, Canada. His research has investigated how people draw inferences, make plans, generate predictions, and pursue goals under conditions of uncertainty.

Evan F. Risko is an Assistant Professor and Canada Research Chair in the Cognitive area of the Department of Psychology at the University of Waterloo, Canada. Current research includes work on cognitive offloading, effort-based decision making, everyday attention, and multimedia learning.

Authors' addresses:

Timothy L. Dunn, University of Waterloo, Waterloo, ON Canada.

Derek J. Koehler, University of Waterloo, Waterloo, ON Canada.

Evan F. Risko, University of Waterloo, Waterloo, ON Canada.

DOI: 10.1002/bdm