The role of peripheral cues in judgments of difficulty

by

Lisa Vangsness

B.A., University of Iowa, 2012
B.S., University of Iowa, 2012
M.S., Kansas State University, 2017

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Judgments of difficulty (JODs) can be used to inform effort allocation strategies and subsequent performance. This dissertation integrates several models to advance specific hypotheses regarding the role that feedback, a performance-based peripheral cue, plays in these processes. These predictions are tested in two experiments that manipulate when and how feedback information is made available. In Experiment 1, people alternated between observing and performing a visual search task; performance-based peripheral cues informed JODs to a lesser degree until people had a chance to perform the task themselves, suggesting that receiving feedback on one's performance informs self-efficacy beliefs. In Experiment 2, people learned about the incentive structure of the environment at different times. Incentives changes peoples' effort allocation strategies, but the way in which it did so depended on when this information was made available. People who learned of the incentives in advance used this information to engage in preventative effort allocation strategies, while those who learned of the incentives through feedback alone engaged in compensatory effort allocation strategies. Together, these results disambiguate when and how people use performance-based peripheral cues to make JODs, and provide information about how the environment can be structured to facilitate learning and behavioral change.

The role of peripheral cues in judgments of difficulty

by

Lisa Vangsness

B.A., University of Iowa, 2012
B.S., University of Iowa, 2012
M.S., Kansas State University, 2017

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Michael E. Young

Copyright

Abstract

Judgments of difficulty (JODs) can be used to inform effort allocation strategies and subsequent performance. This dissertation integrates several models to advance specific hypotheses regarding the role that feedback, a performance-based peripheral cue, plays in these processes. These predictions are tested in two experiments that manipulate when and how feedback information is made available. In Experiment 1, people alternated between observing and performing a visual search task; performance-based peripheral cues informed JODs to a lesser degree until people had a chance to perform the task themselves, suggesting that receiving feedback on one's performance informs self-efficacy beliefs. In Experiment 2, people learned about the incentive structure of the environment at different times. Incentives changes peoples' effort allocation strategies, but the way in which it did so depended on when this information was made available. People who learned of the incentives in advance used this information to engage in preventative effort allocation strategies, while those who learned of the incentives through feedback alone engaged in compensatory effort allocation strategies. Together, these results disambiguate when and how people use performance-based peripheral cues to make JODs, and provide information about how the environment can be structured to facilitate learning and behavioral change.

Table of Contents

List of Figures

List of Tables

Acknowledgements

It is difficult to put into words the intellectual and emotional supports that are required to write a dissertation. I hope that the reader will forgive my omissions, both of the people and moments that were pivotal to the creation of this work.

Phil, thank you for your unending patience with my work schedule and preoccupation with statistics. I imagine it is particularly difficult to support someone as they complete a project that produces few tangible outcomes and is unlikely to earn any accolades, yet you have always encouraged me in your own way. Many of the ideas that appear in this document occurred as we rode our bikes, working together to fight the unbearable winds and heat of rural Kansas. Just as we look out for one another on our bike rides, you are always perceptive of how my research affects me: you know when to encourage me to work hard and when to force a break from writing. What little sense-making remains at the end of this endeavor I attribute to you.

Mike, I can still remember the day I sat across from you at Happy Valley and told you that I couldn't remember anything about statistics. I didn't know a t-test from a Tukey's, but I did know that I wanted to *learn*, knew that statistics was the secret sauce that could help me to discover the unknown and go where no researcher had gone before… well, something like that, anyway. When you invited me to pursue my PhD, you opened the door to the most interesting and worthwhile endeavor I've had the privilege to undertake. Thank you for teaching me the inside baseball, and for sending me on a journey that would forever change the course of my life.

Mom, thanks for letting me keep nine mice in the upstairs bedroom. I don't know how Dad convinced you, but it seems to have been the earliest beginnings of a (hopefully) long research career. I know I was a ridiculous kid, always coming home covered in pine sap or exploring the roof or taking apart my sister’s alarm clock. Hopefully, you find that your patience has paid off after all these years.

Dad, thank you for nurturing my love for science and appreciation for the unknown. The memories that strike me while researching are often those of time spent with you. I remember keeping you company as you coded to Russ Freeman & The Rippingtons and the time we used a coffee warmer to surface mount the resistors to my Altoids-tin ham radio. There are a lot of initiatives to support young women in science; I could not have found better sponsorship than the support and insistence of my father. While I know that you will always be proud of me, no matter the outcome, I still dedicate this dissertation to you.

Chapter 1 - Introduction

One hundred years ago, the earliest psychology researchers noted that peoples' perceptions of difficulty seemed to be unique (James, 1890; Titchener, 1908); people emphasized different aspects of a task when judging its difficulty. For this reason, peoples' efforts did not always align with a task's demands. Some people underestimated difficulty and performed poorly, while others overestimated difficulty and allocated more time and effort than was necessary to perform well. Recent research has confirmed this general relationship across different domains including procrastination (Hensley, 2014; Mitchell, 2017), safety compliance (Sigurdsson, Taylor, & Wirth, 2013), tool use (Liu & Wickens, 1994; Vallières, Hodgetts, Vachon, & Tremblay, 2016), and task performance (Cohen, Purdie-Vaughns, & Garcia, 2016; Ortner, Weißkopf, & Gerstenberg, 2013). However, this recent work focuses narrowly on the relationship between peoples' judgments of task difficulty (JODs) and their subsequent performance, rather than on the intervening variables and processes that might bias judgment and lead to pronounced changes in behavior (Cain, 2007; Gopher & Donchin, 1986; Hart & Staveland, 1988). Consequently, the degree to which intervening variables and processes influence JODs and task performance remains unclear.

This dissertation identifies these relationships by measuring the sources of information that are used to make JODs during task completion. This approach has been used successfully in the past (Vangsness & Young, 2018); when combined with mixed-effects modeling it can capture the cognitive and perceptual limitations that lead to poorly-calibrated judgements (Dixon, 2008; Higham, Zawadzka, & Hanczakowski, 2016; Maniscalco & Lau, 2014). To ensure that the findings of this dissertation are broadly generalizable, the experiments will employ a simple visual search task. Performance in visual search is reflective of basic cognitive and perceptual processes (Chan & Hayward, 2013; Wolfe, 1994, 2007) and is frequently used to study difficult tasks such as cancer detection (Evered, 2005) and security search (Wolfe, Horowitz, & Kenner, 2005). However, the theoretical concepts and relationships described in this dissertation use a variety of examples to illustrate the many situations in which JODs are relevant.

To this end, I will first review current approaches to the study of task difficulty and use a non-linear model to illustrate how traditional approaches confound task difficulty with an individual's skills and abilities, which are affected by intervening variables such as metacognitive skill and perceptual sensitivity. Next, I will discuss the challenges in operationalizing task difficulty and draw clear delineations between associated constructs such as difficulty, effort, and workload. Finally, I will advance a cohesive framework for testing the relationships between these constructs and outline three experiments that selectively tested different relationships within this framework.

Current Approaches to the Study of Task Difficulty

Assessing Task Difficulty, Effort, and Performance

Traditionally, task difficulty is assessed through performance (Boksem, Meijman, & Lorist, 2005; Cain, 2007; Libedinsky et al., 2013; Schouppe, Demanet, Boehler, Ridderinkhof, & Notebaert, 2014) or by asking participants to provide a subjective workload assessment (Borg, 1998; Cain, 2007; Hart & Staveland, 1988; Reid & Nygren, 1988; Tsang & Velazquez, 1996). In these experiments, researchers compare participants' performance or retrospective ratings of perceived effort across versions of a task that produce changes in performance. This approach requires researchers to make four implicit assumptions. First, researchers must assume that changes in performance are only driven by the difficulty of the task. In doing so, the researchers also assume that task difficulty cannot be objectively defined. Finally, researchers assume that task difficulty shares a consistent, linear relationship with performance and ratings of perceived effort. Yet, these assumptions contradict well-established theories regarding the relationship between task difficulty and performance.

At a basic level, task performance is determined by effort allocation (Gopher & Donchin, 1986; Wickens, Hollands, Banbury, & Parasuraman, 2016). A person who puts forth more effort will achieve a higher level of performance than will a person who does not. The decision to allocate effort involves a person's task-related skills, their goals, and the difficulty level of the task (see Figure 1). This can be illustrated by a simple example: a student studying for an upcoming exam. First, the student must determine how well they know the material that will appear on the exam, a judgment of skills and abilities that is known as self-efficacy (Bandura, 1977). Next, they must make a JOD about the exam. Finally, the student must estimate how much effort to put forth in order to achieve their goal successfully (R. Ackerman, 2014; Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Kornell, 2005).

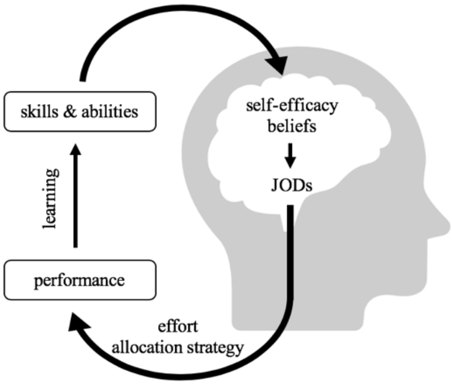


Figure 1. Two metacognitive judgments (self-efficacy, difficulty) inform effort allocation strategies and subsequent task performance.

Ideally, effort allocation decisions involve comparisons between accurate judgments of skills, abilities, and task difficulty. When they do, the relationship between effort allocation and performance can be modeled using a logistic function:

$$f(x) = \frac{L}{1+\ e^{k(x-x_0)}} \tag{1}$$

Here, *L* represents a person's ceiling performance, restricted by the task's objective difficulty. The *k* parameter represents the efficiency (i.e., slope) of their effort allocation strategy. Finally, x_0 represents the average level of performance that can be achieved, given a person's skills and abilities. Figure 2 illustrates the performance curves of three students. A well-practiced student (dot-dashed purple) does not need to invest much effort to score 100% on an assignment. This is due to both skill and resource efficiency. Because this student is highly skilled, their performance curve is shifted relative to other students' – they can achieve strong performance with less effort. In addition, their experience allows small investments of effort to produce greater performance gains, as evidenced by a steeper slope. This pattern can be contrasted with that of a less-practiced student (solid purple) who must invest more effort to achieve a similar score and whose effort returns gains at a lesser rate. Finally, consider a situation in which the less-skilled student is faced with a different assignment, one that is objectively more challenging and, perhaps, impossible to successfully complete (dashed dark blue). In this case, task difficulty lowers ceiling performance and the curve asymptotes at a lower value. This change affects the relationship between the students' skills and the task's difficulty (i.e., median performance), but leaves resource efficiency (i.e., the slope) intact.

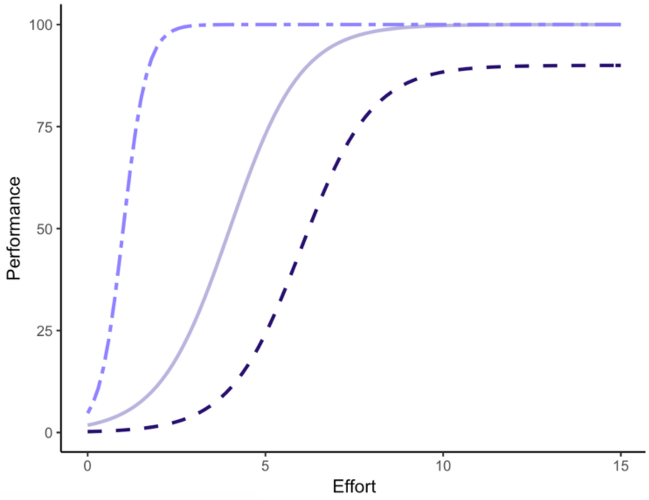


Figure 2. The relationship between effort and performance can be modeled by a logistic regression. The figure depicts the performance curves of three individuals described in the body text.

The effort/reward trade-off. While ceiling performance is possible, it is not necessarily desirable because it involves a trade-off between effort and reward (Boksem & Tops, 2008; Kool

& Botvinick, 2014; Kurzban, 2016; Kurzban, Duckworth, Kable, & Myers, 2013). That is, effort requires resources that are limited in some way (e.g., attention, time; (Bruya & Tang, 2018; Kahneman, 1973); limited resources cannot be reallocated without affecting performance on the primary task (Wickens, 2002). For example, driving and checking a text message both require visual attention. Young drivers who allow themselves to become temporarily distracted by a text message take their attention off the road and exhibit poorer performance (Caird, Johnston, Willness, Asbridge, & Steel, 2014). This performance decrement does not represent a change in the difficulty of the driving task; rather, it is indicative of a shift in cognitive resources that was driven by the relative value or utility of concurrent tasks.

The degree to which the effort/reward trade-off affects performance is a function of both skill and ability (Kiesel et al., 2010; Monsell, 2003; Pashler, 2000). Figure 3 depicts the hypothetical performance curves of three individuals before and after they chose to invest their cognitive and perceptual resources in checking a text message. The driving performance of the less-skilled driver (Figure 3A) is worse than that of a driver of average ability (Figure 3B) both when the cell phone is present and when it is not. These negative consequences can be mitigated by improving either drivers' overall skill (i.e., median performance) or by increasing their resource efficiency (i.e., slope) through additional practice. For example, task-switching is less detrimental for a person with greater resource efficiency (Figure 3C) because their increased processing capacity (i.e., the steepness of the performance curve) mitigates the negative effects of task-switching, albeit to a lesser degree than would an improvement of skill. Even if these drivers were involved in a situation that presented identical levels of difficulty, their performance will still impaired to different degrees. In other words, their performance is partially due to the difficulty level of the task, but is also affected by their resource allocation, skills, and the incentive structure of the environment.

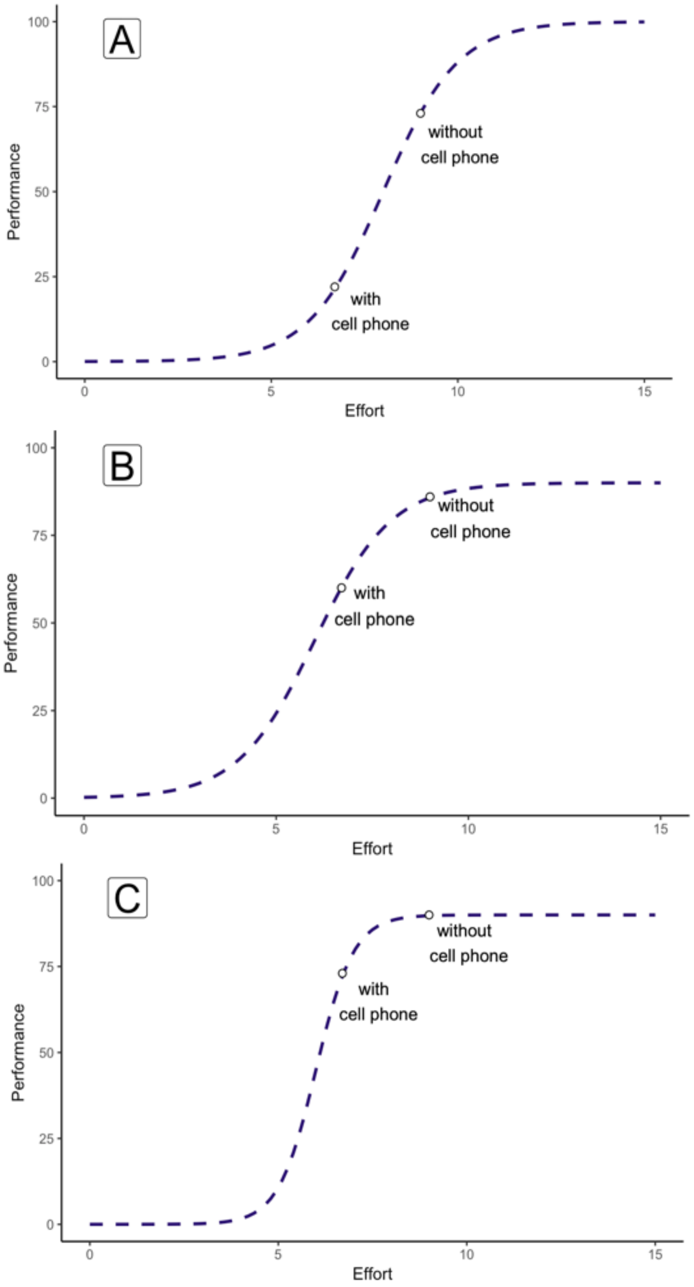


Figure 3. Hypothetical performance curves for drivers operating a vehicle with and without the distraction of a cell phone. The degree to which task-switching affects performance is affected by an individual's skills and resource efficiency: highly-skilled (panel B) and highly-efficient (panel C) individuals experience fewer performance decrements than less-skilled, less-efficient counterparts (panel A).

The moderating effects of perceived effort and reward are obvious in task-switching situations (Wickens, Gutzwiller, et al., 2016; Wickens, Gutzwiller, & Santamaria, 2015); however, the effort/reward trade-off can also affect peoples' engagement in a single task (Ariel et al., 2009; Floresco, Tse, & Ghods-Sharifi, 2008; Hull, 1943; Minamimoto, Hori, & Richmond, 2012; Mitchell, 2017; Nishiyama, 2014; Prévost, Pessiglione, Météreau, Cléry-Melin, & Dreher, 2010; Schouppe et al., 2014; Walton, Kennerley, Bannerman, Phillips, & Rushworth, 2006). Consider a student who must participant in seven research experiments to pass General

Psychology. This student can choose to schedule research appointments at any time during the semester; however, their decision about when to participate is influenced by incentive magnitude and ease of access. Studies that can be completed from home (i.e., online studies) are, on average, scheduled 15 days earlier in the semester than are those that require an in-person appointment, which is more effortful to attend. Students also tend to complete alternative experiments, summary research papers, around 50 days later in the semester than online studies (see Figure 4). Students also tend to prioritize studies that are worth more points (see Figure 5), although this effect is smaller due to restriction of range (most research studies are between 0.5 and 1 credits; Vangsness, unpublished data).

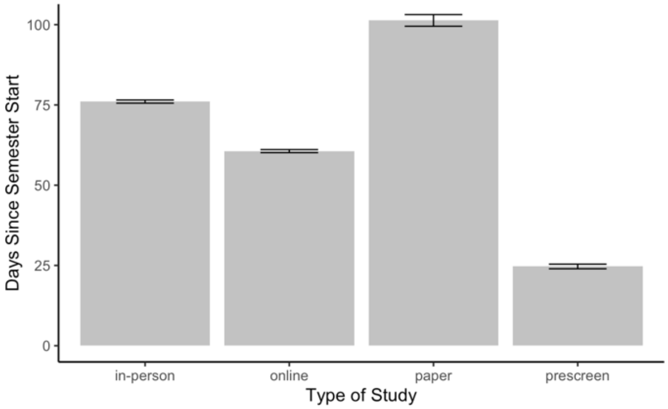


*Figure 4. T*he nature of a research study affects when students complete their research credits. Students complete online studies sooner than they do in-person studies or alternative assignments (research papers). The prescreen can only be completed during the first month of the semester.

Research findings cast in the traditional framework attribute these changes in performance and task engagement to difficulty alone. These preliminary findings illustrate that while difficulty is a strong driver of behavior, performance is also affected by the incentive structure of the environment. Performance decrements due to task difficulty and those due to

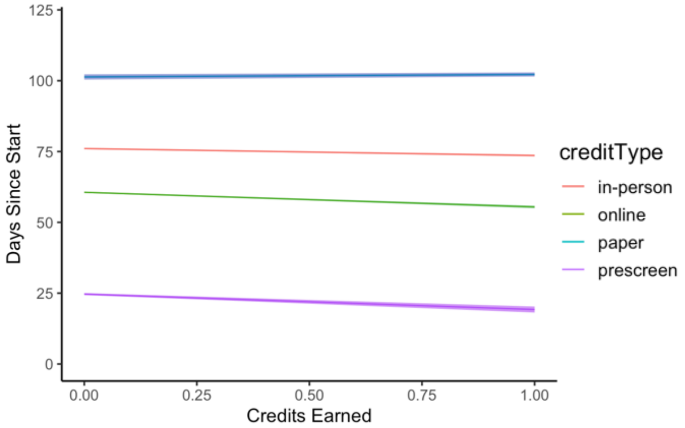


Figure 5. The size of a research credit reward affects when students participate in a research study. Students tend to prioritize studies that provide larger credit rewards, although this effect is attenuated by a restriction of range problem.

incentive structure demand different interventions. To return to the driving example, reducing the difficulty of a driving task will not reduce the likelihood that someone will check their cell phone. Likewise, removing unwanted distractions will not improve the performance of an unskilled driver. Differentiating sources of variance such as these will indicate how targeted interventions (e.g., skill improvement, reducing distractions) affect performance and enhance the broader understanding of task difficulty.

Misestimations of Skill, Efficiency, and Task Difficulty. Although the relationships described in Figures 2 and 3 appear straightforward, this is likely the exception rather than the rule. These figures assume an ideal situation in which accurate judgments of skill, efficiency, and task difficulty inform effort allocation decisions. More likely, these judgments and the decisions that they elicit are prone to biases and misestimations (Gigerenzer, Todd, & The ABC Research Group, 1999; Simon, 1955).

Empirical examples of biases abound, especially in the field of education. Novice students are biased towards overestimating their skills, while more expert students tend to underestimate them (Dunning, 2011; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). Students also rely upon poor sources of information, such as fluency (Karpicke, Butler, & Roediger, 2009) and familiarity (Oppenheimer, 2008) when making effort allocation decisions. Although reading over a textbook many times increases fluency and familiarity, making content seem easier-to-learn, research has shown that the opposite is true: these invalid cues lead students to overestimate their resource efficiency and underestimate the difficulty of study materials (Karpicke et al., 2009). Errors such as these can be modeled using a logistic regression.

Consider the hypothetical (solid black line) and estimated (dashed teal line) performance of a B-average student who is capable of obtaining an 85% on an exam. Poor metacognitive ability might lead this student to overestimate their skills (Figure 6A) or their resource efficiency (Figure 6B), leading to worse performance than expected. Similar outcomes would arise if the student underestimates the difficulty of the exam (Figure 6C). In each case, the student believes their investment of effort will produce a higher score than they obtained. These errors arise from different mechanisms and must be addressed in different ways: the first two examples require improved metacognitive skills, while the latter requires an improved ability to judge exam difficulty.

Underestimation is also problematic. When a student underestimates their skills (Figure 7A) or resource efficiency (Figure 7B), they invest more effort than is required to achieve the desired level of performance. While these errors are less likely to lead to poor performance on the exam, they may negatively impact performance on competing tasks – every hour the student spends studying is one fewer hour they can spend doing something else. Finally, it is unclear how overestimation of task difficulty affects performance (Figure 7C). If a student overestimates the difficulty of the exam, they may believe it is impossible to achieve their desired score. While it is possible that the student will overinvest resources to reflect anticipated task difficulty, they are just as likely to disengage from the task and stop studying (E. L. Bjork & Bjork, 2014; Harris, 1986).

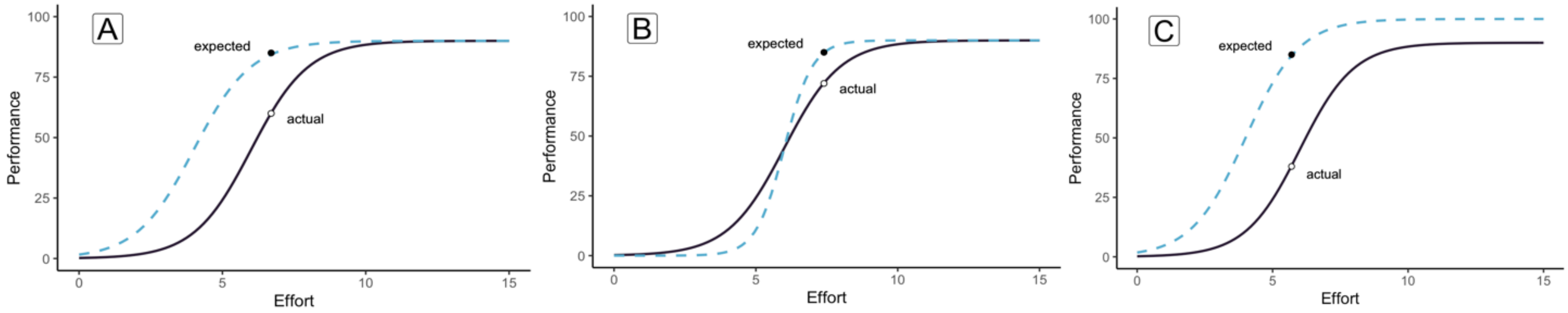


Figure 6. Graphs of actual (solid black lines) and estimated (dotted teal line) performance curves for a student capable of obtaining an 85% on an exam. Misestimations involving skill (panel A) have more serious impacts on performance than do misestimations of efficiency (panel B). Similarly, underestimations of difficulty (panel C) can strongly impact performance.

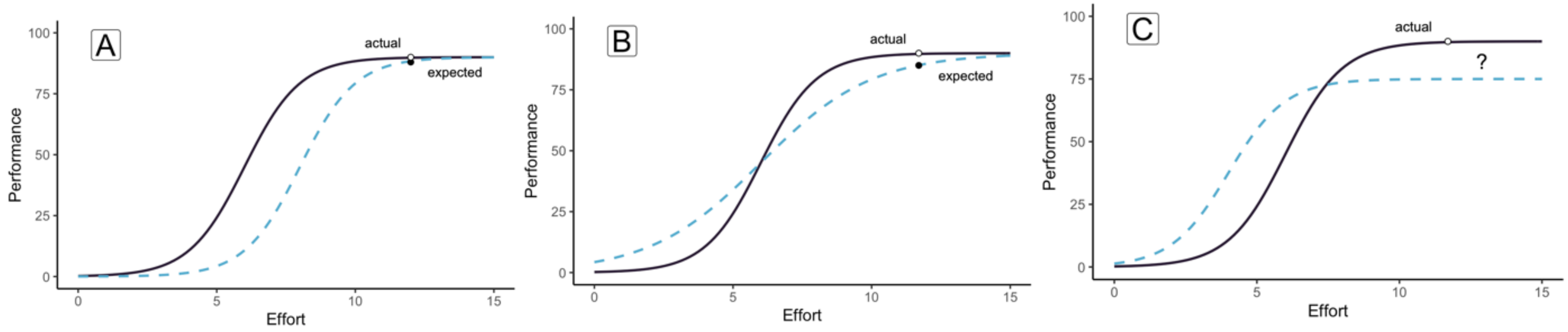


Figure 7. Graphs of actual (solid black lines) and estimated (dotted teal line) performance curves for a student capable of obtaining an 85% on an exam. Underestimations of skill (panel A) and resource efficiency (panel B) have less serious consequences than do overestimations (see Figure 5). The consequences of overestimation of task difficulty (panel C) are unknown.

These examples illustrate the first challenge in operationalizing task difficulty: performance-based measures are confounded with incentive magnitude, metacognitive ability, and perceptual sensitivity. That is, poor performance may occur because an individual:

1. does not value the outcome associated with strong performance;
2. is unaware of the effort than is required to succeed; and/or
3. misestimates the difficulty level of the task.

Although these moderators are related to task difficulty, they involve cognitive and perceptual processes that must be separated from the objective difficulty of the task. Assessing the influence of cognitive and perceptual processes on performance will enhance the theoretical understanding of task difficulty and outline ways in which environments can be restructured to facilitate learning.

Defining Difficulty, Effort, and Associated Constructs

People define task difficulty and effort in many ways (Bruya & Tang, 2018; Cain, 2007). As Hart and Staveland (1988) note, “people are unaware of the fuzziness of their own definitions [of difficulty and effort] or the possibility that theirs might be different than someone else’s.” This also appears to be true of scientists who are interested in understanding how changes in task difficulty affect subjective workload, effort allocation, performance, and downstream decisions. Just as laypeople tend to confuse difficulty, effort, and workload (Fisher & Oyserman, 2017), so too do researchers (Bruya & Tang, 2018; Cain, 2007). Furthermore, empirical definitions of difficulty tend to employ circular logic: people perform poorly when tasks are hard, therefore hard tasks will always produce poorer performance. Likewise, people rate challenging tasks as hard, therefore challenging tasks will always give rise to higher subjective ratings of difficulty. However, research on motivation (Covington, 2000) and fatigue (Åhsberg, 2000) suggest that many factors contribute to peoples' performance and subjective ratings of difficulty. For these reasons, definitions of task difficulty should not be grounded in peoples' performance or subjective ratings.

Moving forward, verbal theories must separate task difficulty from effort allocation and operator workload. Additionally, researchers must acknowledge the diverse factors that affect performance (e.g., reward magnitude, operator skill), isolate them from task difficulty, and identify and measure their influence on JODs and subsequent performance. To this end, I will first contrast difficulty with the related concepts of effort and workload. I will next discuss common approaches to measuring these constructs and describe the challenges that limit these approaches. Finally, I will conclude by describing methods for minimizing confounds and nuisance variance that can affect measures of task difficulty. This discussion will inform both the methodology of the dissertation and the proposed relationships between task difficulty and performance.

Contrasting difficulty, effort, and workload. Commonly, difficulty is defined as an interaction between an operator and the task (i.e., the allocation of available physical and cognitive resources (Curry, Jex, Levison, & Stassen, 1979; Eggemeier, 1991; Gopher & Donchin, 1986) and is operationalized using measures of subjective workload (i.e., metacognitive judgments) or performance. Defined in this way, objective task difficulty is confounded with individual sensitivity to reward and effort allocation. Alternatively, task difficulty can be defined as an objective construct that is based on and quantified by measurable task characteristics. Consider

Box 1. Difficulty, effort, and workload can be defined as separate and distinct constructs.

Difficulty. An objective construct based on and quantified by measurable task characteristics.
Effort. A measure of the amount of cognitive or physical resources allocated towards a task relative to a person's current workload.
Workload. A point estimate of the demands that have been placed on a person's cognitive or physical resources.

the word pairs that are frequently used in metacognition research involving judgments of difficulty and ease-of-learning (EOLs). Low-frequency words in the English language may be objectively more difficult to learn relative to high-frequency words, as evidenced by their lower probability of recall (Hulme et al., 1997) and ease of learning ratings (Jönsson & Lindström, 2010). It is possible to calculate word frequency - large-scale databases exist for many languages and dialects (e.g., the Corpus of Contemporary American English; COCA) - and use this value as an index of task difficulty. For example, the word "big" appears in written and spoken word 250 times more often than "calamity" (COCA frequencies of 258911 and 1033, respectively). If word frequency is a dimension of difficulty that affects performance, people should be more likely to remember "big" than they are to remember "calamity." Furthermore, it should be possible to determine the functional nature of this relationship by measuring recall performance for many words of varied frequency.

Difficulty should be contrasted with effort (see Box 1), which refers to the proportion of available resources an individual allocates towards a task. While harder tasks should warrant the allocation of more resources, this is not always the case nor is doing so always optimal (R. A. Bjork, 1999). Instead, it is important to note that effort may be over- or under-allocated depending on individual goals and task rewards. For example, people allocate more study time to word pairs and logic tasks that are incentivized with higher rewards, even when they are not any more challenging than lower-reward items (Koriat, Ma'ayan, & Nussinson, 2006). However, differential incentives are only effective in encouraging additional effort when an individual perceives success as being attainable or desires the rewards associated with task completion (Hensley, 2014). It is likely that individuals weigh estimates of difficulty and reward before engaging in a task (Kurzban et al., 2013), and that performance is a product of many factors beyond task difficulty.

Difficulty can also be contrasted with workload (see Box 1), which refers to the current demands that are placed on a person's cognitive and physical resources (Cain, 2007; Gopher & Donchin, 1986). Thus, workload is a function of a person's effort allocation as well as the level of difficulty presented by the task. The more resources that are allocated towards a task, the more significant a person's workload. Likewise, task difficulty is frequently correlated with workload because difficult tasks require more resources to complete successfully. To illustrate these principles, consider the situations faced by a payload operator, who must determine and adjust the weight and fuel specifications of aircraft. It is more difficult to perform this task with many aircraft than with few. Consequently, an operator who successfully manages five aircraft is under greater workload than one who manages only two. Both operators can choose to allocate additional effort towards managing their aircraft; however, the resources they can invest to do so are limited by the existing workload. Although workload is chiefly measured through subjective assessments (e.g., NASA-TLX, SWAT, WP; (Hart & Staveland, 1988; Reid & Nygren, 1988; Tsang & Velazquez, 1996), there is evidence to suggest that the aggregation involved in calculating these workload indices does not effectively capture individual differences in resource efficiency (McKendrick & Cherry, 2018) that protect against changes in performance under conditions of increasing workload. This metric is best captured by performance and the performance-effort slope, which represents the degree of resource efficiency (see Figures 6B and 7B).

Measuring difficulty, effort, and workload. Despite the ease with which difficulty can be objectively defined along continuous dimensions, researchers frequently assess difficulty as a function of an operator’s performance or subjective ratings, and dichotomize this metric following ad hoc procedures. Consider the trials of a flanker task[1]: it might be easier to indicate the direction of an on-screen arrow if you’ve received a brief clue (e.g., four additional arrows pointed in the same direction) about which way it will be pointing, just as it may be harder to indicate the direction of this arrow when the clue is misleading (i.e., the flanking arrows are pointed in the opposite direction; (Schouppe et al., 2014). Comparing the proportion of errors that participants make on these “easy” and “hard” trials provides a coarse estimate of how resources were allocated under each circumstance. That is, it is possible to attribute relative changes in performance to difficulty – here, the direction of the central arrow – because the other task and sample characteristics that might affect performance have been averaged away. However, it is incorrect to suggest that the variability in performance is solely due to the difficulty of the task itself or to state that task difficulty alone was responsible for participants’ performance in the task. Additional sources of variance hide in the error term and cannot be disambiguated from one another. Specifically, aggregated measures of task performance include changes due to task difficulty, a person’s effort allocation strategy, and a person’s workload.

Although it is impossible to disambiguate these sources of variance directly, they can be modeled by observing the effects of different variables on performance. For example, changing the negative consequences of an incorrect identification (e.g., randomly selected values between -5 and -50 points) may affect a person’s resource allocation strategy but will not affect the objective difficulty of the flanker task or that person’s current workload. Similarly, requiring a

[1] Participants in a flanker task are instructed to indicate the direction of an arrow that appears in the middle of a computer screen. This arrow is flanked by four additional arrows, two on each side. Sometimes, all of the arrows point in the same direction (congruent trials); other times, the flanking arrows point in the opposite direction (incongruent trials).

person to concurrently perform a demanding n-back task should affect their current workload, but would not necessarily affect their effort allocation strategy or the difficulty of the flanker task, which is independent of the n-back task. Thus, changes to incentive magnitude and workload should affect performance independently of changes in task difficulty.

Minimizing confounds and nuisance variance. The current practice of dichotomizing difficulty into “easy” and “hard” conditions masks the sources of variance in performance. Unless a dimension of difficulty is truly dichotomous (e.g., writing a sentence with one’s non-dominant hand), its continuous nature should be maintained in the form of objective task difficulty. Maintaining the continuous nature of each dimension allows mathematical models to capture the functional relationship between unique facets of task difficulty and performance, as well as the interactions that occur between difficulty and other variables, such as incentive magnitude or time.

Consider the congruent (easy) and incongruent (hard) trials in a flanker task. Comparing participants’ performance on these two trial types provides a coarse estimate of the effects of this dimension of difficulty (see Figure 8, left panel); however, it limits how effectively the

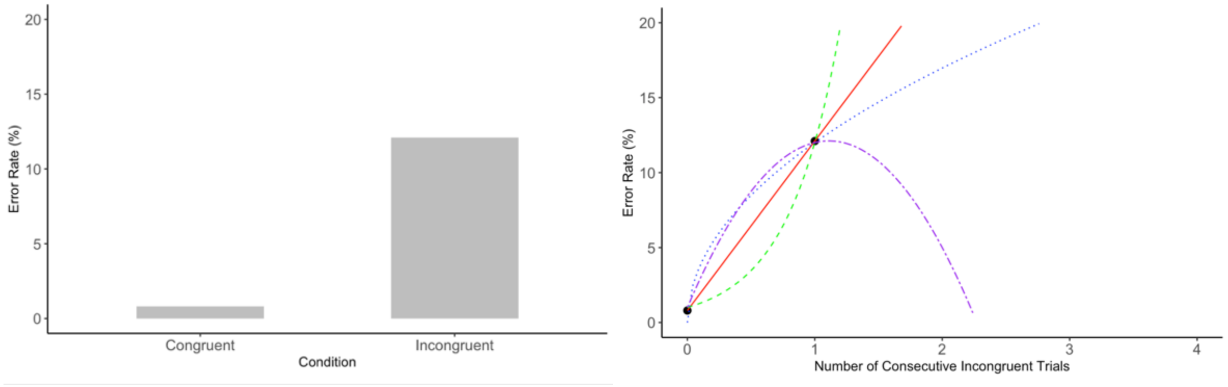


Figure 8. Error rates (i.e., percent incorrect responses) across congruent and incongruent flanker trials (left panel; Schouppe et al., 2014). Measures of confidence and variability were not originally reported. The graph on the right illustrates various underlying relationships that could produce this pattern of performance.

relationship between this dimension of difficulty and relevant outcomes can be modeled (Young, 2016).

An alternative way to assess the relationship between flanker congruence and task performance is to model how a person's error rate changes as a function of the number of consecutive presentations of incongruent trials (see Figure 8, right panel). The error rate could increase linearly (solid red line), exponentially (dashed green line), or logarithmically (dotted blue line); alternatively, performance could deteriorate but then improve as incongruent arrows become a clue in themselves – a quadratic relationship (dashed purple line). These meaningful distinctions can only be made when task difficulty is defined objectively along continuous dimensions.

Disaggregating task difficulty along continuous dimensions improves the quality of model predictions, but also creates challenges in the form of stimulus selection. Specifically, participants must experience similar rates of change in task difficulty. Consider two participants tasked with answering timed multiplication problems. Both participants have 15 s to complete their first math problem. However, they differ in how long they have to complete the second: one participant is given 5 s, while the other participant receives 14 s. The change in task difficulty that has occurred will be more noticeable to the first participant than to the second because the degree of change is larger. Thus, the first participants' JODs will be more accurate than those of the second (Jemstedt, Kubik, & Jönsson, 2017).

One explanation for this effect is that metacognitive accuracy is improved by stimulus variability (Jemstedt et al., 2017). However, it is equally likely that these improvements are due to physiology or to a statistical artifact. Increasing stimulus variability decreases the perceptual similarity between any two items or tasks, which may improve inter-item discrimination

(Stevens, 1957). Similarly, maximizing stimulus variability also expands the scale of a particular dimension; standardizing with conditions or dimensions that do not exhibit the same degree of variability will artificially enhance the slope of some conditions while reducing that of others.

For illustration of this latter point, consider Figure 9, which illustrates changes in performance across levels of task difficulty in a videogame. Here, task difficulty was standardized across conditions so that 0 represented the easiest level and 1 represented the hardest level that a person encountered. In this analysis, individual differences in skill, effort allocation, and learning were accounted for by the random effect structure of a multi-level model; the remaining variance in performance was due to the difficulty level of the task. Although the performance slopes for three conditions (Damage, Line-of-Sight, and Strength) are roughly equivalent across levels of difficulty, two conditions (Population, Speed) differ drastically in this regard. These two conditions contained a wider range of difficulty levels that, when standardized, appear to represent a meaningful effect (Vangsness & Young, 2018).

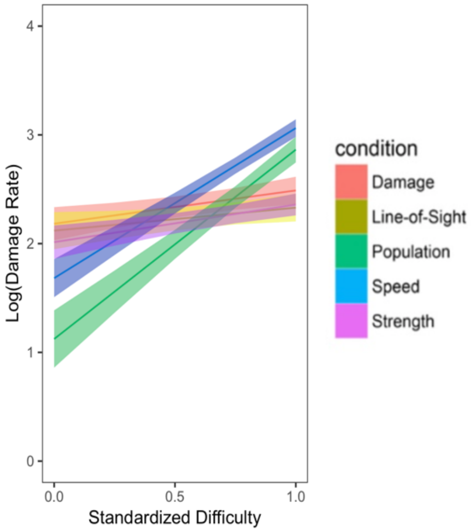


Figure 9. Participants' performance (log-transformed Damage Rate) was affected by task difficulty, which was standardized so that 0 represented the easiest and 1 represented the hardest level that participants would encounter. Individual differences in skill, effort allocation, and learning were captured in the random effect structure of this model; the remaining variance in performance is due to the difficulty level of the task, nuisance error, and the confounding variable of condition difficulty.

In summary, research on task difficulty must address the limitations of past research in three respects:

1. provide clear definitions of difficulty that are grounded in the characteristic(s) responsible for making a task more difficult;

2. disaggregate task difficulty along its original, continuous dimension(s); and
3. minimize confounds and nuisance variance through rigorous pilot testing and the use of covariates.

Addressing these challenges will make it possible to determine the effects of task difficulty on performance and JODs.

A Cohesive Framework for Testing Relationships

Systematically addressing the relationships between task difficulty, effort, and performance requires a cohesive framework that includes both task-related and operator characteristics. That is, research efforts must acknowledge that effort allocation and, in turn, performance are affected both by task characteristics as well as by how well a person perceives and responds to those characteristics (Gopher & Donchin, 1986; Meshkati, Hancock, Rahimi, & Dawes, 1995). For example, people may over- or underestimate their abilities relative to task demands (Ehrlinger et al., 2008; Kruger & Dunning, 1999), or may differ in the degree to which they notice changes in characteristics that are valid indicators of task difficulty (Stevens, 1957). Furthermore, the degree to which these characteristics influence JODs may depend upon previous experiences (Löffler, von der Linden, & Schneider, 2016). Finally, people may be more or less sensitive to performance-based feedback; individual tolerance for poor performance may differ from person to person, and larger violations of performance expectations may be required to trigger an updated representation of one's abilities.

These stages and the intervening variables that affect them are outlined in Figure 10 and can be illustrated by a simple example: attempting to parallel park a vehicle. Before parking, a driver must judge the difficulty of the parking task. An initial estimate can be made using self-efficacy beliefs and does not require any information about the parking situation itself: a person who is learning to drive and is aware of their lack of experience is likely to provide higher estimates of task difficulty than a seasoned driver or a novice with unrealistic expectations of their skill. This initial estimate can then be modified by central cues to difficulty, task characteristics that provide information about difficulty. Even an experienced driver will indicate that it is more difficult to park in a narrow spot on a busy street than it is to park along an empty curb. Once a JOD is made, the driver must decide how to allocate resources to the task by considering the incentive structure of the environment and the costs of resource allocation. In the case of parallel parking, a driver may contrast the costs and benefits of seeking an alternative spot with those of successfully parking in the given location. If success is unlikely or an individual is poorly calibrated to task difficulty, fewer resources may be allocated than are necessary to maximize the probability of success. Information about the incentive structure of the environment may be used to make JODs (e.g., some people believe that difficult tasks are highly rewarded; (Jönsson & Lindström, 2010), or inform the allocation of resources after a JOD has been made. If the task is attempted, the driver will learn whether their resource allocation was appropriate, given their skills and the objective difficulty of the task. That is, performance-based (e.g., inability to perform the task; a fender bender) and proxy peripheral cues (e.g., time-on-task) will become available. This feedback can facilitate learning and be used to update self-efficacy beliefs.

The model presented in Figure 10 provides a cohesive framework for elements and concepts that have been discussed separately in the context of metacognition (R. Ackerman, 2014; Nelson & Narens, 1990) and workload allocation (Anderson, 1996; Kahneman, 1973; Wickens, Gutzwiller, et al., 2016). In addition, it provides an iterative mechanism by which past experiences can be used to inform current judgments and decisions related to task engagement.

This framework will serve as a basis for the current section, where I will review theoretical and empirical evidence relating to each stage of task assessment and resource allocation.

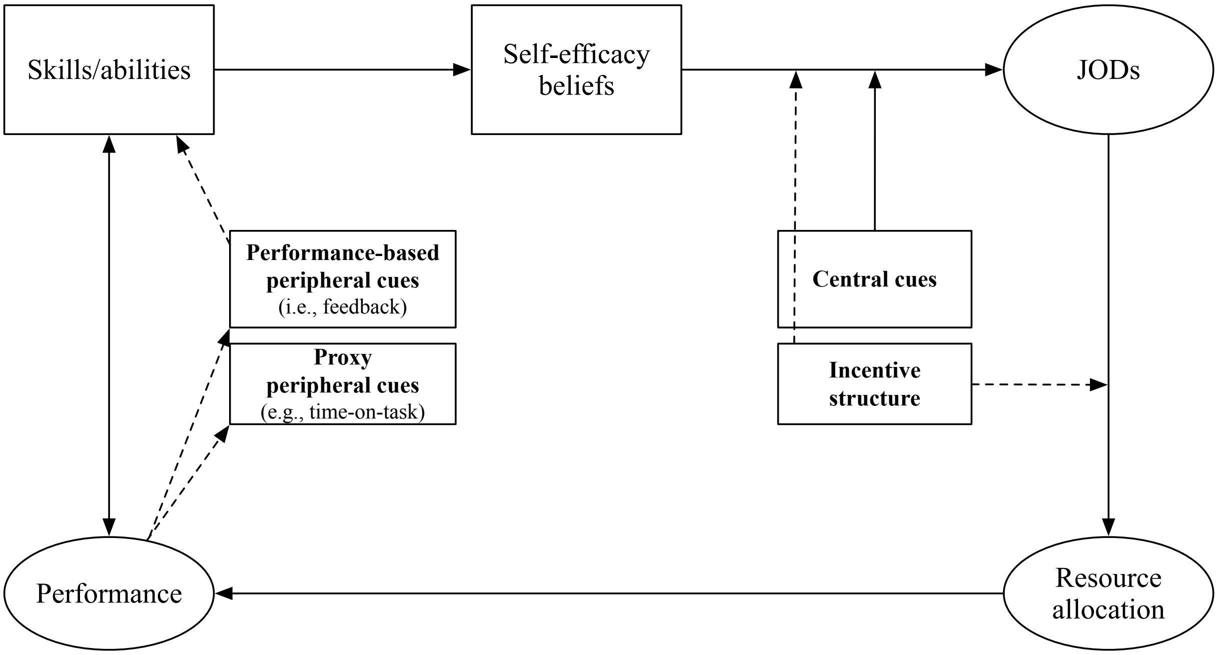


Figure 10. A cohesive framework illustrating the proposed relationships between JODs, resource allocation, and performance.

Skills and Competency Beliefs Inform JODs

Even when task-specific information is absent or ambiguous, a person can use their skills and abilities to make a JODs (see Figure 11). For example, a person might use their overall fitness to predict the difficulty of going for a run. In this context, a self-proclaimed couch potato might predict that the run would be more difficult than would someone who bikes to work every day. Although these JODs could be improved by knowing the distance of the run, the kind of terrain on the route, or the amount of fatigue one would experience, this information is not

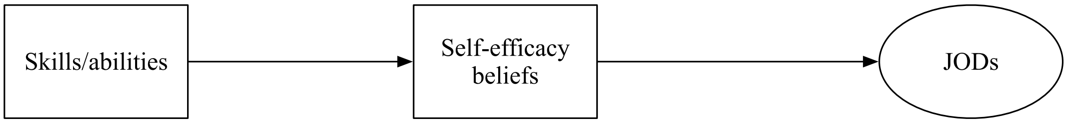


Figure 11. Self-efficacy beliefs can inform JODs in the absence of task-specific information.

required to make a judgment; JODs can be made independently of or be enhanced by task engagement (Bandura, 1977).

When task-specific information is absent or ambiguous, self-efficacy beliefs are predictive of effort allocation and task completion (Bandura, 1986, 1989). For example, job seekers who believe they are highly skilled apply for more jobs and are more likely to complete applications than those who do not feel as confident in their abilities (Ellis & Taylor, 1983). Similarly, research scientists' self-efficacy beliefs are predictive of the number of papers they publish and the number of times they are cited by other researchers (Taylor, Locke, Lee, & Gist, 1984). Neither employers nor journal editors provide much information about how difficult it will be to apply and may offer little to no feedback when an applicant is rejected. In situations such as these, self-efficacy is the strongest predictor of task performance (Stajkovic & Luthans, 1998).

Although self-efficacy can be a strong predictor of task performance, it may not always produce accurate JODs. Two kinds of errors can occur: a person can misjudge their skills and abilities relative to the task (i.e., misestimate the upper asymptote) or fail to recognize how differences in task-specific characteristics affect the difficulty of the task (i.e., misestimate the slope). The couch potato in the previous example (purple solid line) may recognize that a run will be more challenging for them than for their fit friend (teal solid line),

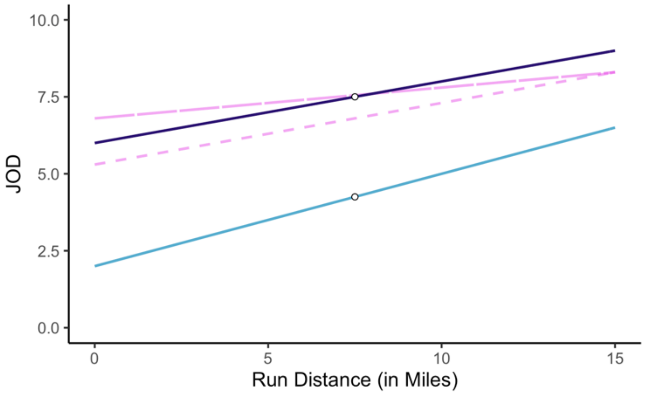


Figure 12. A person's JODs (purple solid line) can be accurate relative to a reference group (teal solid line) while still exhibiting inaccuracy. Although this figure assumes a linear relationship, it is possible that a nonlinear (e.g., exponential) function better describes the relationship.

regardless of distance. That is, their average JOD is accurate as compared to a reference group, as represented by the open points in Figure 12. However, the couch potato may still misestimate the difficulty of the run relative to their skills and abilities: they may perceive themselves as fitter than they actually are (i.e., misestimating the intercept; pink dotted line) or may lack an understanding of how distance can affect the difficulty of a run (i.e., the slope; pink dashed line). Thus, self-efficacy beliefs and task characteristics both contribute to the accuracy of JODs.

Accuracy in estimating overall skills and abilities. It is well-established that poor performers overestimate their performance relative to peers, while high performers exhibit slight underestimation (Dunning, 2011; Dunning et al., 2003; Ehrlinger et al., 2008; Larrick, Burson, & Soll, 2007). One explanation for this reversal is regression to the mean. That is, a ceiling effect artificially constrains the error in top performers' estimates while poor performers' estimates exhibit a greater degree of error. This hypothesis offers a simple explanation for the results of many experiments (P. L. Ackerman, Beier, & Bowen, 2002; Kruger & Dunning, 1999), and can help explain why many drivers think they are above average.

On the other hand, it is possible that poor performers are "unskilled and unaware." That is, poor performers' estimates may be undermined by their inexperience: they do not know what information can be used to determine whether a task will be difficult or easy for them to complete, nor can they accurately assess their performance in the absence of overt feedback. High performers, on the other hand, may fall prey to a false consensus effect and incorrectly assume that others are similarly talented (Kruger & Dunning, 1999). This problem may be exacerbated when individuals' reference group changes. A transition of this sort affects the skill level of the reference group (Krajč & Ortmann, 2008) as well as the predictive power and availability of different task characteristics (Ryvkin, Krajč, & Ortmann, 2012).

Consider a student who graduates from high school and attends a university. Most universities have some performance threshold for acceptance, which inflates the skill level of the reference pool - most of the students who are accepted to a university are likely to be in the top half of their graduating class. Thus, a once above-average student could now be part of the lower quartile of the incoming class. Until the student recognizes the situation has changed, they will likely overestimate their average academic skill and/or resource efficiency relative to their peers. This misestimation may be exacerbated because the information that could be used to correct this judgment is provided infrequently (e.g., official grades, informal assessments). Instead, the student may rely upon readily available but less-accurate surface information from lectures and textbooks, such as the ease with which they can read the materials (i.e., processing fluency; (Oppenheimer, 2008). Under these conditions, inaccurate self-efficacy beliefs may be caused by changes in the reference group, in the quality of task characteristics, and in the frequency of feedback.

Although longitudinal research suggests that the accuracy of peoples' self-efficacy beliefs improves with experience (Ryvkin et al., 2012), it is unclear whether this improvement is due to increased familiarity with the reference group, with the task, or with performance-based feedback. The mechanism underlying this improvement is important because poorly-calibrated self-efficacy beliefs have immediate and long-lasting effects on learning and downstream performance. When people overestimate their abilities relative to task demands, they fail to engage in strategies that compensate for their weaknesses (Azevedo, Moos, Greene, Winters, & Cromley, 2008) and engage in tasks even when success is unlikely (Corbalan, Kester, & van Merriënboer, 2008; Kostons, van Gog, & Paas, 2012; Ross, Morrison, & O'Dell, 1989). Conversely, those who underestimate their skills are unlikely to select tasks that challenge and improve their abilities, which is particularly detrimental in self-paced learning situations (Azevedo et al., 2008; Deci, Ryan, & Williams, 1996; Goforth, 1994; Lawless & Brown, 1997; Niemiec, Sikorski, & Walberg, 1996). Understanding the relationship between self-efficacy beliefs, task characteristics, and performance-based feedback will eliminate competing hypotheses and improve learner interventions.

The Moderating Effect of Cues to Difficulty

Cues to difficulty can be used to determine which skills will be required to complete a task. When many cues are available, a person must identify which are valid predictors that provide information about the skills that will be required to successfully complete the task. Tasks with many cues require multi-attribute judgments (Bandura, 1977), which are prone to higher degrees of error than single-attribute judgments (Nickerson, 1967). Errors are introduced when people incorporate invalid cues or inappropriately weight valid cues to difficulty.

Consider a marathoner tasked with completing a short run. In the absence of task-specific information, the runner may rate their athletic skills as being fairly high. However, once additional information becomes available, their skill assessment may change: a 5-mile trail run is shorter but more technically challenging than a road race. While marathoners are highly skilled

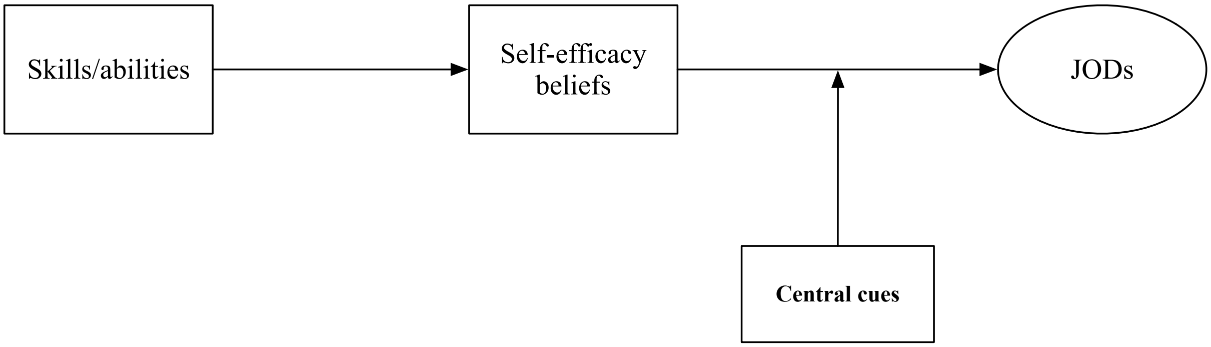


Figure 13. Task characteristics (i.e., cues to difficulty) can moderate the relationship between self-efficacy beliefs and JODs.

at running long distances, they are less skilled at maintaining balance and control in unpredictable environments. Over- or under-weighting these pieces of information can weaken the relationship between self-efficacy and performance (see Figure 13).

Cues used to evaluate difficulty. I have proposed that effort allocation occurs at task onset in response to judgments of difficulty, reward, and the probability of successful task completion. That is, an individual must assess task difficulty relative to their ability and allocate resources in such a way that they achieve their goal and are rewarded for their investment. The process of effort allocation must be repeated in response to external (e.g., task characteristics) and internal (e.g., fatigue) changes that occur during the process of task completion (Koriat, 1997; Sweller, 1994; Sweller, Ayres, & Kalyuga, 2011). For example, a tactical coordinator charged with evaluating the threat level of nearby aircraft must allocate more cognitive resources during high-traffic occasions and fewer when there are no aircraft approaching. Regardless of the circumstances, the coordinator may increase resource allocation in response to an order from superiors or may decrease it in anticipation of a much harder task. When monitored properly, external and internal changes allow the coordinator to balance resource allocation with successful task completion while mitigating the effects of fatigue.

An individual's ability to monitor changes to external and internal characteristics depends on perceptual sensitivity and cognitive processing limitations (Simon, 1955; Stevens, 1957). For example, a tactical coordinator is unlikely to notice the difference between a training scenario that contains 54 airplanes and one that contains 55 (Kaufman, Lord, Reese, & Volkmann, 1949). But even obvious changes in task difficulty may be overlooked by a coordinator who must complete multiple tasks at the same time or is distracted by cues that are not indicative of task difficulty, such as the presence of a decision aid (Vallières et al., 2016). In this way, subjective perceptions of workload, effort allocation, and performance are constrained by the factors that affect judgment and decision-making processes.

Additionally, the cues that signal external and internal changes during task completion occur at different times. Some cues are available to all people and exist independently of skill or resource allocation, while others are affected by individual experience with the task. I refer to these as central and peripheral cues, respectively (see Table 1). Generally, central cues are available at the onset of the task or become known very early in task completion: a tactical coordinator can estimate the number of approaching aircraft (information complexity) and would know that failure to identify an enemy could be catastrophic (reward magnitude). During the process of task completion, peripheral cues become available: the coordinator may receive automated feedback (a performance-based cue) or may note the amount of time it takes to identify the aircraft (a proxy cue). Thus, central cues can be used to make initial and ongoing judgments of task difficulty and resource allocation, while peripheral cues serve to monitor one's progress in addressing the challenges presented by the task (Vangsness & Young, 2018).

Although central cues are likely the most valid predictors of difficulty, they are task-specific and may be hard to identify. Consequently, people may find it easier to track peripheral cues that are valid under many circumstances and serve as a proxy for task difficulty (Potts, Pastel, & Rosenbaum, 2018). For example, while rescue pilots are familiar with inhospitable flight conditions, they are unlikely to have experience landing at the exact site of a future mission. When confronted with a novel situation, they may use feelings of fatigue to inform their ongoing assessment of task difficulty. Because fatigue frequently correlates with task difficulty, it often serves as a valid cue; however, it may be misleading under certain circumstances (e.g., a person suffering from insomnia).

Table 1.
Central and peripheral cues differ in their availability and consistency during a task. The early availability and consistency of central cues lend themselves to prediction, while peripheral cues can be used to monitor performance during a task.

Central Cues	Peripheral Cues	
• Available at task onset • Independent of skill and resource allocation	• Available after task interaction • Vary with skill and resource allocation	
Task-based	**Performance-based**	**Proxy**
e.g., stimulus complexity; target speed; time deadline	e.g., accuracy; reward; score	e.g., completion time; feelings of fatigue

Central and peripheral cues can also be contaminated by the context surrounding the task. For example, women perform more poorly on difficult math tests when they complete them in a mixed-gender group; this effect disappears when women receive instructions that dispel myths about gender differences in math ability (Cohen et al., 2016; Spencer, Steele, & Quinn, 1999). In this example, central cues to difficulty are identical across exam contexts; however, the gender composition of the classroom increases anxiety (a proxy cue) for some test takers, which in turn impairs performance. Now, some students' performance-based cues are reflective of task difficulty as well as their physiological responses to an invalid cue. If invalid proxy cues inform their judgments of difficulty, these students may poorly allocate resources and further impair their performance (Nelson & Narens, 1990).

Sources of bias in self-efficacy beliefs. Stereotype threats are one example of a situation in which cues to difficulty moderate self-efficacy beliefs to produce inaccurate JODs and poor performance; however, many such tasks exist. Broadly speaking, when cues to difficulty suggest that a task will be easy, people tend to overestimate their skills relative to those of others (Alicke, 1985; Marottoli & Richardson, 1998; Zenger, 1992). However, people tend to underestimate their skills when cues suggest that a task will be difficult (Hoelzl & Rustichini, 2005; Kruger & Dunning, 1999; Kruger, Windschitl, Burrus, Fessel, & Chambers, 2008). This is true even when cues to difficulty are invalid predictors of performance: participants presented with seemingly easy and hard trivia questions develop this pattern of behavior when questions are equated for difficulty across groups (Arkes, Wortmann, Saville, & Harkness, 1981).

The degree to which a person's self-efficacy beliefs are susceptible to these biases depends on their actual abilities and performance. Poor performers strongly overestimate their performance relative to peers when tasks are perceived as easy, while high performers exhibit slight underestimation. The magnitude of these errors reverse when tasks appear difficult: top performers underestimate their performance to a greater degree while poor performers become more accurate in their assessments (Burson, Larrick, & Klayman, 2006).

The importance of a unified cue framework. Although cues to difficulty and self-efficacy beliefs share a close relationship, individual differences are often minimized by averaging across levels of difficulty or skill. This averaging allows researchers to determine how people evaluate task difficulty in general but masks the unique effects that self-efficacy beliefs and cues to difficulty have on JODs. If an insensitivity to cues (i.e., the "unskilled and unaware" problem) drives the above average effect, JODs should be informed by cues to difficulty, especially among those who are experienced. That is, the intercept should be affected but the slope should not. Similarly, gaining familiarity with appropriate task-specific characteristics should affect sensitivity without affecting average judgments relative to a reference group (see Figure 14B). Additionally, observing others' performance should improve peoples' understanding of the reference class and affect the average accuracy of metacognitive judgments without affecting sensitivity to task-specific characteristics (see Figure 14A). Finally, receiving extensive performance-based feedback should reduce the variability of individuals' estimates above and beyond what is warranted by either an improved understanding of the reference class or cues to difficulty (see Figure 14C). If the above average effect is simply due to ceiling and floor effects (i.e., regression to the mean), changes in JODs will occur as participants learn or gain experience with a task, but will be insensitive to other changes in parameters.

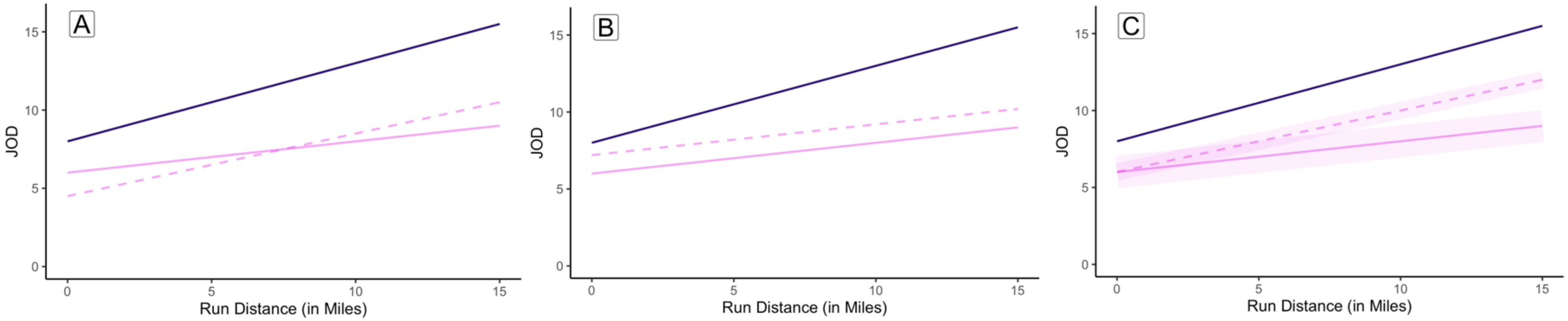


Figure 14. If metacognitive biases are caused by task-specific experience, pre- (solid pink line) and post-task (dotted pink line) should share a systematic relationship with actual performance (purple line). Specifically, improving one's knowledge of the reference class should affect the intercept but not the slope (panel A); improving sensitivity to cues to difficulty should affect the slope but not the intercept (panel B); and receiving performance-based feedback should significantly affect the variability of people's JODs (panel C).

Judgments of Difficulty and Effort Allocation

Once a person has made a JOD, they can decide how much effort to invest in completing that task. Although systematic research on the relationship between JODs and effort allocation is lacking (Jemstedt et al., 2017), insights can be leveraged from metacognitive research on Judgments of Learning (JOLs), Ease of Learning (EOLs), and study time. This research supports a direct relationship between JODs and effort allocation; however, the nature of this relationship remains unclear. In general, it seems that although students invest more effort in studying challenging topics (Nelson & Leonesio, 1988; Son & Metcalfe, 2000) this relationship occasionally inverts (Koriat & Ackerman, 2010). Competing hypotheses have arisen to explain the inconsistent relationship between JOLs, EOLs, and study time; several of these models suggest that the incentive structure of the environment has both a direct and moderating role in effort allocation. That is, the incentive structure can provide information regarding how difficulty a task will be – a 10-point essay is often harder than a 2-point multiple true/false question ((Jönsson & Lindström, 2010); see Figure 13). The incentive structure may also moderate the relationship between peoples' JODs and their effort allocation strategies (see Figure 15): both incentives (e.g., bonus points) and disincentives (e.g., a late penalty) can change behavior. Although the incentive structure can involve both positive and negative outcomes, I refer to these as the "incentives" of the environment for conciseness.

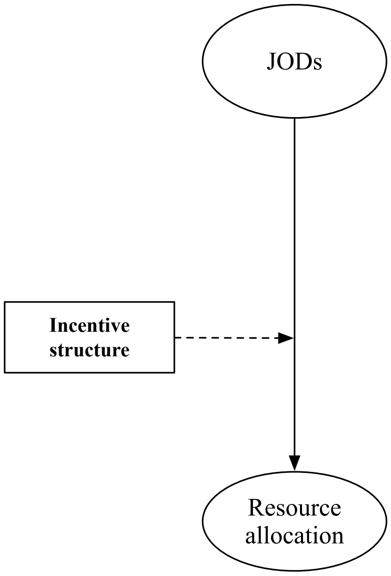


Figure 15. Incentive value, a central cue to difficulty, may moderate the relationship between JODs and effort allocation.

Competing hypotheses have arisen to explain the inconsistent relationship between JOLs and study time; however existing research designs do not allow researchers to determine when the incentive structure plays a moderating role. In the section that follows, I will briefly review these competing hypotheses and their predictions regarding the relationship between JODs, central cues, and effort allocation.

Diminishing Criterion Model. The Diminishing Criterion Model (DCM; (R. Ackerman, 2014) suggests that effort is allocated after a person compares their current state to their goal state. In the context of study time, a student might estimate their current level of content knowledge and compare it to what would be required to reach their desired level (e.g., a desired course grade). Thus, the less well-known the content, the more time the student will spend studying it now (Son & Metcalfe, 2000) and in the future (Nelson & Leonesio, 1988; Thiede, Wiley, & Griffin, 2011). Similar behavior should emerge within the context of a dynamic task, such as running a race. A runner might use their current pace to estimate their finishing time. The further the current pace from the goal pace, the more effort the runner will invest to run faster and reach their goal.

The DCM's assertion that effort allocation strategies are developed by comparing desired states to current states, restricts the moderating role of the incentive structure. Specifically, peoples' JODs should be sensitive to changes in task incentives; however, incentives should not explain any of the variance in task performance above and beyond JODs. Furthermore, this relationship should be positive in nature: tasks that are perceived as being more challenging should warrant the allocation of additional effort, regardless of the reward associated with their completion.

Proximal Zone of Learning Theory. The DCM can be contrasted with the Proximal Zone of Learning Theory, which predicts that people will allocate more effort towards tasks that lie within the level of their skills and abilities (Dunlosky, Kubat-Silman, & Hertzog, 2003; Metcalfe, 2009) because they are likely to be most successful in completing them. In other words, the Proximal Zone of Learning Theory predicts that people will maximize their payouts: students should spend the most time studying content that they are likely to master, given the amount of time they have before the exam. This perspective is consistent with effort allocation theories that emphasize resource conservation (Boksem & Tops, 2008; Kool & Botvinick, 2014; Kurzban et al., 2013) in that students' success rate for the most difficult items will almost always be lower than their success rate for easier items. When items are worth the same number of points, an easy-item investment strategy is optimal.

The Proximal Zone of Learning Theory predicts that a person's effort allocation strategy will be determined by their current level of skills and abilities, relative to the challenges presented by the task. In this sense, Proximal Zone of Learning Theory is similar to the DCM: peoples' JODs should be sensitive to changes in the incentive structure, but incentives should not explain variance in task performance. However, this relationship should be negative in nature: tasks that are perceived as being easier should warrant the allocation of additional effort, regardless of the reward associated with their completion.

Moderation models of effort allocation. Unlike the DCM and Proximal Zone of Learning Theory, moderation models suggest that peoples' effort allocation strategies are affected by the incentives associated with task completion (Koriat et al., 2006; Metcalfe & Kornell, 2005; Son & Metcalfe, 2000; Undorf & Ackerman, 2017). Specifically, people will invest more effort in tasks that are associated with greater rewards or lesser consequences. For example, students might spend extra time studying topics that will be worth more points on an exam, regardless of how difficult those topics are. This effect is lessened when people are distracted (Sobel, Gerrie, Poole, & Kane, 2007) or when working-memory load is high (Barrett, Tugade, & Engle, 2004), suggesting that attending to and remembering the incentive structure contributes to workload itself.

Although the relationship between incentives and effort is well-established (Hull, 1943; Mitchell, 2017), moderation models of effort allocation such as Agenda-Based Regulation (Ariel et al., 2009) and the Strategic Task Overload Model (Wickens, Gutzwiller, et al., 2016) offer specific predictions regarding JODs. These models suggest that people will allocate greater effort towards tasks that seem easier to complete, but that this preference will be affected by task incentives: harder tasks that are strongly incentivized will encourage greater efforts, as well. Framed another way, JODs will affect the average amount of effort allocated towards a task, but this will change as a function of the incentive's magnitude (i.e., an interaction effect will be present).

Evaluating Task Performance

When a person allocates resources towards a specific task, they engage in the process of performance. A runner does not need to complete a 5K, but if they choose to do so they must engage their physical resources and begin to take steps towards the finish line. As these resources are recruited in pursuit of the task, performance – running or perhaps walking – occurs. These cumulative efforts inform the runner's skills and abilities. The more frequently the runner practices, the better they will become. In this way, current efforts affect future performance (see solid arrows, Figure 16).

Task performance also gives rise to peripheral cues that can be used to make JODs. Some of these peripheral cues are directly related to performance itself. For example, an athlete may begin to feel the physical effects of running. The harder they work, the more they will experience these effects. This performance-based peripheral cue is a product of the factors that contribute to effort allocation (self-efficacy beliefs, objective task difficulty, incentive value, and effort allocation), and can be used to inform subsequent resource allocation. Task performance also gives rise to proxy cues, which frequently (but not always) correlate with task difficulty. For example, difficult running routes frequently take longer to complete; however, time-on-task can also be affected by situations unassociated with task difficulty, such as pausing on the trail for an unexpected conversation with a friend. These proxy cues may inform peoples' JODs or effort allocation strategies; however, these relationships are beyond the scope of this dissertation and do not appear in the model.

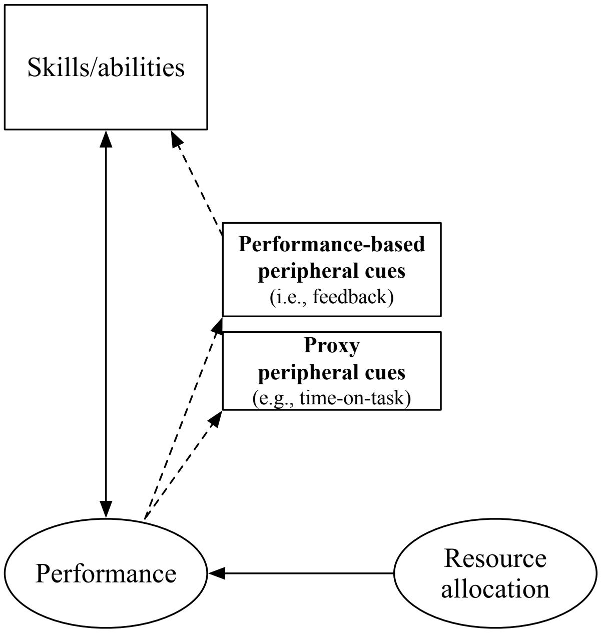


Figure 16. Resources can be allocated towards task performance, which in turn affects skills and abilities through learning. Task performance also gives rise to peripheral cues – performance-based and proxy cues – that can be used to make JODs.

Establishing the relationship between peripheral cues and task completion would help explain anecdotal examples of human behavior. Consider the relationship between student performance and motivation. Perhaps students become discouraged and disengage from tasks because heightened levels of fatigue lead them to doubt their self-efficacy. Conversely, students may withdraw because peripheral cues lead them to believe a task is more challenging than it is.

However, it is also possible that students disengage because performance-based feedback alters the incentive structure of the environment by providing a positive punishment. Although it is clear that peripheral cues to difficulty play an important role in learning (Butler & Winne, 1995; Gaeth & Shanteau, 1984; Gonzalez, 2005; Kluger & DeNisi, 1996; Shanteau, 1992) and student engagement (R. A. Bjork, 1999), it is unclear when these cues become integrated during the process of task evaluation and completion (see Figure 17).

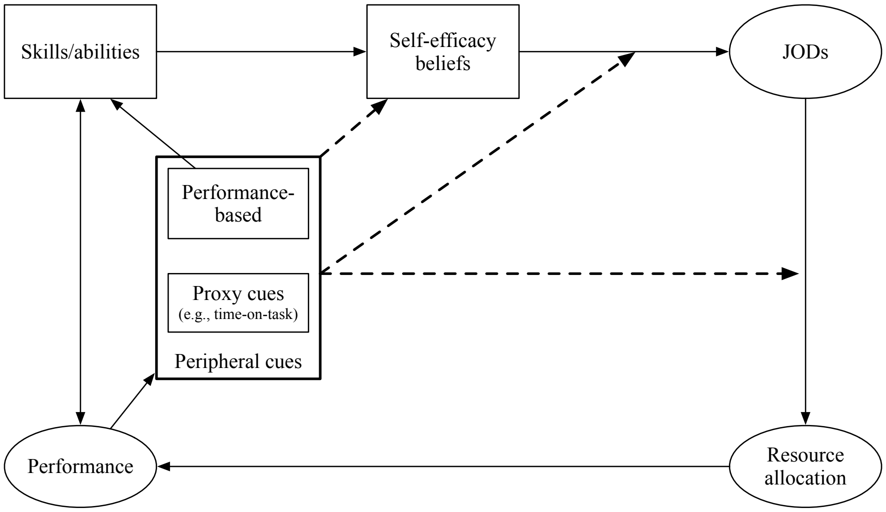


Figure 17. Peripheral-cues may influence effort allocation in a variety of ways: they may negatively impact self-efficacy beliefs, inform JODs, or change in the incentive structure of the environment.

The Dissertation

This dissertation narrowly focuses on the relationship of peripheral cues to difficulty and JODs, being mindful of the fact that this relationship exists within the larger framework of task completion. Specifically, this dissertation will:

1. Assess the degree to which metacognitive judgments negatively affect task completion (Pilot Study);
2. Determine whether performance-based peripheral cues must be experienced to inform JODs (Experiment 1);

3. Evaluate the relationship between performance-based peripheral cues and the incentive structure of a task (Experiment 2).

These aims will be addressed within the context of visual search, a task in which participants must find a target item that is hidden among a set of similar distractor items. The amount of time participants have to identify the target is limited; at the end of each search trial, participants receive performance-based feedback that indicates whether their search strategy was successful or not. Because this task is relatively straightforward, it provides a controlled context in which peripheral cues can be studied independently of other factors that may influence task completion.

Chapter 2 - Pilot Study

The primary goal of the pilot study was to calibrate the task. Specifically, the pilot was conducted to ensure that the difficulty levels presented within each condition had similar effects on participants' performance. This step ensured that any JOD or performance differences that emerged across difficulty conditions was due to an inherent property of those dimensions of difficulty rather than the level of difficulty they presented.

The secondary goal of the pilot study was to determine the effects of frequent JOD assessment on participants' task performance. Many experiments require participants to alternate between completing a task and making metacognitive judgments (Brewer & Sampaio, 2006; El Saadawi et al., 2010; Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013; Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014; Koriat, 2008; Maclaverty & Hertzog, 2009; Souchay, Isingrini, Clarys, & Taconnat, 2004) For example, a person taking a test may be asked to rate their confidence in each answer before proceeding to the next question (Koriat, 2008). The rationale for this procedure is simple: a person must perform a task before they can report the cognitions they had while completing it. It is difficult to rate your confidence in an answer before one has been provided or to estimate the degree to which you "know" something if you haven't tried to recall it. Research confirms that participants' metacognitive judgments are more accurate after performing a task (Siedlecka et al., 2018); however, this traditional experimental design implicitly assumes that alternation has little effect on participants' task performance or metacognitive accuracy.

Although there is little research on whether alternation affects task performance or metacognitive accuracy, evidence from task-switching experiments suggests that alternating metacognitive judgments with task completion may represent a form of task interruption that interferes with the performance of both tasks (Kiesel et al., 2010; Monsell, 2003; Pashler, 2000; Wickens, Gutzwiller, et al., 2016; Wickens et al., 2015). That is, participants must disengage from the primary task (e.g., completing an exam; recalling word pairs) to make a metacognitive judgment and vice versa. This procedure is similar to the fixed sequences found in some task-switching paradigms, during which participants must alternate between performing two tasks. Alternating tasks leads participants to perform worse in both (Allport, Styles, & Hsieh, 1994; Spector & Biederman, 1976), even when the switches are predictable and both tasks are relatively simple (Rogers & Monsell, 1995). If metacognitive judgments represent a form of task switching, task performance and metacognitive accuracy should deteriorate the more frequently the switch is performed during an experiment.

On the other hand, making repeated metacognitive judgments may improve participants' metacognitive skill and task performance (Jemstedt et al., 2017; Schwartz, Boduroglu, & Tekcan, 2016) by providing them with more accurate scale anchors (Colle & Reid, 1998). For example, participants who only provide feelings-of-knowing (FOKs) for items that they can't answer may not know what a strong FOK is like. When participants provide metacognitive judgments for all answers, they are better able to gauge their FOKs and can provide more accurate judgments on subsequent items. Furthermore, frequent metacognitive evaluations may allow participants to engage in self-paced study or training that improves their task performance (R. Ackerman, 2014; Koriat et al., 2006; Tullis & Benjamin, 2011). Thus, it is equally possible that task performance and metacognitive accuracy will improve with more frequent assessments.

In summary, three competing hypotheses arise from the literature:

> H1 (*traditional paradigm*): the frequency of metacognitive assessment has no effect on task performance or metacognitive accuracy.

H2 (*self-regulated study*): as the frequency of metacognitive assessments increase, participants' metacognitive accuracy improves (i.e., stronger slope effect) and their task performance improves (i.e., higher median performance).

H3 (*task switching*): as the frequency of metacognitive assessments increase, participants' metacognitive accuracy deteriorates (i.e., shallower slope effect) and their task performance is capped by increased workload (i.e., lower performance asymptote).

The pilot experiment addressed these hypotheses by manipulating the frequency of metacognitive judgments (here, JODs) between-subjects. Specifically, some participants made a JOD following each trial while other participants made a JOD after every five trials. Importantly, the level of difficulty that participants experienced remained the same for five trials, regardless of condition. This allowed the comparisons made across conditions to reflect the relative effects of frequency on performance and metacognitive accuracy.

Method

Participants

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Kansas State University. The experimental task was completed by 59 participants (44 female) who received 1 hr of research credit as compensation.

The Conjunctive Visual Search Task

Participants completed 320 trials of a conjunctive visual search task that was programmed with the Unity (Unity Technologies, 2019) game engine (see Figure 18). Participants identified a target (blue circle) item from among non-target distractors that shared

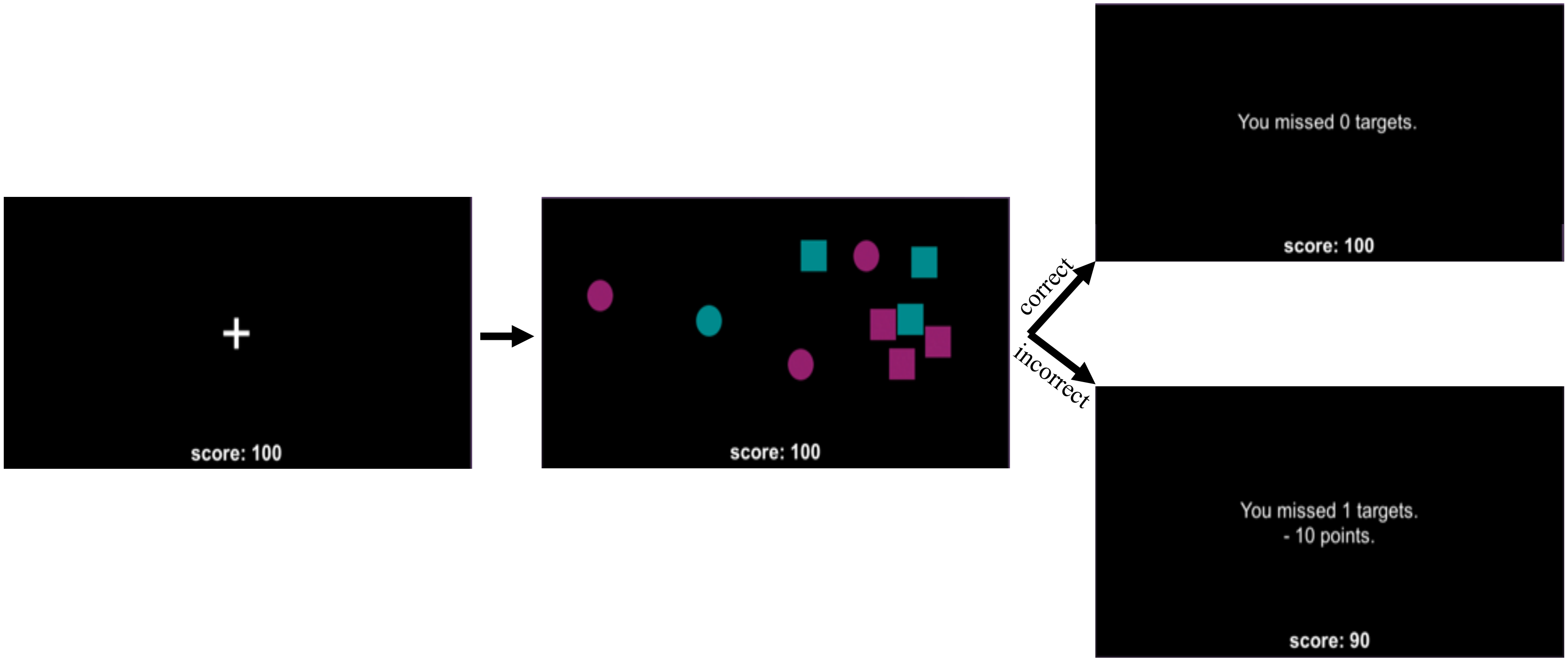


Figure 18. A schematic diagram of the conjunctive visual search task.

the same color (blue squares), shape (pink circles), or that had no overlapping features (pink squares). To successfully identify the target item, participants needed to click on it with the computer mouse. Once the target was identified, it disappeared from the screen; the stimulus array remained visible for the full duration of the trial to ensure that time-in-trial was decorrelated from task difficulty.

Stimulus array. Each stimulus array contained one target item and several non-target items that were evenly assigned to the array without replacement. That is, non-target types were equally represented in the stimulus array; when the number of non-targets was not evenly divisible by three, the remainder were randomly selected to ensure that non-target items were evenly represented throughout the experiment. These items were presented at random, non-overlapping on-screen locations.

Trial timing. Although aspects of the task differed across conditions, all participants experienced the same sequence of events. Each trial began with a fixation cross, presented for 500 ms, followed by the presentation of the stimulus array. Once the trial elapsed, participants received on-screen feedback that was visible for 750 ms. These screens were separated by a 500 ms delay, jittered by ±150 ms. Given these parameters, participants completed the experiment in about 50 mins.

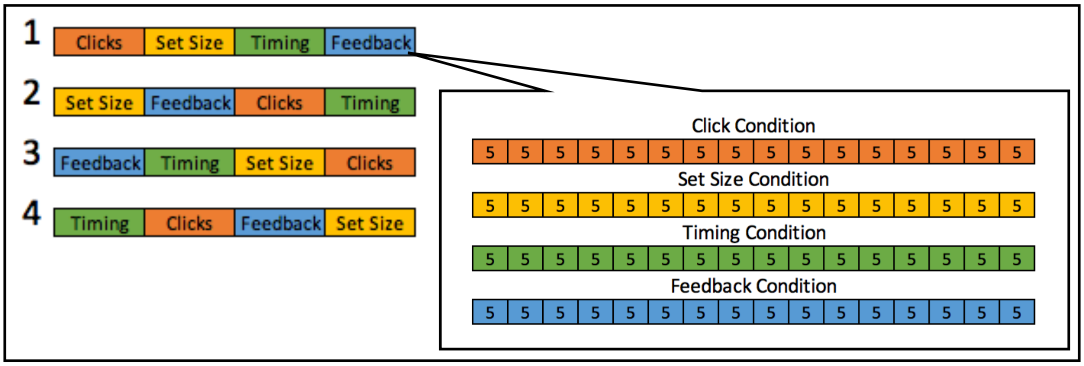
Task difficulty. Because this task was developed to assess participants' sensitivity to changes in difficulty, difficulty was manipulated as a within-subject variable. Specifically, the difficulty of the task changed along a single dimension after every five trials; the remaining task dimensions were fixed at a value that represented an average level of difficulty (see Table 2). The values of the changing dimension were selected by semi-random algorithm (a Halton sequence; (Halton & Smith, 1964) from a sampling distribution that represented floor and ceiling performance. This ensured that participants experienced levels of difficulty that reflected the full range of performance. Finally, sampling values along the same dimension were further randomized using the Fisher-Yates shuffle algorithm (Black, 2005) before assignment. This additional step ensured that task difficulty was decorrelated from time-on-task and was not informed by participants' performance.

Table 2.
Both the dimension and level of difficulty were manipulated within-subjects. Difficulty condition was Latin Square counter-balanced, and the level of difficulty was randomly assigned after every five trials. The remaining dimensions were held at fixed values when they were not manipulated.

difficulty condition	description	sampling values	fixed values
Click	The number of times the target had to be clicked before it disappeared.	1 - 6	3
Feedback	The number of points lost after failing to identify the target.	2 – 45	25
Set Size	The number of non-target items in the stimulus array.	2 – 45	25
Timing	The amount of time the stimulus array appeared on screen.	1.04 s – 4.46 s	2.10 s

Dimension of difficulty. The dimension along which the task was made more or less difficult was also manipulated as a within-subject variable (difficulty condition). Specifically, participants experienced changes along four different dimensions of difficulty (see Table 2). These changes

Box 2. *A schematic diagram of the pilot study's trial counterbalancing. Each rectangle at left represents 80 trials of each difficulty condition, which were blocked so that consecutive changes occurred across a single dimension of difficulty. For illustrative purposes, a single participant's trials are illustrated at right.*



were blocked so that participants performed 80 consecutive trials for each dimension of difficulty; counterbalancing was conducted using a Williams Latin Square design (see Box 2).

Participants learned of these difficulty conditions at the beginning of the experiment and were informed of "changes in the nature of the task's difficulty" at the start of each block; however, they did not receive information about presentation order. This omission allowed the experiment to more closely resemble realistic tasks in which the nature of difficulty must be inferred through past or current experience.

Feedback. Participants began each five-trial segment with 100 points, a portion of which were lost each time the participants failed to identify the target. Participants were informed of their performance, in terms of target identification and points, after every trial. When participants' points dropped below zero, they had to wait 30 s while the level "reloaded." This waiting period served as an aversive consequence that motivated participants to engage with the game, a practice that has been adopted successfully across several applications153. After the waiting period, participants' points were restored to 100 and the search task resumed.

JODs. On some trials, the program paused and requested that participants make a JOD after receiving feedback. Participants were asked to indicate whether the task was easier or harder than before by clicking on one of two buttons. Framing JODs in this way increased reliability by ensuring that all participants anchored judgments to an objective task experience (Böckenholt, 2004). Once participants selected an option, the buttons disappeared, and the task resumed.

JOD frequency manipulation. The frequency with which participants made JODs was manipulated as a between-subjects variable: half of the participants made a JOD after each trial (frequent condition), while the other half made a JOD after every five trials (infrequent condition).

Results

Defining Variables

Two outcome variables were used to conduct hypothesis tests and calibration analyses. Participants' *metacognitive accuracy* was coded as a dichotomous variable that indicated whether participants were correct (1) or incorrect (0) in their JODs. Task performance was also coded as a dichotomous variable that indicated whether participants correctly identified the target on a trial (1) or not (0).

Task difficulty was quantified by *standardized difficulty*, a within-subject predictor that equated difficulty across conditions by subtracting the lowest possible sampling value of a condition and dividing by the sampling range. This produced a variable where 0 represented the easiest level that participants encountered in a given condition and 1 represented the hardest; timing trials were reverse-coded to reflect that providing participants with more time made the task easier. *Difficulty condition*, a four-level categorical predictor, was included to model how task difficulty differed across dimensions (clicks, feedback, set size, or timing).

Both *experimental trial* and *trial in block* were considered as within-subject predictor that could model the changes in behavior that occurred as people completed the task. The first of these predictors captured the number of trials that had elapsed since the start of the experiment (1-320), while the second captured the number of trials that had elapsed since the start of a difficulty condition block (1-80). Finally, *JOD frequency*, a two-level categorical predictor, indicated whether participants made a JOD after every trial (frequent) or after each block (infrequent).

All predictors were median-centered or effect-coded prior to analysis. This ensured that the model intercept represented participants' average performance or metacognitive accuracy. Thus, significant effects could be interpreted as the degree by which a variable affected participants' average performance or average metacognitive accuracy. With respect to task calibration, this allowed algebraic equations to be used to determine the appropriate values for the difficulty parameters in subsequent experiments (see Appendix A for additional information about task calibration).

Selecting A Random Effect Structure

The random effect structure was informed by a previous publication that used a similar preparation (Vangsness & Young, 2018) and included the intercept, standardized difficulty slope, and trial slope. This allowed the model to capture individual differences in skill (intercept, standardized difficulty slope) and rate of learning (trial slope). Unlike the earlier preparation, this experiment manipulated difficulty condition as a within-subject variable. Thus, it was unclear whether learning effects would occur across the entire experiment or within each block (i.e., for each difficulty condition). AIC comparisons (Akaike, 1973) were used to determine which predictor – experimental trial or trial in block – was most likely to have produced the data. These analyses indicated that experimental trial best modeled individual differences in participants' performance and JODs over time (see Table 3).

Table 3.
AIC comparisons of random effect structures modeling participants' performance and metacognitive accuracy.

Random effect structure	AIC	ΔAIC
Performance		
standardized difficulty + total trials	22283.66	
standardized difficulty + block trials	22730.44	446.78
Metacognitive Accuracy		
standardized difficulty + total trials	4650.39	
standardized difficulty + block trials	4653.14	2.75

Note. Models with lower AIC values are more likely to have generated the data; differences larger than ±3 are strong justification to select the best-fitting model.

Competing Hypotheses - Performance

A multi-level logistic regression was used to calibrate the task and evaluate competing hypotheses regarding the influence of metacognitive judgments on performance. This model included the main effects of standardized difficulty, difficulty condition, and experimental trial, as well as the Standardized Difficulty × Difficulty Condition interaction. These terms were included to aid in calibration. The model also contained the main effect of JOD condition, as well as its two-way interactions with standardized difficulty and experimental trial. Intercept, standardized difficulty slope, and experimental trial slope were allowed to vary across participants to model individual differences in skill and rate of learning. The results of this model are presented in Table 4.

Participants' performance was affected by task difficulty (standardized difficulty) such that they were more likely to identify the target on easier trials than on harder ones; however, substantial performance differences emerged across difficulty condition, confirming that the task required calibration (see Figure 19 and Tables 5-6). Regression weights were calculated using the emmeans package in R (Lenth, Singmann, Love, Buerkner, & Herve, 2019) and were used to calibrate the task. Additional details about this procedure can be found in Appendix 1.

Planned contrasts indicated that participants' performance did not differ as a function of JOD frequency, $B_{difference} = 0.16$, $SE = 0.21$, $z = 0.75$, $p = .45$, nor did JOD frequency affect participants' rate of learning in the task, $B_{difference} = 0.12$, $SE = 0.10$, $z = 1.18$, $p = .24$ (see Figure 20). These results provide partial support for the traditional paradigm (H1) and partially refute the self-regulated study (H2) and task switching (H3) hypotheses.

Table 4.
Unstandardized regression weights from a multi-level logistic model predicting the likelihood that participants would correctly identify the target on a given trial.

predictor	B	*SE*	*z*	*p*
intercept	0.57	0.11	5.35	< .001
standardized difficulty	-5.64	0.13	-42.29	< .001
click condition	-1.22	0.04	-27.56	< .001
feedback condition	-0.55	0.04	-13.78	< .001
set size condition	-0.43	0.04	-10.43	< .001
experimental trial	0.39	0.05	7.84	< .001
frequent JODs	0.10	0.10	0.92	.36
Standardized Difficulty × Click Condition	0.16	0.16	0.95	.34
Standardized Difficulty × Feedback Condition	5.73	0.15	38.89	< .001
Standardized Difficulty × Set Size Condition	3.80	0.15	25.68	< .001
Standardized Difficulty × Frequent JODs	-0.01	0.08	-0.18	.86
Experimental Trial × Frequent JODs	0.06	0.05	1.18	.24

Note. Standardized difficulty (Mdn = 0.5) and total trials (Mdn = 160) were median-centered prior to analysis. Difficulty condition and JOD frequency were contrast-coded, with the timing condition and infrequent JODs serving as {-1, -1, -1} and {-1} baseline conditions, respectively.

Table 5.
Predicted performance intercepts for each difficulty condition.

Condition	B	*SE*	$95\%_{CI}$
click	-0.77	0.11	[-0.99, -0.56]
feedback	-0.10	0.11	[-0.31, 0.11]
set size	0.02	0.11	[-0.20, 0.23]
timing	2.63	0.14	[2.36, 2.91]

Table 6.
Predicted standardized difficulty slopes for each difficulty condition.

Condition	B	*SE*	$95\%_{CI}$
click	-5.48	0.16	[-5.79, -5.17]
feedback	0.09	0.11	[-0.13, 0.31]
set size	-1.84	0.12	[-2.07, -1.61]
timing	-15.32	0.46	[-16.21. -14.42]

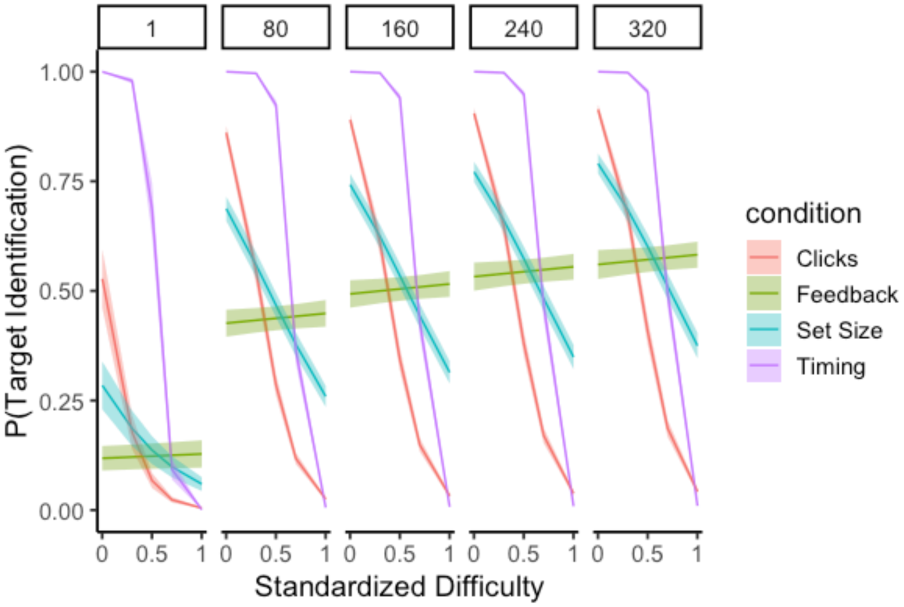


Figure 19. Standardized difficulty, difficulty condition, and experimental trial significantly predicted visual search performance. Experimental trial is depicted across panels; error ribbons represent ±1SE.

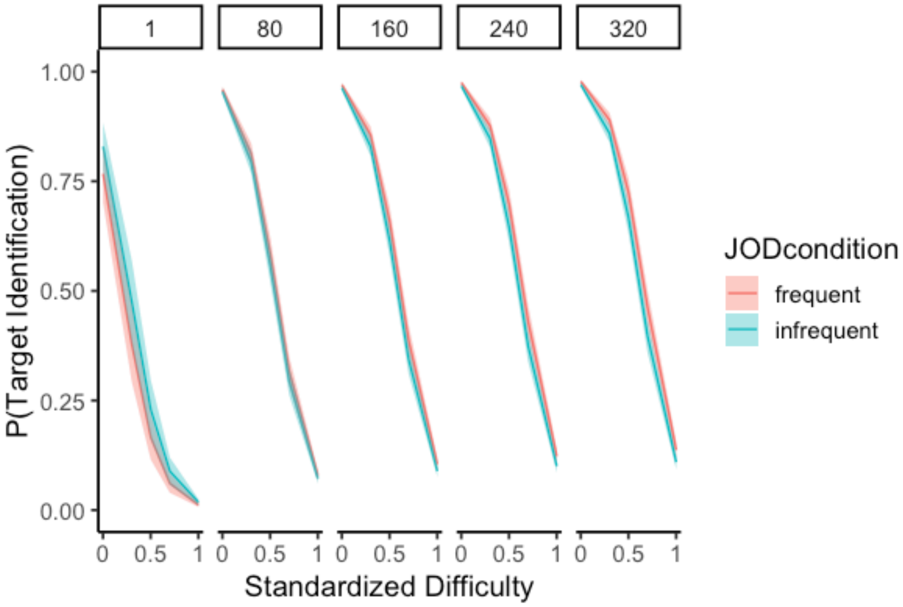


Figure 20. Participants' performance, learning rate, and sensitivity to changes in task difficulty did not differ as a function of JOD condition. Experimental trial is depicted across panels; error ribbons represent ±1SE.

Predicting Metacognitive Accuracy

A second multi-level logistic regression was used to assess whether the frequency of metacognitive judgments affected the accuracy of participants' JODs. The fixed effect structure of the model included the main effects of standardized difficulty, difficulty condition, and experimental trial, as well as the Standardized Difficulty × Difficulty Condition interaction. These terms were included to determine the degree to which condition differences and workload affected metacognitive accuracy. The model also included the main effect of JOD frequency, as well as its two-way interactions with standardized difficulty and total trials. The results of this model are presented in Table 7.

Table 7.
Unstandardized regression weights from a multi-level logistic model predicting the accuracy of participants' metacognitive JODs.

predictor	B	*SE*	*z*	*p*
intercept	0.75	0.05	14.18	< .001
standardized difficulty	-0.22	0.15	-1.44	.15
click condition	0.19	0.06	2.89	.004
feedback condition	-0.41	0.06	-6.74	< .001
set size condition	-0.07	0.06	-1.13	.26
experimental trial	0.04	0.05	0.89	.38
frequent JODs	-0.16	0.05	-2.94	.003
Standardized Difficulty × Click Condition	0.14	0.21	0.67	.50
Standardized Difficulty × Feedback Condition	-0.18	0.20	-0.89	.38
Standardized Difficulty × Set Size Condition	-0.58	0.21	-2.76	.01
Standardized Difficulty × Frequent JODs	-0.11	0.15	-0.73	.46
Experimental Trial × Frequent JODs	0.04	0.04	0.98	.33

Note. Standardized difficulty (Mdn = 0.5) and total trials (Mdn = 160) were median-centered prior to analysis. Difficulty condition and JOD frequency were contrast-coded, with the timing condition and infrequent JODs serving as {-1, -1, -1} and {-1} baseline conditions, respectively.

In general, participants tended to make less accurate metacognitive judgments as the difficulty of the task increased; however, the strength of this relationship differed across difficulty conditions (see Figure 21 and Tables 8-11). Participants' metacognitive accuracy was highest in the Timing condition, followed by the Clicks and Set Size conditions. Additionally, participants were most sensitive to changes in Timing, followed by Clicks and Set Size (see Figure 21).

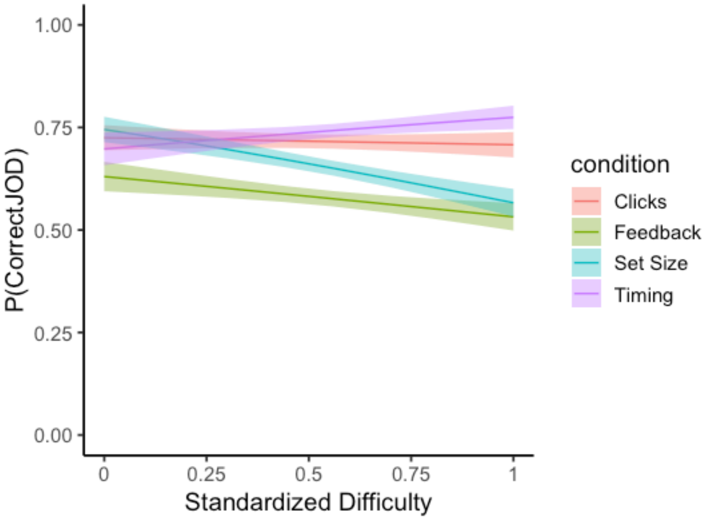


Figure 21. Metacognitive accuracy differed as a function of task difficulty and difficulty condition. Error ribbons represent ±1SE.

Table 8.

Predicted averages (i.e., intercepts) for metacognitive accuracy in each difficulty condition.

Condition	B	*SE*	95%$_{CI}$
click	0.93	0.08	[0.77, 1.09]
feedback	0.34	0.07	[0.18, 0.49]
set size	0.67	0.08	[0.52, 0.83]
timing	1.04	0.09	[0.87, 1.21]

Table 9.

Planned comparisons of participants' average metacognitive accuracy (i.e., intercept) in each difficulty condition.

Condition	B	*SE*	*z*	*p*
Click – Feedback	0.60	0.10	5.83	< .001
Click – Set Size	0.26	0.10	2.48	.06
Click – Timing	-0.11	0.11	-1.01	.75
Feedback – Set Size	-0.34	0.10	-3.42	.003
Feedback – Timing	-0.70	0.10	-6.81	< .001
Set Size – Timing	-0.37	0.11	-3.43	.003

Table 10.
Predicted standardized difficulty slopes for metacognitive accuracy in each difficulty condition.

Condition	B	*SE*	$95\%_{CI}$
click	0.93	0.08	[0.77, 1.09]
feedback	0.34	0.07	[0.18, 0.49]
set size	0.67	0.08	[0.52, 0.83]
timing	1.04	0.09	[0.87, 1.21]

Table 11.
Planned comparisons of standardized difficulty slopes for metacognitive accuracy in each difficulty condition.

Condition	B	*SE*	*z*	*p*
Click – Feedback	0.32	0.33	0.98	.76
Click – Set Size	0.72	0.34	2.15	.14
Click – Timing	-0.48	0.38	-1.28	.58
Feedback – Set Size	0.40	0.33	1.21	.62
Feedback – Timing	-0.80	0.37	-2.15	.14
Set Size – Timing	-1.20	0.38	-3.17	.01

Additionally, the model indicated that participants' metacognitive accuracy differed as a function of JOD condition. Participants who made frequent JODs were 64% likely to make a correct judgment on a given trial. In contrast, those who made infrequent JODs were 71% likely to make a correct judgment. Interestingly, these condition differences were unaffected by standardized difficulty, $B_{difference} = -0.22$, $SE = 0.30$, $z = -0.73$, $p = 0.46$, or experimental trial, $B_{difference} = 0.09$, $SE = 0.09$, $z = 0.98$, $p = 0.33$ (see Figure 22), suggesting that metacognitive accuracy was not further affected by the workload demands of the primary task. This provides partial support for the task switching hypothesis (H3) and refutes the traditional paradigm (H1) and self-regulated study (H2) hypotheses.

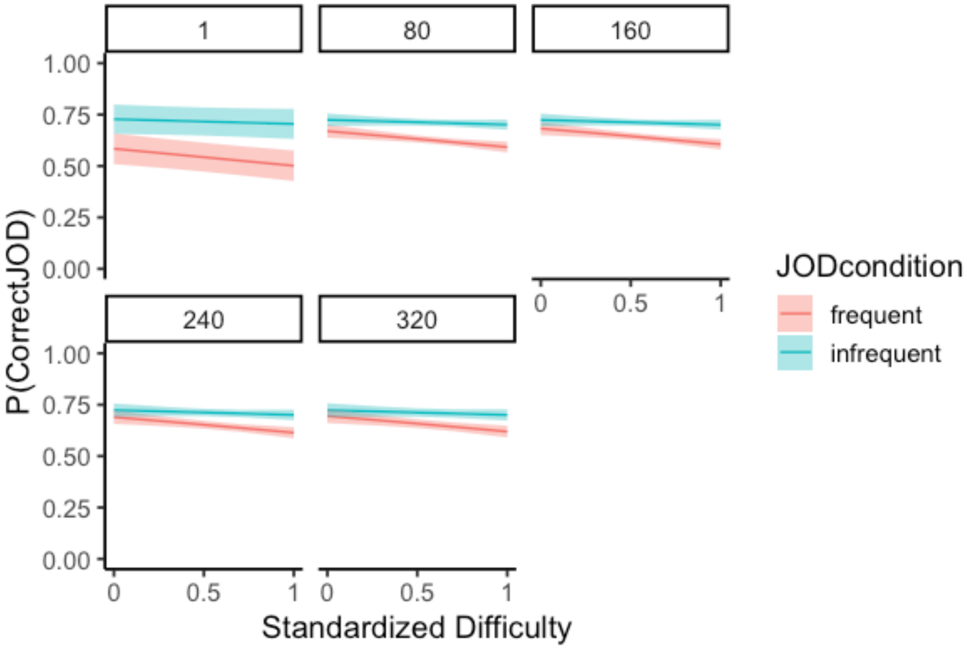


Figure 22. Metacognitive accuracy was affected by JOD condition, but not standardized difficulty or experimental trial. Experimental trial is depicted across panels; error ribbons represent ±1SE.

Discussion

The results of this experiment confirmed that task calibration is an important step for researchers to take if they are interested in understanding the cognitive processes that inform effort allocation and JODs. Participants performed substantially better in some difficulty conditions than in others, indicating that they encountered different ranges of difficulty levels in each condition. This finding was used to calibrate the experimental task for subsequent studies.

Additionally, this experiment addressed competing hypotheses regarding the relationship between metacognitive JODs and performance. Participants' performance in the visual search task was not adversely affected by making JODs. In contrast, the accuracy of participants' JODs was influenced by the frequency with which they made these judgments. Participants who made

JODs after every trial made judgments that were, on average, 7% less accurate than their peers who made JODs at the end of each block (five trials). Although this effect was not moderated by task difficulty or experimental trial, it provides support for the task switching hypothesis (H3) while refuting alternative hypotheses.

Unlike the traditional paradigm (H1) or self-regulated study (H2) hypotheses, the task switching hypothesis predicts that participants' metacognitive accuracy will be adversely affected by cognitive load. Here, cognitive load arises from two sources: the difficulty of the visual search and the process of making a JOD. Performance was not adversely affected by JOD frequency, suggesting that participants considered the visual search their primary task. Consequently, performance decrements appeared in the accuracy of participants' JODs.

One interpretation is that frequent metacognitive assessments increase workload and produce poor judgments. However, primary task workload (task difficulty) did not affect metacognitive accuracy, raising the potential for an alternative hypothesis. In the frequent JOD condition, participants were required to indicate whether the task was easier or harder than before on every trial, even those for which task difficulty did not change (i.e., trials 1-4). It is possible that these uninformative JODs led participants to down-weight information from certain cues to difficulty. This hypothesis lies beyond the scope of this dissertation; however, future follow-up analyses should be conducted to rule out this competing hypothesis.

Chapter 3 - Experiment 1

The primary goal of Experiment 1 was to determine whether the cues that people used to make JODs while they performed a task were the same as those they used when they observed someone else perform the same task. In doing so, this experiment addressed competing hypotheses about the underlying mechanisms responsible for mindreading and metacognition (understanding other's or one's own experiences, respectively). The three competing hypotheses advanced by the metacognitive literature differed in their predictions about the relationship between performance-based peripheral cues and JODs.

The simulation theory proposes that observation allows a person to imagine how they might perform a task. Imagined performance can be used to guide metacognitive judgments, such as estimates of task difficulty. This can be contrasted with performance settings in which a person may receive feedback. In these settings, feedback serves as a performance-based peripheral cue that can be used to calibrate metacognitive judgments (Dimaggio, Lysaker, Carcione, Nicolò, & Semerari, 2008). This suggests that that people weigh cues to difficulty differently depending on whether they are performing or observing a task. In concrete terms, the simulation theory suggests that performance-based peripheral cues only inform JODs when participants are engaged in the task; when trials are observed, self-efficacy beliefs guide metacognitive judgments. Although simulation is sometimes viewed as a conscious process (Koriat & Ackerman, 2010), it may also occur without our knowledge (e.g., the activation of mirror neurons while watching others perform a task; (Gallese & Goldman, 1998).

This perspective can be contrasted with Carruther's (2009) assertion that mindreading precedes metacognition. The mindreading-as-metacognition hypothesis asserts that cue use is unaffected by role. Both central and performance-based peripheral cues, such as feedback, inform a person’s JODs whether they complete a task or watch someone else perform it. In the event that they are observing, performance-based peripheral cues reflect another person’s skill and effort allocation strategy. In this sense, metacognition is simply a specific form of mindreading in which the subject is oneself (i.e., all metacognition is mindreading; (Carruthers, 2009). Thus, JODs should be informed by the same cues to difficulty, regardless of whether a task is performed or observed.

The third perspective suggests that peoples’ metacognitive judgments are more accurate when they have access to internal cues that are unavailable while observing others perform a task (Koriat, 1997, 2000). These mnemonic cues (e.g., how hard it feels to retrieve a memory) or subjective “feelings” are only available when actively engaged in task completion. Although introspective feelings cannot be measured, metacognitive judgments do become more accurate after people gain personal experience with a task (Koriat & Ackerman, 2010) and can be biased by personal knowledge of task outcomes (Arkes et al., 1981; Christensen-Szalanski & Willham, 1991; Kelley & Jacoby, 1996; Nussinson & Koriat, 2008). However, these same outcomes would also be expected given simulation theory. Because the key differentiating feature of this introspective theory is immeasurable (mnemonic cues), it will be captured by the error variance of statistical models. Thus, introspective theory predicts that central and peripheral cues to difficulty will become less predictive of JODs as people spend more time performing the task; changes in the weighting of central and peripheral cues will not occur when people are observing others.

Experiment 1 will address these hypotheses by manipulating participants’ role in the visual search task. Specifically, participants will perform half of the trials in the experiment and will observe as another person (the computer) completes the other half of the trials. The order in which participants serve in these roles will be manipulated between subjects. Participants will either perform half of the experimental trials and then observe the second half, observe half of the trials and perform the second half, or alternate between roles every 5 trials.

In the context of this task, the three competing hypotheses are:

> H1 (*simulation theory*): The degree to which performance-based peripheral cues inform JODs will increase when participants perform the task and decrease when they observe someone else completing the task.
>
> H2 (*mindreading theory*): The degree to which performance-based peripheral cues inform JODs will increase over time across all role conditions.
>
> H3 (*introspective theory*): The degree to which performance-based peripheral cues and central cues inform JODs will decrease when participants perform the task. Additionally, there will be greater modeling error during performance trials.

Method

Participants

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Kansas State University. The experimental task was completed by 65 participants (52 females) who received 1 hr of research credit as compensation. Two participants were omitted for failing to understand and follow the directions. Due to experimenter error, demographic information was not collected from two participants.

The Conjunctive Visual Search Task

Participants completed a visual search task that was similar to that of the pilot study. The task differed in four important respects. First, task difficulty was equated across conditions using performance data from the pilot study. Second, participants made JODs after every trial. Third, participants' role was manipulated to test the critical hypothesis. Finally, an additional demographic measure was added for exploratory purposes. Additional details are provided below; readers interested in learning more about specific task parameters are directed to the pilot study.

Task difficulty. Participants' performance during the pilot study informed the difficulty level of the task. Specifically, a multi-level modeling approach generated estimates of the performance intercept and standardized difficulty slope for all four conditions; adjustments were made to equate participants' performance (see Table 6). Unfortunately, attempts to equate difficulty across dimensions were unsuccessful; a more thorough discussion of this issue is provided in Appendix A. As before, the difficulty level of the task changed along a single dimension every 5 trials. The remaining task dimensions were fixed at the same level participants encountered during the pilot study to ensure that performance estimates were accurate.

Table 12.
Both the dimension and level of difficulty were manipulated within-subjects. Difficulty condition was Latin Square counter-balanced and the level of difficulty was semi-randomly assigned after every five trials. The remaining dimensions were held at fixed values when they were not manipulated.

difficulty condition	description	sampling values	fixed values
Click	The number of times the target had to be clicked before it disappeared.	5 - 7	3
Feedback	The number of points lost after failing to identify the target.	2 – 45	25
Set Size	The number of non-target items in the stimulus array.	1 – 42	25
Timing	The amount of time the stimulus array appeared on screen.	4.25 s – 5.21 s	2.10 s

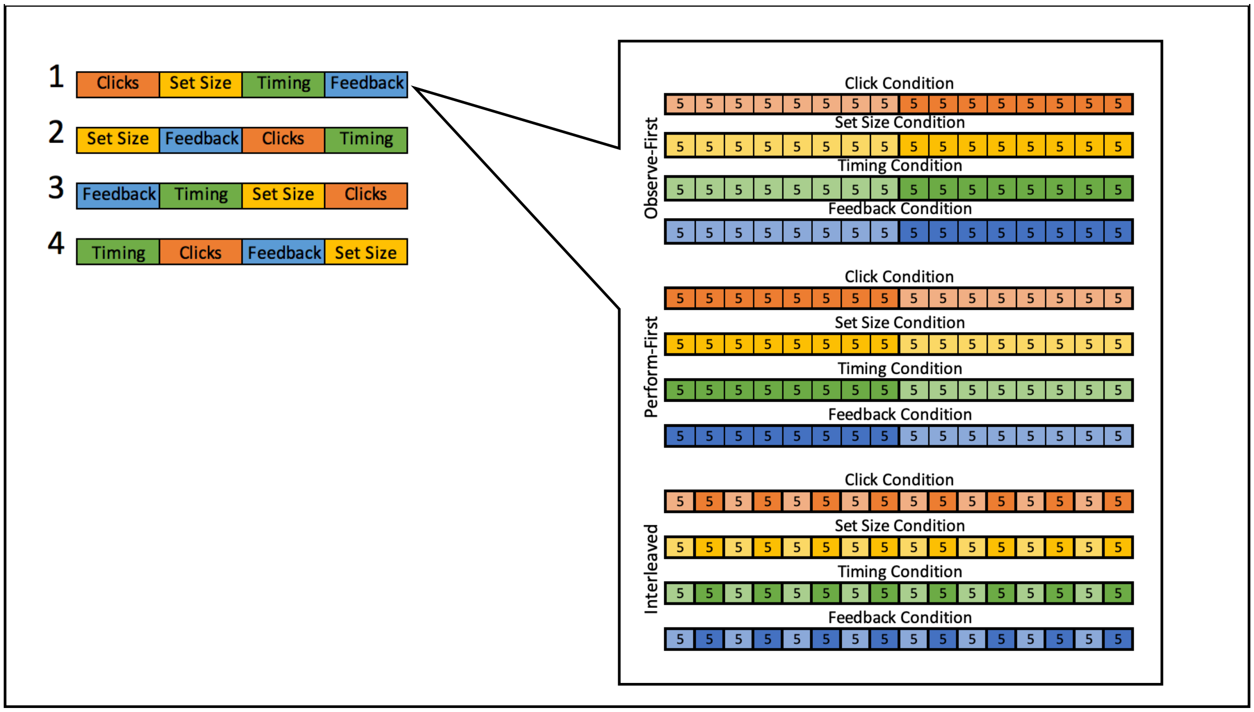
JODs. Participants' performance during the pilot study informed the frequency of JOD assessment. Although requesting a JOD after every trial reduced participants' accuracy by 7%, it provided greater precision with which to track changes in cue use that occur as people learn about a task. Thus, the program paused after participants received feedback on each trial to allow them to make a JOD. Participants indicated whether the task was easier or harder than before by clicking one of two buttons. Once participants selected an option, the buttons disappeared, and the task resumed.

Role manipulation. At the beginning of each session, participants were automatically assigned to one of three role conditions. Participants in the *perform-first* condition completed the first 40 trials of each difficulty block themselves and watched "a video of participants' average performance in the task" during the second 40 trials. Those in the *observe-first* condition watched the video during the first 40 trials and completed the second 40 trials of each block. Finally, participants assigned to the *interleaved condition* alternated between performing and observing every 5 trials.

Participants were instructed that the observer trials would feature a video that depicted other peoples' average task performance. In reality, these trials were programmed with a predictive modeling equation that mimicked participants' performance during the pilot study. Although participants did not see cursor movements during these trials, they experienced all other aspects of the task including post-trial feedback and 30s time-outs. To ensure that participants experienced the same levels of difficulty that they observed, the difficulty levels of the first 40 trials were saved and used to construct the trials in the second half of each block (see Box 3).

Demographic Information. At the end of the experimental session, participants indicated their sex and completed the 16-item Situational Motivation Scale (SIMS; Guay, Vallerand, & Blanchard, 2000), a survey developed to measure individuals' motivations for engaging in a task, including their level of intrinsic motivation.

Box 3. A schematic diagram of difficulty condition counterbalancing. Each rectangle at left represents 80 trials of each difficulty type, which was blocked so that consecutive changes occurred across a single dimension of difficulty. The between-subject role manipulation is depicted at right, where dark-colored squares represent observation trials and light-colored squares represent performance trials.



Results

Defining Variables

Participants' use of central cues to difficulty was quantified by *standardized difficulty*, a variable that equated difficulty across conditions by subtracting the lowest possible sampling value of a condition and dividing by the sampling range. This produced a variable where 0 represented the easiest level that participants could encounter in a given condition and 1 represented the hardest (trial time was reverse-coded). Participants' use of performance-based peripheral cues was quantified by *target identification*, a dichotomous variable that indicated whether participants correctly identified the target on a trial (1) or not (0).

Trial number was also included as a predictor to model the behavioral changes that occur as people gain experience with and learn from a task. Analyses from the pilot study indicated that performance tended to improve over the course of the experiment, while changes in JODs tended to occur within block. Log-transformed *experimental trial* and *trial in block* were used to model these changes in behavior; these variables are shortened for succinctness.

Role condition, a three-level categorical predictor, indicated the between-subject manipulation a participant received (observe first, perform first, or interleaved). Finally, *intrinsic motivation* was used to quantify participants' average score on the intrinsic motivation subscale of the SIMS. All variables were median-centered or effect coded prior to analysis. Performance data was subset to exclude observation trials, and JOD data was subset to exclude trials on which the difficulty of the task did not objectively change.

Performance Analysis

A multi-level logistic regression was used to determine the degree to which difficulty condition, standardized difficulty, experimental trial, and role condition predicted successful target identification. The model also included the Standardized Difficulty × Difficulty Condition interaction to assess task calibration. Intercept, standardized difficulty slope and experimental trial slope were allowed to vary across participants to model individual differences in skill and rate of learning. The results of this model are presented in Table 13.

Task difficulty (standardized difficulty) affected participants' performance in a meaningful way (see Figure 23). Participants were more likely to identify the target on easier trials than on harder trials. The strength of this relationship depended upon the difficulty condition. Despite steps taken to calibrate the task, performance intercepts and slopes differed across conditions. While this is disappointing, it is not critical: difficulty condition was manipulated within-subject to ensure that participants experienced similar tasks. A more thorough discussion of these effects is provided in Appendix B.

Table 13.
Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

parameter	B	*SE*	z	*p*
intercept	-0.10	0.15	-0.64	.52
standardized difficulty	-1.25	0.22	-5.79	< .001
clicks	-4.81	0.20	-24.45	< .001
feedback	-0.07	0.09	-0.79	.43
set size	0.23	0.09	2.61	.01
experimental trial	0.25	0.06	4.21	< .001
perform-first	-0.24	0.19	-1.25	.21
interleaved	0.10	0.21	0.48	.63
Standardized Difficulty × Clicks	-1.77	0.48	-3.70	< .001
Standardized Difficulty × Feedback	1.21	0.24	4.95	< .001
Standardized Difficulty × Set Size	-0.53	0.25	-2.14	.03

Note. Standardized difficulty (Mdn = 0.5) and experimental trial (Mdn = 180) were median-centered prior to analysis. The observe-first and timing conditions served as the {-1, -1} and {-1, -1, -1} baselines, respectively.

Importantly, participants displayed evidence of learning in that they improved as they gained experience with the task (see Figure 23). In addition, performance did not differ across role conditions (see Figure 24 and Table 14), a between-subject manipulation. This provides

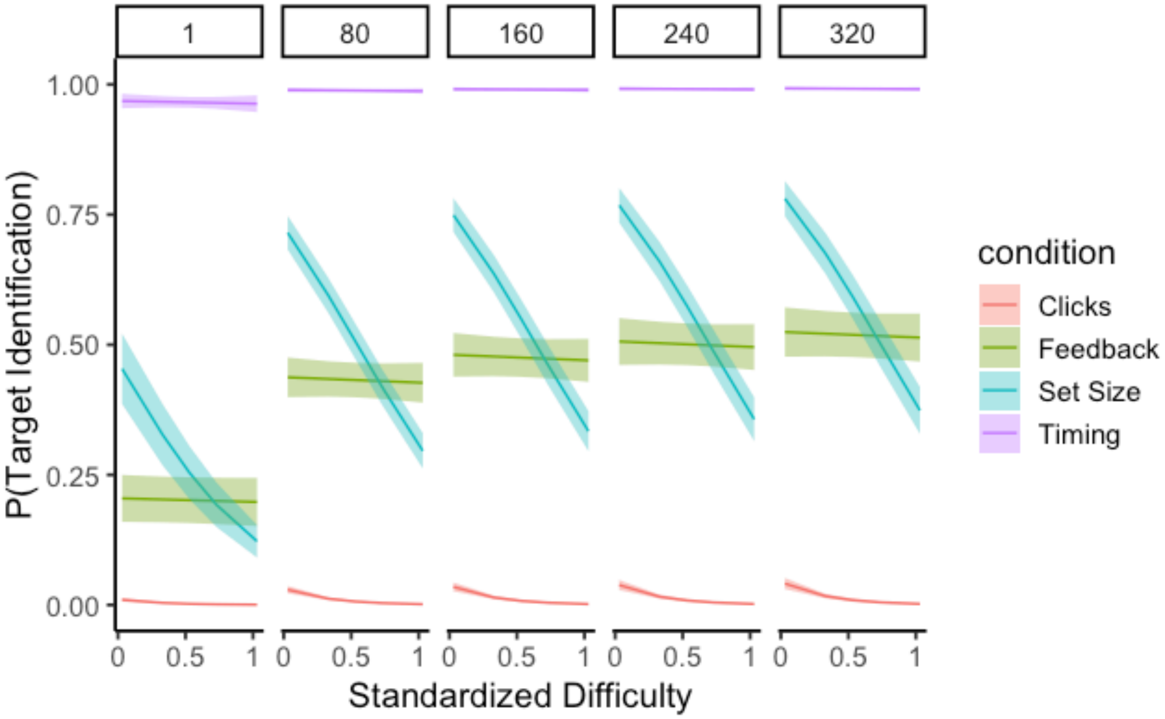


Figure 23. Standardized difficulty, condition, and experimental trial significantly predicted visual search performance. Error ribbons represent ±1SE.

assurance that any differences in JODs that emerge across role conditions are not due to task difficulty or performance.

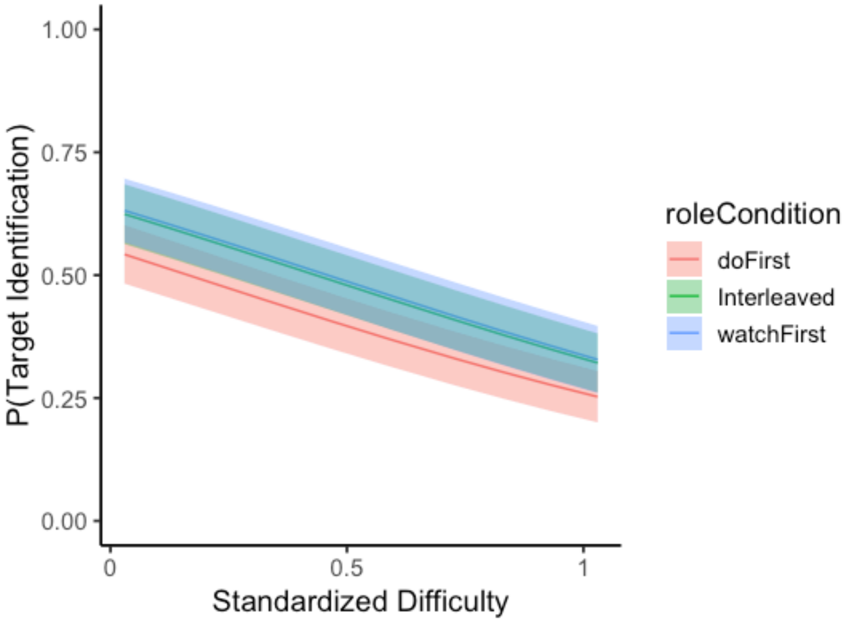


Figure 24. Role condition did not significantly predict visual search performance. Error ribbons represent ±1SE.

Table 14**.**
Performance estimates across role conditions.

Condition	B	*SE*	95%$_{CI}$
interleaved	-0.14	0.25	[-0.62, 0.34]
observe first	-0.11	0.28	[-0.65, 0.43]
perform first	-0.48	0.24	[-0.94, -0.02]

Predicting JODs

A multi-level logistic regression was used to determine the degree to which central (standardized difficulty) and performance-based peripheral (target identification) cues to difficulty informed participants' JODs. The model also included the Target Identification × Role Condition × Trial in Block interaction to test the competing hypotheses regarding cue use and role. Intercept, standardized difficulty slope, and target identification slope were allowed to vary across participants to model individual differences in cue use. The results of this model are presented in Table 15.

Table 15.
Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would say the task was harder than before.

parameter	B	*SE*	*z*	*p*
intercept	-0.23	0.09	-2.65	.01
standardized difficulty	0.65	0.14	4.53	< .001
identified target	1.24	0.09	13.85	< .001
trial in block	-0.01	0.04	-0.14	.89
do first	0.21	0.12	1.73	.08
interleaved	-0.16	0.12	-1.29	.20
Identified Target × Trial in Block	0.004	0.04	0.10	.92
Identified Target × Do First	0.04	0.12	0.30	.77
Identified Target × Interleaved	0.26	0.12	2.11	.03
Trial in Block × Do First	-0.13	0.06	-2.21	.03
Trial in Block × Interleaved	-0.01	0.06	-0.23	.82
Identified Target × Trial in Block × Do First	-0.10	0.06	-1.66	.10
Identified Target × Trial in Block × Interleaved	0.001	0.06	0.03	.98

Note. Standardized difficulty (Mdn = 0.5) and trial in block (Mdn = 37) were median-centered prior to analysis. Missed target and the observe-first role condition served as the {-1} and {-1, -1} baselines, respectively.

Both central and performance-based peripheral cues informed participants' JODs (see Figure 25). Although participants were more likely to say the task was harder than before when the task became objectively challenging, the strongest contributing factor to participants' JODs was their performance. Participants were much more likely to indicate that the task was harder than before when they received feedback that they had missed the target. This contrasted a previous finding in which central cues to difficulty contributed most strongly to JODs (Vangsness & Young, in press). This study involved a very different task (a videogame), and it is possible that task characteristics affect when and how cues to difficulty inform JODs.

Planned contrasts indicated that participants' use of performance-based peripheral cues depended on their role condition and trial in block. Participants' JODs did not differ on average across role conditions (see Table 16), nor did differences in cue use emerge when participants correctly identified the target (see Table 17). Condition and trial differences only emerged when

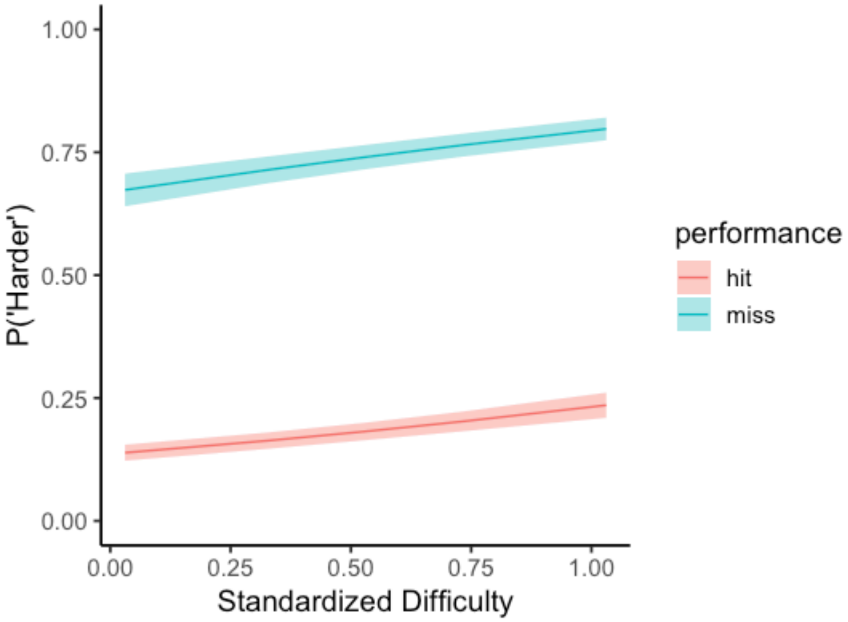


*Figure 25. C*entral (standardized difficulty) and performance-based peripheral cues significantly predicted participants' JODs. Error ribbons represent ±1SE.

participants received feedback that they did not successfully identify the target, as illustrated by Figure 26. At the beginning of a difficulty block, participants differed in their willingness to endorse the task as being harder than before. Those in the perform-first condition were 89% likely to make this estimate. Similarly, those in the interleaved condition were 77% likely to do so. In contrast, participants in the watch-first condition were only 46% likely to indicate that the task was harder than before. This suggests that performance-based feedback most strongly informs the JODs of those performing a task, as compared to those observing. This difference gradually attenuated as participants gained experience and started performing the task – by the end of each block, participants in the interleaved, observe-first, and perform-first conditions used peripheral cues to a similar degree. Furthermore, the standard error decreased over time, regardless of condition (see Table 18). Together, these comparisons provide direct support for the simulation hypothesis (H1) and refute the metacognition-as-mindreading (H2) and mnemonic (H3) hypotheses.

Table 16.
Planned comparisons of role condition intercept, as a function of target identification.

Condition	B	*SE*	*z*	*p*
Missed Target				
do first – interleaved	0.20	0.30	0.67	.78
do first – watch first	0.72	0.33	2.20	.07
interleaved – watch first	0.52	0.32	1.62	.24
Hit Target				
do first – interleaved	0.59	0.28	2.14	.08
do first – watch first	-0.05	0.30	-0.17	.98
interleaved – watch first	-0.64	0.30	-2.14	.08

Table 17.
Slope estimates for the critical Role Condition × Trial in Block × Performance interaction.

Condition	B	*SE*	95%$_{CI}$
Missed Target			
interleaved	-0.01	0.10	[-0.19, 0.16]
observe first	0.23	0.09	[0.04, 0.42]
perform first	-0.22	0.10	[-0.41, -0.03]
Hit Target			
interleaved	-0.02	0.10	[-0.22, 0.18]
observe first	-0.03	0.10	[-0.17, 0.24]
perform first	-0.04	0.10	[-0.24, 0.16]

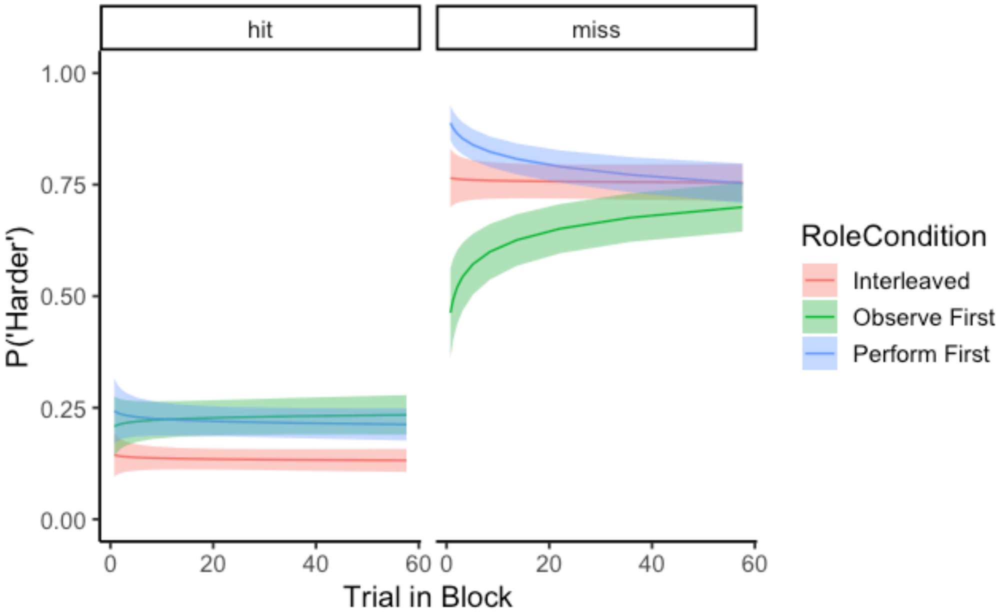


Figure 26. Role condition significantly changed participants' use of performance-based peripheral cues. Error ribbons represent ±1SE.

Table 18.
Planned comparisons of each role condition's trial in block slopes.

Condition	B	*SE*	*z*	*p*
Missed Target, trial = 1				
Perform First – Interleaved	0.90	0.53	1.69	.21
Perform First – Watch First	2.23	0.56	3.96	< .001
Interleaved – Watch First	1.33	0.54	2.45	.04
Missed Target, trial = 40				
Perform First – Interleaved	0.13	0.30	0.42	.91
Perform First – Watch First	0.56	0.33	1.71	.20
Interleaved – Watch First	0.44	0.32	1.35	.37
Missed Target, trial = 80				
Perform First – Interleaved	0.04	0.32	0.13	.99
Perform First – Watch First	0.38	0.34	1.10	.51
Interleaved – Watch First	0.34	0.34	1.00	.58

Discussion

This experiment demonstrated the conditions under which performance-based peripheral cues inform peoples' JODs. While performance-based peripheral cues always informed JODs to some extent, the strength of this relationship differed as a function of role. Participants weighed performance-based peripheral cues more strongly when they were performing a task, whether that was at the beginning (perform-first condition) or the end (observe-first condition) of a block. When participants alternated between performing and observing the task, they did not change the weight they assigned to performance-based peripheral cues. Furthermore, the variability of participants' JODs decreased over time. Together, these results discriminate between the competing hypotheses raised by the metacognitive literature and have important implications for learning environments.

These results supported the simulation hypothesis, which asserts that people base their metacognitive judgments on their imagined performance of a task. This suggests that people have trouble evaluating the difficulty of novel tasks because they cannot estimate their performance; not because other people "make it look easy." To this end, participants did not weigh peripheral cues to difficulty as strongly when they were observing a task as they did while performing it. These results directly contradict the metacognition-as-mindreading hypothesis (H2), which predicts that cues should be used equally regardless of role. In addition, participants' JODs converged as they gained experience of a task, regardless of condition. This decrease in variability could be occurring at the group (role condition) or subject level. That is, participants' JODs may converge due to increased agreement between-subjects or to decreased variability at the individual subject level. Sample size limitations prevent the inclusion of trial in block interaction terms in the random effect structure. As such, the mnemonic hypothesis (H3) is only weakly refuted.

Recently, researchers demonstrated that people underestimate task difficulty when they observe someone else perform a task (Kardas & O'Brien, 2018); however, these experiments did not identify the mechanism that underlies peoples' misestimation. The results from Experiment 1 clearly attribute misestimation to differences in the weighting of performance-based peripheral cues and illustrate that peoples' bias is reduced over time, even as they continue to observe a task. This finding furthers the basic understanding of metacognitive processes and highlights several situations in which learners may be especially at risk. Consider a person driving their car in a snowstorm. Performance-based peripheral cues, such as abandoned vehicles, that provide valid cues to difficulty may be ignored in such a context. Future research should investigate the relationship between these cues and peoples' risk mitigation strategies.

Future research projects can also explore situations in which others' performance can be observed as it occurs. Teachers, for example, have access to performance-based peripheral cues (students' grades) as well as to learning behaviors that unfold in the classroom (i.e., performance). While this experiment focused narrowly on the former, it is possible to adapt its structure to determine how proxy cues to difficulty, such as the fluidity of movements, contribute to JODs. A simple example might be to yolk observers to a performer and allow them to observe their on-screen cursor movements. These cue-rich situations remain an important avenue for future study.

Chapter 4 - Experiment 2

Experiment 1 determined that performance-based peripheral cues are both experienced and estimated, depending on whether a person performs or observes a task. However, Experiment 1 did not determine how the information obtained through performance-based peripheral cues informs effort allocation strategies and JODs. Consider a student who must study for an upcoming exam. Past performance is a strong indicator of future performance, and the student may use performance-based peripheral cues (i.e., feedback) to inform their study strategies. One approach might be to invest extra effort in the material they've gotten wrong on previous assessments, as they are most likely to answer these questions incorrectly on the exam. Alternatively, they might cut their losses and invest less effort in the material they frequently get wrong. This would allow them to focus their resources on strengthening their understanding of the material they are most likely to answer correctly. A third approach would be to integrate performance-based peripheral cues with information about the incentive structure of the environment and to spend the most time studying unfamiliar material that has been assigned a high point value. Experiment 2 sought to disambiguate these competing hypotheses regarding peoples' effort allocation strategies.

The Diminishing Criterion Model (DCM; (R. Ackerman, 2014) proposes that people compare past to desired performance and adjust their effort allocation strategies to align with their goals. If a person's past performance aligns with their goals, they will not invest additional effort in a task; however, if they perform more poorly than expected, they will invest additional effort to meet their goal. In this way, the DCM predicts that people will invest the most effort in challenging tasks – those for which they will perform the poorest. Thus, the DCM predicts that future performance will be informed by performance-based peripheral cues (performance feedback) and central cues (task difficulty), but little else.

The DCM can be contrasted with the Proximal Zone of Learning Theory, which predicts that people will allocate more effort towards tasks that lie within the level of their skills and abilities (Dunlosky et al., 2003; Metcalfe, 2009). This theory suggests that people compare the difficulty of a task to their current skills and abilities, and invest effort only in those tasks that they believe can be achieved. When a person's skills and abilities are accurately estimated, this strategy prioritizes resource efficiency by allocating effort towards tasks that are achievable and avoiding those that are not. In this way, the Proximal Zone of Learning Theory predicts that people will invest the most effort in easier tasks – those that are most easily accomplished, given their skills and abilities. Thus, future performance will be informed by performance-based peripheral cues (performance feedback) and central cues (task difficulty), but little else.

By contrast, moderation models of effort allocation such as Agenda-Based Regulation (Ariel et al., 2009) and the Strategic Task Overload Model (Wickens, Gutzwiller, et al., 2016) suggest that people will allocate greater effort towards tasks that seem easier to complete, but that this preference is affected by the incentive structure of the environment: people will also allocate greater effort toward more challenging tasks that are strongly incentivized. In other words, peripheral cues, central cues, and incentives will affect the amount of effort that is allocated towards a task, and performance will change as a function of incentive value.

In the context of this task, the three competing hypotheses are:

H1 (*Diminishing Criterion Model*): future performance (i.e., effort allocation) should be sensitive to past performance-based peripheral and central cues.

Specifically, performance will improve after participants receive negative feedback on the previous trial. This effect will be stronger for harder trials than for easier ones.

H2 (*Proximal Zone of Learning Theory*): future performance (i.e., effort allocation) should be sensitive to past performance-based peripheral and central cues.

Specifically, performance will improve after participants receive negative feedback on the previous trial. This effect will be stronger for easier trials than for harder ones.

H3 (*Agenda-based Regulation Theory*): future performance (i.e., effort allocation) should be sensitive to past performance-based peripheral and central cues. This relationship will be moderated by incentive value such that performance will improve the most for easy tasks assigned the highest incentive value.

Method

Participants

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Kansas State University. The experimental task was completed by 53 participants (36 females) between the ages of 18-65, who received $10 compensation. Two participants were omitted for failing to comply with instructions.

The Conjunctive Visual Search Task

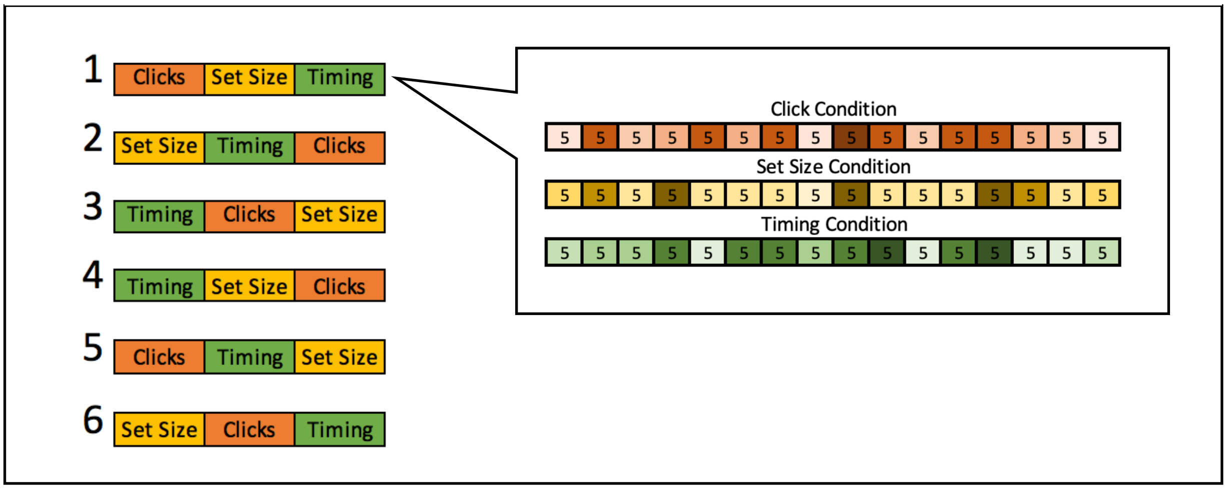
Participants completed a visual search task that was similar to that of the pilot study. The task differed in several respects. First, task difficulty was concurrently manipulated along two dimensions. The primary dimension (clicks, set size, or timing) involved central cues that affected performance in the pilot study, while the secondary dimension (feedback) changed the incentive structure of the environment. As before, these cues to difficulty were manipulated every five trials, and the remaining task dimensions were fixed at the same level participants encountered during the pilot study to ensure that performance estimates were accurate (see Table 6).

This task also differed in when information about the incentive structure became available: half of the participants learned the consequences for failing to identify the target in advance of each five-trial block; the other half only received information about each block's incentive value through the feedback provided after each trial. Finally, participants made JODs after every trial, and an additional demographic measure was added for exploratory purposes. Details are provided below; readers interested in learning more about specific task parameters are directed to the pilot study.

Task difficulty. In contrast to the pilot study, the difficulty level of the task changed both along a primary (clicks, set size, timing) and a secondary (feedback) dimension. The ranges of difficulty encountered in each dimension were informed by participants' performance in the pilot study. Specifically, a multi-level modeling approach generated estimates of the performance intercept and standardized difficulty slope for all three conditions; adjustments were made to equate performance across dimensions (see Appendix A and Table 6). Changes in task difficulty occurred every five trials, and the remaining task dimensions were fixed at the same level participants encountered during the pilot study to ensure that performance estimates were accurate. These changes are illustrated by the schematic diagram in Box 4.

Feedback timing manipulation. At the beginning of each session, participants were automatically assigned to one of two feedback conditions. Participants in the pre-trial condition learned of the consequences for failing to identify the target at the start of each trial and again when they received performance-based feedback. Participants in the post-trial condition only learned of the block's incentive value when they failed to identify the target.

Box 4. A schematic diagram of difficulty type counterbalancing. Each rectangle at left represents 106 trials of each difficulty type, which will be blocked so that consecutive changes occur across a single dimension of difficulty. Concurrent changes in incentive value, the number of points lost for failing to identify the target, are depicted at right. Light-colored squared represent fewer points than do dark-colored squares.



JODs. Participants' performance during the pilot study informed the frequency of JOD assessment. Although requesting a JOD after every trial reduced participants' accuracy by 7%, it provided greater precision with which to track changes in cue use that occur as people learn about a task. Thus, the program paused after participants received feedback on each trial Participants indicated whether the task was easier or harder than before by clicking one of two buttons. Once participants selected an option, the buttons disappeared, and the task resumed.

Demographic Information. At the end of the experimental session, participants indicated their sex and age. They also completed the 16-item Situational Motivation Scale (SIMS; Guay, Vallerand, & Blanchard, 2000), a survey developed to measure individuals' motivations for engaging in a task, including their level of intrinsic motivation.

Results

Defining Variables

Participants' use of central cues to difficulty was quantified by *standardized difficulty*, a variable that equated difficulty across conditions by subtracting the lowest possible sampling value of a condition and dividing by the sampling range. This produced a variable where 0 represented the easiest level that participants could encounter in a given condition and 1 represented the hardest (trial time was reverse-coded). Participants' use of performance-based peripheral cues was quantified by *target identification*, a dichotomous variable that indicated whether participants correctly identified the target on a trial (1) or not (0). Finally, *time-lagged target identification* (*TTI*) was used to determine the degree to which performance on the previous trial informed effort allocation on the subsequent trial.

Trial number was also included as a predictor to model the behavioral changes that occur as people gain experience with and learn from a task. Analyses from the pilot study indicated that performance tended to improve over the course of the experiment, leading to changes in JODs. Log-transformed *experimental trial* and *block trial* were used to model these changes in behavior; the variables are shortened for succinctness.

Feedback condition, a two-level categorical predictor, indicated the between-subject manipulation a participant received (pre-trial or post-trial feedback), and *incentive value* indicated the number of points that were at stake on a given trial. Finally, *intrinsic motivation* was used to quantify participants' average score on the intrinsic motivation subscale of the SIMS. *Age* was included as a model covariate to control for differences in performance due to age-related decline. All variables were median-centered or effect coded prior to analysis. JOD data was subset to exclude trials on which the difficulty of the task did not objectively change.

Performance Analysis

An initial exploratory analysis confirmed that the calibration challenges encountered in Experiment 1 were also present in Experiment 2. Namely, participants' performance was only affected by changes in set size. Performance in the timing and clicks conditions was at ceiling and floor, respectively. While this is disappointing, it is not critical: difficulty condition was manipulated within-subject to ensure that participants experienced similar tasks. The consequences of this challenge are discussed briefly in the analysis section, while a more thorough discussion is provided in Appendix B.

A multi-level logistic regression was used to determine the degree to which standardized difficulty, TTI, feedback condition, incentive value, experimental trial, and age predicted successful target identification. The model also contained the Incentive Value × Feedback Condition × Standardized Difficulty and TTI × Standardized Difficulty interactions to test the critical hypotheses. Lower-order interactions were also included. Intercept, standardized difficulty slope and experimental trial slope were allowed to vary across participants to model individual differences in skill and rate of learning. The results of this model are presented in Table 19.

Table 19.
Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

Parameter	B	*SE*	*z*	*p*
intercept	0.17	0.05	3.67	< .001
standardized difficulty	-0.50	0.08	-6.58	< .001
missed TTI	-1.31	0.02	-55.91	< .001
incentive value	-0.001	0.002	-0.55	.58
post-trial feedback	0.05	0.04	1.20	.23
trial in experiment	0.25	0.29	0.86	.39
age	-0.01	0.003	-4.51	< .001
Standardized Difficulty × Missed TTI	-0.29	0.07	-4.01	< .001
Standardized Difficulty × Incentive Value	-0.003	0.01	-0.53	.60
Standardized Difficulty × Post-trial Feedback	-0.10	0.07	-1.33	.18
Incentive Value × Post-trial Feedback	-0.003	0.002	-1.97	.05
Standardized Difficulty × Incentive Value × Post-trial Feedback	0.01	0.01	2.63	.01

Note. Standardized difficulty (Mdn = 0.5), trial in experiment (Mdn = 160), and incentive value (Mdn = 24) were median-centered prior to analysis. Correct TTI and pre-trial feedback served as the {-1} and {-1} baselines, respectively.

As before, participants' performance was affected by task difficulty (standardized difficulty) such that they performed more poorly on trials that were objectively harder. This was particularly true when they had failed to identify the target on the previous trial. In this way, past performance was a strong predictor of future performance (see Figure 27). As anticipated, performance deteriorated as a function of age (see Figure 28). Surprisingly, there was insufficient evidence to suggest that participants' performance improved over time (see Figure 29), perhaps due to the floor effect present in the clicks condition.

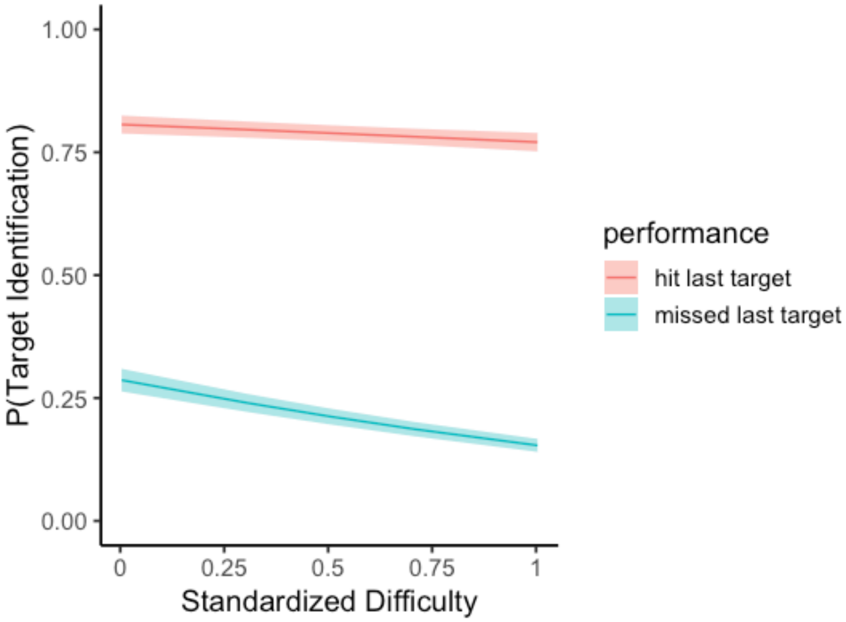


Figure 27. Central (standardized difficulty) and peripheral (past performance) cues significantly predicted participants' performance. Error ribbons represent ±1SE.

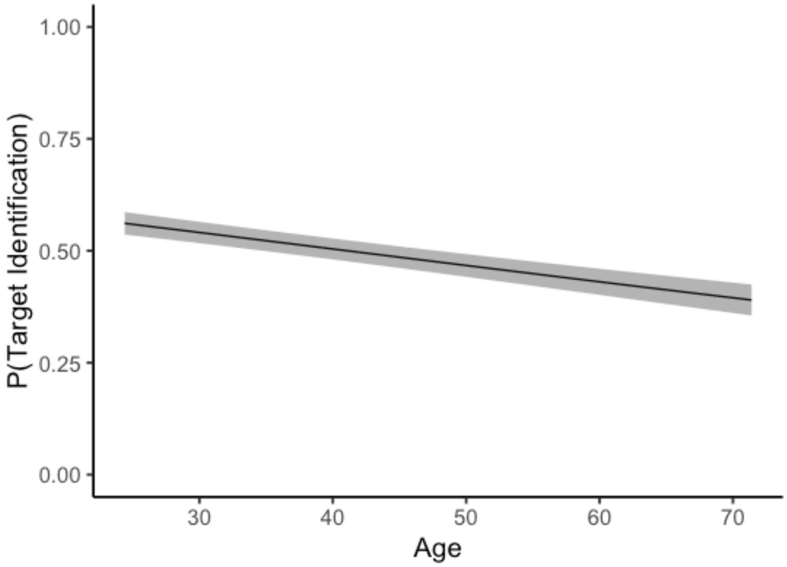


Figure 28. Older participants were less likely to identify the target on a given trial. Error ribbons represent ±1SE.

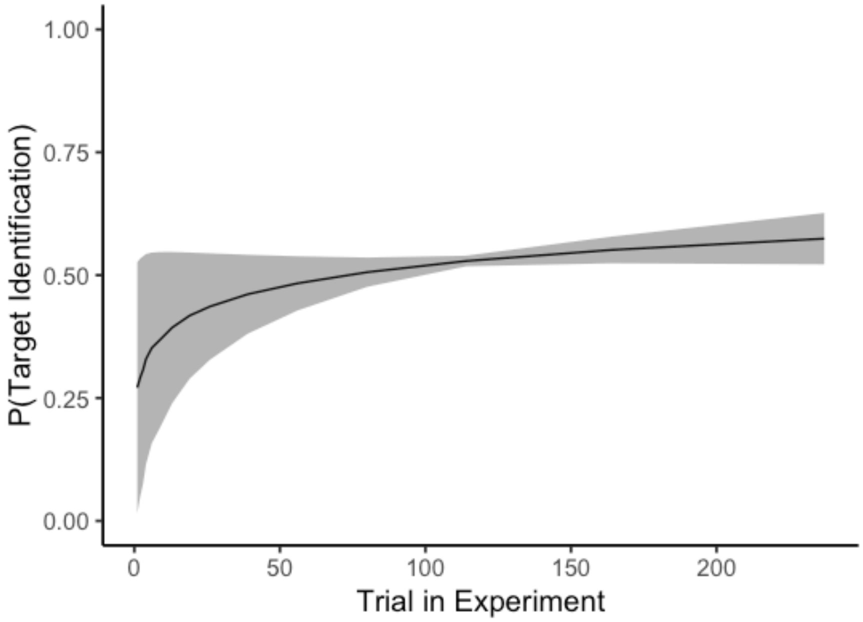


Figure 29. Participants did not become better at identifying the target over time. Error ribbons represent ±1SE.

Planned contrasts indicated that participants' performance depended on the difficulty level of the task and their knowledge of the incentive structure (see Table 20). Participants who knew of the incentives in advance (pre-trial feedback condition) incorporated this knowledge into their effort allocation strategy. When incentives were low, pre-trial participants performed more poorly than did those in the post-trial condition, regardless the difficulty level of the task. As incentives increased, pre-trial participants allocated more effort towards easy trials and less effort towards harder trials. By contrast, post-trial participants' performance was sensitive to task difficulty when incentives were low. As incentives increased, they became less likely to identify the target regardless the difficulty level of the task. This effect is illustrated by the changes in standardized difficulty slope that occur across panels in Figure 30. Slope estimates are available in Table 21. Together, these results provide support for agenda-based regulation (H3) and directly refute the DCM (H1) and Proximal Zone of Learning (H2) theories.

Table 20.
Planned comparisons of each feedback condition's standardized difficulty slopes at three points along the critical feedback magnitude continuum.

Condition	B	*SE*	*z*	*p*
Feedback magnitude = 1				
Post-trial – Pre-trial	-0.80	0.28	-2.85	.004
Feedback magnitude = 24				
Post-trial – Pre-trial	-0.20	0.15	-1.33	.18
Feedback magnitude = 48				
Post-trial – Pre-trial	0.46	.28	1.64	.10

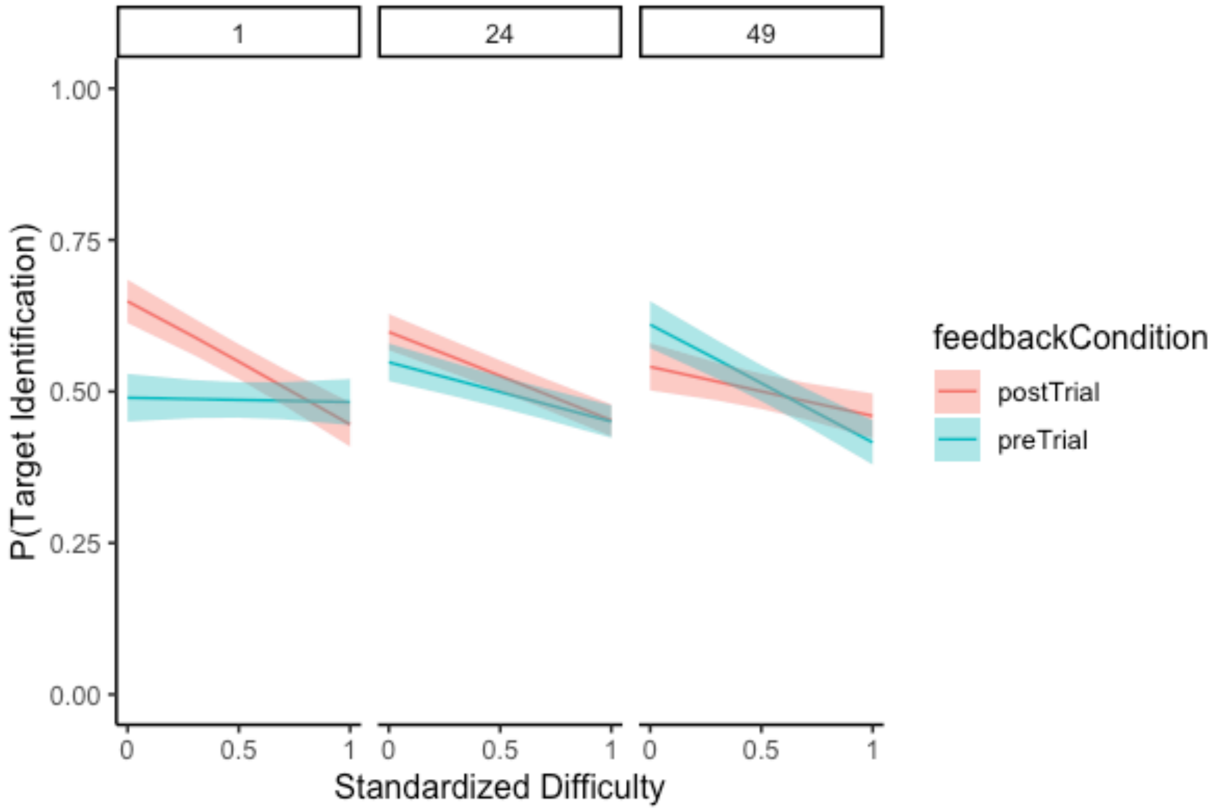


Figure 30. Effort allocation strategy was affected by central (standardized difficulty) cues and feedback magnitude. The direction of this effect depended on when participants received information about feedback magnitude. Feedback magnitude is depicted across panels; error ribbons represent ±1SE.

Table 21.
Slope estimates at three points along the critical Feedback Condition × Standardized Difficulty × Feedback Magnitude interaction.

Condition	B	*SE*	95%$_{\mathrm{CI}}$
Feedback magnitude = 1			
pre-trial	-0.04	0.20	[-0.43, 0.36]
post-trial	-0.84	0.20	[-1.23, -0.46]
Feedback magnitude = 24			
pre-trial	-0.40	0.11	[-0.61, -0.19]
post-trial	-0.60	0.11	[-0.81, -0.39]
Feedback magnitude = 48			
pre-trial	-0.80	0.21	[-1.20, -0.40]
post-trial	-0.34	0.20	[-0.72, 0.05]

Predicting JODs

A multi-level logistic regression was used to determine the degree to which central (standardized difficulty) and performance-based peripheral (target identification) cues to difficulty informed participants' JODs. The model also included the Standardized Difficulty × Difficulty Condition interaction to control for calibration issues, and the Target Identification × Role Condition × Trial in Block interaction to test competing hypotheses regarding cue use and role. Intercept, standardized difficulty slope, and target identification slope were allowed to vary across participants to model individual differences in cue use. The results of this model are presented in Table 22.

Both central and peripheral cues informed participants' JODs (see Figure 31). Participants were more likely to say the task was harder than before when the task became objectively challenging and when the incentives were higher. Still, the strongest contributing factor to participants' JODs was their performance. Participants were much more likely to indicate that the task was harder than before when they received feedback that they had missed the target. This aligned with the results of Experiment 1.

Table 22.
Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would say the task was harder than before.

parameter	B	*SE*	*z*	*p*
intercept	-0.03	0.12	-0.28	.78
standardized difficulty	0.59	0.16	3.72	< .001
incentive value	0.01	0.003	3.11	.002
missed target	1.66	0.13	13.26	< .001
post-trial feedback	0.14	0.11	1.29	.20
clicks condition	-0.52	0.10	-5.22	< .001
set size condition	0.17	0.07	2.52	.01
trial in block	-0.04	0.05	-0.83	.41
age	0.01	0.01	0.83	.40
Standardized Difficulty × Incentive Value	-0.01	0.01	-1.00	.32
Incentive Value × Missed Target	-0.0004	0.003	-0.14	.89
Missed Target × Post-trial Feedback	-0.08	0.10	-0.81	.42
Incentive Value × Post-trial Feedback	0.001	0.003	0.24	.81
Standardized Difficulty × Clicks Condition	-0.75	0.19	-4.06	< .001
Standardized Difficulty × Set Size Condition	1.17	0.23	5.12	< .001
Incentive Value × Missed Target × Post-trial Feedback	0.004	0.003	1.31	.19

Note. Standardized difficulty (Mdn = 0.5), trial in block (Mdn = 53), and incentive value (Mdn = 24) were median-centered prior to analysis. Identified target, pre-trial feedback, and the timing condition served as the {-1}, {-1}, and {-1, -1} baselines, respectively.

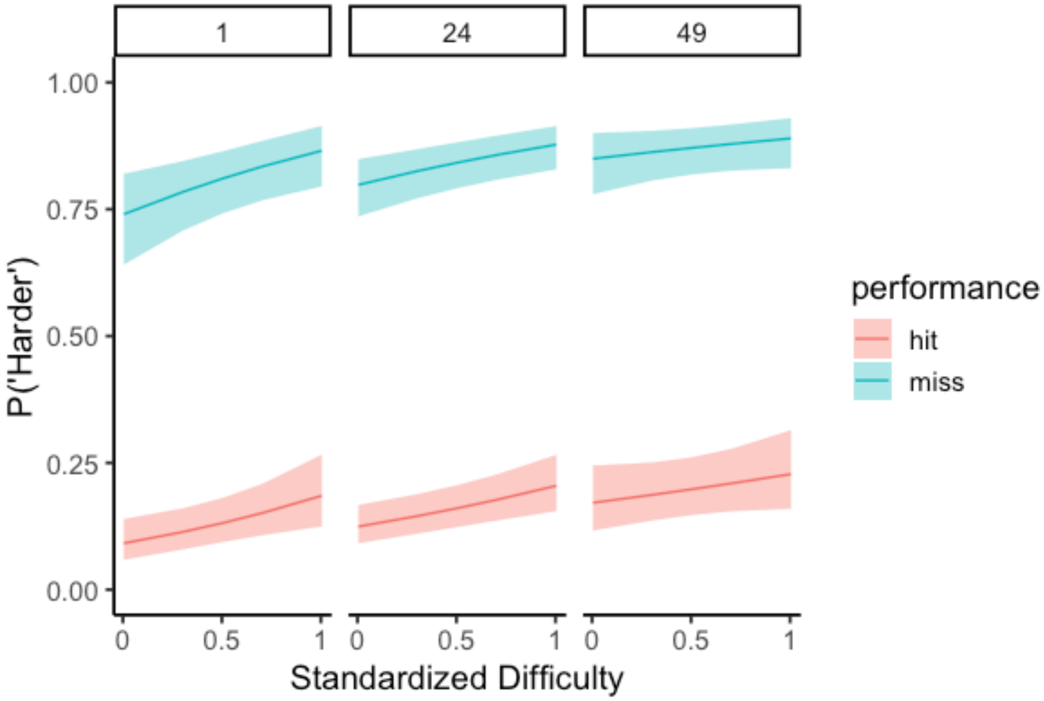


Figure 31. JODs were informed by central (standardized difficulty) and peripheral (performance, feedback magnitude) cues to difficulty. Feedback magnitude is depicted across panels; error ribbons represent ±1SE.

Due to the calibration of the task, some difficulty conditions were objectively more challenging than others. Model estimates indicate that participants recognized these differences, as illustrated by the intercept and slope differences present in Figure 32. A Tukey's HSD confirmed that the differences across difficulty condition were significant (see Tables 23-24). Feedback condition did not affect participants' JODs, $B_{\text{difference}} = 0.27$, $SE = 0.21$, $z = 1.29$, $p = .20$, nor did feedback condition affect participants' sensitivity to changes in task difficulty or incentive value (see Figure 33 and Table 25).

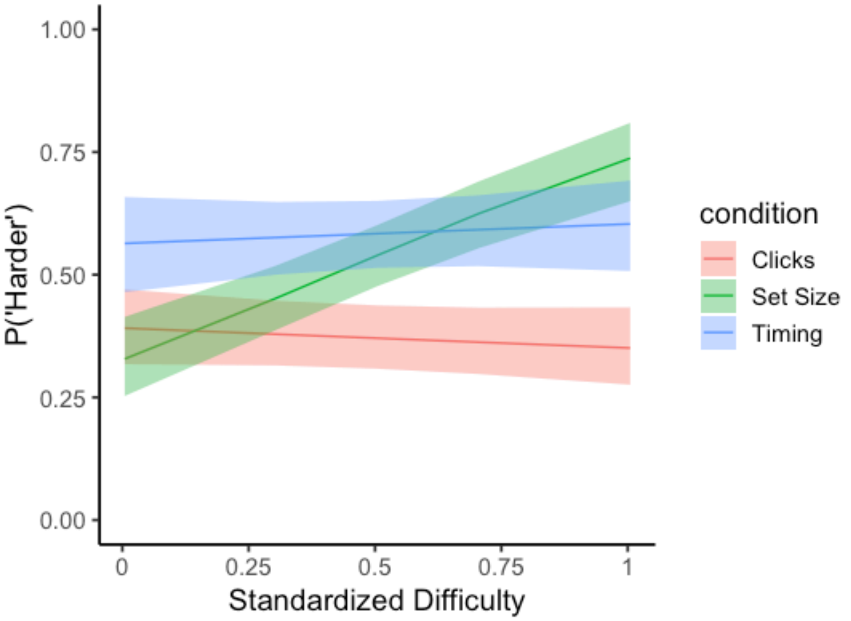


Figure 32. JODs differed across difficulty condition. Error ribbons represent ±1SE.

Table 23.
Comparisons of each difficulty condition's standardized difficulty slope.

Condition	B	*SE*	*z*	*p*
Clicks – Set Size	-1.92	0.35	-5.45	< .001
Clicks – Timing	-0.34	0.34	-1.00	.58
Set Size – Timing	1.59	0.41	3.89	< .001

Table 24.
Standardized difficulty slopes for each difficulty condition.

Condition	B	*SE*	$95\%_{CI}$
Clicks	-0.17	0.19	[-0.55, 0.20]
Set Size	1.75	0.31	[1.15, 2.35]
Timing	0.16	0.28	[-0.39, 0.71]

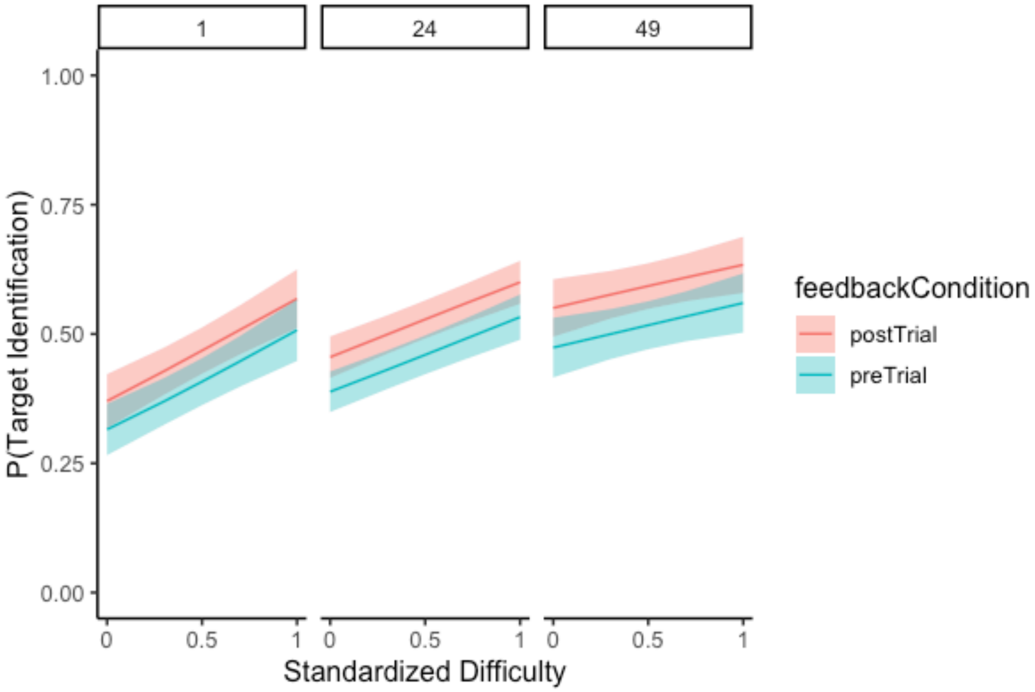


Figure 33. JODs were not affected by feedback condition, nor did the manipulation affect participants' sensitivity to changes in task difficulty (standardized difficulty) or incentive value. Incentive value is depicted across panels; error ribbons represent ±1SE.

Table 25.
Standardized difficulty and incentive value slopes for each feedback condition.

Condition	B	*SE*	$95\%_{CI}$
standardized difficulty slope			
pre-trial	0.58	0.16	[0.27, 0.89]
post-trial	0.58	0.16	[0.27, 0.89]
feedback magnitude slope			
pre-trial	0.01	0.004	[0.001, 0.02]
post-trial	0.01	0.005	[0.03e^{-3}, 0.02]

Discussion

This experiment determined how people use performance-based peripheral cues to inform effort allocation strategies and identified an important way in which this differed from how they make JODs. Effort allocation strategies and JODs were both informed by central and peripheral cues to difficulty, including past performance and feedback magnitude. However, effort allocation strategies differed as a function of when participants learned of the incentive structure of the environment. When participants learned of this information in advance, it informed their effort allocation strategies: participants allocated the most effort to easy tasks that were highly incentivized. When participants learned of this information through feedback, they allocated less effort towards easy tasks, especially those that were highly incentivized. These feedback condition differences did not emerge with respect to JODs, suggesting that participants' effort allocation strategies did not change because they perceived the tasks to be easier or harder. Rather, effort allocation seemed to be driven by the task's incentives. Together, these results disambiguate the results of Experiment 1 and discriminate between competing hypotheses regarding peoples' effort allocation.

These results supported moderation models of effort allocation, which propose that peoples' effort allocation decisions are informed by both task difficulty and incentives. Participants who knew of the incentive structure in advance performed better on easy trials that were highly incentivized. When these same easy trials were poorly incentivized, performance dropped and was similar to that of more difficult trials. These results directly contradict the predictions of the DCM (H1), which predicts that people moderate their effort to improve their performance on the most difficult tasks. The results also contradict the Proximal Zone of Learning Theory (H2), which predicts that people moderate their effort to improve performance on easier tasks that are within the level of their skills and abilities. While people were more likely to improve following poor performance on an easy trial, the degree to which they did was affected by the task's incentives.

Participants also used task incentives to inform their JODs, even though the two variables were decorrelated in this context. This illustrates how ecological adaptations (e.g., (Gigerenzer et al., 1999)) can lead people to make seemingly illogical judgments. Often, difficult tasks are more highly rewarded. Consider that teachers assign more points to term papers than to weekly quizzes, and that there is more prestige associated with climbing Mount Everest than with climbing the stairs to your office. The relationship between difficulty and reward is so pronounced that it defines some peoples' understanding of difficulty (difficulty-as-importance; (Fisher & Oyserman, 2017). Although the inverse is not always true, as in this case, it likely takes time for people to adapt in their judgments.

These findings further advance the basic understanding of effort allocation and illustrate how future performance can be affected by task difficulty and incentives. Consider a student who must decide whether to allocate study time to a 5-point paper or a 50-point exam. These results suggest that they are likely to study for the exam, but only if they believe that it lies within the level of their skills and abilities. If the exam seems insurmountably hard, the student may allocate just as much time towards writing their paper. This interplay between task difficulty and incentives can be used to structure environments in ways that advance learning. Future research should seek to confirm this relationship within a naturalistic context and with direct measures of effort allocation (e.g., EEG).

Chapter 5 - General Discussion

Performance-based peripheral cues to difficulty arise from many circumstances. For example, a student may use their semester quiz scores to determine how much time to invest in studying for the final exam. In a similar way, performance-based peripheral cues can inform behaviors in dynamic environments where feedback is frequent. For example, a pilot might use moment-by-moment data to make decisions about whether to make a route adjustment or to radio for assistance. In both cases, past performance is used as a predictor of future behavior and as a guide for behavioral change. To make educated decisions, these individuals must have an accurate understanding of the task's difficulty and how well they can perform under those circumstances.

There has been a great deal of effort invested in determining what makes task difficult, but considerably less attention to when or why tasks are perceived as being so. This dissertation addressed the latter by manipulating the conditions under which performance-based peripheral cues to difficulty were made available. In Experiment 1, participants received performance-based peripheral cues as they performed a task as well as while they observed someone else complete it. In Experiment 2, some participants received information that foreshadowed the magnitude of the performance-based peripheral cues, while others did not. Together, these experiments disambiguated competing hypotheses and clarified the relationships between task difficulty, resource allocation, and performance.

Performance-based Peripheral Cues are Down-weighted in Observational Settings

Frequently, observation is used as a tool to familiarize people with a task. For example, surgeons are trained by watching other people in the operating theater and football players often watch tape-recorded games before playing a new opponent. Common wisdom suggests that observation allows trainees to “learn through experience;” however, recent research has shown that people tend to under-estimate the difficulty of tasks they observe. This bias is evidenced across many contexts (Arkes et al., 1981; Hoelzl & Rustichini, 2005; Kardas & O’Brien, 2018; Kruger & Dunning, 1999; Kruger et al., 2008); however, its underlying cause is unclear. Identifying the factors that lead to observers’ under-estimation of task difficulty will improve the specifications of observational contexts and, in turn, the quality of training.

Experiment 1 used a role-based manipulation to determine how people incorporate performance-based peripheral cues into JODs. The results illustrated the importance of hands-on experience: people who engaged in early observation underestimated the difficulty of the task relative to their peers who received full or partial hands-on experience. Once observers received hands-on experience, their estimation dramatically improved and aligned with their peers. This change was driven by observers’ weighting of performance-based peripheral cues, which contributed less to their early JODs. This suggested that observers’ misestimation was caused by their inexperience with the task, rather than a willingness to treat another person’s performance as a stand-in for their own. In other words, difficult tasks seem easy because we’ve never tried them; not simply because other people make them look easy.

These results align with the simulation hypothesis (Dimaggio et al., 2008), which suggests that peoples’ JODs are based upon their imagined performance of a task. On the surface, this tendency to down-weight observed performance seems helpful: people are unlikely to use an expert’s performance as an indicator of their own. However, it also means that people are unlikely to learn from watching others and may assume that they “could do it better,” especially when they lack experience with a task. Providing hands-on experience seems to reduce the degree of peoples' mis-estimation relative to peers; however, additional research is necessary to understand the mechanism responsible for this change.

It is also important to note that alignment does not necessarily imply accuracy – participants may exhibit bias in their JODs. Previous research indicates that biased JODs are prevalent across many contexts (Burson et al., 2006) - for example, most people rate their driving skills as above average, relative to their peers (Marottoli & Richardson, 1998). Previous research has found that novices tend towards under-estimation of task difficulty relative to experts, a result that is partially explained by Experiment 1. However, a lack of experience with performance-based peripheral cues does not explain over-estimation by experts, implying that additional mechanisms may be at work. Future research should address this concern by determining the degree to which different cues influence peoples' discrimination and bias (Maniscalco & Lau, 2012).

Updating the Model of Task Difficulty and Performance

The specific circumstances under which performance-based peripheral cues informed JODs provides insight to the relationship between task difficulty and performance. Performance-based peripheral cues were down-weighted when people first observed a task. This led observers to underestimate task difficulty relative to those who had performed the task. However, their JODs came into alignment once they received hands-on experience. This suggests that the performance-based peripheral cues that arise through a person's experiences have enduring effects on their JODs, lending support to a model where performance-based peripheral cues inform self-efficacy beliefs (see Figure 34).

Disambiguating the role of performance-based peripheral cues

In addition to serving as a valid cue to difficulty, performance-based peripheral cues provide information about risks and rewards. Researchers have suggested that people use this information to plan their resource allocation strategies (Kurzban et al., 2013). For example, a student may allocate more study time to topics they've previously failed to accurately recall for past assignments, especially when those topics are likely to be worth many points on an upcoming exam. The score a student receives could inform their future resource allocation (e.g., study time) in two ways. Perhaps lost points lead students to doubt their self-efficacy and over-estimate the difficulty of the material. They, in turn, allocate more (R. Ackerman, 2014) or less (Metcalfe & Kornell, 2005) effort to studying it. Conversely, a student's self-efficacy beliefs could remain intact, and point-value information could be used to plan future study strategies, given how difficult the material is perceived to be (Ariel et al., 2009). Determining the nature of this relationship will reveal the ways in which incentives impact effort allocation behavior and facilitate the development of learning contexts that prompt task engagement.

Experiment 2 used a feedback manipulation to determine how information about the incentive structure of the environment informs JODs and effort allocation decisions. Some participants completed a task with clearly defined incentives that were presented at the start of each five-trial block. Other participants only learned of task incentives through performance-based peripheral cues (i.e., post-trial feedback). Although the participants within these groups made similar JODs, they differed in their performance. Participants who learned of incentives in advance used this information to develop a resource allocation strategy: these participants performed the best on easy trials that were strongly incentivized. In contrast, participants who only learned of incentives through performance-based peripheral cues exhibited the opposite pattern of behavior: they performed worse on easy trials that were strongly incentivized. Because JODs did not differ across the two groups, it is likely that incentive information moderates the relationship between JODs and effort allocation. However, the degree to which it does so appears to differ depending on how that information is received.

Incentive information that is provided at the start of a task does not change the likelihood of a negative outcome. However, when incentives are learned through performance-based peripheral cues, the riskiness of the situation has changed. Consider a student studying for an exam. A high-performing student may feel more motivated to review the material they've missed on previous exams because their grade is at stake. However, it is not worthwhile for a student with a D in the class to invest the same amount of effort if it is unlikely to impact their grade. Thus, both risks and rewards contribute to effort allocation decisions, supporting the hypotheses of agenda-based regulation (Ariel et al., 2009).

It is interesting to note that when performance is incentivized, resource allocation becomes a preventative risk mitigation strategy that can be used to avoid negative consequences. Previous research has shown that people are less likely to implement preventative strategies after experiencing losses unless those preventative strategies are made easier-to-use (Vangsness & Young, 2017). The results of Experiment 2 would suggest that people are also more likely to implement preventative strategies when they are aware of consequences in advance. Future research should confirm this relationship in contexts that involve other risk mitigation strategies.

Updating the Model of Task Difficulty and Performance

In Experiment 2, participants' performance-based peripheral cues informed JODs, replicating the results of Experiment 1. Specifically, participants' accuracy on the previous trial strongly informed their JODs. This result provides additional support for a relationship between performance-based peripheral cues and self-efficacy beliefs. Interestingly, the incentive value of a given trial also informed participants' JODs: trials on which participants stood to lose more points were judged to be harder than those on which participants lost few. This relationship was present among all participants, regardless of whether the incentive structure was known in advance or learned through feedback (i.e., embedded within performance-based peripheral cues). In addition, participants' knowledge of the incentive structure informed their JODs separately from their use of central cues to difficulty. This indicates that incentive structure uniquely moderates the relationship between self-efficacy beliefs and JODs.

Finally, participants' knowledge of the incentive structure affected their performance and changed their resource allocation strategies. Participants who knew the incentive structure in advance used it to plan their resource allocation; they performed strongly on easy trials that were highly incentivized and invested less effort on trials that were poorly incentivized or particularly difficult. Those who learned of the incentive structure through feedback employed the opposite strategy and tried to compensate for losses they'd already incurred: they performed well on easy trials that were poorly incentivized and invested less effort on trials that were highly incentivized or particularly difficult. These results indicate that knowledge of the incentive structure moderates the relationship between JODs and resource allocation. These results also indicate that this relationship is further moderated by central cues to difficulty (see Figure 34).

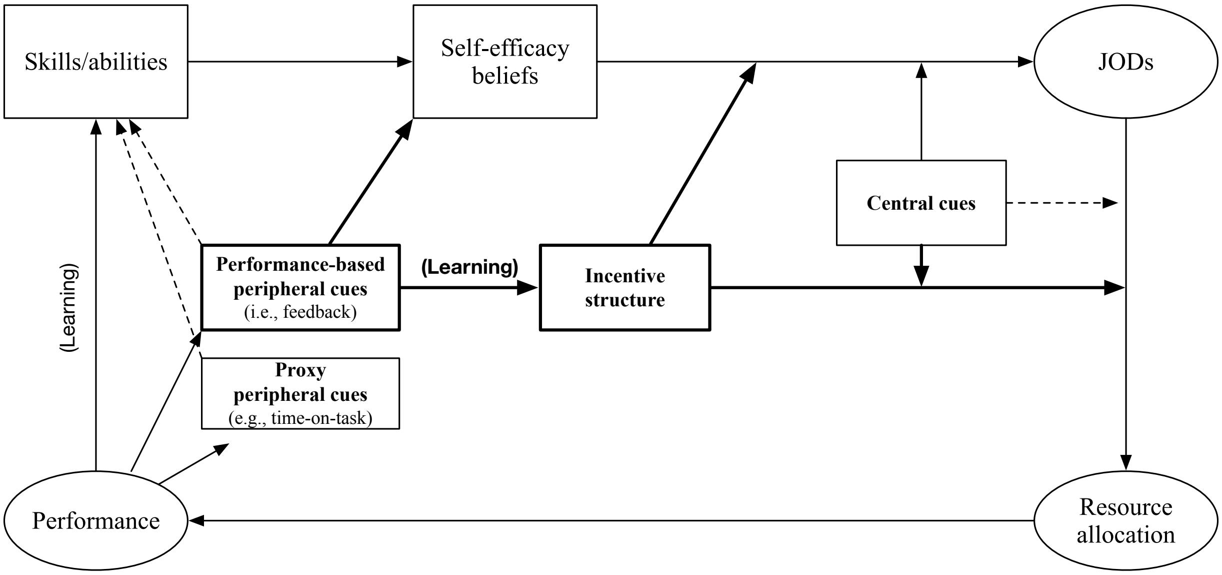


Figure 34. A revised model of the relationships between JODs, resource allocation, and performance.

Although this dissertation clarifies the relationships between performance-based peripheral cues, JODs, and resource allocation, important questions remain. For example, the relationship between the incentive structure of the environment and participants' performance was strongly affected by when this information was made available. It is possible that context (e.g., gain/loss framing (Tversky & Kahneman, 1991)) moderates this relationship and changes how people allocate their resources in response to incentives. It is also possible that separate mechanisms are responsible for peoples' resource allocation strategies in each context. This question could be answered through a future modification that manipulates how frequently the incentive structure changes. This manipulation changes how easily the incentive structure can be learned when this information is provided through feedback alone. If fundamentally different mechanisms are at work, participants in the post-trial condition should exhibit different behavioral patterns even when the incentive structure of the environment is constant (i.e., is easily learned).

Another important question involves the role of JODs in effort allocation and risk mitigation decisions. In a previous study, domain-specific experience (i.e., self-efficacy beliefs) affected participants' JODs. Risk mitigation decisions were informed by performance-based peripheral cues (Vangsness & Young, 2017). These findings provide further support for the relationships between performance-based peripheral cues and self-efficacy beliefs, as well as the moderational relationship between learned incentives and resource allocation. They also suggest that risk assessment arises from peoples' learned experiences with a task (see FIGURE 35), and that a person's risk tolerance may moderate their allocation of resources.

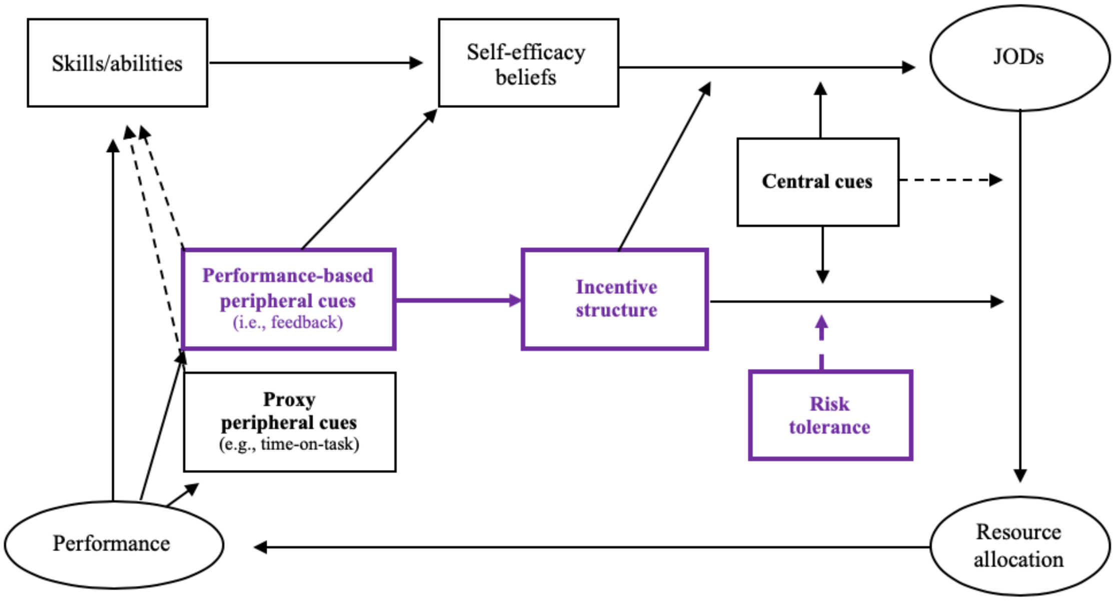


Figure 35. Performance-based peripheral cues provide information about a situations' risks; resource allocation may be moderated by a person's risk tolerance.

Implications and Broader Impacts

While adages occasionally undervalue the importance of hands-on experience[2], personalized feedback provides information about a person's skill level, as well as the incentives of the task. This information allows people to make accurate JODs and to invest their effort in ways that maximize reward and performance. When this information is lacking, people tend to overestimate their skills and underestimate the difficulty level of the task. Such judgments produce poor outcomes and deteriorate performance.

Although these experiments were conducted in a controlled environment, it is not difficult to generalize their results to an applied context. While writing this dissertation, I also taught indoor cycling classes at a local gym. One of these classes required riders to create a bike profile that automatically calculated the physical effort they should put forth to reach different levels of intensity. During class, the bike computer provided feedback that signaled whether riders were meeting the level of intensity that I was coaching. Early in the semester, I discovered that new riders frequently overestimated their fitness level and were unable to keep up with the group. Often, these riders did not return to class. In light of this, I began encouraging new riders to underestimate their fitness level and scale up as they felt comfortable. This allowed riders to gain experience with the class and their abilities. Although I still saw new riders, fewer dropped the class after making this change.

The challenges encountered in cycling studios, college classrooms, and other training environments differ from the visual search task in many ways. For example, in some contexts, learners who are feeling overwhelmed can select alternative courses of study rather than to

[2] e.g., Smart people learn from their own mistakes, while wise people learn from others.

withdraw from a task completely. It is unknown whether the relationships observed in this dissertation will generalize directly to applied domains. However, the model of task difficulty, resource allocation, and performance depicted in Figure 34 provides a framework to test the generalizability of these findings to other measures (e.g., EEG) and circumstances. Doing so will improve the validity of these findings and provide insights that can be used to improve training environments and to encourage task completion.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences*, *33*(4), 587–605. doi:10.1016/S0191-8869(01)00174-X

Ackerman, R. (2014). Metacognitive regulation of time investment. *Journal of Experimental Psychology. General*, *143*(3), 1349–1368.

Åhsberg, E. (2000). Dimensions of fatigue in different working populations. *Scandinavian Journal of Psychology*, *41*, 231–241.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. doi:10.1037/0022-3514.49.6.1621

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and Performance 15: Conscious and Nonconscious Information Processing* (pp. 421–452). Cambridge, MA: The MIT Press.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. *Journal of Experimental Psychology. General*, *138*(3), 432–447. doi:10.1037/a0015928

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *The Journal of Applied Psychology*, *66*, 252–254.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development*, *56*(1), 45–72. doi:10.1007/s11423-007-9067-0

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, *4*, 359–373.

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology*, *25*(5), 729–735. doi:10.1037/0012-1649.25.5.729

Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, *130*(4), 553–573. doi:10.1037/0033-2909.130.4.553

Bjork, E. L., & Bjork, R. A. (2014). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher & J. R. Pomerantz (Eds.), *Psychology and the Real World: Eassys Illustrating Fundamental Contributions to Society* (2nd ed., pp. 59–68). New York, NY: Worth.

Bjork, R. A. (1999). Assessing Our Own Competence: Heuristics and Illusions. In Daniel Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application* (pp. 435–459). Cambridge, MA: The MIT Press.

Black, P. E. (2005). Fisher-Yates shuffle. Retrieved February 21, 2019, from https://xlinux.nist.gov/dads/HTML/fisherYatesShuffle.html

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *9*(4), 453–465. doi:10.1037/1082-989X.9.4.453

Boksem, M. A. S., Meijman, T. F., & Lorist, M. M. (2005). Effects of mental fatigue on attention: an ERP study. *Brain Research. Cognitive Brain Research*, *25*(1), 107–116. doi:10.1016/j.cogbrainres.2005.04.011

Boksem, M. A. S., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain Research Reviews*, *59*(1), 125–139. doi:10.1016/j.brainresrev.2008.07.001

Borg, G. (1998). *Borg's Perceived Exertion and Pain Scales*. Champaign, IL: Human Kinetics.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, *14*, 540–552.

Bruya, B., & Tang, Y.-Y. (2018). Is attention really effort? revisiting daniel kahneman's influential 1973 book attention and effort. *Frontiers in Psychology*, *9*, 1133. doi:10.3389/fpsyg.2018.01133

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77. doi:10.1037/0022-3514.90.1.60

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245–281.

Cain, B. (2007). *A review of the mental workload literature* (No. RTO-TR-HFM-121-Part-II). Defence Research and Development Canada Toronto Human System Integration Section.

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident; Analysis and Prevention*, *71*, 311–318. doi:10.1016/j.aap.2014.06.005

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(2), 121–38; discussion 138. doi:10.1017/S0140525X09000545

Chan, L. K. H., & Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews. Cognitive Science*, *4*(4), 415–429. doi:10.1002/wcs.1235

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*, 147–168.

Cohen, G. L., Purdie-Vaughns, V., & Garcia, J. (2016). An identity threat perspective on intervention. In M. Inzlicht & T. Schmader (Eds.), *Stereotype Threat: Theory, Process, and Application* (pp. 280–296). New York, NY: Oxford University Press.

Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors*, *40*(4), 591–600. doi:10.1518/001872098779649283

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology*, *33*(4), 733–756. doi:10.1016/j.cedpsych.2008.02.003

Covington, M. V. (2000). Goal theory, motivation, and school achievement: an integrative review. *Annual Review of Psychology*, *51*, 171–200. doi:10.1146/annurev.psych.51.1.171

Curry, R., Jex, H., Levison, W., & Stassen, H. (1979). Final report of control engineering group. In N. Moray (Ed.), *Mental Workload* (pp. 235–252). New York, NY: Springer.

Deci, E. L., Ryan, R. M., & Williams, G. C. (1996). Need satisfaction and the self-regulation of learning. *Learning and Individual Differences*, *8*, 165–183.

Dimaggio, G., Lysaker, P. H., Carcione, A., Nicolò, G., & Semerari, A. (2008). Know yourself and you shall know the other... to a certain extent: multiple paths of influence of self-reflection on mindreading. *Consciousness and Cognition*, *17*(3), 778–789. doi:10.1016/j.concog.2008.02.005

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. doi:10.1016/j.jml.2007.11.004

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Training monitoring skills improves older adults' self-paced associative learning. *Psychology and Aging*, *18*(2), 340–345. doi:10.1037/0882-7974.18.2.340

Dunning, D. (2011). The Dunning-Kruger Effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology* (1st ed., Vol. 44, pp. 247–296). San Diego, CA: Academic Press.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. doi:10.1111/1467-8721.01235

Eggemeier, F. T. (1991). Performance-based and subjective assessment of workload in multi-task environments. In D. L. Damos (Ed.), G. F. Wilson (Trans.), *Multiple-task Performance* (pp. 217–278). Washington, DC: Taylor & Francis.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. doi:10.1016/j.obhdp.2007.05.002

El Saadawi, G. M., Azevedo, R., Castine, M., Payne, V., Medvedeva, O., Tseytlin, E., … Crowley, R. S. (2010). Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education : Theory and Practice*, *15*(1), 9–30. doi:10.1007/s10459-009-9162-6

Ellis, R. A., & Taylor, M. S. (1983). Role of self-esteem within the job search process. *Journal of Applied Psychology*, *68*(4), 632–640. doi:10.1037/0021-9010.68.4.632

Evered, A. (2005). What can cytologists learn from 25 years of investigations in visual search? *British Journal of Biomedical Science*, *62*(4), 182–192.

Fisher, O., & Oyserman, D. (2017). Assessing interpretations of experienced ease and difficulty as motivational constructs. *Motivation Science*, *3*(2), 133–163. doi:10.1037/mot0000055

Floresco, S. B., Tse, M. T. L., & Ghods-Sharifi, S. (2008). Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. *Neuropsychopharmacology*, *33*(8), 1966–1979. doi:10.1038/sj.npp.1301565

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, *33*, 263–282.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501. doi:10.1016/S1364-6613(98)01262-5

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.

Goforth, D. (1994). Learner control = Decision making + Information: A model and meta-analysis. *Journal of Educational Computing Research*, *11*, 1–26.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Processes*, *96*(2), 142–154. doi:10.1016/j.obhdp.2004.11.002

Gopher, D, & Donchin, E. (1986). Workload: An Examination of the Concept. In *Handbook of Perception and Human Performance* (Vol. 2, pp. 1–49). New York, NY: Wiley.

Halton, J. H., & Smith, G. B. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, *7*(12), 701–702.

Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. *Journal of Memory and Language*, *69*(3), 368–383. doi:10.1016/j.jml.2013.04.005

Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic Bulletin & Review*, *21*(6), 1617–1622. doi:10.3758/s13423-014-0619-0

Harris, K. R. (1986). Self-monitoring of attentional behavior versus self-monitoring of productivity: effects on on-task behavior and academic response rate among learning disabled children. *Journal of Applied Behavior Analysis*, *19*(4), 417–423. doi:10.1901/jaba.1986.19-417

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Hensley, L. C. (2014). Reconsidering active procrastination: Relations to motivation and achievement in college anatomy. *Learning and Individual Differences*, *36*, 157–164. doi:10.1016/j.lindif.2014.10.012

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). *Internal mapping and its impact on measures of absolute and relative metacognitive accuracy*. (J. Dunlosky & S. (Uma) K. Tauber, Eds.) (Vol. 1, pp. 39–63). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199336746.013.15

Hoelzl, E., & Rustichini, A. (2005). Overconfident: do you put your money on it? *The Economic Journal*, *115*(503), 305–318. doi:10.1111/j.1468-0297.2005.00990.x

Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. (R. M. Elliott, Ed.). New York, NY: Appleton-Century-Crofts, Inc.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, M., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(5), 1217–1232.

James, W. (1890). *The Principles of Psychology*. New York, NY: Henry Holt & Company.

Jemstedt, A., Kubik, V., & Jönsson, F. U. (2017). What moderates the accuracy of ease of learning judgments? *Metacognition and Learning*, *12*(3), 337–355. doi:10.1007/s11409-017-9172-3

Jönsson, F. U., & Lindström, B. R. (2010). Using a multidimensional scaling approach to investigate the underlying basis of ease of learning judgments. *Scandinavian Journal of Psychology*, (51), 103–108.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kardas, M., & O'Brien, E. (2018). Easier seen than done: merely watching others perform can foster an illusion of skill acquisition. *Psychological Science*, *29*(4), 521–536. doi:10.1177/0956797617740646

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. doi:10.1080/09658210802647009

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*, 498–525.

Kelley, C. M., & Jacoby, L. L. (1996). Adult Egocentrism: Subjective Experience versus Analytic Bases for Judgment. *Journal of Memory and Language*, *35*(2), 157–175. doi:10.1006/jmla.1996.0009

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching--a review. *Psychological Bulletin*, *136*(5), 849–874. doi:10.1037/a0019842

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology. General*, *143*(1), 131–141. doi:10.1037/a0031048

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. doi:10.1037/0096-3445.126.4.349

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*(2 Pt 1), 149–171. doi:10.1006/ccog.2000.0433

Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(4), 945–959. doi:10.1037/0278-7393.34.4.945

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. doi:10.1016/j.concog.2009.12.010

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

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*(2), 121–132. doi:10.1016/j.learninstruc.2011.08.004

Krajč, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal Of Economic Psychology*, *29*(5), 724–738. doi:10.1016/j.joep.2007.12.006

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121

Kruger, J., Windschitl, P. D., Burrus, J., Fessel, F., & Chambers, J. R. (2008). The rational side of egocentrism in social comparisons. *Journal of Experimental Social Psychology*, *44*(2), 220–232. doi:10.1016/j.jesp.2007.04.001

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

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679. doi:10.1017/S0140525X12003196

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76–94. doi:10.1016/j.obhdp.2006.10.002

Lawless, K. A., & Brown, S. W. (1997). Multimedia learning environments: Issues of learner control and navigation. *Instructional Science*, *25*, 117–131.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Package emmeans (1.3.4). Computer software, R.

Libedinsky, C., Massar, S. A. A., Ling, A., Chee, W., Huettel, S. A., & Chee, M. W. L. (2013). Sleep deprivation alters effort discounting but not delay discounting of monetary rewards. *Sleep*, *36*(6), 899–904. doi:10.5665/sleep.2720

Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: An evaluation of subjective and time estimation metrics. *Ergonomics*, *37*, 1843–1854.

Löffler, E., von der Linden, N., & Schneider, W. (2016). Influence of domain knowledge on monitoring performance across the life span. *Journal of Cognition and Development*, *17*(5), 765–785. doi:10.1080/15248372.2016.1208204

Maclaverty, S. N., & Hertzog, C. (2009). Do age-related differences in episodic feeling of knowing accuracy depend on the timing of the judgement? *Memory*, *17*(8), 860–873. doi:10.1080/09658210903374537

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. doi:10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d′, Response-Specific Meta-d′, and the Unequal Variance SDT Model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-45190-4_3

Marottoli, R. A., & Richardson, E. D. (1998). Confidence in, and self-rating of, driving ability among older drivers. *Accident Analysis & Prevention*, *30*(3), 331–336. doi:10.1016/S0001-4575(97)00100-0

McKendrick, R. D., & Cherry, E. (2018). A deeper look at the NASA TLX and where it falls short. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 44–48. doi:10.1177/1541931218621010

Meshkati, N., Hancock, P. A., Rahimi, M., & Dawes, S. M. (1995). Techniques in mental workload assessment. In *Evaluation of Human Work: A Practical Ergonomics Methodology* (pp. 749–782). Philadelphia, PA: Taylor & Francis.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science : A Journal of the American Psychological Society*, *18*(3), 159–163. doi:10.1111/j.1467-8721.2009.01628.x

Metcalfe, J., & Kornell, N. (2005). A Region of Proximal Learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463–477. doi:10.1016/j.jml.2004.12.001

Minamimoto, T., Hori, Y., & Richmond, B. J. (2012). Is working more costly than waiting in monkeys? *Plos One*, *7*(11), e48434. doi:10.1371/journal.pone.0048434

Mitchell, S. H. (2017). Devaluation of outcomes due to their cost: extending discounting models beyond delay. In J. R. Stevens (Ed.), *Impulsivity* (Vol. 64, pp. 145–161). Champaign, IL: Springer International Publishing. doi:10.1007/978-3-319-51721-6_5

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi:10.1016/S1364-6613(03)00028-7

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the “labor-in-vain effect”. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *14*(4), 676–686.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 125–173). San Diego, CA: Academic Press.

Nickerson, R. S. (1967). Same-different response times with multi-attribute stimulus differences. *Perceptual and Motor Skills*, *24*, 543–554.

Niemiec, R. P., Sikorski, C., & Walberg, H. J. (1996). Learner-control effects: A review of reviews and a meta-analysis. *Journal of Educational Computing Research*, *15*, 157–174.

Nishiyama, R. (2014). Response effort discounts the subjective value of rewards. *Behavioural Processes*, *107*, 175–177. doi:10.1016/j.beproc.2014.08.002

Nussinson, R., & Koriat, A. (2008). Correcting experience-based judgments: the perseverance of subjective experience in the face of the correction of judgment. *Metacognition and Learning*, *3*(2), 159–174. doi:10.1007/s11409-008-9024-2

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. doi:10.1016/j.tics.2008.02.014

Ortner, T. M., Weißkopf, E., & Gerstenberg, F. X. R. (2013). Skilled but unaware of it: CAT undermines a test taker’s metacognitive competence. *European Journal of Psychology of Education*, *28*, 37–51.

Pashler, H. (2000). Task switching and multitask performance. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes* (pp. 275–307). Cambridge, MA: The MIT Press.

Potts, C. A., Pastel, S., & Rosenbaum, D. A. (2018). How are cognitive and physical difficulty compared? *Attention, Perception & Psychophysics*, *80*(2), 500–511. doi:10.3758/s13414-017-1434-2

Prévost, C., Pessiglione, M., Météreau, E., Cléry-Melin, M.-L., & Dreher, J.-C. (2010). Separate valuation subsystems for delay and effort decision costs. *The Journal of Neuroscience*, *30*(42), 14080–14090. doi:10.1523/JNEUROSCI.2752-10.2010

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, *52*, 185–218.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. doi:10.1037/0096-3445.124.2.207

Ross, S. M., Morrison, G. R., & O’Dell, J. K. (1989). Uses and effects of learner control of context and instructional support in computer-based instruction. *Educational Technology Research and Development : ETR & D*.

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal Of Economic Psychology*, *33*(5), 1012–1031. doi:10.1016/j.joep.2012.06.003

Schouppe, N., Demanet, J., Boehler, C. N., Ridderinkhof, K. R., & Notebaert, W. (2014). The role of the striatum in effort-based decision-making in the absence of reward. *The Journal of Neuroscience*, *34*(6), 2148–2154. doi:10.1523/JNEUROSCI.1214-13.2014

Schwartz, B. L., Boduroglu, A., & Tekcan, A. İ. (2016). Methodological concerns: The feeling-of-knowing task affects resolution. *Metacognition and Learning*, *11*(3), 305–316. doi:10.1007/s11409-015-9152-4

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, *53*, 252–266.

Siedlecka, M., Skóra, Z., Paulewicz, B., Fijałkowska, S., Timmermans, B., & Wierzchoń, M. (2018). Responses improve the accuracy of confidence judgements in memory tasks. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(4), 712–723. doi:10.1037/xlm0000608

Sigurdsson, S. O., Taylor, M. A., & Wirth, O. (2013). Discounting the value of safety: Effects of perceived risk and effort. *Journal of Safety Research*, *46*, 127–134. doi:10.1016/j.jsr.2013.04.006

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*, 99–118.

Sobel, K. V., Gerrie, M. P., Poole, B. J., & Kane, M. J. (2007). Individual differences in working memory capacity and visual search: the roles of top-down and bottom-up processing. *Psychonomic Bulletin & Review*, *14*(5), 840–845.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(1), 204–221.

Souchay, C., Isingrini, M., Clarys, D., & Taconnat, L. (2004). Executive functioning and judgment-of-learning versus feeling-of-knowing in older adults. *Experimental Aging Research*, *30*, 47–62.

Spector, A., & Biederman, I. (1976). Mental set and mental shift revisted. *The American Journal of Psychology*, *89*(4), 669–679.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. doi:10.1006/jesp.1998.1373

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*(2), 240–261. doi:10.1037/0033-2909.124.2.240

Stevens, S. S. (1957). On the psychophysical law. *The Psychological Review*, *64*, 153–181.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295–312.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer.

Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. E. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance*, *34*(3), 402–418. doi:10.1016/0030-5073(84)90046-1

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *The British Journal of Educational Psychology*, *81*(Pt 2), 264–273. doi:10.1348/135910710X510494

Titchener, E. B. (1908). Lectures on the Elementary Psychology of Feeling and Attention. New York, NY: Macmillan.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. doi:10.1080/00140139608964470

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118. doi:10.1016/j.jml.2010.11.002

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*, 1039–1061.

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

Unity Technologies. (2019). Unity Game Engine (2018.3.5). Computer software, Unity Technologies.

Vallières, B. R., Hodgetts, H. M., Vachon, F., & Tremblay, S. (2016). Supporting dynamic change detection: using the right tool for the task. *Cognitive Research: Principles and Implications*, *1*(1), 32. doi:10.1186/s41235-016-0033-4

Vangsness, L., & Young, M. (2017). The role of difficulty in dynamic risk mitigation decisions. *Journal of Dynamic Decision Making*, *3*(5).

Vangsness, L., & Young, M. (2018). Central and peripheral cues to difficulty in a dynamic task. *Human Factors*, 18720818809877. doi:10.1177/0018720818809877

Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. S. (2006). Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural Networks*, *19*(8), 1302–1314. doi:10.1016/j.neunet.2006.03.005

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. doi:10.1080/14639220210123806

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human-Computer Studies*, *79*, 79–84. doi:10.1016/j.ijhcs.2015.01.002

Wickens, C. D., Gutzwiller, R. S., Vieane, A., Clegg, B. A., Sebok, A., & Janes, J. (2016). Time sharing between robotics and process control: validating a model of attention switching. *Human Factors*, *58*(2), 322–343. doi:10.1177/0018720815622761

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2016). *Engineering Psychology and Human Performance* (4th ed.). New York, NY: Routledge.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. doi:10.3758/BF03200774

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. *Integrated Models of Cognitive Systems*, 99–119.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. doi:10.1038/435439a

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural Processes*, *123*, 43–53. doi:10.1016/j.beproc.2015.09.009

Zenger, T. R. (1992). Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, 198–219.

Appendix A - Calibrating the Visual Search Task

Task Calibration

Standardizing and means-centering difficulty allows comparisons to be made across conditions. These cross-condition comparisons can also be used to calibrate the task by estimating the parameter values that will produce equivalent performance in all four conditions. Specifically, the condition intercepts and Condition × Standardized Difficulty fixed effect slopes can be used to center and set the range of parameter values, respectively. Unfortunately, post-hoc analyses of participants' performance in Experiments 1 and 2 (see Appendix B) indicated that the calibration adjustments were made in the wrong direction.

Calibrating average difficulty

Because incentive magnitude does not change the difficulty level of the task itself, the feedback condition was used to calibrate difficulty across conditions. The emmeans package in R (Lenth et al., 2019) was used to determine condition intercepts, which were uncentered and back-transformed to the original scale using the inverse link (logistic) function. These intercepts could be compared and unstandardized to determine how much to shift the range of each parameter (see Table A.1). Algebraic calculations indicated that the click condition needed to be made more challenging by shifting the mid-point upwards by 2 clicks, while the set size and timing conditions needed to be slightly easier by shifting their ranges downwards by 3 non-target items and upwards by 1.5 s, respectively.

Table A. 1
Intercept parameter estimates from a multi-level logistic model predicting the likelihood that participants would correctly identify the target on a given trial. Estimates are un-centered and back-transformed to the original scale for ease of comparison.

condition	intercept (B_0)	*SE*	original scale
Clicks	-0.79	0.11	0.43
Feedback	-0.05	0.11	0.61
Set Size	0.03	0.11	0.63
Timing	2.58	0.14	0.96

Calibrating the range of difficulty

Because screen dimensions restricted the number of non-target items that could be presented on the screen at one time, the set size condition was used to calibrate the range of difficulty across conditions. The emmeans package in R was used to determine the standardized difficulty slope in each condition. New ranges could be calculated by multiplying this ratio by the existing range (see Table A.2). Algebraic calculations indicated that the clicks condition needed to be narrowed to 2 clicks and the timing condition needed to be narrowed to 0.48 s. The feedback condition was not included in these comparisons because changes to incentive magnitude did not significantly affect participants' performance.

Table A. 2
Slope estimates from a multi-level logistic model predicting the likelihood that participants would correctly identify the target on a given trial.

condition	B_{stdDiff}	*SE*	new parameter range
Clicks	-5.48	0.16	$\frac{-1.81}{-5.48} \times 5 = 1.65$
Set Size	-1.81	0.12	n/a
Timing	-15.00	0.44	$\frac{-1.81}{-15.00} \times 4 = 0.48$

Appendix B - Supplemental Analyses

Experiment 1

Performance Calibration

The emmeans package (Lenth et al., 2019) was used to generate and compare 95% confidence intervals for the estimates that involved difficulty condition. Participants' average performance differed considerably across conditions (see Table B.1). In addition, changes to task difficulty affected performance more strongly in some conditions than in others: only changes to set size affected participants' performance (see Table B.2).

Table B. 1.
Condition estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

condition	B	SE	$CI_{95\%}$
click	-5.10	0.29	(-5.67, -4.52)
feedback	-0.28	0.13	(-0.54, -0.02)
set size	-0.02	0.13	(-0.28, 0.24)
timing	4.45	0.22	(4.02, 4.88)

Table B. 2
Standardized difficulty slope estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

condition	B	SE	$CI_{95\%}$
click	-3.00	0.59	(-4.15, -1.85)
feedback	-0.04	0.16	(-0.34, 0.27)
set size	-1.79	0.17	(-2.13, -1.45)
timing	-0.21	0.64	(-1.46, 1.05)

Performance differed substantially from that of the pilot study, which can be illustrated by comparing human performance with computer performance during the observation trials (see Figure B.1). Performance intercepts (the central points of each line) and standardized difficulty slopes differed during the performance trials (left panel), but were similar across the observation trials (right panel). These differences likely emerged as a result of improper calibration of the

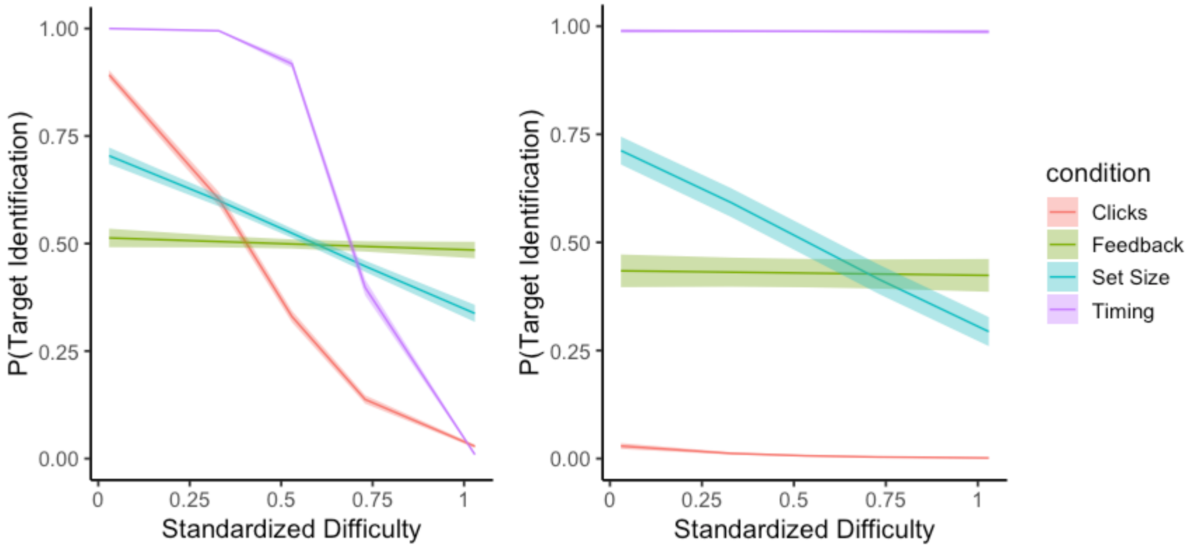


Figure B. 1. Participants' performance in the visual search task (right) differed substantially from the pilot experiment predictions, which are illustrated by the computer performance trials at left. Error ribbons represent ±1*SE*.

task, which was discovered post-hoc. Specifically, the timing and clicks conditions should have been made harder and easier, respectively. Instead, they were made the opposite. Although this error prevents conclusions from being drawn regarding the effects of dimension of difficulty, the within-subject design and counterbalancing techniques that were employed preserve the main findings of the experiments.

Exploratory Analysis – Intrinsic Motivation

Two exploratory analyses were conducted to determine whether intrinsic motivation affected participants' performance or JODs. These analyses involved the same model specifications as before but included the main effect of intrinsic motivation in the fixed effect structure. Intrinsic motivation did not significantly impact performance or JODs, and the reported effects were unchanged (see Tables B.3 and B.4).

Table B. 3

Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

parameter	B	*SE*	z	*p*
intercept	-0.08	0.16	-0.51	.61
standardized difficulty	-1.25	0.22	-5.79	< .001
clicks	-4.81	0.20	-24.43	< .001
feedback	-0.07	0.09	-0.80	.42
set size	0.23	0.09	2.59	.01
experimental trial	0.25	0.06	4.25	< .001
perform-first	-0.23	0.19	-1.18	.24
interleaved	0.07	0.22	0.34	.74
intrinsic motivation	-0.04	0.09	-0.39	.69
Standardized Difficulty × Clicks	-1.78	0.48	-3.71	< .001
Standardized Difficulty × Feedback	1.21	0.24	4.96	< .001
Standardized Difficulty × Set Size	-0.53	0.25	-2.14	.03

Note. Standardized difficulty (Mdn = 0.5), experimental trial (Mdn = 180), and intrinsic motivation (Mdn = 2.75) were median-centered prior to analysis. The observe-first and timing conditions served as the {-1, -1} and {-1, -1, -1} baselines, respectively.

Table B. 4

Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would say the task was harder than before.

parameter	B	*SE*	*z*	*p*
intercept	-0.20	0.09	-2.24	.02
standardized difficulty	0.65	0.14	4.53	< .001
identified target	1.24	0.09	13.82	< .001
trial in block	-0.01	0.04	-0.13	.89
do first	0.23	0.12	1.91	.06
interleaved	-0.20	0.13	-1.61	.11
intrinsic motivation	-0.07	0.06	-1.19	.24
Identified Target × Trial in Block	0.004	0.04	0.10	.92
Identified Target × Do First	0.04	0.12	0.30	.77
Identified Target × Interleaved	0.26	0.12	2.13	.03
Trial in Block × Do First	-0.13	0.06	-2.20	.03
Trial in Block × Interleaved	-0.01	0.06	-0.22	.83
Identified Target × Trial in Block × Do First	-0.10	0.06	-1.67	.10
Identified Target × Trial in Block × Interleaved	0.001	0.06	0.02	.99

Note. Standardized difficulty (Mdn = 0.5), trial in block (Mdn = 37), and intrinsic motivation (Mdn = 2.75) were median-centered prior to analysis. Missed target and the observe-first role condition served as the {-1} and {-1, -1} baselines, respectively.

Experiment 2

Performance Calibration

The emmeans package (Lenth et al., 2019) was again used to generate and compare 95% confidence intervals for the estimates that involved difficulty condition. Participants' average performance differed considerably across conditions (see Table B.5). In addition, changes to task difficulty affected performance more strongly in some conditions than in others: only changes to set size affected participants' performance (see Table B.6).

Table B. 5

Condition estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

Condition	B	SE	$CI_{95\%}$
Clicks	-4.03	0.17	(-4.36, -3.70)
Set Size	-0.04	0.15	(-0.33, 0.24)
Timing	5.81	0.24	(5.33, 6.29)

Table B. 6

Standardized difficulty slope estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

Condition	B	SE	$CI_{95\%}$
Clicks	-1.52	0.23	(-1.97, -1.06)
Set Size	-2.29	0.15	(-2.57, -2.00)
Timing	-0.63	0.55	(-1.71, 0.44)

As before, performance differed substantially from what was predicted by the pilot study, due to the improper calibration of the task. Performance intercepts (the central points of each line) and standardized difficulty slopes significantly differed across difficulty conditions (see Figure B.2).

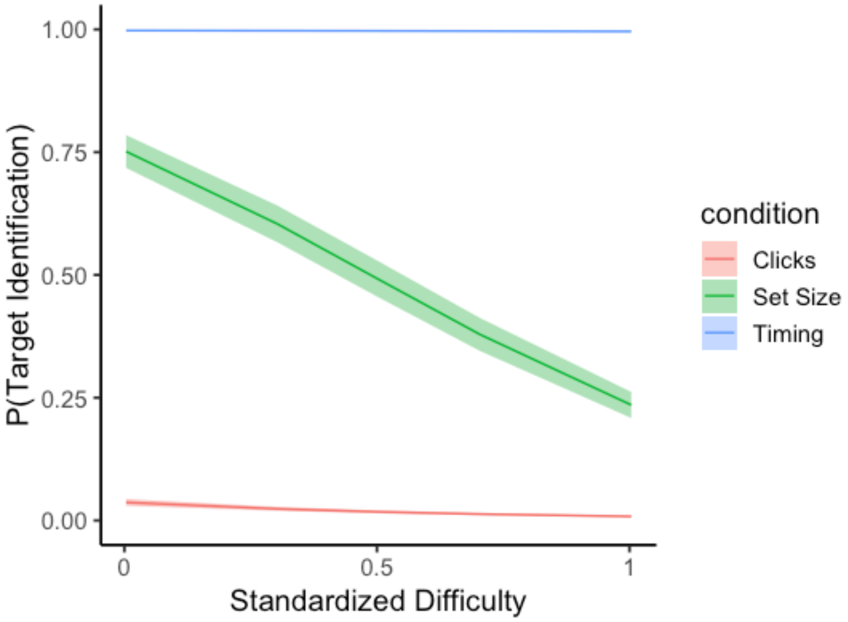


Figure B. 2. Participants' performance in the visual search task (left) differed substantially from the pilot experiment predictions. Error ribbons represent ±1*SE*.

Exploratory Analysis – Intrinsic Motivation

Two exploratory analyses were conducted to determine whether intrinsic motivation affected participants' performance or JODs. These analyses involved the same model specifications as before but included the main effect of intrinsic motivation in the fixed effect structure. Intrinsic motivation did not significantly impact performance or JODs, and the reported effects were unchanged (see Tables B.7 and B.8).

Table B. 7

Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would identify the target on a given trial.

Parameter	B	*SE*	*z*	*p*
intercept	0.17	0.05	3.68	< .001
standardized difficulty	-0.50	0.08	-6.60	< .001
missed TTI	-1.31	0.02	-55.91	< .001
feedback magnitude	-0.001	0.002	-0.57	.57
post-trial feedback	0.06	0.05	1.26	.21
trial in experiment	0.25	0.29	0.86	.39
age	-0.01	0.003	-4.50	< .001
intrinsic motivation	0.01	0.04	0.37	.71
Standardized Difficulty × Missed TTI	-0.29	0.07	-4.00	< .001
Standardized Difficulty × Feedback Magnitude	-0.003	0.01	-0.52	.60
Standardized Difficulty × Post-trial Feedback	-0.10	0.07	-1.32	.19
Feedback Magnitude × Post-trial Feedback	-0.003	0.002	-1.97	.05
Standardized Difficulty × Feedback Magnitude × Post-trial Feedback	0.01	0.01	2.63	.01

Note. Standardized difficulty (Mdn = 0.5), trial in experiment (Mdn = 160), and intrinsic motivation (Mdn = 3.5) were median-centered prior to analysis. Correct TTI and pre-trial feedback served as the {-1} and {-1} baselines, respectively.

Table B. 8

Parameter estimates from a multi-level logistic regression predicting the likelihood that participants would say the task was harder than before.

Parameter	B	*SE*	*z*	*p*
intercept	-0.03	0.12	-0.28	.78
standardized difficulty	0.59	0.16	3.72	< .001
feedback magnitude	0.01	0.003	3.11	.002
missed target	1.66	0.13	13.26	< .001
post-trial feedback	0.14	0.11	1.29	.21
click condition	-0.52	0.10	-5.22	< .001
set size condition	0.17	0.07	2.52	.01
trial in block	-0.04	0.05	-0.83	.41
age	0.01	0.01	0.83	.41
intrinsic motivation	0.01	0.09	0.12	.90
Standardized Difficulty × Feedback Magnitude	-0.01	0.01	-1.00	.32
Standardized Difficulty × Missed Target	-0.0004	0.003	-0.14	.89
Missed Target × Post-trial Feedback	-0.08	0.10	-0.81	.42
Feedback Magnitude × Post-trial Feedback	0.001	0.003	0.24	.81
Standardized Difficulty × Click Condition	-0.75	0.19	-4.07	< .001
Standardized Difficulty × Set Size Condition	1.17	0.23	5.12	< .001
Feedback Magnitude × Missed Target × Post-trial Feedback	0.004	0.003	1.31	.19

Note. Standardized difficulty (Mdn = 0.5), trial in block (Mdn = 53), and intrinsic motivation (Mdn = 3.5) were median-centered prior to analysis. Identified target, pre-trial feedback, and the timing condition served as the {-1}, {-1}, and {-1, -1} baselines, respectively.