UNIVERSITY OF CALIFORNIA
RIVERSIDE

Cross-Modal Task Difficulty Comparison

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Iman Feghhi

September 2021

Dissertation Committee:
Dr. David A. Rosenbaum, Chairperson
Dr. John M. Franchak
Dr. Lawrence D. Rosenblum

The Dissertation of Iman Feghhi is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This dissertation would not have been possible without the support and guidance of my dissertation committee, my lab mates, my research assistants, and my friends.

Dedication

To my wonderful wife, Marjan Rahimpour, and children, Nila and Nick.

ABSTRACT OF THE DISSERTATION

Cross-Modal Task Difficulty Comparison

by

Iman Feghhi

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, September 2021
Dr. David A. Rosenbaum, Chairperson

We usually strive to do tasks in the easiest way possible. What makes a task easy or difficult is poorly understood, though. I started this line of research by arguing that the prerequisite to study a construct – including the perception of task difficulty - is to measure it. I developed a method to measure the perceived difficulty of a task. I also showed that people can reliably compare the difficulty of different kinds of tasks. Next, I tested different hypotheses that could explain this ability. Given that I did not find compelling support for any of the examined hypotheses, I then proposed that difficulty is represented at a more abstract level than what scientists have been searching for. Finally, I tested the effect of task difficulty on task scheduling and showed that people prefer doing easy tasks first, unless the tasks are physical.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 – INTRODUCTION

"Choices are the hinges of destiny."
- Pythagoras

The epigraph above from the ancient Greek thinker Pythagoras offers deep insight into the importance of decision-making in our lives. Every day, we make all sorts of decisions. Some are strategic, like accepting or declining a job offer, while some are more mundane, like using a ramp or stairs on the way to a parking lot. Some are conscious, like stopping at an elevator door for a few seconds and then convincing yourself to use the stairs instead, while some are unconscious, like shopping the middle items in an aisle more often than the end items (Christenfeld, 1995). Christenfeld showed that even when different options seem to have similar demands, choices are not random. Without realizing it, we routinely make alternative plans and then choose among them.

On what basis does the system choose between the alternative options? A promising answer was suggested by Tversky (1972), who advocated an approach based on elimination by aspects. According to this hypothesis, which was picked up for modeling physical action selection by Rosenbaum et al. (2001), choice relies on hierarchies. Among alternatives, the ones that fail to satisfy the most important need are eliminated first, the ones that fail to satisfy the second-most important need are eliminated second, the ones that fail to satisfy the third-most important need are

eliminated third, and so on. At the end, if more than one option is left, the system picks a survivor randomly.

Is such an approach actually used for physical actions or other mundane decisions in everyday life? Elimination by aspects makes sense for deciding whom to hire or which job to take – the kinds of examples that Tversky (1972) discussed – but does it make sense for deciding which physical act to carry out in the kitchen or supermarket?

Consider evidence from experiments in university students were asked to pick up a beach bucket on a table and carry it to one of two target positions (Rosenbaum et al., 2011). The participants could go to the right side of the table, grab the bucket, and carry it to a right stool beyond the table; or they could go to the left side of the table, grab the bucket, and carry it to a left stool beyond the table. Situations like this are ubiquitous in our lives. Whenever you want to clean a table, for example, you need to decide the easiest way to reach an item, and if you are not in the most comfortable position, is it worth walking around the table, or would it be easier to lean over? Or when you are cleaning your car's windshield, should you lean over to reach all the spots, or should you do half of it from the right side and the rest from the left side? You need to solve problems like these by comparing different costs, and some costs may be more important than others. Through modeling, it was shown that the behavioral choices shown by participants in the Rosenbaum et al. (2011) study were well predicted by a decision model in which each meter of reaching was implicitly judged to be approximately as hard as 11 meters of walking. Therefore, reaching was implicitly judged to be more important than walking by a factor of 11.

The first scientist who pointed to the importance and even necessity of difficulty evaluation in decision making was an Italian philosopher named Guglielmo Ferrero (1871–1942). He maintained that whenever possible, people opt for the least effortful action (https://en.wikipedia.org/wiki/Guglielmo_Ferrero). Fifty years after Ferrero articulated this principle of least effort, Zipf (1949) argued that the principle explains word frequency. Zipf showed that the frequency of a word in natural language is a logarithmic function of its rank. The second most frequent word occurs half as often as the most frequent word, the third most frequent word occurs half as often as the second most frequent word, and so on, this principle is called Zipf's Law (https://en.wikipedia.org/wiki/Zipf%27s_law).

Considerations of minimizing effort come up in other domains as well. In the field of library and information science, it is well known that information seekers use the least effortful method in their searches and that they stop their searches as soon as minimally accepted results are found (Mann, 2015). In surgery rooms, Yang et al. (2015) showed that surgeons are likely to check patients' records if they need to walk less 5 meters to get that information but are very unlikely to check the records if they need to walk more than 5 meters. How many lives could be saved if surgery room designers knew about the effect of perceived difficulty on the probability of consulting patients' records?

The effect of the perceived difficulty in our lives can be even more dramatic. It is the case when a deficit in evaluating the difficulty of a task leads to unnecessarily risky behavior. Traffic accidents are often due to drivers underestimating risk (Penmetsa & Pulugurtha, 2017), and underestimating the difficulty of driving has been reported for

sleepy drivers and drunk drivers (Bazilinskyy et al., 2020; Watling et al., 2016). According to the United States Centers for Disease Control and Prevention, these risky behaviors are the main contributors to more than 1.35 million fatal accidents that have occurred worldwide (https://www.cdc.gov/injury/features/global-road-safety/index).

Another case of reckless behavior due to underestimating task demands is texting while driving, which is most common among young drivers (Watters, & Beck, 2016). While both texting and driving might be easy to perform independently, the difficulty of doing them simultaneously exceeds the sum of the two difficulties (e.g., Caird et al., 2014). Knowing that traffic crashes are the leading cause of death among teens and young adults in the U.S., we have a powerful reason to better understand the factors that affect perceived difficulty. Part of that understanding should be about overestimation of task difficulty, not just underestimation. If one overestimates the difficulty of a task, one is more likely to disengage in the task. An example of that kind of avoidance behavior can be seen in girls who shy away from math. Although their grades suggest that, if anything, they perform slightly better than boys (https://www.nytimes.com/interactive/2018/06/13/upshot/boys-girls-math-reading-tests.html), the underrepresentation of women in STEM fields shows that girls perceive math and science to be more difficult than boys do. Because social factors seem to be the main contributor to this overestimation (Wieselmann, et al., 2020), society can benefit from knowing how social factors influence the perception of task difficulty.

Given the importance of the perception of task difficulty in different aspects of personal and social lives, it makes sense that it has been studied in diverse fields,

including philosophy, sport science, psychology, language, education, and robotics (André et al., 2019; Burgess & Jones, Larry, F, 1997; Cos, 2017; Fisher & Steele, 2014; Halperin & Emanuel, 2020; Montero, 2016; Pageaux, 2014; Shenhav et al., 2017; Song et al., 2019; Steele, 2020). Regardless of all the attention that has been paid to this topic, no consensus has been reached on how we perceive task difficulty. The way we tackled this question was to build a foundation that could be used in various disciplines. Inspired by advances in physics, we realized that the first step would be to devise a way to measure task difficulty. Much as Fechner (1966) introduced psychophysics to open a new window into the study of perception, we think being able to measure perceived difficulty can provide a foundation for reconciling scattered efforts in different fields and help better understand this enormously important construct.

The next chapter is my first paper on this topic, in which I introduced a method to reliably measure the perceived difficulty. The chapters after that are about the work I have done to replicate and extend this method, test different hypotheses about task difficulty, introduce the common code hypothesis for the perception of task difficulty, and finally to test the effect of perceived difficulty on task scheduling. There is a clear bottom-line message from all this work: Significant progress can be made on understanding task choice by quantifying perceived task difficulty.

References

André, N., Audiffren, M., & Baumeister, R. F. (2019). An integrative model of effortful control. *Frontiers in Systems Neuroscience*, 13, 79.

Bazilinskyy, P., Eisma, Y. B., Dodou, D., & De Winter, J. C. F. (2020). Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention*, 21, 347-353.

Burgess, P. R., & Jones, Larry, F. (1997). Perceptions of effort and heaviness during fatigue and during the size-weight illusion. *Somatosensory & Motor Research*, *14*, 189–202. https://doi.org/10.1080/08990229771051

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 & Prevention*, 71, 311-318.

Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, *6*, 50-55.

Cos, I. (2017). Perceived effort for motor control and decision-making. *PLoS Biology*, *15*, e2002885.

Fechner, G. (1966). *Elements of psychophysics. Vol. 1*. Holt, Rinehart and Winston: New York.

Fisher, J., & Steele, J. (2014). Questioning the resistance/aerobic training dichotomy: A commentary on physiological adaptations determined by effort rather than exercise modality. *Journal of Human Kinetics*, *44*, 137–142.

Goldberg, G., & Bloom, K. K. (1990). The alien hand sign. Localization, lateralization and recovery. *American Journal of Physical Medicine & Rehabilitation*, *69*, 228-238.

Halperin, I., & Emanuel, A. (2020). Rating of perceived effort: Methodological concerns and future directions. *Sports Medicine*, 1–9.

Mann, T. (2015). *The Oxford guide to library research*. Oxford University Press.

Montero, B. G. (2016). *Thought in action: Expertise and the conscious mind*. Oxford University Press.

Pageaux, B. (2014). The psychobiological model of endurance performance: An effort-based decision-making theory to explain self-paced endurance performance. *Sports Medicine*, *44*, 1319.

Penmetsa, P., & Pulugurtha, S. S. (2017). Risk drivers pose to themselves and other drivers by violating traffic rules. *Traffic Injury Prevention*, 18, 63-69.

Rosenbaum, D. A., Meulenbroek, R. J., Vaughan, J., & Jansen, C. (2001). Posture-based motion planning: applications to grasping. *Psychological Review*, *108*, 709.

Rosenbaum, D. A., Brach, M., Semenov, A. (2011). Behavioral ecology meets motor behavior: Choosing between walking and reaching paths. *Journal of Motor Behavior*, 43, 131–136.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental

Effort. *Annual Review of Neuroscience*, *40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Song, J., Kim, S., & Bong, M. (2019). The more interest, the less effort cost perception and effort avoidance. *Frontiers in Psychology*, *10*, 2146.

Steele, J. (2020). *What is (perception of) effort? Objective and subjective effort during task performance* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kbyhm

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281.

Watling, C. N., Armstrong, K. A., Smith, S. S., & Obst, P. L. (2016). Crash risk perception of sleepy driving and its comparisons with drink driving and speeding: Which behavior is perceived as the riskiest? *Traffic Injury Prevention*, 17, 400-405.

Watters, S. E., & Beck, K. H. (2016). A qualitative study of college students' perceptions of risky driving and social influences. *Traffic Injury Prevention*, 17, 122-127.

Wieselmann, J. R., Roehrig, G. H., & Kim, J. N. (2020). Who succeeds in STEM? Elementary girls' attitudes and beliefs about self and STEM. *School Science and Mathematics*, *120*, 297-308.

Yang, X. J., Wickens, C. D., Park, T., Fong, L., & Siah, K. T. (2015). Effects of information access cost and accountability on medical residents' information retrieval strategy and performance during prehandover preparation: Evidence from interview and simulation study. Human Factors, 57, 1459 –1471. http://dx.doi.org/10.1177/0018720815598889

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

CHAPTER 2 - JUDGING THE SUBJECTIVE DIFFICULTY OF DIFFERENT KINDS OF TASKS

Iman Feghhi and David A. Rosenbaum

Published in Journal of Experimental Psychology, Human Perception and Performance (2019)

Abstract

People judge the relative difficulty of different kinds of tasks all the time, yet little is known about how they do so. We asked university students to choose between tasks that taxed perceptual-motor control and memorization to different degrees. Our participants decided whether to carry a box through a wide (81 cm) or narrow (36 cm) gap after memorizing 6, 7, or 8 digits. The model that maximized the likelihood of observing the choice data treated the extra physical demand of passing through the narrow gap as functionally equivalent to memorizing an extra .55 digits. Substantively, the model suggested that participants judged the difficulty of the compound tasks in terms of separate resources. The approach introduced here may help inter-relate different kinds of task difficulty.

Keywords: Action, Decision Making, Effort, Memory, Meta-cognition

Public Significance Statement

If a doctor must take a long walk to get easy-to-understand information about a patient or a short walk to get hard-to-understand information, how should the doctor decide? Little is known about people's judgments concerning the difficulty of different kinds of tasks. This article introduces a method for addressing this question. It offers a way of expressing the difficulty of one kind of task in terms of another, and provides a way of determining whether the resources for the two kinds of task are treated as independent or dependent.

Introduction

A core aim of experimental psychology is to characterize relations between external events and their internal representations. One such relation that has been studied in detail is between intensities of external stimuli and intensities of their internal analogs. Starting with Weber and Fechner in the 1800's, such relations have been studied psychophysically. One of the psychophysical methods that has been used is to ask participants to adjust the intensity of a stimulus of one modality (e.g., the loudness of a sound) to match the intensity of a stimulus of another modality (e.g., the brightness of a light). The orderliness of the data has been taken to suggest that there may be an amodal representation of stimulus intensity (Marks et al., 1986; Pitts et al. 2016).

We sought to extend this approach to the perception of task difficulty. We reasoned that if people have access to some metric of task difficulty, or if they can map the difficulty of one kind of task to another, they should be able to compare the difficulty of different kinds of tasks. We were especially interested in tasks that draw on cognitive abilities and perceptual-motor abilities. All tasks rely on both kinds of abilities, of course. The two sorts of abilities share common resources, as shown by the fact that acquisition of intellectual skills and perceptual-motor skills have much in common (Rosenbaum, Carlson, & Gilmore, 2001; Rosenbaum, 2017; Schmidt & Bjork, 1992), that aspects of working memory are linked to aspects of performance (Baddeley, 1976; Logan & Fischman, 2011, 2015; MacDonald, 2016; Weigelt et al., 2009), and that ostensibly intellectual tasks rely on body representations (Barsalou, 2008; Beilock, 2015; Goldin-

Meadow, 2003; Witt, 2011). Nevertheless, perceptual-motor abilities and intellectual abilities come on-line at different phylogenetic and ontogenetic stages; children learn to walk before they learn to count, for example. In addition, perceptual-motor abilities and intellectual abilities are treated as different in our culture. People who excel in chess matches do not necessarily excel in wrestling matches, and vice versa. Given these broad considerations, we wondered how people judge the difficulty of tasks that tax perceptual-motor abilities and intellectual abilities to different degrees. We sought to develop a measure of one sort of task difficulty relative to the other, and we sought to determine whether the two kinds of difficulty are independent or interactive when it comes to making choices about them and to measuring how well they are done.

To investigate the subjective difficulty of different kinds of tasks, we used the two-alternative forced choice (2AFC) procedure. Our lab has used this procedure extensively in the past to relate the subjective difficulty of different kinds of physical activities, namely, walking and reaching (Rosenbaum et al., 2013). In the present experiments, we asked participants to choose between tasks that had a perceptual-motor component (what we call the "physical" component) and a more cognitive component (what we call the "mental" component). The physical component was carrying an empty cardboard box through a wide (81 cm) or narrow (36 cm) gap. The mental component was memorizing 6, 7, or 8 digits. The design let us determine how variation of one task's demands is quantitatively related to the other's.

Our deeper theoretical aim was to see if the two task demands were independent or interactive, both with respect to errors and choices. Regarding *errors*, following the logic of Sternberg (1969), we reasoned that if physical and mental performance were affected by a single resource, then errors of either type would be affected by demands of that type and the other type as well. Physical errors would depend on physical demands *and* memorial demands, or memory errors would depend on memorial demands *and* physical demands. In short, there would be an interaction. Conversely, if physical and mental performance were affected by distinct resources, then errors of either type would be affected only by that type of demand. The effects would be additive.

Regarding *choices*, the same logic applied. If choices of physical task were made only with respect to physical demands, or if choices of memorial task were made only with respect to memorial demands, then choice probabilities would be additively affected by the two kinds of demands. On the other hand, if choices of physical task were made with respect to physical *and* memorial demands, or if choices of memory task were, similarly, made with respect to physical *and* memorial demands, then choice probabilities would be interactively affected by the two kinds of demands.

A particularly interesting extension of these lines of thought was to check for congruity of incongruity of errors and choices. As seen in the table below, we could distinguish among four possibilities. One was that error probabilities were independent and choice probabilities were as well (cell *a*). Another was that error probabilities were dependent and choice probabilities were, too (cell *d*). Either of these outcome would

comprise evidence for *congruity* of errors and choices. Conversely, another possibility was that error probabilities were independent and choice probabilities were not (cell *c*), or that error probabilities were dependent and choice probabilities were not (cell *b*). Either of these outcome would comprise evidence for *incongruity* of error and choice.

	Error Probabilities	
Choice Probabilities	Independent	Dependent
Independent	*a*	*b*
Dependent	*c*	*d*

To the best of our knowledge, no prior study has addressed this full set of issues.

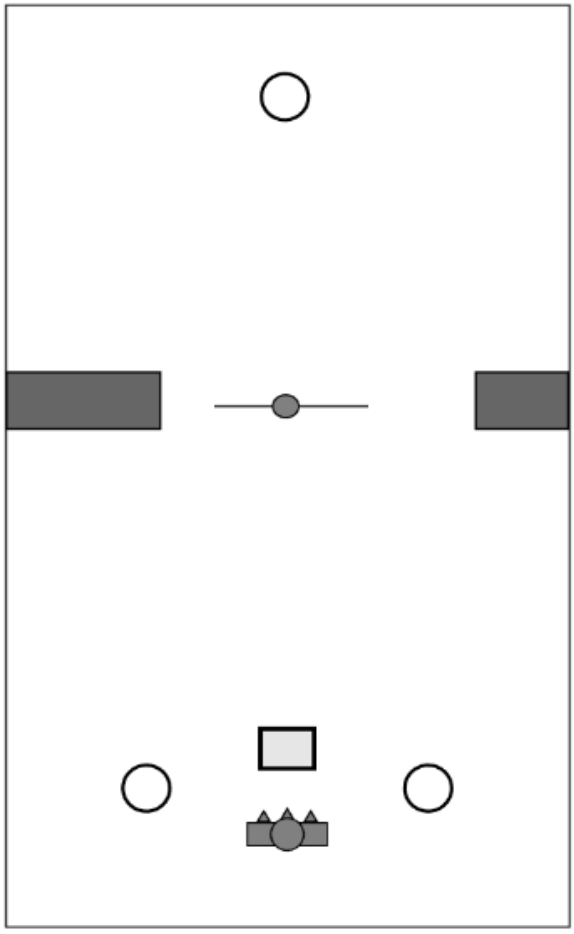
Experiment 1

The participants in our first experiment chose between two routes (Figure 2.1), both of which required picking up and carrying an empty cardboard box from a central start position through a gap to a table on the other side. We used an empty box in this experiment, figuring that we might add weight to it in later studies. The box was carried through a wide (81 cm) gap on the left or right, or a narrow (36 cm) gap on the other side. Associated with each gap was a list of 6, 7, or 8 random digits to be memorized. Subjects were asked to do whatever seemed easier: (a) memorize the digits associated with the *wide* gap, carrying the box through the wide gap, and then trying to recall those digits; or

(b) memorize the digits associated with the *narrow* gap, carrying the box through the narrow gap, and then trying to recall *those* digits.

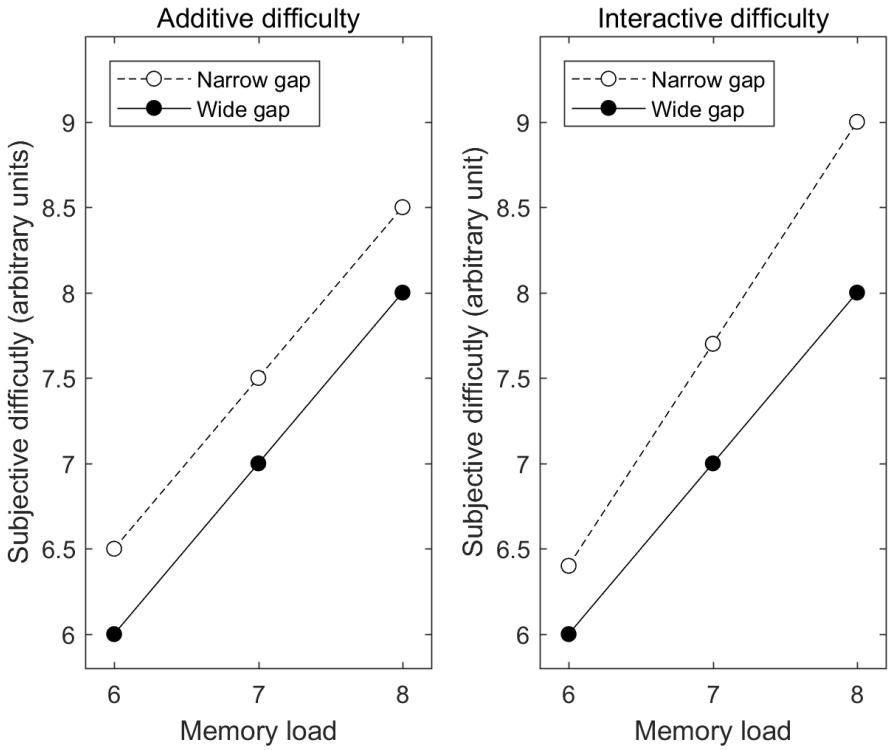
We assumed, as shown in Figure 2.2, that the difficulty of the cognitive task would increase as more digits were memorized (Baddeley, 1976). We also assumed that the difficulty of the perceptual-motor task would be greater for passing through the narrow gap than for passing through the wide gap (Franchak, van der Zalm, & Adolph, 2010). The combination of these factors could be expressed as two lines representing the subjective difficulty of memorizing 6, 7, and 8-digit lists when the navigation challenge was either small (the wide-gap case) or large (the narrow-gap case). According to one model (Figure 2.2, left panel), the physical and mental tasks would contribute *independently* to task difficulty, as shown by the fact that the two lines are parallel. According to the other model (Figure 2.2, right panel), the physical and mental tasks would contribute *interactively* to task difficulty, as shown by the fact that the two lines are not parallel. The procedure for distinguishing the models is explained in a later section.

Figure 2. 1. Schematic overhead view of the experimental setup.



Note: participant stood at a waist-high table with a box on it which the participant would pick up and carry. Two waist-high stools stood to either side, each with a card (not shown) on which appeared a list of 6, 7, or 8 random digits that were easily readable from the participant's standing position. Beyond the box was a pointer that could turn if the participant touched it while attempting to pass through the gap on one side or the other. In this example, the left gap was narrow and the right gap was wide. Alternatively, the narrow gap could be on the right and the wide gap could be on the left. A waist-high stool at the end of the alley was where the box would be deposited and where the participant would stand while attempting to recall the digits for the chosen side.

Figure 2. 2. Additive vs. interactive models.



Note: Hypothetical relation between subjective difficulty and number of digits to be recalled when the physical task was passing through a wide or narrow gap when subjective difficulty was an additive function of the two demands (left panel), or an interactive function of the two demands (right panel). The lines in the left panel are actually from the best-fitting model, so the left panel shows the main result of this study, namely, that the greater challenge of passing through the narrow gap is equivalent to the challenge of memorizing an extra .55 digits.

Method

Participants

Forty Penn State University undergraduates (22 women and 18 men, average age 19.8 years), took part for course credit after giving informed consent. The sample size was based on previous 2AFC studies in our lab where multiple choices were obtained per participant. In the present experiment, the number of choices per participant was 18, so there were 720 observations altogether. This number exceeds the value of n>500 recommended for evaluation of logistic regression models (Cohen et al., 2013; Hosmer et al., 1997). The experiment was approved by the Penn State Institutional Review Board.

Apparatus

Each participant stood at a home position and faced the empty cardboard box referred to above (35.56 cm × 35.56 cm × 35.56 cm), which rested on a 76 cm high platform that stood 33 cm in front of the participant, who stood at a mark on the floor. To the left of the platform was a stool 63 cm high and 100 cm from the base of the box-bearing platform. A card lay on the stool, with a sequence of 6, 7, or 8 random, distinct digits. This was the digit list the subject was supposed to memorize if he or she chose that side. An identical stool to the right of the box-bearing platform (also 63 cm high and 100 cm from the base of the box-bearing platform) had a card with a different sequence of 6, 7, or 8 random distinct digits for possible memorization. The digits were large enough to be read from the participant's starting point.

Beyond the box on the home platform, the subject saw two gaps, one 81 cm (the wide gap), the other 36 cm (the narrow gap). Between the gaps was a 98 cm high stand, 184 cm from the starting position, on which was mounted a light wooden stick (pointer) which turned if it was touched. The pointer extended 23 cm into the gaps. The gap sizes given above ignored the pointer length. We chose the wide and narrow gap sizes based on pilot work to get a clear difference in the physical demands of the two options.

At the start of each trial, the pointer was set perpendicular to the straight line going from the start platform to the goal platform (described below). The purpose of the pointer was to register gap-clearance failures. The pointer was easy to see, had low friction, and was easily jostled. For participants to avoid touching the pointer, they had to turn while passing through the gap and hold the box above the plane of the pointer. We recorded whether the pointer was jostled, but we did not record by how much, although a typical pointer rotation caused by touching the pointer was about 45 degrees.

The goal platform mentioned above stood at the end of the alley, with its center 368 cm from the start position. The goal platform was 63 cm high and 44 cm wide and served as the station where the box was set down after being carried through the wide or narrow gap. After the box was set down, the participant stood at the platform and attempted to recall the digits.

Procedure and Design

The subject's first task was to decide which list to memorize and which gap to traverse based on which side seemed easier. Prior to doing the choice trials, the subject was told that the correct digits had to be recalled in the original order, that after recall, the experimenter would tell the subject whether the recall was correct and, if it was not, that the trial would have to be repeated, in which case the subject would have to go back to the start position, pick up and carry the box through the same gap, and recall the list again. The subject was also told that the trial would have to be repeated if the pointer was moved. In the actual trials, the experimenter said nothing until the subject reached the goal, set down the box, and recalled the digits, so even if the subject bumped into the pointer, he or she still had to recall the digits. If a subject made a mistake in a repeat trial, the trial was not repeated again. Performance in the repeat trials was not analyzed.

A random half of the participants had the narrow gap on the left in the first nine trials and the wide gap on the left in the next nine trials. The other half of the subjects got the opposite assignment. There were nine trials per gap setup because there were three digit lengths for the narrow side, crossed with three digit lengths for the wide side. The digit sequences were random except for the constraints that a digit sequence never started with zero, had no repeated numbers, never had three or more successive numbers in successive positions, and was never repeated per subject.

Results

Table 1 shows the probability, *p*(Wide), of choosing the wide side given the corresponding wide-gap and narrow-gap memory loads. As seen in the table, *p*(Wide), whose mean value was .62, was greater than .5, $t(39) = 2.99$, $p < .01$, indicating that there was a preference for the wide gap. In addition, *p*(Wide) increased as the narrow-gap list length increased relative to the wide-gap list length.

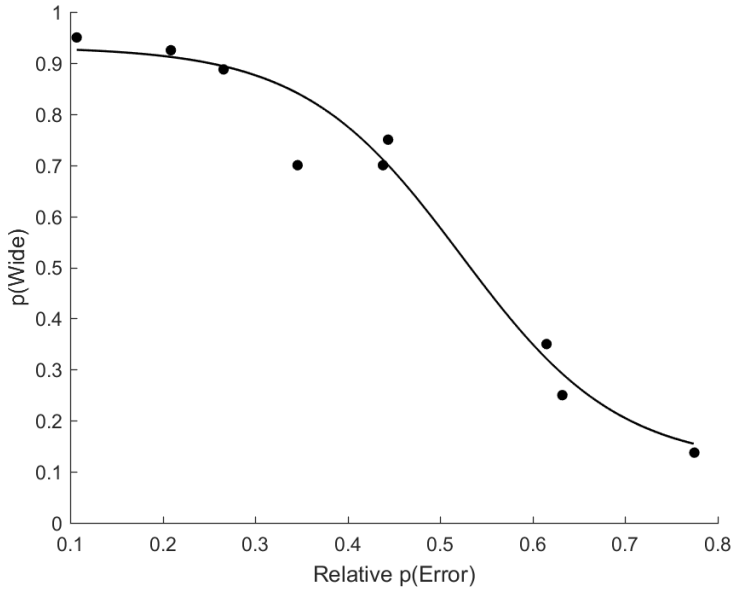
Table 2. 1. Probability, *p*(Wide), of choosing the wide gap in Experiment 1 when the wide gap had 6, 7, or 8 memory digits and the narrow gap had 6, 7, or 8 memory digits.

	Narrow Gap		
Wide Gap	6	7	8
6	.70	.92	.95
7	.35	.75	.89
8	.14	.25	.70

Regarding errors, Figure 2.3 shows *p(*Wide*)* as a function of the relative probability of error of any kind in the wide gap relative to the narrow gap. As seen in Figure 2.3, *p(*Wide*)* decreased as the relative probability of error increased. In general, subjects picked tasks that reduced errors. They did so in a manner consistent with Luce's choice axiom (Luce, 1959), according to which, as applied here, the probability of picking an easier task was 1 minus that task's difficulty divided by the sum of that task's difficulty plus the other task's difficulty. Consistent with the Luce choice axiom, the probability, *p*(Wide), of picking the easier navigation task (passing through the wide gap), decreased as the memory load for the wide gap grew relative to the memory load for the narrow gap. This statement follows from the fact that narrow-gap difficulty and

wide-gap difficulty were constant (or were assumed to be constant to arrive to Figure 2.3), so the only terms that distinguished the values along the abscissa were the wide-gap and narrow-gap memory loads. Overall, the average value of *p*(Wide) was .62, so the narrow gap was chosen, on average, on 38% of the trials. In terms of the individual participants' contributions, of the 40 participants, 28 (70%) made choices consistent with gradations of difficulties. Another set of 10 participants (25%) always picked the gap with the lower memory load when such a choice was possible (i.e., when the two memory loads differed). For these 10 participants, when two memory loads were the same, 8 of them chose the wider gap more often than the narrow gap and 2 of them chose the wide gap as frequently as the narrow gap. The remaining 2 participants (5%) always picked the wide gap.

Figure 2. 3. Probability, *p*(Wide), of selecting the wide gap as a function of relative probability of error of any kind in Experiment 1.



Note: Relative *p*(Error) was defined as *p*(Error) in the wide gap divided by the sum of *p*(Error) in the wide gap and *p*(Error) in the narrow gap.

The graph in Figure 2.3 did not break down the two kinds of errors. These are listed in Table 2, which shows, for the present experiment and also for the experiment to come, the number of times, *N*, subjects chose a combination of gap width and list length, as well as two additional probabilities: the probability, *p*(R), that subjects made a recall error, and the probability, *p*(B), that subjects bumped into the pointer. Note that the six conditions in Table 2 reflect the choices drawn from the nine conditions in Table 1. In other words, Table 2 shows how often each combination of gap width and digit length was chosen, irrespective of the *pair* of gap widths and digit lengths from which it was chosen.

Table 2. 2. Number of times, N, the wide gap was chosen, the probability, *p*(R), of recall error, and the probability, *p*(B), of gap-clearance error (bumping into the pointer) in Experiment 1 (with choice, total N = 720) and Experiment 2 (without choice, total N = 480) when the wide or narrow gap had 6, 7 or 8 digits to be memorized.

	Experiment 1			Experiment 2		
Condition	N	*p*(R)	*p*(B)	N	*p*(R)	*p*(B)
Wide-6	206	.05	.00	80	.05	.00
Wide-7	159	.17	.00	80	.20	.00
Wide-8	87	.37	.00	80	.35	.00
Narrow-6	145	.06	.05	80	.12	.05
Narrow-7	86	.22	.01	80	.23	.03
Narrow-8	37	.48	.02	80	.38	.03

As seen in Table 2, *N* was larger for wide gaps than for narrow gaps and decreased as list length grew. Table 2 also shows that *p*(R) grew with list length and was larger for the narrow gap than for the wide gap.

To evaluate the latter outcome statistically for Experiment 1, we conducted a Generalized Estimating Equations (GEE) analysis that tested the effects of list length (6, 7, or 8) and gap width (wide or narrow) on *p*(R), with $\alpha = .05$. We also used GEE to test the effects of list length (6, 7, or 8) and gap width (wide or narrow) on p(B), again with $\alpha = .05$, as described below.

Before turning to these analyses, we offer a few remarks about GEE in general because it is not widely used in our field. GEE was attractive for the analyses of *p*(R) and *p*(B) for several reasons. First, it can handle missing data better than traditional ANOVAs (see Duenas, et al., 2016). We had missing data in the case of participants who entirely avoided combinations of gap widths and digit lengths. Second, GEE, unlike ANOVA,

does not assume normality of residuals or homogeneity of variance. Both of these assumptions were violated here, where our *p*(R) and *p*(B) values were based on aggregates of binary data; for each trial, participants, in effect, contributed a 0 or 1 to each choice option. Third, GEE circumvents technical problems surrounding co-variances in repeated measure ANOVAs (see Ballinger, 2004).

Given this backdrop and based on statistical guidelines for how to conduct GEEs offered by the above-named authors, we analyzed the *p*(R) and *p*(B) data using IBM SPSS Statistics 22, which includes GEE as an option. Links must be set when running GEE in SPSS and we used GEE's binary logistic link for the *p*(R) data and GEE's linear link for the *p*(B) data. We could not use the binary logistic link function for *p*(B) because *p*(B) equaled zero for the wide gap and a logistic function can never reach, though it can approach, 0 or 1. Using the linear link function made the analysis of binary data less likely to capture the S-shaped logistic data pattern, so when we talk about the independence of physical and mental resources, we will rely more on the effect of physical demands on *p*(R) than cognitive demands on *p*(B).

The GEE results were as follows. For *p*(R), the analysis yielded a significant main effect of list length, Wald Chi-Square = 78.08, $p < .001$. The estimated marginal mean for the 6-digit list (M=.06, 95% CI [.04 .09]), was lower than for the 7-digit list (M=.20, 95% CI [.14 .26]), which in turn was lower than for the 8-digit list (M=.44, 95% CI [.34 .53]). This GEE analysis also yielded a non-significant main effect of gap width, Wald Chi-Square = 1.19, $p = .275$. The estimated marginal means for the narrow and wide gaps

were .21 (95% CI [.15 .28]) and .17 (95% CI [.13 .22]), respectively. The interaction between list length and gap width was not significant, Wald Chi-Square = .384, p = .825. These results are consistent with the hypothesis that physical and mental demands had independent (additive) effects on p(R).

For p(B), the values were lower than for p(R). In fact, p(B) was zero in all of the wide-gap conditions, as seen in Table 2. Clearly, avoiding the pointer in the wide gap condition was very easy for our participants, as we fully expected based on pilot work, and avoiding the pointer in the narrow gap condition was easier than avoiding recall errors, as we also expected from pilot work. Nothing in our approach required that the overall level of task difficulty be the same for the physical and mental aspects of the task, though we understood from the outset that whatever conclusions we draw would necessarily be limited to the ranges of physical and memory performance we tested.

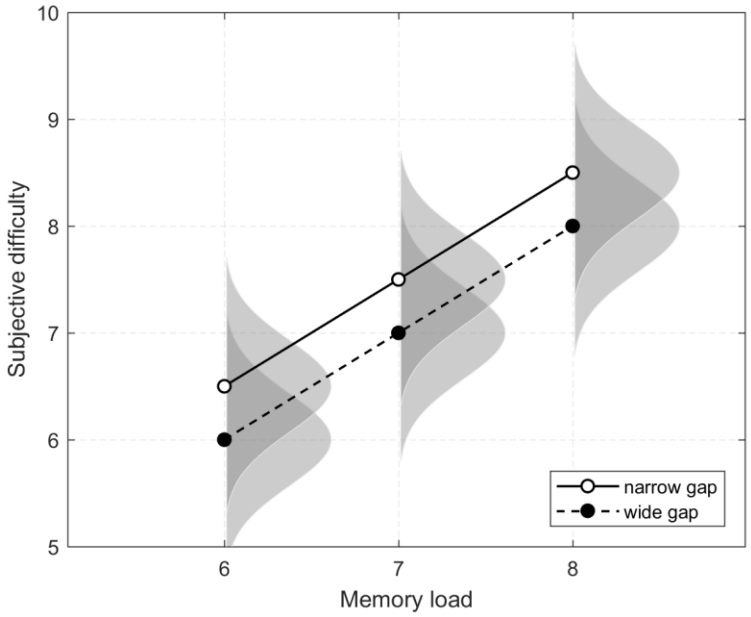
The GEE analysis for p(B), with α = .05, showed that p(B) was significantly affected by gap width, such that p(B) was larger for the narrow gap (M=.04, 95% CI [.01 .07]) than for the wide gap (M=0.00, 95% CI [.00 .00]), Wald Chi-Square = 5.37, p = .02, but p(B) was not affected by list length, Wald Chi-Square = 3.31, p = .19, and the interaction between list length and gap width was not statistically significant, Wald Chi-Square =3.30, p = .19. These results are consistent with the hypothesis that physical and mental demands had independent (additive) effects on p(B).

The analyses just reported used GEE to evaluate error probabilities. Next, we describe a further analysis to evaluate choice probabilities. The aim was to determine the

likelihood of the choice-probability data given different possible scenarios for how the choices may have arisen. Specifically, we sought to test the two hypotheses shown in Figure 2.2 concerning additivity (left panel of Figure 2.2) or interactivity of the choices (right panel of Figure 2.2).

The conceptual basis for the test is shown in Figure 2.4. Here, for sake of illustration, we focus on the additive case. For this case but also for the interactive one, we assumed that the subjective difficulty of any given task (i.e., any given combination of digit length and gap width) could be characterized by a normal distribution with mean μ and standard deviation σ such that the probability of perceiving one task as easier than the other was based on the degree of overlap between the two tasks' distributions. We assumed that μ increased linearly with the number of digits and that for the wide gap, μ increased with a slope of 1 and an intercept of 0. We made these assumptions about the wide-gap line just to have the wide-gap difficulty function serve as the reference line for testing competing models about the narrow-gap function.

Figure 2. 4. Basis in the model for judging the relative difficulty of the tasks.



Note: Subjective difficulty for any given task is subject to normally distributed variability. The overlap between the distributions determines the probability of choosing one task or the other. This figure illustrates the additive model because the two subjective difficulty lines are parallel. It happens that the separation between the narrow-gap and wide-gap lines, which is the same for all memory loads, corresponds to the separation associated with the model that maximizes the likelihood of the observed choice data.

According to the additive model, the slope, θ, of the narrow-gap line would be the same as for the wide-gap line (i.e., $\theta = 1$), but the zero-intercept would be greater by an amount $\pi > 0$. According to the interactive model, θ would be different from the wide-gap slope (i.e., $\theta \neq 1$), subject to the constraint that the resulting intercept was non-negative. We also assumed that σ was a linear function of θ, $\sigma = k_1 + k_2\mu$, with $k_1 \geq 0$ and $k_2 \geq 0$. It was important to check k_2 because if we obtained evidence for the additive

model (i.e., $\theta = 1$), by allowing k_2 to vary, we could say that the support of the additive model was not just an artifact of forcing σ to stay constant. We varied π, θ, k_1, and k_2 to see which values allowed for the best simulation of the choice probability data. The predicted choice probabilities were based on the signed differences between the theoretical subjective difficulty values. All of the predicted choice probabilities occupied the range $\{0, 1\}$.

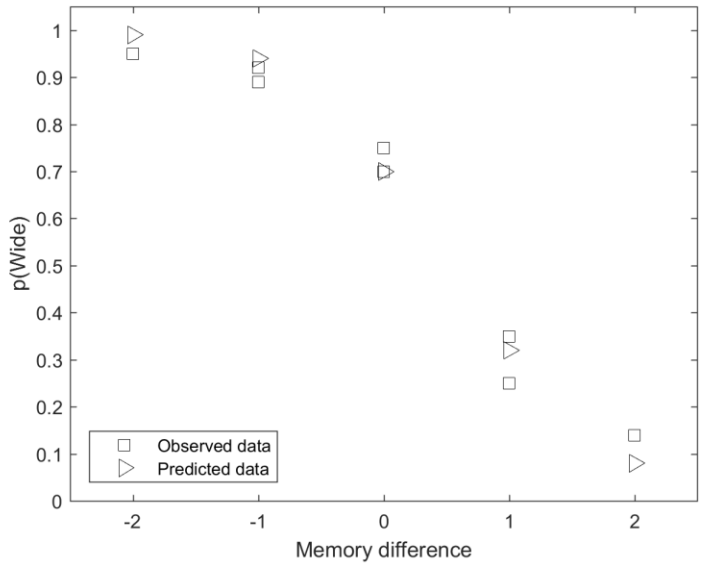
The model did not take into account the random effect of subjects. Instead, we pooled the subjects' data. Such pooling is justified if the data over subjects are homogeneous. In fact, if the subjects' data are homogeneous, then taking subject random effects into account increases the chance of overfitting, and it is advisable to pool the individuals' data (Smith & Batchelder, 2008). Following this advice, we checked the homogeneity of our subjects' data by calculating the homogeneity in all nine conditions using the method that Smith and Batchelder (2008) recommended, namely, counting the number of times each participant chose the wide gap and the narrow gap and put these numbers in the two columns of an $N \times 2$ contingency table. Here, the N rows were for the N=40 subjects. We did so for each of the 9 conditions of Table 1 separately. To test the homogeneity among participants in the contingency tables, we applied Fisher's Exact Test, setting α to .05/9 = .006 as per Bonferonni correction. Only one of the conditions was non-homogeneous by this criterion, so we concluded, for the sake of our data-analysis approach, that our participants were mainly homogeneous in their choice strategy. The fact that one case was not is a limitation of our study.

Figure 2.5 shows the best-fitting results, where the associated model was the one that minimized deviance, defined as $-2 \times \log(likelihood)$ of the data given the model. The likelihood,

$$L = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}$$

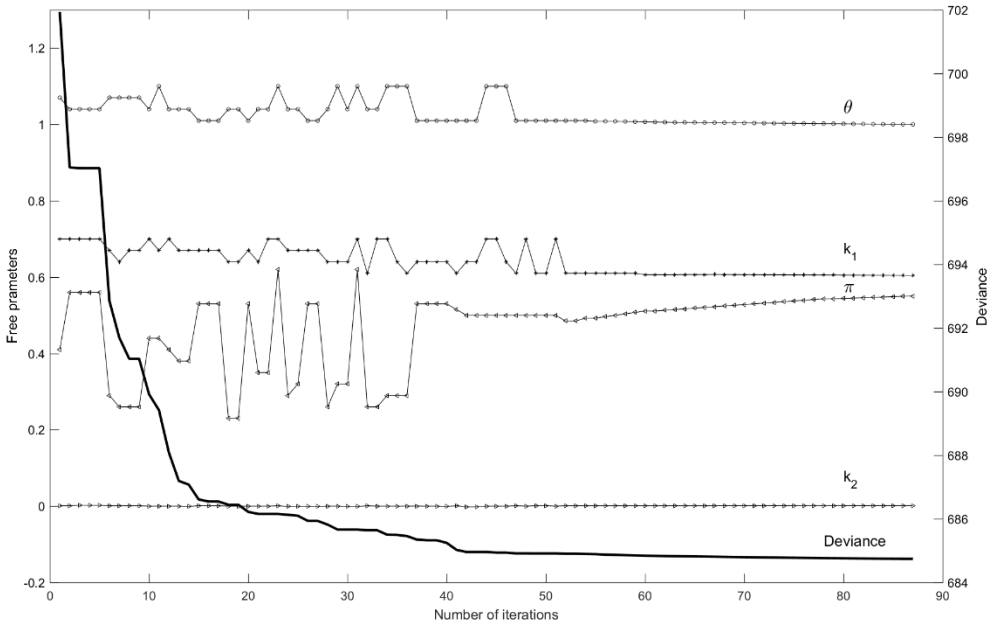
was the product over the N=720 trials of the predicted probabilities, *p*, of choosing the wide side in the *i*th trial, where y_i was 1 or 0 when the choice made in the *i*th trial was either wide or narrow, respectively. To find the best parameters we used grid search (see Figure 6) with the initial values of $\pi_0 = .41$, $\theta_0 = 1.07$, $k_1 = .7$, and $k_2 = .003$ with the step of $\pi_{step} = .03$, $\theta_{step} = .03$, $k_{1\_step} = .03$, $k_{2\_step} = .001$. The initial values were selected in a range that the model would not give an infinitely large deviance. When the deviance did not improve more than .0001 in 10 successive iterations the model stopped searching. Based on this approach, the parameter values for the best (smallest deviance) model were $k_1 = .60$, $k_2 = 0.00$, $\pi = .55$, and $\theta = 1.00$.

Figure 2. 5. Observed and predicted choice probabilities.



Note: The predicted values come from the model that maximized the likelihood of the data.

Figure 2. 6. Step search to find the values for the free parameters that minimized the deviance.



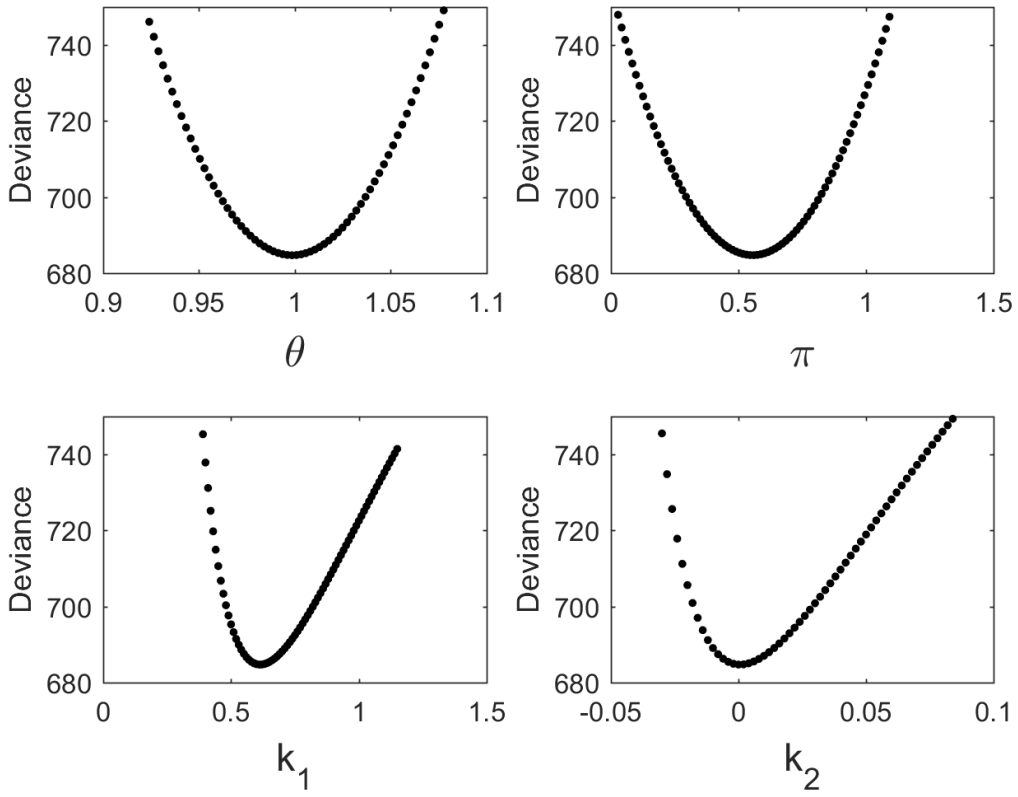
Note: iterations that led to a better deviance are shown.

To compare the additive vs interactive models we compared the deviance of a model with four free parameters (π, θ, k_1, k_2) to the deviance of a model with three free parameters (π, k_1, k_2). The first represents the interactive model. The second represents the additive model. The deviance for the model with four free parameters was 684.74 and so was the deviance for the model with three free parameters. More specifically, the deviance for interactive and additive models were 684.743383 and 684.743914 respectively. Therefore, the deviances for the two models were equivalent, at least to three significant figures. Accordingly, the simpler model, the one with three parameters

(the additive model) was preferable to the more complicated model, the one with four parameters (the interactive model). Note that we also have greater confidence in the additive model because the best value of k_2 was zero even though we let k_2 vary. Had we prevented σ from growing with μ, (i.e., locking k_2 at zero rather than letting it vary) that could have favored the interactive (super-additive) model.

Figure 2.7 provides still more information related to the model fit. Here we show deviance as a function of changes in π, θ, k_1, and k_2. Our aim in showing the deviance curves is to document the robustness of our model fit (Young & Holsteen, 2017). The values with minimum deviance are the ones mentioned above.

Figure 2. 7. Deviance as a function of values in the model's free parameters



Note: Deviance as a function of values in the model's free parameters. The minima are associated with θ = 1, π = .55, k_1 = .62, and k_2 = 0. The curves were obtained by first finding the best values when all four parameters could freely vary and then by letting the parameter named in each panel vary systematically while the three of the other parameters were fixed at their best values.

Discussion

Much as previous research has shown that people can provide orderly data when they make cross-modal perceptual judgments (Marks et al., 1986; Pitts et al. 2016), the experiment just reported has shown that people can provide orderly data when they make cross-modal task-difficulty judgments. Our participants made rational choices with respect to task difficulty, as seen in the inverse relation between *p*(Wide) and probability of error (Figure 2.3). In terms of the bases for the errors, the two kinds of error apparently arose from independent sources—one giving rise to cognitive errors, *p*(R), and one giving rise to physical errors, *p*(B). A way to characterize the choice process in descriptive terms is to say that the data were most likely to have stemmed from a choice process in which the extra difficulty of passing through the narrow gap compared to the wide gap was the same as the extra difficulty of memorizing an extra $\pi = .55$ digits. This was true for all of the memory loads used here. The choice data appear to have been based on consideration, explicit or implicit, of the two sources treated as independent rather than interactive factors.

Three further remarks are worth making in connection with Experiment 1. One concerns the role of practice. It is possible that participants underestimated or overestimated the difficulty of the tasks before they experienced the tasks and then changed their evaluations after the tasks were done. If this were true, the lack of exposure or benefit of exposure might have biased the outcome one way or another. To test this possibility, we divided the data into two halves (the first nine trials and the second nine

trials) and ran a GEE analysis on *p*(Wide) as a function of this factor as well as the nine conditions of Table 1. The analysis showed that the *p*(Wide) pattern was not significantly affected by practice, Wald Chi-Square = 5.72, $p = .67$. Therefore, we doubt that the choice results were mainly affected by practice.

Second, having shown that people can relate mental and physical task difficulty and make choices of mental-physical task pairs that reduced the sum of the two kinds of error, does it follow that they had some amodal representation of task difficulty? Our answer is no. It is possible that some mapping existed (or exists) between the difficulty values on the two putative dimensions of mental and physical difficulty. Such a mapping could be posited without positing an extra "box" with amodal representations. There is likely to be a common abstract code for perception and action (Hommel et al., 2001; Prinz, 1990), but we have no basis for asserting (or denying) that there is a common abstract code for mental and physical difficulty. We simply remain agnostic on this issue.

Third and related to the last point, Witt and Sugovic (2013) have suggested that perception is independently influenced by four action-related factors: (1) the likelihood of success; (2) the energetic cost of action; (3) the benefits that accompany successful performance; and (4) the penalties associated with failure. Our main finding, that errors and choices were independently shaped by cognitive and physical factors, aligns with this perspective.

Experiment 2

The conclusions reached based on Experiment 1 were based on the assumption that participants in that experiment made choices that reduced errors. It is possible, however, that the direction of causation was the reverse. Rather than choosing sides based on trying to reduce errors, participants may have gotten better on tasks they chose more. Our estimates of the error probabilities in Experiment 1 may have been biased by differences in observed task frequencies.

To check this possibility in Experiment 2, we tested another group of participants who were told which task to do. Each task that was tested in Experiment 1 was now tested an equal number of times in Experiment 2. This let us get equal numbers of observations per task.

Method

There were six possible conditions: the three memory lengths for the narrow gap and the three memory lengths for the wide gap.

Participants

Forty new Penn State University undergraduates (31 women and 9 men, average age 19.3 years), took part for course credit after giving informed consent. The sample size was based on the number of participants in Experiment 1. The experiment was approved by the Penn State Institutional Review Board.

Procedure

The experimenter identified the gap and list of digits by saying “left” or “right.” Then the subject memorized the list, carried the box through the corresponding gap, set the box down on the target table, and tried to recall the list. A random half of the participants started with the narrow gap on the left, and vice versa for the other participants. The six conditions per left-right/wide-narrow arrangement were tested in a random order per subject. After completing the six trials for the six conditions, the experimenter reversed the left-right arrangement of wide and narrow gaps and the procedure was repeated. As in Experiment 1, if an error was made, the trial was rerun at most once. Only the data from the first passage were analyzed.

Results

The probabilities of error for Experiment 2 are shown in the right column of Table 2. A Generalized Estimating Equations analysis for these data showed a main effect of digit length on *p*(R), Wald Chi-Square = 50.08, $p < .001$, such that the estimated marginal mean for a 6-digit list (M=.08, 95% CI [.05 .13]) was lower than the estimated marginal mean for a 7-digit list (M=.22, 95% CI [.17 .28]), which was lower than the estimated marginal mean for an 8-digit list (M=.37, 95% CI [.27 .46]). The estimated marginal means for the narrow and wide gaps were .23 (95% CI [.17 .31]) and .16 (95% CI [.11 .23]), respectively. The main effect of gap width, Wald Chi-Square = 3.18, $p = .74$, and the interaction between memory load and gap width, Wald Chi-Square = 1.35, $p = .50$, were not significant. This result accords with the hypothesis that physical and mental demands had independent effects on cognitive error.

The same analysis on *p*(B) showed a main effect of gap width, Wald Chi-Square = 8.69, $p = .003$, such that the estimated marginal mean for the narrow gap (M=.04, 95% CI [.01 .06]) was bigger than the wide gap (M=0, 95% CI [0 0]). The main effect of digit length was not significant, Wald Chi-Square = .15, $p = .92$. The estimated marginal mean for a 6-digit list was M=.02, 95% CI [.00 .04]. It was M=.01, 95% CI [.00 .03] for a 7-digit list and M=.01, 95% CI [.00 .03] for an 8-digit list. The interaction between gap width and list length was not significant, Wald Chi-Square = .15, $p = .92$. This result accords with the hypothesis that physical and mental demands had independent effects on physical error.

Finally, we compared the overall error rates in the two experiments. A 2 (experiments) × 2 (gap width) × 3 (memory load) factorial GEE analysis showed that the main effect of experiment on *p*(R) was not significant, Wald Chi-Square = .35, *p* = .81, and neither was the main effect of experiment on *p*(B), Wald Chi-Square = .28, *p* = .59. None of the interactions for *p*(R) and *p*(B) was significant. The breakdowns were as follows. Regarding *p*(R), the Wald Chi-Square and the significance level for memory demand × physical demand, memory demand × group, physical demand × group, and the three-way interaction of memory demand × physical demand × group were .34 (*p* =.84), 3.12 (*p* =.21), .29 (*p* =.59), and 1.69 (*p* =.42) respectively. Regarding *p*(B), the Wald Chi-Square and the significance level for memory demand × physical demand, memory demand × group, physical demand × group, and the three-way interaction of memory demand × physical demand × group were 2.38 (*p*=.30), .97 (*p*=.63), .012 (*p*=.91), .96 (*p*=.61), respectively. The lack of interaction among these factors accords with the hypothesis that there were physical and mental demands on error rates that were unaffected by the context in which the errors were obtained (with the opportunity for choice, in Experiment 1, or without the opportunity for choice, in Experiment 2).

Discussion

The purpose of the second experiment was to check whether subjects in Experiment 1 made choices that reflected accurate knowledge of error probabilities. The alternative hypothesis was that the error probabilities in Experiment 1 reflected

differences in the frequencies of the chosen tasks. To test this hypothesis, we removed the element of choice in Experiment 2 and asked participants to perform every gap/list combination. All the gap-list combinations used in Experiment 1 were tested an equal number of times in Experiment 2. The error probabilities were statistically indistinguishable from those of the first experiment. This result suggests that subjects in Experiment 1 made choices that reflected accurate estimates of the difficulty of the tasks as indexed by the tasks' error probabilities, not that the error probabilities reflected differences in the frequencies of the chosen tasks.

General Discussion

Judging the subjective difficulty of different kinds of tasks is something people do all the time, yet there is scant knowledge about how such judgments are made. One might imagine that there is some common currency for the comparisons. Perhaps some metabolic or physiological resource indexes physical and mental effort (e.g., build-up of lactic acid in the muscles or depletion of glucose in the brain). The problem with this hypothesis is that after decades of research in pursuit of the physiological "holy grail" of what effort is, no single resource or small set of resources has been found (e.g., Cos, 2017; Morel et al, 2017). Even for physical tasks, where one might expect there to be a clear physiological index of fatigue, attempts to find it have failed (e.g., Enoka & Duchateau, 2008; Schoenmarklin & Marras, 1989).

Knowing this, we took a psychophysical approach to the problem, taking inspiration from earlier studies of cross-modal intensity estimation. We focused on the subjective difficulty of tasks that drew, to varying degrees, on cognitive and perceptual-motor skills. Instead of asking participants to match the difficulty of a physical task to the difficulty of a mental task, as in traditional cross-modal intensity estimation tasks, we used the 2-alternative forced choice (2AFC) procedure, asking participants to choose between tasks based on apparent task difficulty. This approach let us ask new questions, the most basic of which was whether people can provide systematic data when they choose between tasks varying with respect to more than one dimension. Beyond that initial question, we could ask whether participants would be able to reduce errors on both dimensions of interest and, if so, whether the participants would be able to do so in a way that suggested the errors were treated as independent or dependent. We found that the errors were treated as independent. The model of choice that we were led to was one that treated the subjective difficulty of cognition (memorization) and perceptual-motor skill (navigating a gap while carrying a box) as additive rather than interactive. In terms of the 2 × 2 table shown in the introduction, the cell that was best supported was cell *a*.

If the cognitive and physical resources were indeed independent, as just stated, then how could participants choose between the task options? Is there a problem saying that choices can be made though the underlying cost dimensions are independent? The answer is no.

Consider choosing between two hybrid cars. The cars might have two independent sources of fuel (electricity and gas), but the cars can still be compared with respect to factors that are mainly due to each of these resources—durability on one hand (mainly related to electric fuel nowadays) and pollution (mainly related to gas these days). The fact that choices can be made between the cars need not be taken to imply that the resources they rely on are dependent.

What factor or factors did our participants actually use to make their choices in Experiment 1? Would it suffice to say, for example, that error probability was the one factor that participants relied on? *A priori*, this is possible. The orderliness of the data in Figure 2.3 suggests that it could be. Considering the neural consequences of error, it is conceivable that error avoidance was the main determinant of choice, for it is known that, upon committing an error, event-related potentials show a fast-negative deflection, the Error Related Negativity (ERN). The ERN might signal the need for effortful control processes (Westbrook & Braver, 2016) or it might reflect negative affective responses to errors (Maier, et al., 2016). The greater the strength of ERNs, the greater the avoidance of acts that elicit them (e.g., Frank et al., 2005). Relatedly, Dunn et al (2017) showed that *anticipated effort* (effort perception before doing a task) can be well explained by expected chance of error. Dunn et al (2017) proposed a heuristic reasoning process for such anticipatory meta-cognitive evaluations.

Notwithstanding these arguments, we think an error-only account is unlikely to be correct. First, previous research has shown a dissociation between effort and error rate

(e.g., Kool et al., 2010). Second, a thought experiment indicates why it is questionable that error alone is unlikely to be the determinant of subjective difficulty. Consider two tasks people might consider: rolling a boulder up a hill, or pitching pennies to get the pennies to fall through remote holes whose diameters only slightly exceed the penny widths. The probability of error in both cases would be close to 1 but because rolling a boulder up a hill requires more energy and has a higher chance of injury, the subjective difficulty would probably be greater for boulder rolling. As this example shows, error probabilities alone don't fully capture effort.

What else could be used? One possibility is that more subjective weight might be attached to one kind of error than another. Perhaps in Experiment 1 more weight was attached to physical effort than to mental effort, or vice versa. From the observed error probabilities, it is impossible to distinguish between these alternatives.

If one thought that there is a straightforward way to resolve this issue by paying subjects to assign different weights to different tasks (Westbrook & Braver, 2015), one would still have the question of why the costs would need to differ to yield comparable task-choice probabilities (i.e., what the underlying difficulty metric is). Furthermore, if one thought that one could resolve the issue by simply asking subjects to estimate physical effort or mental effort, it turns out that people have a very hard time giving such estimates (Rosenbaum & Gregory, 2002). Indeed, this was the main reason why our lab went to the 2AFC procedure, which has proven extremely useful to us in over a decade of

work on action choices. For a review, see Rosenbaum, Chapman, Coelho, Gong, and Studenka (2013).

Yet another possibility is that more attention is needed for one task than the other (e.g., Dunn, Lutes, & Risko, 2016; Hasher & Zacks, 1979; Westbrook & Braver, 2015). Such a claim would need independent confirmation. The challenge would be to develop measures of attention (or cognitive resources) that are neutral with respect to the modalities of the task. Meeting that challenge is very difficult (e.g., Luck & Vecera, 2002). Adducing attention or cognitive resource raises as many questions as it answers and might be viewed, from at least one theoretical perspective (ecological psychology), as taking out on a loan on intelligence (Turvey, Shaw, Reed, & Mace, 1981).[1]

Could it be, to raise still another possibility, that the *sine qua non* of task difficulty is task completion time? This hypothesis has been proposed for some tasks (Gray et al., 2006; Potts, Pastel, & Rosenbaum, 2017), but has been ruled out for others (Kool et al., 2010). We are sympathetic to the challenge of Kool et al (2010), notwithstanding our lab's endorsement of time (or subjectively modulated time) as a

[1] An idea related to the attention hypothesis is that effort reflects the degree of cognitive control needed to mediate between available resources and task performance. The more cognitive control that must be allocated to a task, the more effortful the task is perceived to be (Shenhav et al. 2017). Shenhav et al. conceded that a challenge for this account is to have a measure of the cost of cognitive control that is independent of performance. The lack of an independent measure of effort or cognitive control – the mediating variable in their model – is what prevented us from pursuing a mediation analysis of our data, as one of the reviewers suggested. To pursue a mediation analysis (MacKinnon, Fairchild, & Fritz, 2007), one needs a variable M to mediate variables X and Y, but one also needs an independent way of measuring M, X, and Y. No such measure exists for effort. The lack of such a measure was the main driver of this investigation.

possible index of perceived task difficulty in another study (Potts, Pastel, & Rosenbaum, 2017). However, time reduction was certainly not the basis for decisions made by subjects in another experiment (Rosenbaum, 2012), where subjects walked long distances to avoid long reaches; their task completion times actually grew greatly as a result. Time may have been a factor in the two experiments reported here, as it was in the experiment of Potts, Pastel, and Rosenbaum (2017), but because we did not measure times in the present experiments we cannot rule out this possibility. Still, we doubt that time can be viewed as the only or main determinant of subjective difficulty in general, for the reasons given above.

A final possibility is that subjective task difficulty boils down to utility: task benefit minus task cost (e.g., Kurzban et al., 2013; Westbrook & Braver, 2015). We find this hypothesis intriguing but are hesitant to endorse it because, as with the other putative bases for subjective task difficulty, it is hard to independently say what the cost is.

These considerations lead us to endorse a more modest approach to the characterization of subjective task difficulty. The more modest approach is to pursue a descriptive rather than explanatory model (Lewandowsky & Farrell, 2011). A descriptive model can be used to provide estimates of functionally important parameters, which in turn can be tested in new experiments. Being able to make and test new predictions is coming to be recognized as an important, though surprisingly unappreciated, priority for experimental psychology (Yarkoni & Westfall, 2017), just as it always has been in physics and other "hard" sciences. Descriptive models can be used to posit and test

alternative hypotheses such as the two of primary interest here – one that had mental and physical difficulty treated as independent, and another that had mental and physical difficulty treated as interactive. Insofar as interactive models are classically taken to reflect a shared resource whereas independent models are classically taken to reflect distinct resources (Sternberg, 1969), deciding between the two models at a descriptive level sheds light on a deep theoretical question. From the data we have, we can say that the two subtasks appear to have been treated as independent, both in choosing between them and in their contributions to errors.

Whatever the theoretical basis for comparing physical and mental task difficulty, understanding how the comparisons are made or what choices are made about them is important for practical purposes. We end by pointing to an example that we found gripping when we learned about it. Yang et al. (2015) showed that doctors were loath to walk 5 meters to retrieve patients' accurate information from a computer. Instead, the doctors preferred to recall the information from memory. The chance of ending up with wrong information would have been much lower had the doctors been willing to walk, but the doctors made the choices they did (relying on memory) for whatever reason. This is a particularly dramatic example of the way that comparisons of physical and mental task may have important practical consequences. In the most extreme case, the consequences may literally be a matter of life or death.

We know too little at this stage to be able to make strong recommendations about how such decisions should be made, especially when life-or-death consequences may

arise. However, we hope the approach we have introduced here, which boils down to asking participants to choose between tasks of different kinds and modeling the underlying choice process, will enable others do the same and perhaps come away with confidence about the following four claims we wish to make from the present pair of experiments:

1. Systematic cross-modality difficulty comparisons can be made by human participants.

2. By collecting task preferences using the 2-AFC method along with measures of performance quality, one can model people's meta-cognitive beliefs about the ease or difficulty of performance.

3. In the conditions tested here, participants' choices and actual performance were congruent: Physical and mental demands appeared to be independent, both in their contributions to performance accuracy and in the decisions made about the relative ease or difficulty of the various physical/mental task combinations.

4. Finally, at the descriptive level and also at the phenomenological level, as that term is used in physics – see https://en.wikipedia.org/wiki/Phenomenology_(physics) – it appeared that the greater difficulty of passing through the narrow gap compared to the wide gap was subjectively equal to memorizing an extra $\pi = .55$ digits on average.

The last point leads to our closing remark. With the figure of $\pi = .55$ digits, one can go on to test hypotheses about the possible change in that estimate (and the other parameters of the model) accompanying other task variations, the nature of the

individuals making and carrying out the choices, training regimens, rehabilitation programs, drug treatments, and so on. The first step in doing good science is getting good measurements. We hope the new question raised in this article and the method developed to address it will help advance good science in the study of human perception and performance.

Author Notes

The experiments were conducted while both authors were at Penn State University. Assistance with data collection was provided by Connor Corrente, Yiyi Dai, Amanda Koch, Skylar Korek, Jennifer Norris, Veronika Onischenko, Emily Wolfskill, and Jenny Zhao. Analysis of the data and writing were supported by a UCR Committee on Research grant to the second author. The authors are indebted to those who reviewed the manuscript upon its submission for publication. Correspondence should be directed to Iman Feghhi (iemanifk@gmail.com) or David A. Rosenbaum (david.rosenbaum@ucr.edu), both of whom are at the Department of Psychology, University of California, Riverside, CA 92521.

References

Baddeley, A. D. (1976). *The psychology of memory*. New York, NY: Basic Books.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*, 127-150.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617-645.

Beilock, S. (2015). *How the body knows its mind*. New York, NY: Simon & Schuster.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Cos, I. (2017). Perceived effort for motor control and decision-making. *PLoS Biology*, *15*, e2002885.

Duenas, M., Salazar, A., Ojeda, B., Arana, R., & Failde, I. (2016). Generalized Estimating Equations (GEE) to handle missing data and time-dependent variables in longitudinal studies: an application to assess the evolution of Health Related Quality of Life in coronary patients. *Epidemiologia e Prevenzione*, *40*, 116-123.

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

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2017). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological Research*, 1-24

Enoka, R. M., & Duchateau, J. (2008). Muscle fatigue: what, why and how it influences muscle function. *Journal of Physiology, 586*, 11-23.

Franchak, J. M., van der Zalm, D. J., & Adolph, K. (2010). Learning by doing: Action performance facilitates affordance perception. *Vision Research, 50*, 2758-2765.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, *47*, 495-501.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think.* Cambridge, MA: Harvard University Press.

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

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *10*8, 356-388.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679-709.

Hommel, B., Müsseler, J., Aschersleben, G. & Prinz W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-937.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16*, 965-980.

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

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*, 661-679.

Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Los Angeles, CA: Sage.

Luce, R. D. *(1959). Individual Choice Behavior: A Theoretical Analysis.* New York: Wiley.

Luck, S. J., & Vecera, S. P. (2002). Attention. In Pashler, H., & Yantis, S., (Eds.), *Stevens' Handbook of Experimental Psychology: Vol. 1. Sensation and Perception* (pp. 235-286). New York, NY: Wiley.

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science, 25*, 47-53.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology,* 58, 593-614.

Maier, M. E., Scarpazza, C., Starita, F., Filogamo, R., & Làdavas, E. (2016). Error monitoring is related to processing internal affective states. *Cognitive, Affective, & Behavioral Neuroscience*, *16*, 1050-1062.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 517-534.

Morel, P., Ulbrich, P., & Gail, A. (2017). What makes a reach movement effortful? Physical effort discounting supports common minimization principles in decision making and motor control. *PLoS Biology*, *15*(6), e2001323.

Pitts, B., Riggs, S. L., & Sarter, N. (2016). Cross-modal matching: A critical but neglected step in multimodal research. *IEEE Transactions on Human-Machine Systems, 46*, 445-450.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action: Current approaches* (pp. 167-201). Berlin: Springer.

Rosenbaum, D. A. (2012). The tiger on your tail: Choosing between temporally extended behaviors. *Psychological Science, 23*, 855-860.

Rosenbaum, D. A. (2017). *Knowing hands: The cognitive psychology of manual control*. Cambridge University Press.

Rosenbaum, D. A., Carlson, R. A., & Gilmore, R. O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology*, *52*, 453-470.

Rosenbaum, D. A., Chapman, K. M., Coelho, C. J., Gong, L., & Studenka, B. E. (2013). Choosing actions. *Frontiers in Psychology*, Volume 4, Article 273, doi:10.3389/fpsyg.2013.00273.

Rosenbaum, D. A. & Gregory, R. W. (2002). Development of a method for measuring moving-related effort: Biomechanical considerations and implications for Fitts' Law. *Experimental Brain Research, 142*, 365-373.

Schoenmarklin, R. W. & Marras, W. S. (1989). Effects of handle angle and work orientation on hammering: II. Muscle fatigue and subjective ratings of body discomfort. *Human Factors, 31*, 413-420.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207-218.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99-124.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713-731.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276–315.

Turvey, M. T., Shaw, R. E., Reed, E. S., & Mace, W. M. (1981). Ecological laws of perceiving and acting: In reply to Fodor and Pylyshyn (1981). *Cognition, 9*, 237-304.

Weigelt, W., Rosenbaum, D. A., Huelshorst, S. & Schack, T. (2009). Moving and memorizing: Motor planning modulates the recency effect in serial and free recall. *Acta Psychologica, 132*, 68–79.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuro-economic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 395-415.

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron*, *89*, 695-710.

Witt, J. K. (2011). Action's effect on perception. *Current Directions in Psychological Science, 20*, 201-206.

Witt, J. K., & Sugovic, M. (2013). Spiders appear to move faster than non-threatening objects regardless of one's ability to block them. *Acta Psychologica, 143*, 284-291.

Yang, X. J., Wickens, C. D., Park, T., Fong, L., & Siah, K. T. (2015). Effects of information access cost and accountability on medical residents' information retrieval strategy and performance during prehandover preparation: Evidence from interview and simulation study. *Human Factors, 57*, 1459-1471.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100-1122.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*, 3-40.

CHAPTER 3 – EFFORT-AVOIDANCE ISN'T SIMPLY ERROR-AVOIDANCE

Iman Feghhi and David A. Rosenbaum
University of California
Riverside, CA 92521

Published in Psychological Research (2020)

Abstract

Little is known about how effort is represented for different kinds of tasks. Recently, we suggested that it would help to establish empirical benchmarks for this problem. Accordingly, Feghhi and Rosenbaum (2019) estimated how many additional digits to be memorized corresponded to navigating through a narrow gap versus a wide gap. The estimates were based on a study in which participants chose between walking paths with associated memory demands. We found that participants were equally willing to choose to walk through a narrow gap as to walk through a wide gap when the narrow-gap walk required memorization of .55 fewer digits on average than the wide-gap walk. In the present experiment, we sought to replicate and extend this previous finding in two ways: (1) by presenting the memory digits in auditory rather than visual form to test the hypothesis that participants used phonological recoding of the visually presented digits; and (2) by providing a new metric of the relative difficulty of navigation errors compared to recall errors. We provided 36 university students with two action/memorization options per trial and asked them to choose the easier option. Each option had varying degrees of physical demand (walking through a wide or narrow gap) and mental demand (memorizing 6, 7, or 8 digits). We expected performance to be comparable to what we

observed earlier with visually presented digits to be memorized, and this prediction was confirmed. We also used a new metric to show that navigation errors were implicitly judged to be 17% more costly than recall errors. The fact that this percentage was not 0 indicates that reducing percent error was not the only basis for reducing effort.

Keywords: Effort, Metacognition, Subjective difficulty

Introduction

On what basis does one judge the effort of a task? This question has been hard to answer. Different theories have been proposed, positing alternative variables that might be used, including time (Gray et al., 2006; Potts, Pastel, & Rosenbaum, 2018; Rosenbaum & Bui, 2019), energy (Craig, 2013; Job, Dweck, & Walton, 2010), attention (Kool et al., 2010), opportunity-cost (Kurzban et al., 2013), sustainability (Rosenbaum & Bui, 2019), and error avoidance (Dunn, Inzlicht, & Risko, 2019). It would be desirable to identify a single index of effort, but doing so has not occurred yet.

It is worth considering each of the candidates listed above as a prelude to the present investigation. Time is one candidate. Tasks that take longer might be judged to be more difficult or effortful (we use the two terms interchangeably) than tasks that take less time. Data consistent with this view have been reported by Gray et al. (2006), Potts, Pastel, and Rosenbaum (2018), and Rosenbaum and Bui (2019). Still, it is unlikely that time alone can always predict effort. Raising and lowering a heavy weight in a short amount of time (at a higher frequency than the resonant frequency of the mass-spring system) takes more effort than raising and lowering that same heavy weights over the same amplitude in a longer amount time (at a lower frequency closer to the resonant frequency of the system).

Energy may be a viable candidate, especially in view of energetic considerations like those just intimated. Energy has in fact been posited to be a determinant of effort (Craig, 2013; Job, Dweck, & Walton, 2010). Craig (2013) suggested that there may be a part of the brain that registers energy consumption: the anterior insula. Still, the term

"energy" applies more naturally to physical tasks than to more intellectual tasks. For intellectual tasks, the term "energy" has more of a metaphorical than independently verifiable meaning.

The same sort of concern applies to attention or cognitive control (Kool et al., 2010). It is intuitive that more effortful tasks require more attention or cognitive control than less effortful tasks but independently verifying how much attention or cognitive control is needed for a task is often bedeviled by the fact that the means of assessing attention or cognitive control is almost always tied up with the means of carrying out the primary task of interest. Dual task methods are notoriously difficult and the attendant results are correspondingly hard to interpret (Pashler, 1994).

Another hypothesis is that effort amounts to opportunity cost (Kurzban et al., 2013), the reduced utility of one action compared to others. The opportunity cost hypothesis may explain a variety of behaviors, but it is unclear how the cost and benefits contributing to the utility of any given action are measured in the first place. The problem is especially acute when the possible actions are of different sorts, such as navigation versus memorization.

Sustainability is a recently suggested candidate for indexing effort. It was recently proposed that the difficulty of a task could be indexed by how long the task can be continued. The longer it can be continued, the less effortful it will be judged to be (Rosenbaum & Bui, 2019). This hypothesis can potentially provide an explanation of people's ability to compare the difficulty of different kinds of tasks, which they can in fact do reliably (Feghhi & Rosenbaum, 2019; Potts et al., 2018; Rosenbaum & Bui,

2019). The sustainability may also explain how it is possible for people to compare the difficulty of tasks whose costs seem indistinguishable when the possible tasks are done just once or very few times. For example, a short reach to a large target was consistently judged to be easier than a slightly longer reach to the same large target when participants were invited to perform whichever task seemed easier (Rosenbaum & Gaydos, 2008). Relying on expected sustainability could provide a basis for such decisions. Imagining how well each task could be performed many times might amplify whatever tiny difference in difficulty exists between them; more samples could provide more accurate estimates (a basic tenet of sampling theory in statistics).

To test the sustainability hypothesis, Rosenbaum and Bui (2019) asked participants to judge the sustainability of a cognitive task and a physical task. Given a description of the task and the number of times it would, hypothetically, have to be performed, participants indicated whether they thought they could do the task that many times. As the number increased, the probability of saying "yes" decreased, and it did so at different rates for different tasks. Other participants were asked to indicate which of the tasks, done just once, would be easier. The positive finding was that the tasks that were judged less sustainable were also judged to be more difficult in the one-task case. This outcome corroborated the sustainability hypothesis. Nonetheless, Rosenbaum and Bui found that times to perform the tasks did a better job explaining the preferences than the sustainability judgments. The status of the sustainability hypothesis is therefore up in the air.

Last in the aforementioned list of possible determinants of task difficulty is the likelihood of error. According to the error hypothesis, tasks that are more apt to lead to error should be judged harder than tasks that are less apt to lead to error. That expectation has been confirmed (Dunn, Inzlicht, & Risko, 2019).

Of course, not all tasks yield explicit error measures. As long as the subjective registration of error is unclear, so too must be the claim that error alone is the determinant of subjective difficulty. That said, in cases where tasks *do* yield explicit error measures, objectively recorded and reported proportions of errors of any kind should index effort to the same degree. That would be the clear prediction of a hypothesis which states that effort avoidance is error avoidance. (The title of this article indicates where we end up on this.)

We tested this prediction in the present study. The context in which we did so was to extend a series of studies in which we found that judgments about the relative difficulty of intellectual tasks paired with physical tasks are highly reliable (Feghhi & Rosenbaum, 2019; Potts, Pastel, & Rosenbaum, 2018; Rosenbaum & Bui, 2019). In pursuing this line of inquiry and considering the absence of a clear winner among the candidates for a single measure of effort or difficulty, we suggested that it would be worth taking a more descriptive approach, seeking quantitative equivalencies between the subjective difficulty of different tasks. Here we drew inspiration from previous work on multi-modal perception, thinking, for example, of the classic work by Stevens and Marks (1965) comparing the subjective magnitudes of different sorts of stimuli such the lumens of light and the loudness of sounds. We thought that the establishment of such

equivalencies here could undergird predictions for future experiments, including ones designed to test predictions about specific numerical values for factors of significance. For example, the method could be used to equate the difficulty of walking through gaps of different width and the difficulty of memorizing varying numbers of items. Surprisingly, psychological research has seen less work of this kind than one might expect (Yarkoni & Westfall, 2017).

We pursued this approach in a previous study (Feghhi & Rosenbaum, 2019). There we showed that the subjective difficulty of navigating through gaps of different size could be compared to the subjective difficulty of memorizing lists with different numbers of digits. The basis for comparison was not effort ratings, which would be the analogs of magnitude estimates à la Stevens and Marks (1965), but rather 2-alternative force choice tasks. We simply asked participants to do what they thought was easier, one task or the other, similar to what was done by Rosenbaum and Gaydos (2008) and then done in other similar studies; for a review, see Rosenbaum et al. (2013). The general approach in these studies has been to vary features of the two tasks to see how the probability of choosing one depends on the properties of the other.

Participants in the main experiment of Feghhi and Rosenbaum (2019) were asked to choose between passing through a wide gap or a narrow gap. In each trial, each gap had an associated number of random digits to be memorized. The number of such digits was 6, 7, or 8. All possible combinations of gap width and number were given, with the wide or narrow gap appearing equally often on the left or right. The data we obtained included such things as the probability of choosing the narrow gap when 6 digits had to

be memorized rather than choosing the wide gap when 8 digits had to be memorized. We used the full set of such probabilities, pooled over participants because each participant did each choice condition once, to draw inferences about participants' beliefs (explicit or implicit) about task difficulty.

We found that the choice probabilities could be explained with a model that treated the subjective difficulty of passing through the narrow gap rather than the wide gap as equivalent, in terms of subjective difficulty, to memorizing an extra .55 digits. The specific method will be replayed here, so details of its use will be given below. We also found that there was no interaction between error rates for navigation (bumping into a pointer in the middle of the workspace) and error rates for memorizing (misremembering numbers). That result we took to mean that the resources for navigation and memorizing were independent.

Experiment 1

In the present experiment, we sought to replicate and extend the previous study. Whereas in our earlier study, we let participants *see* the memory lists for the two alternative navigation routes and then choose the memory-navigation task that they preferred, in the present experiment, we changed the modality used to present the to-be-memorized digits. Rather than showing the two possible memory lists, we read aloud the memory lists. We told the participants how many digits (6, 7, or 8) would be associated with each navigation task (going through the wide gap or narrow gap), and we asked participants to indicate which navigation/memorization task they thought would be

easier. Once they made their choice, we read aloud the list to be memorized. As before, participants carried an empty box through the wide or narrow gap, having memorized, or attempted to memorize, the digit list for the chosen side; see Figure 1. They were to pass through the gap without touching a pointer that would move to a new orientation if it were bumped, and they were to recall the list once they brought the box to the platform beyond the gaps and centered between their inner edges, directly ahead of the participant's start position for each trial.

We expected to replicate our results with the auditory presentation modality because we hypothesized that participants in the earlier study encoded the digits phonologically both because that typically occurs in reading (Conrad & Hull, 1964; Posner & Mitchell, 1967) and also because it made sense to use phonological coding here insofar as it would reduce the interference effects of relying on a visual representation of the digits to be memorized in the midst of a visuo-spatial navigation (Baddeley, 1976).

In seeking to determine whether we would replicate our previous results, we sought to do so in a way that is more specific than what typically happens in psychological research. Rather than checking that a *qualitative* result was replicated, we sought to determine whether we would replicate a particular *quantitative* result. We sought to determine whether we would get the same numerical estimate of subjective difficulty as before. In our earlier work, we found that going through the narrow gap was functionally equivalent to memorizing an extra .55 digits. We wanted to know whether we would obtain the same estimate here, noting that in the physical and natural sciences, empirical constants are touchstones for the science. In physics, there is Planck's constant,

the gravitational constant on Earth, and so on. Testing for the constancy of empirical parameters is a useful enterprise for a mature science, and this approach has been advocated for psychological research by Yarkoni and Westfall (2017), among others, such as Cavanagh (1972), who showed that the time to search through the full span of short-term memory no matter what the contents, is close to .25 s. Cavanagh demonstrated this by showing that the rate of memory scanning in the Sternberg (1966) item recognition task is proportional to the memory span for the various kinds of items he studied (e.g., words versus colors). Cavanagh's constant is a benchmark for psychological research. Having more such quantities can help ground work in the field.

Method

Participants

Forty Penn State undergraduate students (28 female and 12 male) participated. Four participants had to be dropped because they showed no variability in their choices; they always picked the wide gap. We omitted their data because their data lacked sufficient variability to let us evaluate differential tradeoffs between navigation demands and memory demands.

The ages of the retained participants ranged from 18 to 23 years (M = 20.17 years, SD = 1.10 years). The number of participants was selected to ensure that we would have enough statistical power for the analysis based on recommendations of Cohen et al. (2013) and Hosmer et al. (1997). The experiment was approved by the Penn State

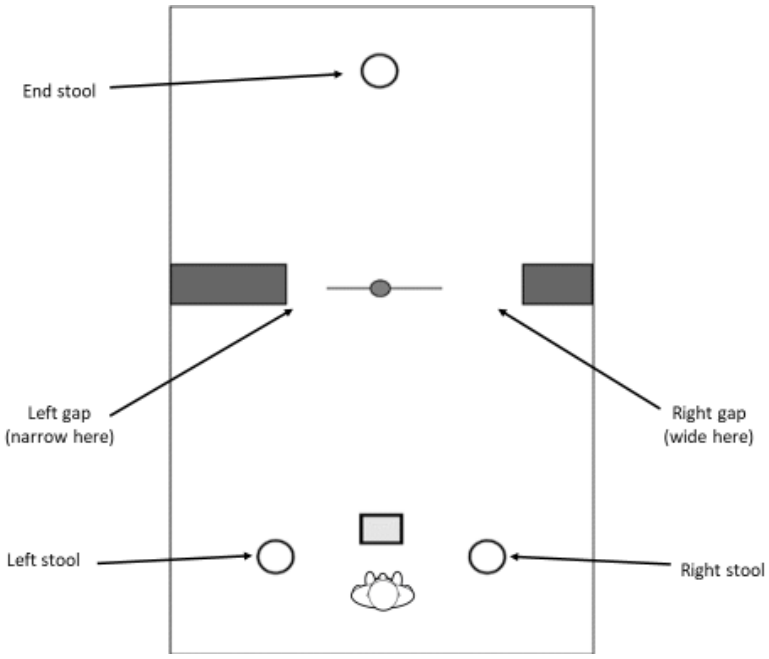
Institutional Review Board. All participants signed the informed consent form before participating in the experiment.

Apparatus

As shown in Figure 1, at the start of each trial, the participant stood at a home position and faced the apparatus. An empty box (35.56 cm × 35.56 cm × 35.56 cm) stood on a platform within reaching distance (33 cm away from the home position) and at waist height for all participants (76 cm high). We used a box because we planned, in future experiments, to explore the effects of carrying a load, and this box or one just like it, would be the container for that load. There was a 63 cm high stool 100 cm, and another 63 cm high stool 100 cm to the left of the box-bearing platform. A card was placed on each of these stools with a number (6, 7, or 8) representing the number of random digits that would have to be memorized if that side were chosen.

The subject could see the gaps, whose centers were 184 cm from the home position. One gap was wide (81 cm). The other was narrow (36 cm). Participants could easily walk through the wide gap, but to clear the narrow gap, they had to turn sideways. Between the gaps was a 98 cm high stand on which was mounted a light, easy-to-see wooden stick. This stick, or pointer as we called it, extended 23 cm into the gaps and turned if it was touched, thereby indicating a physical mistake. The pointer had low fraction and typically turned about 45 degrees when touched. We did not attempt to measure how much the pointer turned if it did.

Figure 3. 1. Schematic overhead view of the experimental setup.



Note: A participant stands at the home position facing the apparatus. Cards with numbers on the left and right stools are not shown.

Procedure and Design

Standing at the home position, participants were able to see an empty box in front of them, two numbers on the stools to their right and left, two gaps (184 cm away from the home position), and a final stool (184 cm away from the gaps). Participants were instructed to choose the side that seemed easier. If they choose the right side, they had to memorize the digit list associated with the right side and carry the box through the right gap. If they choose the left side, they had to memorize the digit list associated with the left side and carry the box through the left gap. We avoided any reference to ease of memorizing versus ease of navigation. Participants were simply told to choose whichever task seemed easier, the one associated with the left side or the right side, and to indicate

their choice by saying left or right. The experimenter then read the memory list for the designated side. If the participant requested, the experimenter read the list again; this repeated as often as wanted by the participant. The experimenter read the digits at a normal conversational pace (about two digits per second). After memorizing the list and carrying the box through the gap on the chosen side, the participant placed the box on the end table and then recalled the memorized list. The participant was told that if s/he made a mistake in recall and/or navigation (hitting the pointer). The participant was told beforehand, in the general instructions for the experiment, that if a mistake was made, the task would have to be repeated. If a mistake was made again, in the repeat trial, the trial was not repeated further. Mistakes in the repeat trial were not analyzed.

A random half of the participants started with the wide gap on the right and the narrow gap on the left for the first nine trials, with that arrangement then reversed for those subjects. The other participants were tested in the opposite order.

Results

Error Rates

To analyze the errors, we used Generalized Estimating Equations (GEE) (Duenas, et al., 2016). This approach does not require equal observations per condition, and it allows for multiple observations per condition with binary data. To take advantage of this approach, we used Generalized Estimating Equations in SPSS. We analyzed the probability, p(Error), of mistakes of any kind, as well as the constituents of the mistakes, namely, the probability, p(R), of recall error and the probability, p(B), of navigation

error; “B” stands for bumping into the pointer. Table 1 shows all three probability measures as well as the total number of trials, N, in each condition.

Table 3. 1. Number of trials, N; probability, *p*(Error), of error of any kind; probability, *p*(R), of recall error; and probability, *p*(B), of navigation error (bumping into the pointer). Data from each of the six conditions of Experiments 1 and 2.

	Experiment 1				Experiment 2			
Condition	N	*p*(Error)	*p*(R)	*p*(B)	N	*p*(Error)	*p*(R)	*p*(B)
Wide-6	223	.10	.10	.00	80	.14	.14	.00
Wide-7	164	.21	.21	.00	80	.25	.25	.00
Wide-8	119	.47	.47	.00	80	.34	.34	.00
Narrow-6	120	.16	.13	.04	80	.21	.14	.09
Narrow-7	68	.32	.28	.08	80	.31	.26	.08
Narrow-8	26	.62	.62	.00	80	.47	.45	.04

Note: Wide and Narrow refer to the gap width. The numbers 6, 7, and 8 refer to the number of digits to be memorized.

The GEE analysis showed that there was a main effect of gap width, Wald Chi-Square = 10.89, $p = .001$, and list length, Wald Chi-Square = 61.94, $p < .001$, with a nonsignificant interaction between the two. With respect to gap width, *p*(Error) had a higher mean value when participants went through the narrow gap ($M = .35$, 95% CI [.27 .44]) than when participants went through the wide gap ($M = .23$, 95% CI [.18 .29]). With respect to list length, *p*(Error) had a higher mean value the greater the length of the memory list, $M_{6\ digit} = .13$, 95% CI [.09 .17], $M_{7\ digit} = .26$, 95% CI [.19 .34], $M_{8\ digit} = .56$, 95% CI [.45 .67].

Choices

Table 2 shows the probability of choosing the wide gap, *p*(Wide), in all of the conditions. In general, by increasing the memory demands of the narrow gap, *p*(Wide) increased. In addition, *p*(Wide) decreased as the memory demand of the wide gap increased.

Table 3. 2. Probability of choosing the wide gap, *p*(Wide), in the nine memory and navigation conditions.

	Narrow gap		
Wide gap	6	7	8
6	.87	.96	.95
7	.35	.83	.86
8	.27	35	.86

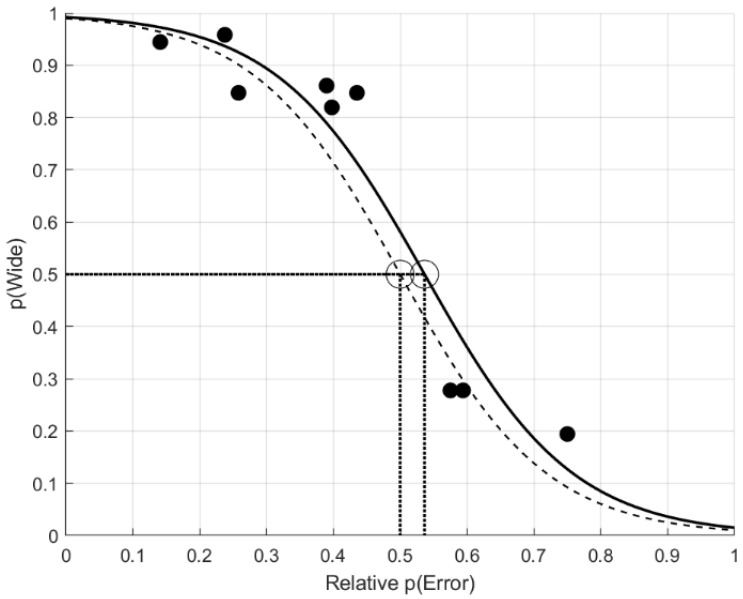
Errors and Choices Together

How were the choice probabilities related to the error probabilities? Figure 2 shows *p*(Wide) as a function of Relative *p*(Error) for the wide gap compared to the narrow gap, where Relative *p*(Error) was defined as

$$\frac{p(Error)wide}{p(Error)wide + p(Error)narrow}$$

As seen in Figure 3.2, *p*(Wide) was systematically related to Relative *p*(Error). The higher the value of Relative *p*(Error), the lower the chance of picking the wide side.

Figure 3. 2. *p*(Wide) as a function of Relative *p*(Error).



Note: The dashed curve shows a hypothetical logit model that would be expected if the only basis for the choices were minimizing errors. Each data point shows the average of *p*(Wide) in each condition. For each participant these values could be either 0, .5. or 1 as we had just 2 observation per condition per participant. The same result would be obtained if we first fitted the psychometric function to each individual subject's data and then average over those fitted functions.

Despite the systematic nature of the relation shown in Figure 3.2, a close look at the graph reveals a problem. The point of subjective equality (PSE) is different from .5; it is at .68. If error-avoidance were the only factor driving participants' choices, the PSE would have been at .5, not at some other value.

A way to accommodate this result is to hypothesize that participants differentially weighed physical errors and mental errors. There is precedent for this idea. Other studies have shown that people attach different weights to different kinds of error in dual-task situations (Bhatt et al., 2016).

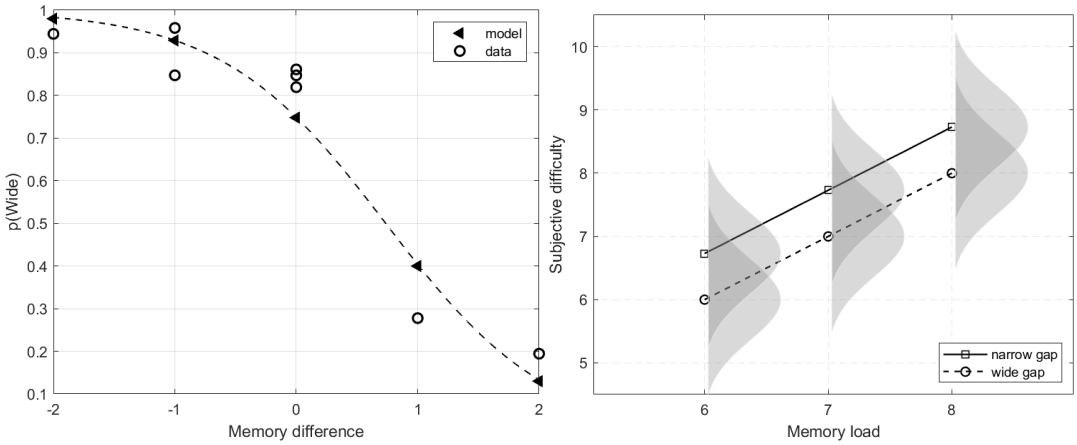
We sought to determine the weights here by finding the coefficient β that, when multiplied by p(Error) in the narrow gap condition, would bring the resulting PSE to .5. Via an iterative procedure, we found that a value of $\beta = 1.17$ yielded a PSE closest to the observed value. By this measure, we could say that making a mistake in the narrow gap was judged to be 1.17 times more costly than making a mistake in the wide gap. We could also say that, insofar as avoiding errors in the narrow gap condition compared to the wide gap entailed extra effort for obstacle-avoidance in the narrow-gap condition, the extra effort was, on average, 17% greater than the effort in the wide gap condition. This statement, and the analytic machinery we have brought to bear to make it, is a new contribution of this article relative to its predecessor article (Feghhi & Rosenbaum, 2019) or, as far as we know, any previous work.

Metacognitive Beliefs About Required Resources

The foregoing analysis suggests that two separate resources were drawn on to perform the tasks. One resource pertained to physical activity; the other pertained to mental activity. To determine how these resources may have been judged, we used a model-fitting approach premiered in our earlier study (Feghhi & Rosenbaum, 2019). We assumed (Figure 3) that the subjective difficulty of a task could be characterized by a normal distribution with mean μ and standard deviation σ. According to the model, when participants compared two tasks in terms of difficulty, the likelihood of picking one task rather than the other would depend on the overlap between the corresponding

subjective-difficulty distributions. We assumed as well that subjective difficulty would increase with the number of digits to be memorized.

Figure 3. 3. Observed data, the best model, and its prediction.



Note: The left panel shows the data and the best fit predicted by the model in the right panel. The right panel shows the model that provided the best fit. According to the model, variability in subjective difficulty could be captured by a normal distribution, subjective difficulty is an additive function of two demands ($\theta = 1$) and going through the narrow gap was subjectively as difficult as memorizing .77 more digits ($\pi = .77$).

To fit the model, we formed a reference line for the wide-gap condition, having it serve as an anchor for different possible subjective-difficulty lines for the narrow-gap condition. For simplicity, the reference line for the wide-gap condition, was simply identity function: subjective difficulty equals memory load. But the reference line could have had any shape or height at all, because we were interested in the differences between the wide-gap line and the narrow-gap line that could provide the best fit to the data. We sought to find the slope, θ, and intercept, π, of the narrow-gap subjective-difficulty line given whatever slope and intercept we used for the wide-gap subjective

difficulty line. The difference between the intercepts of the two lines would indicate how much more difficult the narrow gap seemed to be relative to the wide gap, and the difference between the slopes of the two lines would indicate whether the difficulty of the combined physical and mental tasks was an additive or interactive function of the physical and mental demands. If the function were additive, the two slopes would be the same, but if the function were interactive, the slopes would be different.

There were two other free parameters in the model, k_1 and k_2. These two parameters affected the standard deviation σ of the normal distribution according to $\sigma = k_1 + k_2\mu$, with $k_1 \geq 0$ and $k_2 \geq 0$. The k_1 term was necessary to set the base variability, and the k_2 term was necessary to allow for the possibility that the standard deviation might depend on μ.

The parameter values that maximized the likelihood of the data given the model were $\pi = .77$, $\theta = 1$, $k_1 = .62$, and $k_2 = 0$. The estimate $\theta = .77$ can be taken to mean that participants treated going through the narrow gap as approximately as difficult as memorizing an extra .77 digits on average. The estimate $\theta = 1$ indicates that participants treated the difficulty of physical and mental demands as contributing additively to the overall difficulty of the task. By logic similar to that of Sternberg (1969), this suggests that, from a metacognitive standpoint, the physical and mental task demands contributions to the overall task difficulty in independent stages. The estimate $k_1 = .62$ indicates nonzero variability for the resting state, and the estimate $k_2 = 0$ indicates that the normal distribution had a standard deviation which did not increase with task

difficulty. The images of the distributions in the right panel of Figure 3 reflect these outcomes.

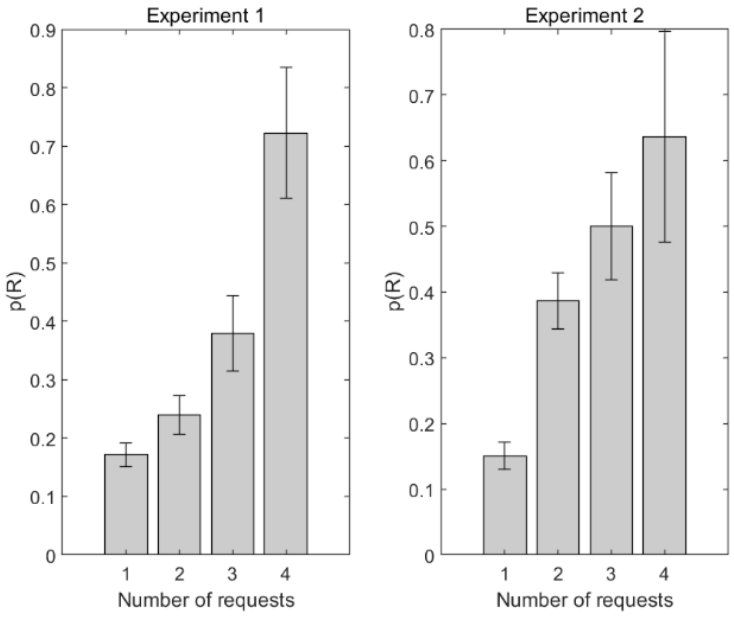
p(R) and Number of Requested Repeats

The major difference between the current experiment and the experiment of Feghhi and Rosenbaum (2019) was in the way the memory items were presented. Feghhi and Rosenbaum (2019) presented their memory items visually. Here we presented the memory items in auditory form. This led to an interesting added facet of the results. When the memory items were presented visually, subjects could read the items however they pleased, going back and rereading the list or parts of it as often as they wished. In the present experiment, where the memory items were presented in auditory form, subjects could only get repeat presentations of the list by asking the experimenter to read it again. The subjects were told in advance that they could do so as often as they wished per trial. This feature of the present experiment afforded us the opportunity to analyze the relation between *p*(R) and how many times a memory list was listened to.

Figure 4 shows how the probability, *p*(R), of recall errors was related to the number of repeat requests in Experiment 1 (the current experiment) and Experiment 2 (the experiment to be presented next, where tasks were assigned rather than chosen). We did not have a prediction about the relation between *p(*R) and number of repeat requests, but were simply curious about it. A one way ANOVA showed that increasing the number of requests was associated with an *increase* in *p*(R), $F(4,646) = 10.70$, $p < .001$, $\eta^2 =$

.057. In the Discussion section, we comment on precedent for this very surprising observation.

Figure 3. 4. Probability, p(R), of making an error in recall as a function of number of times a memory list was repeated upon request in Experiment 1 (left panel), and Experiment 2 (right panel).



Discussion

In the experiment just reported, we used a two-alternative forced choice (2AFC) paradigm in which we asked participants to choose the option that seemed easier – going through a wide gap or a narrow gap when the number of digits to be memorized for either gap was 6, 7, or 8. All combinations of number of digits and wide-narrow gap size were tested. The choice data were orderly, similar to cross-modal intensity-matching research

(Pitts et al., 2016). From the orderliness of the choice data, we infer that participants had a rational basis for making their choices.

What was that rational basis? An obvious candidate is reducing the probability of error (Dunn et al., 2019). Certainly, this hypothesis is plausible considering the clear relation between *p*(Wide) and Relative *p*(Error) shown in Figure 2. On the other hand, closer inspection of that figure brings up a problem, as already noted: The point of subjective equality in Figure 2 is different from .5, so error-minimization per se wasn't all that mattered (cf. Dunn & Risko, 2019; Kool et al., 2010). Using *weighted* error rates made it possible to accept an error reduction account, provided one kind of error was considered more costly than another. In a sense, the extent to which the best-fitting weight deviated from 1 provides an estimate of the extent to which unweighted error reduction alone failed to explain subjects' choices. Our belief is that error reduction was a good proxy for whatever factor was actually used to choose the tasks, but even though Experiment 1 shows that error avoidance is a good proxy for effort avoidance, it isn't the whole story. An additional weighting factor comes into play.

Our last comment about Experiment 1 concerns the result shown in Figure 4, that the probability of error increased with the number of list-repeat requests. It turns out that others have reported the same result. Koriat (2008) and Ackerman (2014) found that participants requested more practice with memory lists when they knew that they had a higher chance of making a mistake. Our result, shown in Figure 4, fits well with Koriat and Ackerman's earlier finding and shows that participants were sensitive to the quality of their memory.

Experiment 2

The second experiment was designed to check that the main results of the first experiment were not artifactual. In Experiment 1, because participants chose some options more often than others, they had more practice for some tasks than others. Therefore, the higher error rates in the less chosen tasks might have simply been an artifact of less practice in those tasks. To test this possibility, in Experiment 2 we controlled for practice. The way we did so was to eliminate the choice element. Feghhi and Rosenbaum (2019) did the same in their study using visually presented memory lists.

Method

Participants

Forty Penn State undergraduate students (31 female and 9 male), none of whom had been in Experiment 1, participated in the second experiment. The participants' age ranged from 18 to 23 years with an average of 19.47 years and a standard deviation of 1.53 years. The sample size was based on the number of participants in Experiment 1. The experiment was approved by the Penn State Institutional Review Board. All participants signed the informed consent form before the experiment.

Apparatus

The apparatus was the same as in Experiment 1 except that there was no card on the right or left stool. In Experiment 2, the experimenter simply said “left” or “right” and then read the digit sequences as often as the participant requested.

Procedure and Design

When the participant was ready to start a trial, the experimenter identified the side by saying “left” or “right.” Then the experimenter read the numbers as often as the participant requested. A random half of the participants started with the wide gap on the left side for the first six trials and then had the reverse arrangement for the next six trials. The other half of the participants had the reverse arrangement. For each participant, the six possible conditions were presented in random order.

Results

Because choices were eliminated in this experiment, we present the results of error rates and, of secondary interest, the relation between error rates and number of times the memory lists were requested.

Error Rates

A 2 (experiments) × 2 (physical) × 3 (cognitive) GEE was conducted on *p*(Error). There was not a main effect of experiment, Wald Chi-Square = .000, $p = .98$; *p*(Error) was .28, 95% CI [.23 .34] for Experiment 1, and .28, 95% CI [.22 .35] for Experiment 2.

There was a main effect of memory load, Wald Chi-Square = 78.66, p <.001. $M_{6\ digit}$ = .15, 95% CI [.12 .20]. $M_{7\ digit}$ = .27, 95% CI [.22 .33]. $M_{8\ digit}$ = .47, 95% CI [.40 .54]. There was a main effect of gap width, Wald Chi-Square = 8.95, p =.003. M_{wide} = .27, 95% CI [.20 .29]. M_{narrow} = .32, 95% CI [.27 .38], but the values for the three-way interaction were Wald Chi-Square = 3.24, p =.198.

p(R) and Number of Requested Repeats

Figure 4 shows p(R) as a function of the number of times a memory list was requested. As seen in the figure, p(R) increased as the number of requests increased. A one way ANOVA confirmed the effect of number of requests on p(R), $F(4,478) = 13.67$, $p < .001$, $\eta^2 = .10$.

Discussion

To check whether error rates in Experiment 1 were affected by the unequal number of opportunities to do each task offered for choice in Experiment 1, in Experiment 2 we removed the choices and had participants do each option an equal number of times. The results showed that the error rates were very similar to those in Experiment 1, indicating that the error rates were not simply an artifact of differential practice.

Experiments 1 and 2 were designed to replicate our previous study (Feghhi & Rosenbaum, 2019). Essentially, they did so. The graphs shown here are remarkably similar to those in the earlier publication. We did find, however, that the probability of making a recall error was somewhat higher in this experiment (M = .235) than in the

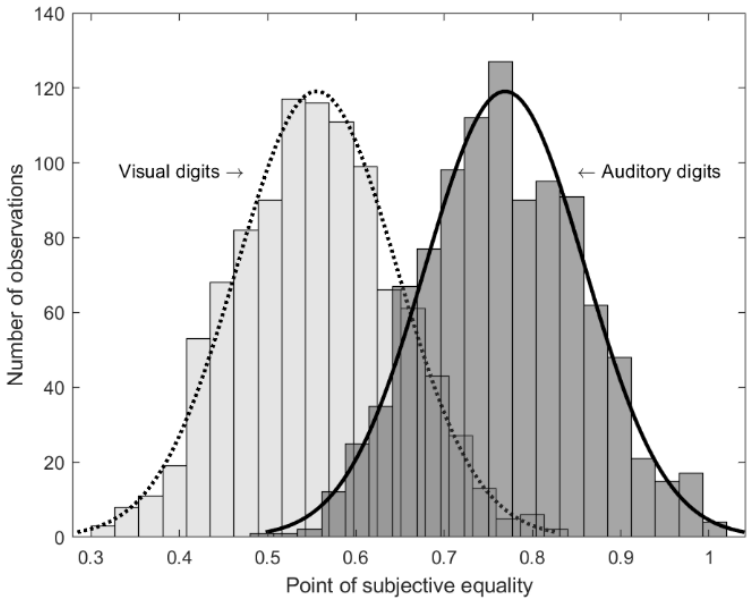
previous one (*M*=.175). In all likelihood, the slightly higher incidence of recall errors here stemmed from the fact that in this experiment, participants had to rely on the experimenter to control the input of the digits to be memorized. If participants were not confident about their ability to recall correctly, they could ask the experimenter to read the digits lists repeatedly, but participants may have been uncomfortable asking for the information multiple times. In fact, no participant ever asked for more than four repetitions (see Figure 4). By contrast, in the visual-presentation case, participants could have read the lists more than four times. We did not monitor this aspect of their performance.

Given the small difference of choice results in the present study and earlier one, we sought to evaluate the difference quantitatively. To compare the choices, we went through three steps. First, we pooled all of the choice data (from both experiments) and ran a binomial regression with the memory difference (wide memory minus narrow memory) as the predictive variable and *p*(Wide) as the dependent variable. The deviance of this model was 1266.07. Next, we added group as a predictive variable. The deviance of this model was 1262.09. Because the difference between the two deviances (3.98) exceeded the critical level of the chi-square, with df =1 (3.84), we concluded, relying on the criterion recommended by Cohen et al., (2013), that there was a significant difference in choices between the two groups.

Our third step was to conduct a separate analysis in which we randomly drew 20 subjects from each group and calculated each group's PSE. By doing this 1,000 times, we got a distribution of PSEs for each experiment (Figure 5). Although there was some

overlap between the distributions, out of 1000 draws there were only four cases in which the mean of the PSE distribution of the earlier study (visual digits) exceeded the minimum PSE of the current study (auditory digits). Based on this result, we have greater confidence of the result of the binomial regression described in the last paragraph. Also notice that we excluded 4 participants who always picked the wide gap. Including those participants makes the difference even more significant.

Figure 3. 5. Distribution of PSEs for the current experiment (auditory digits) and Feghhi and Rosenbaum (2019) (visual digits). Distributions were obtained from 1000 draws of 20 random subjects from each group.



General Discussion

In this article we have described a study of decision-making based on perceived task difficulty. Such decisions are clearly very important both in theoretical and applied

contexts. In the applied realm, if people take on tasks that are more difficult than what they can actually manage, they may run the risk of injury or death. Conversely, if people only take on tasks that are too easy, they may never learn, they may never lose weight, they may never inspire others, and so on.

On the theoretical side and returning to the original motivation for this entire line of research, it has been harder than one might have expected to say what factor (if there is just one) determines task difficulty. Recognizing this difficulty, we suggested, in the predecessor to this article (Feghhi & Rosenbaum, 2019), that a descriptive approach might be useful. For more on this approach, see Lewandowsky and Farrell (2011). In our earlier article, and in this one too, we showed that a descriptive approach can be used to measure the difficulty of one task relative to another even if the two tasks are fundamentally different. We also showed in both studies that this approach can be used to show people's metacognitive beliefs about the interaction or lack thereof between different mental resources. Here, we replicated our previous finding that participants treated physical (navigation) and cognitive (memorization) demands as drawing on separate resources. We appreciate that no task is entirely physical or entirely mental (Rosenbaum, 2017; Rosenbaum and Feghhi, 2019). For example, Cao and Händel (2019) showed that walking and standing still promote distributed styles of attention and focused attention, respectively. Still, looking at error rates, it seems that there were separate modules for the navigation and memorization components of our task.

Another consistency with Feghhi and Rosenbaum's work (2019) was that the higher the anticipated chance of making a mistake, the lower the probability of picking that task option. This brings us back to the question of whether error and effort were the same. We think they were treated as similar but not identical because the point of subjective equality was not at .5; instead, it was at .68. This offset could be compensated by suggesting that navigation errors were judged to be 17% greater than recall errors.

Is there some way to explain this outcome? Saying that it's just that way (a just-so story) is no more satisfying than saying that task difficulty simply depends on the task. We want to do more than kick the core-question can down the road. Here are some possible arguments.

First, even if error minimization were the primary reason for effort minimization, error prediction might be imperfect. Imagine that you were provided with a variety of cognitive tasks. Each time you were presented with two of them, your job would be to pick the one with the lower likelihood of error. Probably, in most cases, your estimation would be accurate, but there would be some occasions when you systematically underestimate or overestimate the likelihood of error of one of the tasks. For example, if one of the tasks were a vigilance task, you might underestimate how difficult it is. In that case, although your intention was to minimize error, if one looked at your choices and then tried to explain them by referring to the obtained error likelihood, there would be some discrepancy.

Second, people may weight different sorts of errors differently, as seemed to be the case here. In real-life, different kinds of mistakes certainly have different

consequences, so they would be expected to have different weights. For example, a mistake in balancing on a tightrope could cause a broken leg or neck, while a mistake in recalling a digit in a list would be much less significant. The fact that consequences can't always be named by an experimenter or a participant doesn't mean those possible consequences play no role in error avoidance. An advance of our general approach, and of the present paper vis a vis our first one, was that here we have suggested a way to quantify the weighing of different kinds of errors.

An important concept in this study and in our previous work (Feghhi and Rosenbaum, 2019) is that the even though the mental and physical tasks relied on different resources (as one can tell from the lack of interaction in physical and mental error rates), their difficulties could be compared systematically. The regularity of the choices enabled us to equate the subjective difficulty of the navigation task to the difficulty of memorizing extra digits.

Our last remark concerns the higher subjective difficulty of the navigation challenge in the current experiment compared to the navigation challenge in Feghhi and Rosenbaum's (2019) study. In our previous study the estimate was .55 digits and here it was .77 digits. Given the higher p(R) in Experiment 1 of the present study compared to p(R) in Experiment 1 (the analogous study) of Feghhi and Rosenbaum (2019), a fair question is whether the subjective difficulty of going through the narrow gap should have been *lower* here than in the previous study, not higher.

A possible answer is that it was not actually the case that the only consequence of the different modality of digit presentation led to a difference in the difficulty of the

memory task. It is also possible that assessing task difficulty was harder in the present choice experiment. It has been argued that evaluating the effort of a task is effortful in itself (Boureau, Sokol-Hessner, & Daw, 2015; Dunn & Risko, 2019). We speculate that deciding which memory list was easier to memorize was easier before, when participants saw the actual digit sequences, than here, when they just saw a number that indicated the length of the list to be memorized. Without as clear an idea of relative memorization difficulty here compared to in the earlier experiment, participants may have placed more weight on navigation in the current choice experiment than did the participants in the earlier choice experiment. If memorization difficulty was itself harder to judge, more weight may have been placed on navigation difficulty here than before insofar as navigation difficulty was easier to assess than memorization difficulty in the present study.

A limitation of this study was the range of memorization difficulty and navigation difficulty. We chose the memorization range to make sure that the memorization task was challenging (doable but not very easy). In the same way, the narrow gap was set in a way that participants needed to turn their body and clear it by going throw the gap sideways. We assumed that difficulty increasing when we increased the number of digits to be memorized from 6 to 8. Out of this specific range, difficulty might increase nonlinearly. This possibility should be tested independently. Still, the general approach taken here could be used regardless of the underlying function.

Finally, in this study we were able to introduce a method to measure the relative cost of making a mistake in different kinds of tasks. For the particular tasks that we used,

we found that navigation error was 17% more costly than memorization error. Based on the differences between the estimation of navigation difficulty in the current study and in our previous article, we have also come to appreciate the importance of the suggestion made by Boureau, Sokol-Hessner, and Daw (2015) and Dunn and Risko (2019) that evaluating the effort of a task is effortful in itself. This raises a question to which future research might be directed: Is difficulty judgment affected by the difficulty of the evaluation itself?

Author Notes

The data were collected while both authors were at Penn State University. The data were analyzed and the manuscript was written while both authors were at the University of California, Riverside. The work was supported by a UCR CoR grant to the second author. Assistance with data collection was provided by Connor Corrente, Yiyi Dai, Amanda Koch, Skylar Korek, Jennifer Norris, Veronika Onischenko, Emily Wolfskill, and Jenny Zhao. We thank Wilfried Kunde, Michael B. Steinborn, and an anonymous reviewer for helpful comments. Correspondence should be directed to Iman Feghhi, Department of Psychology, University of California, Riverside, CA 92521, iemanifk@gmail.com.

Conflict of interest

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

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

References

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*, 1349-1368 https://doi.org/10.1037/a0035098.

Baddeley, A. D. (1976). *The psychology of memory*. New York, NY: Basic Books.

Bhatt, T., Subramaniam, S., & Varghese, R. (2016). Examining interference of different cognitive tasks on voluntary balance control in aging and stroke. *Experimental Brain Research*, *234*, 2575-2584.

Boureau, Y. L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences*, *19*, 700-710.

Cavanagh, J. P. (1972). Relation between immediate memory span and memory search rate. *Psychological Review*, 79, 525-530.

Cao, L., & Händel, B. (2019). Walking enhances peripheral visual processing in humans. *PLoS Biology*, *17*(10), e3000511.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York, NY: Routledge.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology*, *55*, 429-432.

Craig, A. D. (2013). An interoceptive neuroanatomical perspective on feelings, energy, and effort. *Behavioral and Brain Sciences*, *36*, 685-686.

Duenas, M., Salazar, A., Ojeda, B., Arana, R., & Failde, I. (2016). Generalized Estimating Equations (GEE) to handle missing data and time-dependent variables in longitudinal studies: an application to assess the evolution of Health Related Quality of Life in coronary patients. *Epidemiologia e Prevenzione*, *40*, 116-123.

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

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

Feghhi, I., & Rosenbaum, D. A. (2019). Judging the subjective difficulty of different kinds of tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 983-994.

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

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine,* 16, 965–980.

Job, V., Dweck, C. S. & Walton, G. M. (2010) Ego depletion – Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science* 21, 1686–1693.

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

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*, 416-428.

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*, 661-679.

Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Los Angeles, CA: Sage.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*, 220-224.

Pitts, B., Riggs, S. L., & Sarter, N. (2016). Cross-modal matching: A critical but neglected step in multimodal research. *IEEE Transactions on Human-Machine Systems*, *46*, 445-450.

Posner, M. I. & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, 74, 392-409.

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

Rosenbaum, D. A. (2017). *Knowing Hands – The Cognitive Psychology of Manual Control*. New York: Cambridge University Press.

Rosenbaum, D. A., & Bui, B. V. (2019). Does task sustainability provide a unified measure of subjective task difficulty? *Psychonomic Bulletin & Review*, 1-8.

Rosenbaum, D. A., Chapman, K. M., Coelho, C. J., Gong, L., & Studenka, B. E. (2013). Choosing actions. *Frontiers in Psychology*, Volume 4, Article 273, doi:10.3389/fpsyg.2013.00273.

Rosenbaum, D. A. & Feghhi, I. (2019). The time for action is at hand. *Attention, Perception & Psychophysics*, 1-16. DOI 10.3758/s13414-018-01647-7

Rosenbaum, D. A., & Gaydos, M. J. (2008). A method for obtaining psychophysical estimates of movement costs. *Journal of Motor Behavior*, *40*, 11-17.

Rosenbaum, D. A., & Gregory, R. W. (2002). Development of a method for measuring movement-related effort. *Experimental Brain Research*, *142*, 365-373.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences of the United States of America*, *54*, 407-411.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652-654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica,* Amsterdam, 30, 276–315.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12, 1100-1122.

CHAPTER 4 – WHAT MATTERS IN MAKING DEMAND-BASED DECISIONS: TIME ALONE OR EFFORT TOO?

[a]Markus Janczyk, [b]Iman Feghhi, & [b]David A. Rosenbaum

[a] Department of Psychology, University of Bremen, Bremen, Germany

[b] Department of Psychology, UC Riverside, Riverside, CA, USA

Submitted to Psychological Research (as of 7/21/2021)

Abstract

Which is easier, doing 20 single-digit arithmetic problems or moving a wheelbarrow full of rocks over a distance of 10 meters? If it is possible to choose between these "more cognitive" and "more physical" tasks, how are the difficulty levels of the tasks compared? Previous research (Potts, Pastel, & Rosenbaum, 2018, *AP&P*) suggested that subjective time is the underlying common metric. However, the particular task used (counting up to 8, 12, 16, or 20) confounded time requirements with cognitive demand and likelihood of errors. Here we sought to unconfound these variables. We present two experiments where the duration of the cognitive task and the cognitive demands were systematically varied, as was the weight of a to-be-carried bucket. We found that the probability of choosing the bucket task increased when the cognitive task was more demanding, when its duration was longer, and when the bucket was empty. The choice data could be well explained by a model which used a transform of objective times for the alternative tasks as its main elements, and where the transformed times were themselves simple functions of the independent variables for the cognitive and physical

demands. The significance of the model inheres in its incorporation of time and effort into a single value.

Key words: Decision Making; Effort; Motor Control; Numerical Cognition; Perception And Action; Walking

Introduction

A well-established method for studying task difficulty and effort is the Demand Selection Task (DST[2]; e.g., Dunn & Risko, 2019; Botvinick & Rosen, 2009; Gold et al., 2015; Kool et al., 2010). This is a 2-Alternative Forced-Choice procedure in which participants choose between two options based on the options' perceived demands. The DST task has been used in investigations of the effects of different factors on perceived difficulty of tasks, including the effect of task switching cost (Kool et al., 2010), time and error (Dunn, Inzlicht & Risko, 2019), and metacognitive evaluation (Dunn, Gaspar & Risko, 2019; Dunn & Risko, 2019; Dunn, Lutes & Risko, 2016). The same method has been also used to evaluate the physiological (Botvinick & Rosen, 2009) and neurological (McGuire & Botvinick, 2010) consequences of engaging in or anticipating demanding tasks. In typical DST tasks, both options are cognitive in nature However, the same method can be used to investigate the comparison of different kinds of tasks (Potts, Pastel, & Rosenbaum, 2018; Rosenbaum & Bui, 2019; Feghhi & Rosenbaum, 2019,

[2] An alternative method, asking participants to rate their perceived or anticipated effort level, has been also used and might be more suitable in some situations (e.g., Dunn, Koehler, & Risko, 2017), but the DST approach has been preferred in different labs.

2020). For example, if one had to choose between doing ten challenging math problems or moving a wheelbarrow full of rocks back and forth between a couple of locations ten times, what would one choose? On what basis would the choice be made? The choice would surely depend on features of the tasks – how challenging the math problems were, how many of them there were, how far apart the rock locations were, how heavy the rocks were, how many trips were necessary, and so on. The fact that these factors would affect the choice suggests that one can compare task difficulties even when the tasks are of different kinds. How does this happen? What are the core elements for such decisions?

The tasks mentioned above may be said to differentially tax "brain" and "brawn." One task, doing math problems, is "more cognitive." The other, moving rocks, is "more physical." Of course, these terms are intuitive at best, for "mental tasks" also require physical enactment, and "physical tasks" also require thought. If physical tasks only required brawn but no brain, robots would be more capable than they are of complex actions in unpredictable environments, and doing physical tasks would not affect cognitive performance or vice versa, though such interactions have been observed (e.g., Weigelt, Rosenbaum, Huelshorst, & Schack, 2009; Zhang, Wininger, & Rosenbaum, 2014).

As just intimated, the question posed here is, what common currency is used to compare the difficulty of different kinds of tasks and, for that matter, tasks of a given kind? That there might be a common currency is suggested by the fact that when people compare the difficulty of physical and mental tasks – not math and rock-transport tasks,

as in the opening example, but digit memorization and walking through gaps of varying width – their choices are systematic (Feghhi & Rosenbaum, 2019, 2020). Even so, it is possible that the subjective difficulty for each task is measured with a *different* metric. For example, attentional demands might be used to determine mental difficulty, whereas calorie consumption might be used to determine physical difficulty. If mental difficulty values and physical difficulty values were mapped onto one another – say from smallest to largest in both cases – it might be possible to decide between the two kinds of tasks based on the relative positions of their values. No common currency would be needed.

Notwithstanding the latter possibility, several common-currency candidates have been considered in previous literature on this topic. Energy has been suggested (Craig, 2013; Job et al., 2010), though, as far as we can tell, that term has been used metaphorically rather than literally (e.g., Navon & Miller, 2002; Tombu & Jolicoeur, 2003). Likelihood of error has also been considered (Dunn et al., 2019), but Feghhi and Rosenbaum (2019, 2020) showed that error reduction is not the *sine qua non* of task choice, and that point can be reached through a simple thought experiment. The cost of a car crash is much higher than the cost of a math-homework mistake, even though the probabilities of two events could be the same. Finally, time on task has been considered as well (Gray et al., 2006; Potts et al., 2018; Rosenbaum & Bui, 2019), and this candidate will be a focus of the current investigation.

Subjective Time and Choice

One of the studies just cited, Potts et al. (2018), served as the basis for the two experiments reported here. In that study, participants chose between a cognitive task (counting up to target values of 8, 12, 16, or 20) and a physical task (picking up a bucket from a stool and carrying it to a target stool). Two stools stood at the end of an alley, and four other stools stood midway from a starting position to the target stools, with two stools each to the left and to the right (see Fig. 1 in Potts et al. for an illustration). Whether the bucket was on the left or right and whether it required a short or long reach was varied within-participants, as were the target values for the counting task. The combinations of the four count values and four bucket positions resulted in 16 trials per participant. There was also a between-group factor with three levels: the bucket was (1) empty, (2) filled with 3.5 pounds of pennies, or (3) filled with 7 pounds of pennies.

Potts et al. (2018) observed that the probability, *p*(Bucket), of choosing the bucket rather than the counting task increased with the count targets and was larger for short reaches than for long reaches (see Fig. 2 in that study).[3] In contrast, bucket side or bucket weight had no effect on the choices. Most important was the reliable effect of task-completion time: Chosen tasks took less time than unchosen tasks. The authors also observed that the choice probabilities could be better fit if *subjective* time rather than

[3] Preferences for short reaches were also reported by Rosenbaum (2008) and Rosenbaum, Brach, and Semenov (2011), where the alternative task was walking over some distance rather than reaching over some distance.

objective time was input into a model. The subjective time model that Potts et al. developed ascribed 5 extra seconds of subjectively experienced time to long-reach tasks compared to short-reach tasks. The time of 5 extra seconds was found to maximize the goodness of fit of the model to the data.[4]

To test the hypothesis that subjective time was the basis for their *p*(Bucket) data, Potts et al. (2018) ran a second experiment. Here they repeated the first experiment, though with empty buckets only, and replicated the results of Experiment 1. Yet, in addition to collecting choice and performance data, they asked a new group of participants to estimate how long they spent on each of the tasks. These time estimates (see Fig. 6 in that study) were longer for long-reach tasks than for short-reach tasks by a wider margin than the difference in objective times for the long and short reaches. This outcome fits with the model-fitting that Potts et al. did. Participants' time estimates for the counting tasks also differed from the objective counting times. Interestingly, the time estimates exceeded the objective counting times by an amount that grew as the count maximum (the target value at which counting could stop) increased. When the Experiment 2 *p*(Bucket) data were fitted with the obtained subjective times, the data were better fit than when using the objective times.

[4] The model was $p(A) = \frac{T(B)}{T(B)+T(A)}$, where *A* and *B* were the two tasks considered per condition, and *T* was time.

The Present Study

In the present study, we describe two experiments following up on the work by Potts et al. (2018). We embarked on these new experiments, because there were confounds in the earlier study, which – though acknowledged by the authors – were not tracked down by them. While time for counting increased with larger target values, counting to higher target values could have taxed resources in ways that happened only incidentally to be indexed by time, and counting to higher target values could have led to more errors. Participants' decisions could have been driven by any of these factors. We sought to find out which factor(s) really mattered because, as indicated in the title of this article, we wanted to know whether time or effort is the principal basis for choosing actions based on apparent difficulty.

This question was especially important given an earlier influential report by Kool et al., (2010), who reported that for a DST, time on task was not the primary basis for determining subjective difficulty, nor were error rates or, conversely, rates of accumulation of positive feedback. What mattered, the authors concluded, was cognitive effort. Because Potts et al. (2018) reported that subjective time better accounted for task choices than objective time, one could argue that participants in the study of Kool et al.

actually formed estimates of task completion times and relied on those psychologically mediated times to make their choices.[5]

In the experiments reported here, participants decided whether they would do a cognitive task – either adding or multiplying 2 one-digit numbers in Experiment 1, or adding and subtracting 2 or 4 one-digit numbers in Experiment 2 – for a specified duration, or carry a bucket to the end of the alley. Participants were told how long they had to do the math problems, but not how long they had to do the bucket task. The way the math time requirement was implemented was to allow participants to complete the last problem presented to them before the computer-controlled deadline was up.[6] This procedure differed from the one used by Potts et al. (2018) where participants could, in principle, modulate the duration of their counting by varying their counting rates.

Regarding the bucket-carry tasks, we varied bucket weight in a way that Potts et al. (2018) did not. In the first experiment of Potts et al., participants chose between counting to 8, 12, 16, or 20 and reaching over short or long distances to grab a light (empty) or heavy (filled) bucket to be carried to the end of the alley. Bucket weight, however, was a between-participants factor. Potts et al. failed to find an effect of bucket weight on *p*(Bucket), so they did not vary bucket weight in their second experiment, where the other (more positive) innovation was to obtain subjective time estimates.

---

[5] The plausibility of this hypothesis is strengthened by the observation that psychologically experienced time differs from objective time (e.g., Grondin, 2008; Ornstein, 1969).

[6] The actual time ended up being between the predetermined times (9, 18, and 27 s) and + (approx.) 1.2 s.

Rather, the bucket was empty in Potts et al.'s second experiment. The lack of an effect of bucket weight in Experiment 1 was unexpected, as Potts et al. noted. The absence of an effect of bucket weight could have been due to the fact that this factor was varied between-participants (see also Birnbaum, 1999). In the present experiments, we varied bucket weight within-participants to see if we would pick up an effect when variations in that factor became more salient.[7] In other respects, the design and method used here were meant to simulate those of Potts et al.

Experiment 1

Method

Participants

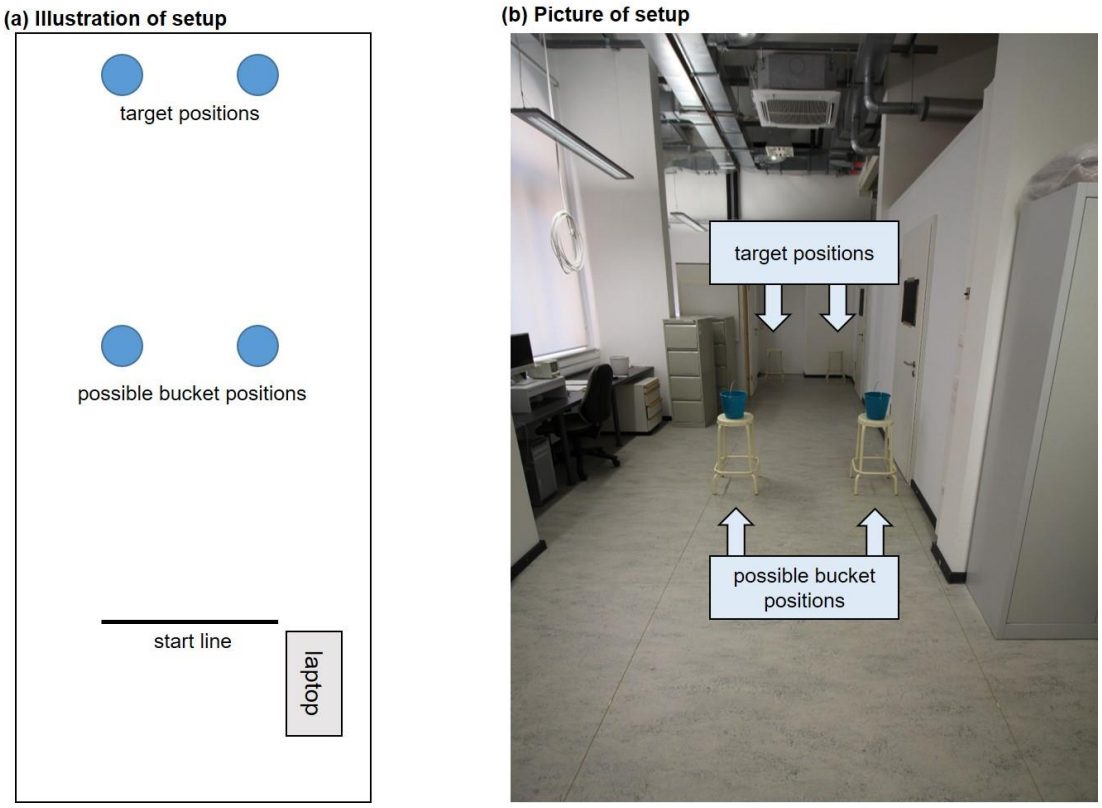
Thirty people (mean age = 25.1 years, 19 female) from the Tübingen area (Germany) participated for money or course credit. A power analysis suggested that this sample size was large enough to detect an effect of size $d \geq .53$ with a power of $1 - \beta = .8$ (two-sided paired *t*-test, $\alpha = .05$). All participants reported normal or corrected-to-normal vision, were naive to the hypotheses, and signed an informed consent form prior to data collection. Data were collected after participants completed an unrelated experiment.

[7] Bucket weight would probably matter in a between-participants design where bucket weights varied enough to strain ethical guidelines regarding the loads that one group could be asked to carry. We were unwilling to impose that physical burden on participants or to impose that ethical burden on ourselves.

Apparatus and stimuli

The setup for the experiment is shown in Figure 4.4.1. In the *bucket task*, four stools (height: 75 cm) were used as platforms (30 cm diameter). Two were closer to the participants' start position and two were farther away. A bucket was placed on one of the closer stools at the start of each trial. One bucket was blue and the other was grey. One of the buckets was empty (0.0 kg condition) and the other was filled with gravel weighing 3.2 kg. Participants were informed of the color-weight mapping (which was counterbalanced across participants) that applied to them and were given a chance before the main experiment began to heft each of the buckets to get a clear haptic sense of their weights. The buckets' handles were fixed in an upright position to facilitate easy grasping and were oriented parallel to the long axis of the walkway. Whereas both panels of Figure 4.1 show both possible bucket positions, in the actual experiment only one bucket was presented per trial, either on the left or right. The distance from the start line to the bucket positions was always 376 cm. The distance from the start line to the target positions was 750 cm. The alley was 90 cm wide.

Figure 4. 1. Illustration of the experimental setup.



Note: (a) Overhead sketch. (b) Photograph of the real setup with the bucket on both stools for illustration only. In the experiment, only one bucket was present, either on the left or right.

The *cognitive task* was administered on a laptop, which was placed to the right of the participant (as in Potts et al., 2018; see Fig. 1a) on a Table 4.100 cm high. The cognitive task was either to add or multiply two digits per trial. We expected multiplication to be judged more difficult than addition (Ashcraft & Guillaume, 2009). In each trial, the participant typed his or her answer, and after hitting the enter button, was shown the next equation. Feedback about accuracy was not provided.

Tasks and Procedure

At the start of each trial, the participant stood behind the starting line, facing away from the area with the stools and bucket, with eyes closed. During this time, the experimenter prepared the upcoming trial and then told the participant to turn around and look at the laptop. The screen informed the participant about the relevant cognitive task for the upcoming trials (i.e., whether they were to add or to multiply digits), and also about its duration, should they choose that task. We provided them with two options and asked them to perform the one they preferred, with the choice between performing the bucket task or the cognitive task being indicated by pressing the left CTRL or right CTRL-key of the keyboard. Depending on the condition, the cognitive task was either addition or multiplication and was to be done for 9, 18, or 27 seconds. Pilot work showed that 18 seconds was the approximate time to walk in a normal pace from the start line, pick up a bucket, place it on the target stool, and return to the start line. The other values were chosen to be shorter and longer than this time by equivalent amounts (± 9 s). If participants opted for the bucket task, they were to embark on the task immediately after hitting the corresponding button on the laptop. That way, the laptop button-press served as a proxy for the start time of the chosen act. The proxy for the end time of the bucket task was when the experimenter hit a button at the moment the participant returned to the starting line. If the participant chose the cognitive task, the first equation appeared immediately. After the participant typed in the sum or product for the problem at hand, their act of pressing the Enter key brought up the next equation unless the duration of 9, 18, or 27 seconds has elapsed.

The entire task had 24 trials based on the combination of 2 bucket locations (left vs. right) × 2 bucket weights (0.0 vs. 3.2 kg) × 2 levels of cognitive demands (addition vs. multiplication) × 3 durations of the cognitive task (9 vs. 18 vs. 27 seconds).

Design and Analyses

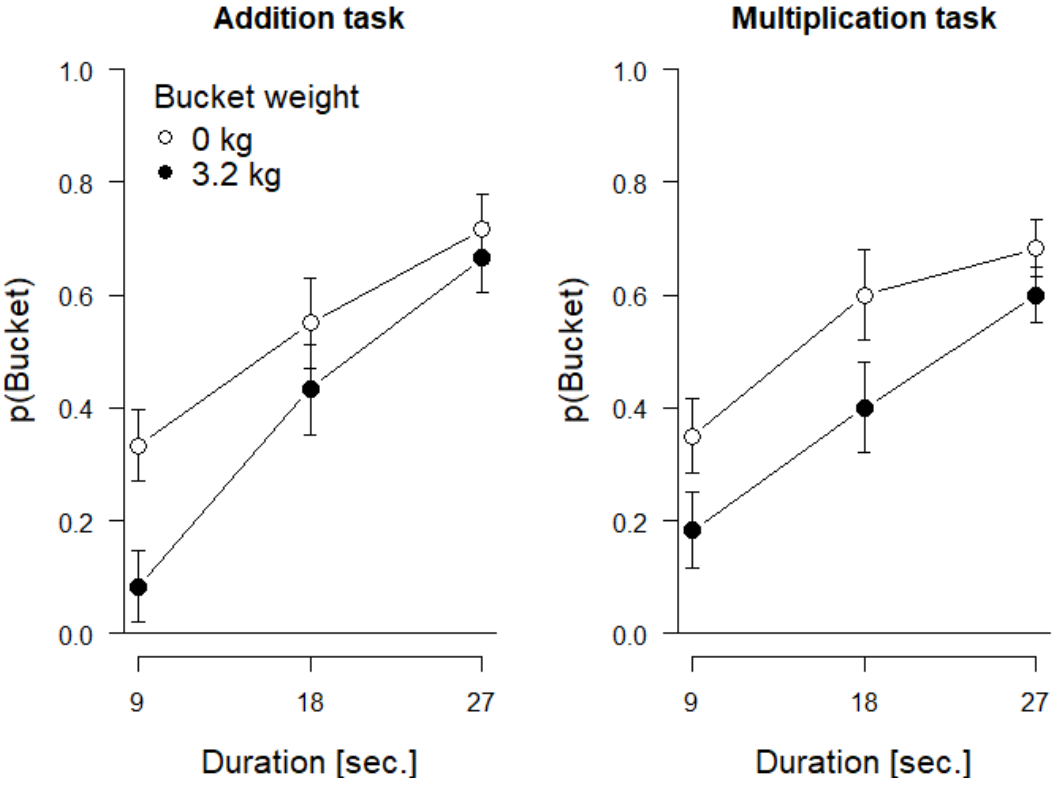
In an attempt to assess task difficulty, we compared (1) the number of the performed calculations between the addition and multiplication tasks and (2) also the error rates in both tasks were compared. In both cases, we averaged over the three durations of the cognitive task. A more difficult task would then be indicated by a smaller number of performed calculations and/or more errors. The main analysis assessed the probability of choosing the bucket, *p*(Bucket), via a repeated measures ANOVA whose independent variables were duration of cognitive task, demands of cognitive task, and bucket weight. Because bucket location had no effect on choices, we aggregated choices regardless of where the bucket stood.[8]

[8] When included in an ANOVA as an additional repeated measure, location produced no significant main effect nor did the variable enter into any interactions.

Results

On average, participants managed to perform 8.46 calculations of the addition task but only 7.47 calculations of the multiplication task, $F(1,29) = 8.23$, $p = .008$, $\eta_p^2 = .22$. Error percentages were 10.92 and 11.17 in the addition and multiplication task, respectively, and this difference was not significant, $F(1,29) < .01$, $p = .966$, $\eta_p^2 < .01$ The probability, p(Bucket), of choosing the bucket as a function of the cognitive task, its duration, and bucket weight is shown in Figure 4.4.2. The impression from the figure is that participants chose the bucket task more often when the bucket was empty than when it was weighted and more often as the duration of the other, cognitive, task increased. Cognitive task demand (addition versus multiplication) did not appear to have a major impact on bucket choices, as confirmed in the ANOVA for these data (see Table 4.1 for the complete results). Only the two main effects of duration and weight were significant, Fs ≥ 20.44, ps $< .001$. All other effects were non-significant, all Fs ≤ 1.99, all ps $\geq .147$.

Figure 4. 2. Probability, *p*(Bucket), of choosing the bucket in Experiment 1, as a function of duration of cognitive task (x-axis), bucket weight (separate lines), and difficulty of task (i.e., addition vs. multiplication) as separate panels



Note: Error bars show 95% within-participants confidence intervals for the difference between the empty and the loaded bucket, calculated separately for each cognitive task and duration.

Table 4. 1. Statistics of the full three-way ANOVA for Experiment 1.

Effect	***F***	***p***	η_p^2
Duration	22.08	<.001	.43
Demand	.01	.923	<.01
Weight	20.44	<.001	.41
Duration × Demand	1.28	.285	.04
Duration × Weight	1.99	.147	.06
Demand × Weight	.07	.791	<.01
Duration × Demand × Weight	1.01	.369	.03

Discussion

Experiment 1 replicated and extended the results of Potts et al. (2018) by showing that when the cognitive task duration was controlled, the longer the cognitive task was to be performed, the more often participants chose the alternative, bucket, task. We also observed a clear effect of bucket weight: Participants chose the cognitive task more often when the bucket was loaded than when it was empty. This outcome accords with our expectation that with a within-participants manipulation, where the same participants got physical tasks with varying demands (buckets with different loads), they would show

greater sensitivity to the demand levels than was the case in the study of Potts et al. where bucket weight was varied between- participants in their Experiment 1.

Unexpectedly, however, in the present experiment, type of cognitive task did not affect choices. Whether the cognitive task was addition or multiplication, it did not matter. Interestingly, the tasks differed in the rate at which problems were solved – more addition problems were solved per unit time than were multiplication problems – but error rates were comparable. Other studies have also reported a lack of differences in error rates between addition and multiplication of two digits, the number of digits per problem used here (e.g., Zhou et al., 2007). However, even though this dissociation may be interesting in itself, the absence of an error-rate difference also suggests that (a) the completion rate was not the determinant of the perceived task difficulty and (b) the intended manipulation of task demand was unsuccessful, or at least not strong enough to influence the choices made by our participants. To address this issue, we manipulated the cognitive demands in a different way in Experiment 2.

Experiment 2

Most aspects of this experiment were the same as in Experiment 1 with the major change relating to the cognitive task. Because it was unclear whether multiplication was substantially harder (objectively and subjectively) than addition, we turned to a different task.

Method

Participants

Forty-eight people (mean age = 23.8 years, 34 female, 15 male) from the Tübingen area (Germany) participated in this experiment for the same criteria as described in Experiment 1. A power analysis with the same parameters as for Experiment 1 indicated that this sample size was sufficient to detect effects of $d \geq .42$.

Apparatus, stimuli, task, procedure, design, and analyses

We used the same material and setup for the *bucket task* as in Experiment 1. The *cognitive task* was changed so that participants were presented with equations involving addition and subtraction of single digit numbers. Depending on task demands, either 1 or 3 successive additions/subtractions were used in an equation (i.e., either 2 or 4 digits occurred on the left side of the equation). For the less demanding condition, the equations were of the form A-B=Z, which we called the 2-digit condition. For the more demanding condition, the equations were of the form A-B+C-D=Z, which we called the 4-digit condition. A solution to each equation was given (on the right side of each equation) and the participants' task was to decide whether the provided answer was correct or incorrect. In the former case, they were to press the right CTRL key; in the latter case, they were to press the left CTRL key. Whether the shown answer was correct or incorrect was

randomly determined in each trial. If the shown answer was incorrect, the displayed result differed from the correct result by +1 or -1, what was determined randomly.

To ensure that all participants had a clear idea of the demands of each task, they started with an exposure period prior to the cognitive task. Both levels (2-digit and 4-digit problems) were administered 40 times, with the order counterbalanced across participants. After this, the main experiment was conducted in the same manner as described for Experiment 1. For the cognitive task, a new equation appeared after participants pressed the response key each time the duration was still in effect.

The task consisted of 24 trials resulting from the combination of the 2 bucket locations (left vs. right) × the 2 bucket weights (0.0 vs. 3.2 kg) × the 2 levels of the cognitive task demand (2 digits vs. 4 digits) × the 3 cognitive task durations (9 vs. 18 vs. 27 seconds). Error rates and RTs during the exposure period were analyzed as a function of difficulty (2 vs. 4 digits) to assess task difficulty objectively. The data analysis protocol followed that of Experiment 1.

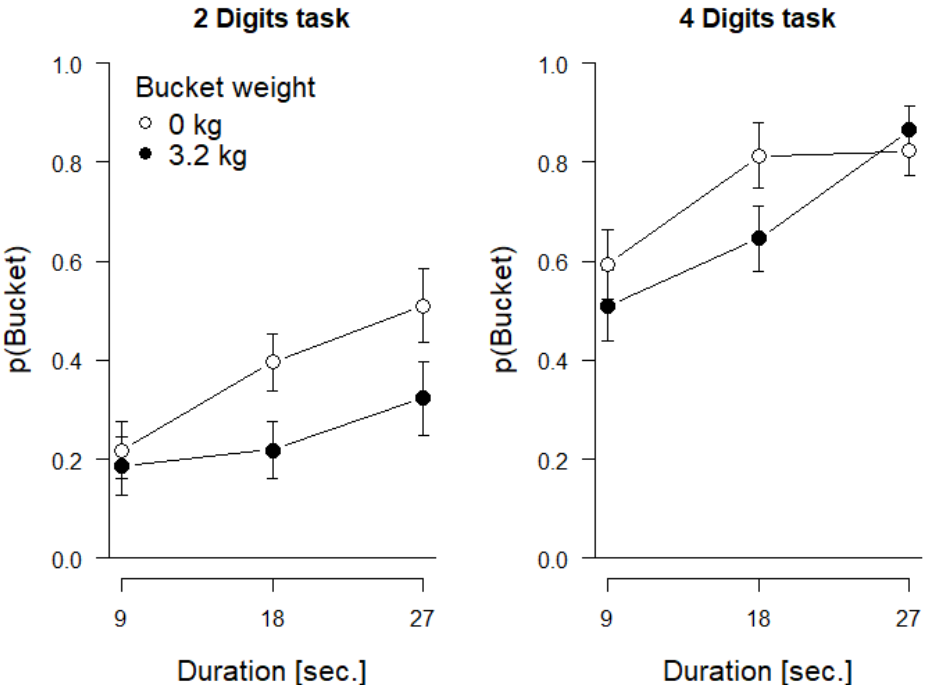
Results

In the exposure block, participants made fewer errors in the 2-digit condition than in the 4-digit condition, (5.68% vs. 11.61%), $t(47) = 5.59$, $p < .001$, $d = .81$, and responses were given faster in the 2-digit condition than in the 4-term condition (1836 vs.

5390 ms), $t(47) = 33.64$, $p < .001$, $d = 4.86$. These outcomes suggest, as expected, that the 2-digit condition would be less demanding than the 4-digit condition.

Figure 4.4.3 shows the probability of choosing the bucket, *p*(Bucket), as a function of the duration of the cognitive task and bucket weight separately for the two levels of cognitive demands. The impression is that participants chose the bucket task more often as the cognitive task duration increased, when the bucket was empty compared to when the bucket was weighted, and when the cognitive task involved 4 digits compared to 2.

Figure 4. 3. Probability, *p*(Bucket), of choosing the bucket task in Experiment 2 as a function of the duration of the cognitive task (x-axis), bucket weight (separate lines), and difficulty of the cognitive task (2 digits vs. 4 digits) as separate panels.



Note. Error bars are 95% within-participants confidence intervals for the difference between the empty and loaded bucket, calculated separately for each cognitive task and duration.

As shown in Table 4.2, this impression was corroborated by the respective three main effects, all of them being significant, assuming $\alpha = .05$ as in Experiment 1. Because the three-way interaction had a *p*-value that, in traditional hypothesis-testing terms would be "just significant," we analyzed the two levels of cognitive task conditions separately with ANOVAs that only had cognitive task duration and bucket weight as repeated measures.

Table 4. 2. Statistics of the full three-way ANOVA for Experiment 2.

Effect	***F***	***p***	η_p^2
Duration	21.13	<.001	.31
Demand	89.32	<.001	.66
Weight	9.36	.004	.17
Duration × Demand	.87	.422	.02
Duration × Weight	2.12	.126	.04
Demand × Weight	2.04	.159	.04
Duration × Demand × Weight	3.08	.050	.06

For the 2-digit task, the main effect of duration was significant, $F(2,94) = 7.36$, $p = .001$, $\eta_p^2 = .14$, and this was also true for the main effect of bucket weight, $F(1,47) = 11.41$, $p = .001$, $\eta_p^2 = .20$. However, the interaction was not significant, $F(2,94) = 2.03$, $p = .137$, $\eta_p^2 = .04$. For the 4-digit task, the main effect of duration was significant, $F(2,94) = 20.29$, $p < .001$, $\eta_p^2 = .30$, whereas the main effect of weight was not, $F(1,47) = 3.02$, $p = .089$, $\eta_p^2 = .06$, and the p-value of the interaction fell just below the traditional α-value for significance, $F(2,94) = 3.21$, $p = .045$, $\eta_p^2 = .06$s.

Discussion

Experiment 2 was similar to Experiment 1 except that we varied the number of digits (2 or 4) in mixed addition and subtraction problems to manipulate objective difficulty. This manipulation differentially taxed participants in ways that the use of addition versus multiplication did not in Experiment 1. Given this outcome, the following conclusions could be drawn. First, there was an effect of the duration of the alternative cognitive task on *p*(Bucket). The longer that alternative-task duration, the higher the value of *p*(Bucket), replicating what was observed in Experiment 1. Second, loaded buckets were chosen less often than empty ones, also replicating what was observed in Experiment 1. Third and finally, there was a clear effect of cognitive demand on the likelihood of choosing the bucket. When only two digits would be dealt with, *p*(Bucket) was higher than when four digits would be dealt with. These results show that all three variables – time, cognitive demands, and bucket weight – affected choices. Perhaps most important given the main question of this study, the decisions about which task to carry out were not just based on time. To our knowledge, this is the first time that there has been a demonstrated dissociation of task time and task demands on task choice.

General Discussion

In this article we have reported two experiments on choosing between a "more cognitive" task and a "more physical" task. Earlier studies (Potts et al., 2018; Rosenbaum & Bui, 2020) suggested that time may be the underlying metric for choosing a less

subjectively difficult task. The present study built on those results and aimed to disentangle the contributions of time, cognitive demands, and physical demands – variables that were confounded in the earlier studies. To this end, we employed two cognitive tasks per experiment to manipulate demands. In Experiment 1, the cognitive tasks were addition versus multiplication of 2 digits. In Experiment 2, the cognitive tasks were mixed addition and subtraction of 2 versus 4 digits. We also varied how long the tasks were to be performed in both experiments. The times were 9, 18, or 27 seconds. These values were based on 18 seconds as the approximate time of the bucket task, with 9 seconds and 27 seconds being 9 seconds shorter and longer, respectively, than that time. We also varied the weight of the bucket, because this variable had unclear effects in earlier studies, where bucket weight was manipulated between-participants. Here we made bucket weight a within-participant factor and used a design with the two levels of physical demand crossed with the two levels of cognitive demand (and duration of the cognitive task) for each single participant.

Two results were clear in both experiments. First, the longer the duration of the cognitive task, the more often participants chose the bucket task. Second, when the bucket was loaded, the bucket task was chosen less often than when the bucket was empty. This latter result contrasts with Potts et al.'s (2018) report that bucket weight didn't matter, though that study used a between-subjects design for bucket weight, whereas here we used a within-a subject design.

With regard to the effect of task demands, choices did not depend on cognitive task in Experiment 1, but did so in Experiment 2. Thus, our expectation that multiplication would be harder than addition was not realized in Experiment 1. Still, an important conclusion could be reached from the fact that participants in Experiment 1 sometimes chose the cognitive task even though it had a higher error rate than the bucket task. The error rate for the bucket task was 0%, but was close to 11% for the math tasks. Participants would have never chosen the math task if the sole criterion for doing so were elimination of errors. Therefore, error elimination was not the sole basis for choosing tasks, a result observed as well by Kool et al. (2010) and Feghhi and Rosenbaum (2020). Where this leaves us that that neither error elimination nor time minimization was the basis for choice, as noted at the end of Experiment 2.

How then can we account for our results? Can we do so with a single metric, as asked in the introduction of this article? Can we say there is a common currency for choosing less demanding tasks? We think we can.

We addressed these questions through modeling. More precisely, we asked whether the contributing factors (i.e., duration of cognitive task, cognitive demand, and bucket weight) could be converted into a single variable. In keeping with the approach taken by Potts et al. (2018), who pursued a subjective time approach, we tried to convert cognitive demand and bucket load into subjective time. Our idea was that, if these variables are indeed convertible to a subjective time variable, then increasing the cognitive demand (e.g., from 2-digits to 4-digits in Experiment 2) should have an effect

similar to increasing the duration of the 2-digit task. Similarly, increasing the duration of the cognitive task should have an effect similar to decreasing the bucket load.

The way we embarked on our modeling was to pursue subjective time as some transform of objective time that would maximize the likelihood of the data. We imagined sliding three of the curves in Figure 4.4.2 horizontally, such that all the points, plus the points along the unshifted curve, would hug a single curve. Similarly, we imagined sliding three of the curves in Figure 4.4.3 horizontally, such that all the points, plus the points along the unshifted curve, would lie on a single curve. Finding the horizontal shifts that achieved the best fit of all the points per figure would amount to transforming the *x*-axis from objective time to subjective time.

We defined two free parameters, denoted *k* (cognitive demand) and *h* (bucket weight) in the following formula, where Φ represents subjective time:

$$\Phi = \begin{cases} duration, & Bucket\ weight = 3.2\ , \quad cognitive\ demand = 2 \\ duration + k, & Bucket\ weight = 3.2\ , \quad cognitive\ demand = 4 \\ duration + h, & Bucket\ \ weight = 0\ , \quad cognitive\ demand = 2 \\ duration + k + h, & Bucket\ \ weight = 0\ , \quad cognitive\ demand = 4 \end{cases}$$

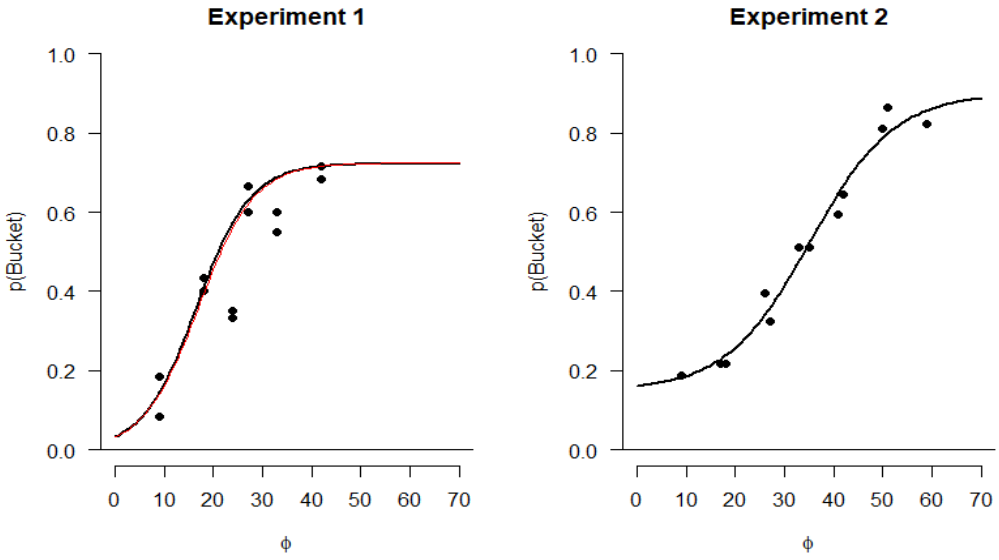
For given values of *k* and *h*, we fitted a logistic regression with four parameters (upper bound, lower bound, inflection point, and steepness) to maximize the coefficient of determination, $R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$. For Experiment 2, the largest value of R^2 was $R^2 = .977$, obtained with *h* = 8 and *k* = 24. The resulting logistic function is shown in the right panel of Figure 4.4. Because we observed no effect of cognitive demand in

Experiment 1, only one free parameter, *h*, was needed to reasonably explain choices in this experiment, and Φ would be

$$\Phi = \begin{cases} duration, & Bucket\ weight = 3.2 \\ duration + h, & Bucket\ \ weight = 0 \end{cases}$$

The largest value $R^2 = .976$ was obtained for $h = 7$ in this case and the resulting logistic function is shown in the left panel of Figure 4.4. When setting $h = 8$, that is, to the same value as obtained when fitting the model in Experiment 2, the fit was negligibly worse with $R^2 = .970$. The resulting logistic function is shown as the red line in the left panel of Figure 4.4.

Figure 4. 4. *p*(Buket) as a function of Φ in both experiments.



Note: Left panel: Probability of choosing the bucket task, *p*(Bucket), in Experiment 1 as a function of a model where cognitive demands are the same in the addition and the multiplication task and bucket load is converted to subjective time. The black line is the logistic function yielding the best fit to the data. The red line is the resulting function when using the same parameter values as obtained with the data from Experiment 2. Right panel: Probability of choosing the bucket task, *p*(Bucket), in Experiment 2.

The foregoing analysis shows that all of the choice data could be modeled by expressing costs of all kinds in terms of a transform of objective time, referred to here as subjective time, though we did not explicitly ask participants to estimate time – a further validation step that can be pursued in future work. Because the model was so simple and successful, we can speculate that it might underlie a simple process model for making the choices: If the participant's estimate of the time to do a task falls below a criterion value, choose that task; otherwise, choose the other task; in case the two estimates are the same, choose at random.

Reflecting on what we have achieved here, we note that while separate studies have compared the effects of each of the three costs that we examined – cognitive demand, physical demand, and time – we have shown that these factors are lawfully convertible to an internal transform of objective time. This outcome confirms and extends the observation of Potts et al. (2018) that choices can be explained with reference to subjective time. The present result also lends support to another study by Rosenbaum and Bui (2019), where it was found that subjective time did a better job of accounting for choices than another hypothesized measure of task difficulty, task sustainability. Task sustainability was not tested in the present experiments, nor was subjective time directly assessed here by gathering subjective time estimates from our participants, so more work is needed to evaluate the role of these hypothesized quantities. Nevertheless, the advance of the present work and the promise of our general approach is clear. By asking people to choose the easier of two tasks, the choices they made turned out to be reliable and systematic. Furthermore, through modeling, we were able to map costs of different kinds onto a single metric.

Characterizing the psychological representation of task difficulty is an important challenge. We are heartened by the progress we have made in addressing this challenge. By building on the work of others as well as previous work by us, we have been able to show that the challenge of explaining perceived task difficulty may not as hard as first imagined.

References

Ashcraft, M. H., & Guillaume, M. M. (2009). Mathematical cognition and the problem size effect. In B. Ross (ed.), *The psychology of learning and motivation, Vol. 51* (pp. 121-151). Burlington, MA: Academic Press.

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

Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research PRPF, 73*, 835-842.

Craig, A. D. (2013). An interoceptive neuroanatomical perspective on feelings, energy, and effort. *Behavioral and Brain Sciences, 36*, 685-686.

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

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

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

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

Dunn, T. L., & Risko, E. F. (2019). Understanding the cognitive miser: Cue-utilization in effort-based decision making. *Acta Psychologica*, *198*, 102863.

Feghhi, I., & Rosenbaum, D. (2019). Judging the subjective difficulty of different kinds of task. *Journal of Experimental Psychology: Human Perception & Performance, 45,* 983-994.

Feghhi, I., & Rosenbaum, D. (2020). Effort avoidance is not simply error avoidance. *Psychological Research.*

Feghhi, I., & Rosenbaum, D. (in preparation). Cross-modal difficulty comparison: Math versus muscle.

Gold, J. M., Kool, W., Botvinick, M. M., Hubzin, L., August, S., & Waltz, J. A. (2015). Cognitive effort avoidance and detection in people with schizophrenia. *Cognitive, Affective, & Behavioral Neuroscience, 15, 145-154.*

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

Grondin, S. (Ed.) (2008). *Psychology of time*. Bingley, U.K, Emerald.

Job, V., Dweck, C. S, & Walton, G. M. (2010). Ego-depletion—Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science, 21,* 1686-1693.

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

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

Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology, 44,* 193-251.

Ornstein, R. E. (1969). *On the experience of time*. Harmondsworth: Penguin.

Potts, C. A., Callahan-Flintoft, C., & Rosenbaum, D. A. (2018). How do reaching and walking costs affect movement path selection? *Experimental Brain Research, 236,* 2727-2737.

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

Rosenbaum. D. A. (2008). Reaching while walking: Reaching distance costs more than walking distance. *Psychonomic Bulletin & Review, 15,* 1100-1104.

Rosenbaum, D. A., Brach, M., & Semenov, A. (2011). Behavioral ecology meets motor behavior: Choosing between walking and reaching paths. *Journal of Motor Behavior, 43,* 131-136.

Rosenbaum, D. A., & Bui, B. B. (2019). Does task sustainability provide a unified measure of subjective task difficulty? *Psychonomic Bulletin & Review, 26,* 1980-1987.

Rosenbaum, D. A., Gong, L., & Potts, C. A. (2014). Pre-crastination: Hastening subgoal completion at the expense of extra physical effort. *Psychological Science, 25,* 1487-1496.

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 3-18.

Weigelt, M., Rosenbaum, D. A., Huelshorst, S., & Schack, T. (2009). Moving and memorizing: Motor planning modulates the recency effect in serial and free recall. *Acta Psychologica, 132,* 68-79.

Zhang, L., Wininger, M., & Rosenbaum, D. A. (2014). Word generation affects continuous hand movements. *Journal of Motor Behavior, 46,* 115-123.

Zhou, X., Chen, C., Zang, Y. et al. (2007). Dissociated brain organization for single-digit addition and multiplication. *NeuroImage, 35,* 871-880.

CHAPTER 5 – TOWARDS A COMMON CODE FOR DIFFICULTY: NAVIGATING A NARROW GAP IS LIKE MEMORIZING AN EXTRA DIGIT

Iman Feghhi, John M. Franchak, and David A. Rosenbaum
Department of Psychology
University of California, Riverside

Accepted for publication in Attention, Perception, and Psychophysics (2021)

Abstract

What makes a task hard or easy? The question seems easy, but answering it has been hard. The only consensus has been that, all else being equal, easy tasks can be performed by more individuals than hard tasks, and easy tasks are usually preferred over hard tasks. Feghhi and Rosenbaum (2019) asked whether task difficulty might reflect a single amodal quantity. Based on their subjects' 2-alternative forced choice data from tasks involving choices of tasks with graded physical and mental challenges, the authors showed that the difficulty of passing through a narrow gap rather than a wide gap was psychologically equivalent to memorizing an extra .55 digits. In the present study, we extended this approach by adding new arguments for the hypothesis that task difficulty might reflect a single amodal quantity (inspired by considerations of physics, economics, and the common code hypothesis for the study of perception and action), and we tested narrower gaps than before to see whether we would find a larger equivalent memory-digit. Consistent with our prediction, we obtained a value of .95. We suggest that our

multi-modal 2-alternative forced choice procedure can pave the way toward a better understanding of task difficulty.

Keywords: decision making, mental effort, metacognition, physical effort, task difficulty

Introduction

What makes a task hard or easy? Electrons have no trouble deciding. They take the path of least resistance, not knowing which way to go, but bunching up in areas of high resistance and veering toward areas of lower resistance. People behave similarly when heading for wide rather than narrow exits in theaters and stadiums. In both cases, the structure of the environment specifies path ease.

Physical systems are replete with such examples: Water tends to flow down steeper slopes, and light travels down least-time paths in accord with Fermat's Principle (https://en.wikipedia.org/wiki/Fermat%27s_principle). Such examples illustrate a foundational principle of physics, the Law of Least Action (https://en.wikipedia.org/wiki/Principle_of_least_action).

The Law of Least Action has been applied to living systems, including human beings. In one of the best known examples, the American linguist/mathematician George Kingsley Zipf (1949) offered the Law of Less Work. According to the Law of Less Work and as expressed here in our words, "The more common a word is, the shorter it is on average." Consistent with the Law of Less Work, word frequency follows a power-function. The second-most common word is half as frequent as the most-common word, the third-most common word is half as frequent as the second-most-common word, and so on. If the Law of Less Work were not operative, information communication would be far less efficient than it is. As Zipf emphasized, communication, like light, minimizes time.

If the Law of Least Action holds for all the elements referred to above – electrons, water, light, and words – then it is natural to ask whether time, the fundamental value in all the cases listed, is sufficient to explain task difficulty? It might be, as several authors have suggested (Gray et al., 2006; Potts et al., 2018; Rosenbaum & Bui, 2019). Clearly, running at top speed for 10 minutes is harder than running at top speed for 5 minutes. However, a problem arises: Running at top speed for 10 minutes is also harder than walking for 11 minutes. Accordingly, time is dissociable from task difficulty (e,g., Kool et al., 2010).

Should the Law of Least Action be repealed, then, for human action? There may be a way to resolve the the problem associated with the walking-for-11-minutes-versus-running-for-10-minutes example. The time to rest and recover from an 11 minute *walk* is less than the time to rest and recover from a 10 minute *run*. Considering the full cycle time to engage and re-engage in the two tasks could explain why the 10 minute run seems harder than the 11 minute walk. If total time is considered, the short-duration run will seem harder than the long-duration walk, consistent with Fermat's Principle and, by extension, the Law of Least Action.

Do people actually think about rest and recovery times when answering questions like this one? We can defer that question because time needn't always be referred to in considerations of task difficulty. For example, time does not arise (in any obvious way) in connection with the shape of a chain suspended between two posts (a catenary). The arc form of the catenary reflects the Law of Least Action and is often used as an example of it. The shape of the catenary is the one that minimizes the difference between potential

energy and kinetic energy (another way of expressing the Law of Least Action) and remains the same over time, provided there is no external disturbance.

These remarks suggest that, more likely than not, the Law of Least Action may underlie perceived task difficulty, even in view of evidence for various specific proposals abut the currency underlying this psychological quantity, including energy (Craig, 2013), opportunity cost (Kurzban et al., 2013), and errors (Dunn et al., 2019). The central claim of this paper is that it may be pointless to try to distinguish among particular alternative accounts of task difficulty even though thinkers from many disciplines have tried to do so, including people working in philosophy, sport science, psychology, language, education, and robotics (André et al., 2019; Burgess & Jones, 1997; Cos, 2017; Fisher & Steele, 2014; Halperin & Emanuel, 2020; Montero, 2016; Pageaux, 2014; Shenhav et al., 2017; Song et al., 2019; Steele, 2020). We think it is very unlikely that one account will be correct in all circumstances because context always matters. For example, people may be willing to pay a lot of money for the most relaxing ride possible (one end of the energy continuum) or for membership in a gym affording the most intense workout imaginable (the opposite end of the continuum). Notwithstanding such circumstantial changes, we hypothesize that within a bounded set of circumstances, a single quantity might be able to explain task ease. We are especially interested in the possibility that the quantity might be abstract and amodal. In much the same way that the difference between potential energy and kinetic energy – an abstract quantity and not one that can be directly or immediately sensed – appears to underlie all of physical efficiency, the true measure of task difficulty might be similarly abstract. We aim here to test for such a quantity. Our pursuit is

motivated not just by physics and related fields, but also by economics, where value is treated as an abstract quantity (https://en.wikipedia.org/wiki/Theory_of_value_(economics)), and, closer to home, the demonstration of a common code for perception and action (Prinz, 1990; Prinz & Hommel, 2002). The common code hypothesis for perception and action has inspired us to hypothesize that there is, likewise, a common code for difficulty.

Lead-Up To The Present Two Experiments

In this article, we will report two experiments based on an earlier pair of experiments by Feghhi and Rosenbaum (2019). These authors inquired into the possibility that task difficulty might reflect a single abstract quantity. They provided university students with two task options, each of which had varying degrees of physical and mental demands. The participants chose between carrying an empty box through a wide gap (81 cm) or a narrow gap (36 cm), having memorized 6, 7, or 8 digits before passing through either gap. The instruction was to do whatever seemed easier – memorizing the list associated with the wide gap and then going through that gap, or memorizing the list associated with the narrow gap and then going through that gap, knowing that the list that had been memorized would have to be recalled upon reaching the other side. Each list length was offered with each gap size in all possible pairs and with the wide or narrow gap on the right or left for all participants. From the obtained 2-alternative forced-choice data, Feghhi and Rosenbaum estimated the point of subjective equality for the wide and narrow gap, expressed in number of digits. They found that

going through the narrow gap was functionally equivalent to memorizing an extra .55 digits.

In their second experiment, Feghhi and Rosenbaum (2019) tested a fresh sample of participants on the same tasks, except that now they dictated to those subjects which task should be done. In that case, the obtained performance data (i.e., the error rates) were virtually identical to what they were in the choice condition. This outcome provided assurance that the results of the first experiment were not biased by unequal numbers of observations in the conditions for which data existed. Beyond that and more importantly, Feghhi and Rosenbaum concluded that having to choose a task or having been told what to do did not affect accuracy; subjects had made wise choices when they could choose.

The two experiments reported here were modeled on the two that Feghhi and Rosenbaum (2019) conducted.[9] In the present Experiment 1, subjects chose walking paths and associated memory loads, as in the earlier study. In the present Experiment 2, subjects were assigned each of the walking paths and associated memory loads of the first experiment, as in the 2019 study. The new feature of the present experiments was that we used a narrower gap than the narrower gap used before. We were motivated to do so because navigation errors (bumping into an obstacle while passing through the narrow gap) were rare in the 2019 experiments.[10] We used a narrower gap here to challenge the

---

[9] Another pair of experiments, by Feghhi and Rosenbaum (2020), replicated the main features of the 2019 results, but aspects of the 2020 procedure were sufficiently different from those of the 2019 report (using auditory inputs rather than visual inputs for the to-be-memorized materials) that we merely mention the replication here in passing.

[10] There were no navigation errors at all while passing through the wide gap in the earlier experiments, and this fact let us use a slightly narrower wide gap here, owing to the pre-existing physical structure of the

perceptual-motor system more than in Feghhi and Rosenbaum's (2019) study. We predicted that by making the narrow gap narrower we would increase the navigation error rate and, more importantly, would elevate the point of subjective equality for the wide and narrow gap, expressed in number of digits. Whereas Feghhi and Rosenbaum found that going through the narrow gap was functionally equivalent to memorizing an extra .55 digits, we predicted that with an even narrower gap, this value would increase. By how much we could not say; too little data exist in this line of work to allow for a more precise prediction.

A further refinement of the method was that we tailored the gap sizes to individual subjects. To do so, we took advantage of the apparatus and expertise of Franchak (2017, 2020) and Labinger et al. (2018), who studied gap clearance using a sophisticated apparatus that had two gaps with sliding doors. This apparatus allowed the widths of the openings to be adjusted with high resolution (0.5-cm increments) to befit, or not, the features of individual subjects. As in the previous work with this apparatus, we were interested in adjusting the width of the aperture to each individual's body size.

---

apparatus for the present experiments. Whereas the wide gap was 81 cm wide before, here it was 70 cm. We expected no navigation errors with the slightly narrower wide gap here.

Experiment 1

Method

Participants

Forty-two undergraduate students (32 female and 10 male) from the University of California, Riverside, participated in this experiment for course credit. The participants ranged in age from 18 years to 24 years, with an average of 19.41 years and a standard deviation of 1.04 years. All participants signed an informed consent form before the experiment. The current sample size was similar to the sample size of Feghhi and Rosenbaum (2019), who tested 40 subjects. That number let us exceed the value of n = 500 observations recommended for evaluation of logistic regression models (Cohen, Cohen, West, & Aiken, 2013; Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997). With 40 subjects, the number of choices per participant was 18, so there were 720 observations altogether. Two more subjects offered their services here via the UCR Psychology subject pool (where students get course credit for participating). We were happy to have 42 subjects rather than 40 subjects in this experiment.

Apparatus

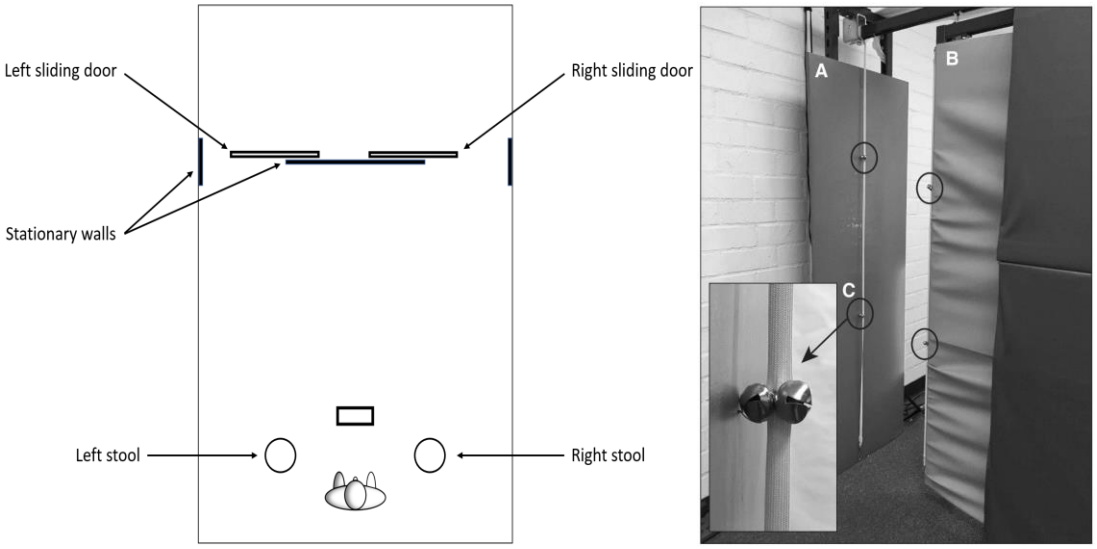
As shown in Figure 5.1, at the start of each trial, participants stood at a home position and saw two lists of digits. One list lay on a stool 90 cm to the left and another list of digits lay on a 90 cm stool to the right. An empty box (48×48×10 cm) stood on a platform (a music stand tilted to be parallel to the ground, 95 cm high above the floor) in front of the subject, who could see two doorways 275 cm away from the home position. The widths of the doorways could be adjusted between 0 and 70 cm (with a resolution of .5 cm) by sliding the doors (185 cm tall × 100 cm wide) along a perpendicular stationary wall (182 cm tall × 62 cm wide). A locking mechanism prevented the doorways from moving when the mechanism was engaged.

On each trial, one of the doorways was kept at a fixed width of 70 cm, which was the widest possible width. We assumed that navigating through the wide doorway provided minimal challenges for all the participants. The other doorway's width was adjusted based on each participant's body size. The narrow doorway's width was based on a calibration process such that each participant had a 50% chance of bumping the edges of the doorway. To detect bumping, two elastic bands were aligned with the edges of the doorways. Small bells were attached to the bands and participants were told to avoid bumping into the elastic bands to prevent the bells from ringing.

After passing through either doorway, the participant attempted to recall the digit list in the presence of an experimenter who awaited the participant's arrival. The experimenter stood between the two doorways and was unable to see the participant until s/he participant entered the post-doorway area. After the recall phase, the experimenter

opened the doorway all the way, so on the way back, the participant did not have to pass through a narrow doorway.

Figure 5. 1. Setup in Experiment 1.



Note: The left panel shows a schematic birds-eye view of the apparatus. The right panel shows the left doorway. In the left panel, the left doorway is narrow and the right doorway is wide. There are three stationary walls, one in the middle and one on each side. The right panel shows the stationary wall (A) parallel to the sliding wall (B) as well as the other stationary wall (C). The magnified inset in the right panel shows one of the four pairs of bells located at the circled areas.

Procedure and Design

After signing the consent form, each participant went through the calibration process that was used to determine the doorway width for each participant that was narrow enough for each of them to have a 50% chance of making a mistake in passing through it (i.e., bumping into an elastic band, causing a bell to ring). Clearly, the participants' body sizes were only one factor in determining the narrow doorway width.

Spatial awareness, controlling body sway, dynamic balance, and practice were also determinative factors. We did not attempt to determine which of these factors contributed to any participant's doorway width.

A single doorway was used during the calibration trials. On each calibration trial, it was set to a width between 35 cm and 60 cm in 0.5-cm increments that participants were requested to attempt to pass through. They were instructed to turn their body and walk sideways to clear the doorway. The doorway width on each trial was set to find the 50% threshold based on the outcome of the previous trial (successfully passing through versus bumping into the side of the doorway). Over the first five trials, a binary search procedure was used, as in Franchak et al. (2010), to find a doorway width close to the 50% point. Another 15 trials was then used to further adjust the doorway width, decreasing it by 1.5 cm or increasing it by 2 cm if the participant succeeded or failed, respectively, on the previous trial. A cumulative Gaussian function was fitted to the data from the calibration trials using the Palamedes Toolbox (Prins & Kingdom, 2018). Ultimately, the doorway width that was used as the narrow width in the main experiment per participant was the width for which that participant could pass through the doorway without causing a bell to sound 50% of the time.

After the calibration process, the participant was asked to stand at the home position. There they saw digit lists (6, 7, or 8-digit random numbers), each of which was printed on a piece of paper and placed on a stool to the subject's left and right. A box (empty rectangle in the left panel of Figure 5.1) stood on another stool (95 cm height) directly in front of the subject and within easy reach. The box was empty and measured

48×48×10 cm. The subject could also see the two doorways (275 cm away), one to the right and one to the left. One of the doorways was always wide (70 cm) and the other was always narrow, set individually for the subject based on the calibration procedure described above. For a random half of the participants, the right doorway was wide for the first nine trials and narrow for the next nine trials. For the other participants, it was the other way around. All participants had 18 trials for the 9 conditions. As a result, the choice data per condition had two observations per participant. With 42 participants, this meant that the possible proportions per condition (i.e., the possible values of *p*(Wide), the probability of choosing the wide gap) were 0/84, 1/84, …, 84/84. The information content was therefore $\log_2(85) = 6.40$ bits.

The participants' task was to do whatever seemed easier, memorize the digit list on the left and then carry the box through the left door, or memorize the digit list on the right and then carry the box through the right door. Participants were told that there was no time limit for memorizing the digit lists or for passing through the door and setting the box down on the target platform. Once they thought they had memorized the lists, they picked up the box and started walking toward the selected doorway. After passing through the chosen door and setting the box down, they tried to recall the digits of the list for the side they had chosen, having been told that order mattered; the digits were to be recalled in the left-right order in which they appeared. Participants were told that if they made a mistake, they would have to redo the trial. A mistake was defined as causing a bell to ring while passing through a door or misrecalling the digits in any way (i.e.,

naming a digit not on the list or recalling the digits in the wrong order). If a mistake was made in a redo trial, the trial did not have to be repeated again.

Results

Number of Choices and Error Rates

Table 5.1 shows the number of times the tasks with different door widths and memory loads were chosen as well as the associated error rates. As seen in Table 5.1, the wide-door option was chosen more often than the narrow-door option, and paths with smaller memory loads were chosen more often than paths with larger memory loads. In addition, error rates of any kind were inversely related to the number of chosen options.

Regarding the two kinds of errors, the probability of recall error, p(R), was inversely related to the number of chosen options. No navigation errors occurred when participants passed through the wide doorway. When participants passed through the narrow doorway, the memory load had little or no effect on the probability of a navigation error, p(N).

Table 5. 1. Main results of Experiments 1 and 2 in the six conditions. The entries are the number of trials, N, in which each door width and memory load combination was chosen; the probability, *p*(Error), of an error of any kind; the probability, *p*(R), of a recall error; and the probability, *p*(N), of a navigation error.

	Experiment 1				Experiment 2			
Condition	N	*p*(Error)	*p*(R)	*p*(N)	N	*p*(Error)	*p*(R)	*p*(N)
Wide-6	236	.11	.11	0	84	.13	.13	0
Wide-7	190	.25	.25	0	84	.19	.19	0
Wide-8	120	.33	.33	0	84	.35	.35	0
Narrow-6	129	.40	.20	.29	84	.38	.22	.31
Narrow-7	73	.47	.30	.29	84	.48	.28	.29
Narrow-8	44	.61	.41	.32	84	.57	.39	.33

To analyze the effect of physical and mental demands on error rate, we conducted a General Estimating Equations (GEE) analysis of the probability of any kind of error, *p*(Error), and the probability of error in recall, *p*(R). We did not conduct a GEE analysis on the probability of error in navigation, *p*(N), because *p*(N) in the wide gap was 0. When a predictive variable perfectly predicts the outcome (in our case, going through the wide gap perfectly), there is a "quasi-complete separation" problem (Albert & Anderson, 1984), which makes the maximum likelihood calculation impossible. With the GEE analysis, using a 2 (wide and narrow doorways) by 3 (6, 7, and 8 digits) design, we found that *p*(Error) showed a main effect of memory load, Wald Chi-Square = 4.32, $p = .03$, such that *p*(Error) with memory load of 6 (.22 95% CI [.17 .28]) was lower than *p*(Error) with memory load of 7 (.35, 95% CI [.27, .44]) and was lower than p(Error) with memory load of 8 (.47, 95% CI [.37, .58]). There was a main effect of doorway width, Wald Chi-Square = 6.38, $p = .01$ such that p(Error) for the wide doorway (.21, 95% CI [.16, .28])

was lower than p(Error) for the narrow doorway (.49, 95% CI [.40, .59]). There was no interaction between door width and memory load, Wald Chi-Square = 2.33, p = .12. A 2 (wide and narrow doorways) by 3 (6, 7, and 8 digits) GEE analysis on p(R) showed a main effect of memory load, Wald Chi-Square = 6.83, p = .009, no main effect of doorway width, Wald Chi-Square = 1.78, p = .18, and no interaction between these factors, Wald Chi-Square = 0.96, p = .32.

Choices

Whereas Table 5.1 showed the total number of times that participants chose a task option with the characteristics listed per row, those numbers do not break down how often each task option was chosen depending on the other task with which it was paired. Table 5.2 shows the relevant data, now expressed in proportions rather than total numbers. The table shows the probability, p(Wide), of choosing the wide doorway depending on the number of digits to be memorized for the wide versus narrow doorway.

Table 5. 2.Probability of choosing the wide gap, p(Wide), in the nine memory load conditions of Experiment 1 (along with 95% confidence intervals).

	Narrow gap		
Wide gap	6	7	8
6	.82 (.74, .90)	.94 (.89, .99)	.92 (.86, .98)
7	.49 (.38, .60)	.77 (.68, .86)	.90 (.83, .96)
8	.23 (.14, .32)	.45 (.35, .56)	.68 (.58, .78)

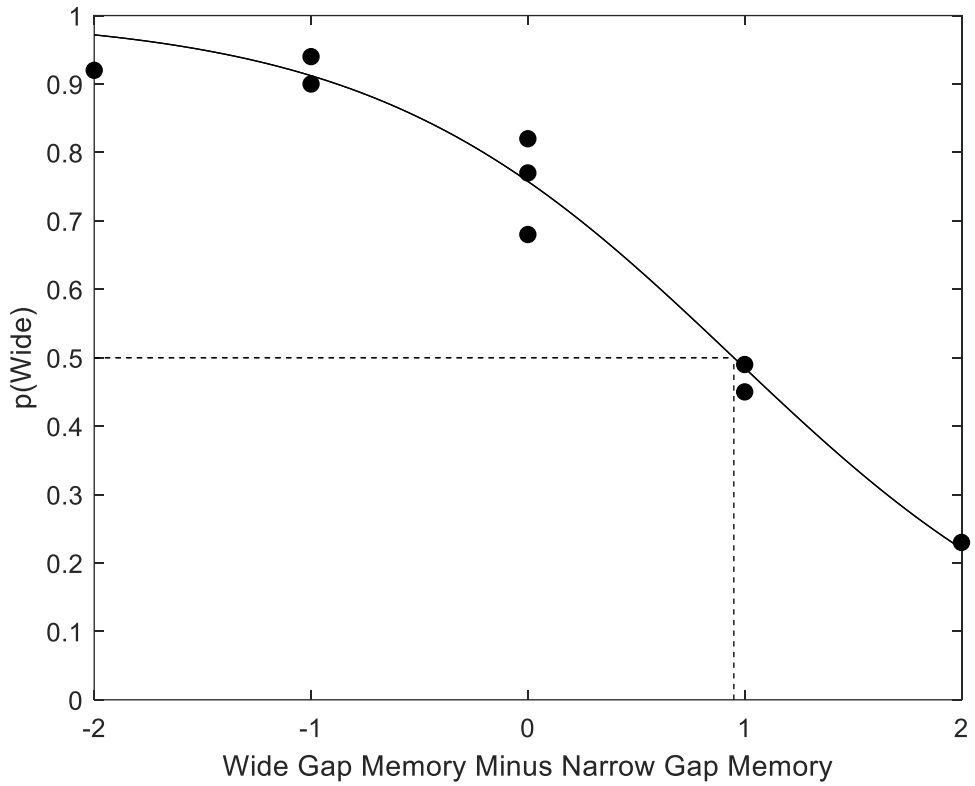
As seen in Table 5.2, the wide gap was chosen less often as its associated memory load increased. The values decreased from the first row down to the third. The wide gap was chosen more as the narrow gap memory load increased. The values increased from the first column to the last.

We sought to put these values together into a single mathematical model whose constructs could be related to the putative steps involved in choosing the task alternatives. We assumed, that by default, participants preferred the wide door, but if the difference between the wide-door memory load and the narrow-door memory load exceeded a threshold, the preference would switch to the narrow door, doing so with increasing probability the greater the difference between the narrow-door memory load and its threshold. To express the model in an equation, we used a logistic function with two free parameters, the critical memory-load difference or switching point, *S*, and the decisiveness of the decision, visualized as the steepness, *K*, of the curve:

$$p(Wide) = \frac{1}{1 + e^{-K(x-S)}}$$

The best fit is shown in Figure 5.2. The parameter values that provided the best fit were $S = 0.95$ and $K = 1.2$. The interpretation of $S = .95$ was that going through the narrow doorway was equivalent, in terms of difficulty, to memorizing an extra .95 digits on average. The model accounted for $R^2 = .97$ of the variance in the observed probabilities.

Figure 5. 2. Probability of choosing the wide gap, *p*(Wide), as a function of the difference between the memory load of the two doorways.



Note: The black dots show the observed probabilities (aggregated single values of 0 or 1 for each participant), and the curve shows the model's best fit. The dashed lines show the switch point. Multiple black dots appear at some horizontal positions because there were multiple conditions with that memory load difference. There were two such conditions for the differences of -1 and 1, and three such conditions for the difference of 0. There was only one condition for which the memory was -2, and only one condition for which the memory was +2.

Discussion

The purpose of Experiment 1 was to replicate the first experiment of Feghhi and Rosenbaum (2019) using a narrower gap than the narrow gap of the 2019 study. Although the original narrow gap yielded a few navigation errors in the 2019 report – subjects bumped into the edge of the narrow gap at most 5% of the time – we predicted that a more challenging navigation task would give rise to more navigation errors and, more interestingly, a rise in the estimated memory-load equivalence. Using an adaptive procedure to set the width of the narrow aperture, we succeeded in increasing the likelihood of navigation errors, though we failed to get the navigation errors up to p(N) = .5. Possibly, practice navigating through the gap during the calibration task helped participants improve subsequent navigation. But more importantly and more interestingly, we obtained a rise in the associated memory-load estimate, from .55 in the 2019 study to .95 in the present study. By considering the unexplained variance of the logistic function when S was set to .95 (the best value in this study) versus .55 (the best value in the 2019 study), and keeping K at 1.2 in both cases, we determined that the present data were 4.97 times more likely to have come from a logistic function whose *S* value was .95 than from a logistic function whose *S* value was .55. The method we used to arrive at this value was the one introduced by Glover and Dixon (2004).

Experiment 2

The second experiment was designed to address the same question as the one addressed in the second experiment of Feghhi and Rosenbaum (2019): Did participants' choices

reflect their actual abilities? A subordinate, less interesting, question was whether the choice data were unduly influenced by unequal numbers of observations in the choices made? It was possible that they could have been.

As in the earlier study, we eliminated choices in Experiment 2 and asked participants do each of the possible tasks that were available to the participants in Experiment 1. We reasoned that if participants' choices reflected their actual abilities and if the choice data were not unduly influenced by unequal numbers of observations in the choices provided, the error data of Experiment 2 would be the same as the error data of Experiment 1.

Method

Participants

Forty-four undergraduate students (33 female and 11 male) from the University of California, Riverside, participated in this experiment for course credit. The participants ranged in age from 18 years to 24 years, with an average age of 19.98 years and a standard deviation of 1.29 years. All participants signed the informed consent form before the experiment. The larger number of subjects in this experiment compared to the last one was simply motivated by wanting to help students get their needed academic credit for their Intro-Psych class. As before, we were happy to test a few more subjects who volunteered than were strictly required or invited.

Apparatus

The apparatus was the same as Experiment 1, but, at the start of each trial, only one of the stools had a set of 6, 7, or 8 random digits on it. The side of the stool indicated the doorway to be traversed.

Procedure and Design

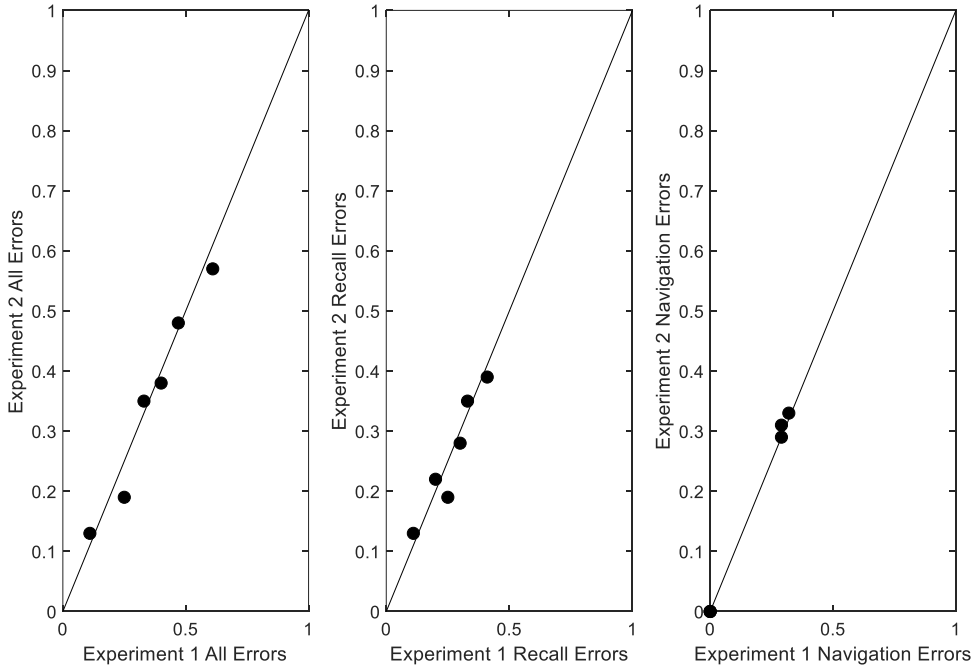
At the start of each trial, the experimenter put a piece of paper on the left or right stool. Participants were asked to memorize the digit list, pick and carry the empty box through the doorway on the corresponding side and then recall the numbers. For a random half of participants, the left doorway was narrow in the first trials and wide in the last trials. For the other half of the participants, the order was reversed. The same calibration procedure for determining the door width per participant was used here as well.

Results and Discussion

Error Rates

The data from this experiment were already shown in Table 5.1. As seen there, the error rates in Experiment 2 were remarkably similar to the error rates in Experiment 1. This is shown in graphical form in Figure 5.3.

Figure 5. 3. Error rates in Experiment 2 plotted as a function of error rates in Experiment 1.



Note: The leftmost graph is for all errors, the middle graph is for recall errors, and the right graph is for navigation errors.

To test the similarity between the two sets of results, we conducted a 2 (wide and narrow doorways) by 3 (6, 7, and 8 digits) by 2 (experiments) GEE analysis on *p*(Error). The results showed a main effect of memory load, Wald Chi-Square = 44.82, $p < .001$, a main effect of doorway width, Wald Chi-Square = 71.082, $p = .01$, but no effect of experiment Wald Chi-Square = 0.12, $p = .73$.

The result was clear. Removing the choices, which forced performance of the indicated tasks an equal number of times per condition, yielded the same pattern of errors in the two experiments. Participants' choices in Experiment 1 reflected their actual abilities.

General Discussion

In this article, we have described two experiments aimed at establishing the relation between two different kinds of variables: the difficulty of a perceptual-motor task, and the difficulty of a mental (memory) task. We reasoned that if these two kinds of variables could not be compared, people would be unable to choose between them in a systematic fashion; their choices would be chaotic, and scientists like us would be unable to make principled predictions about the choice data we obtain. Our results let us reject this hypothesis. We made a specific prediction and we obtained data consistent with it.

We asked participants to choose and perform the easier of two options: memorizing 6, 7, or 8 random digits and going through a wide gap, *or* memorizing 6, 7, or 8 random digits and going through a narrow gap. In an earlier study, Feghhi and Rosenbaum (2019) introduced this task with gaps that were 81 cm wide and 36 cm wide. Feghhi and Rosenbaum found that participants were willing, on average, to memorize .55 more digits to avoid the narrow gap. In the present experiment, we made the narrow gap narrower and found that participants were willing, on average, to memorize .95 more digits to avoid the narrow gap. We reached this estimate by fitting a logistic function to the choice data. According to the process model underlying the logistic function, participants would prefer the wide gap by default but would switch to the narrow gap if the wide-gap memory load exceeded a threshold value. That value turned out to be .95 digits. We could show that our choice data were nearly 5 times more likely to have come from a source in which participants were willing to pass through the narrow gap when its

memory load had .95 fewer items than when its memory load had .55 fewer items, which was the estimate from the previous experiment. Similarly, we could show – and this is a new statistic, not reported earlier in this article – that the choice data from the previous experiment were 8.42 times more likely to have come from a source in which participants were willing to pass through the narrow gap when its memory load had .55 fewer items than when its memory load had .95 fewer items.

In the remainder of this General Discussion, we take up five remaining issues: (1) the relation between *p*(Wide) and *p*(Error); (2) the possibility that mappings between memorial difficulty and physical difficulty may suffice without positing an abstract, amodal representation of difficulty per se; (3) the value of pursuing numerical values in research about action, perception, and psychophysics as well as related fields; (4) the promise of our approach, with special reference to the use of the 2-alternative forced choice procedure; and (5) the limitations of the present study.

Regarding the first issue, the relation between *p*(Wide) and *p*(Error), it is interesting to pursue the possibility that in Experiment 1, these two variables had a simple relation and, moreover, that when *p*(Wide) was plotted as a function of *p*(Error), the point of subjective equality would land squarely on *p*(Error)=.5. Such an outcome would accord with the hypothesis that the decision to go through the wide or narrow gap was based on the desire to minimize error, for at *p*(Error)=.5 the likelihood of error would be indistinguishable for the two gaps. It is certainly plausible that the desire to minimize error could be the sole driver of choice. Dunn et al. (2019) proposed that the more error-prone a task, the more difficult it is perceived to be. It is also known that similar brain

regions are active after making a mistake (Baker & Holroyd, 2011; Miltner et al., 2003) and in value evaluation and effort exertion (Apps et al., 2015; Apps & Ramnani, 2014; Mulert et al., 2005; Shenhav et al., 2013; Walton et al., 2003). These observations indicate that errors can be perceived as costly and therefore to be avoided.

When we fitted a logistic function to data points for *p*(Wide) plotted as a function of *p*(Error), we found that the coefficient of determination, R^2, was comparable to what it was for the logistic fit in Figure 5.2 (very high). However, we found that the point of subjective equality was at *p*(Error)=.39 rather than at *p*(Error) = .50. This result is not consistent with the hypothesis that the choice of gap was solely designed to reduce *p*(Error). Interestingly, analogous results were also found by Feghhi and Rosenbaum (2019) and by Feghhi and Rosenbaum (2020), who placed so much weight on this finding that they entitled their article "Effort avoidance is not simply error avoidance." This conclusion makes sense considering that not all errors are equally costly. Slipping off a stone in one's garden has a very different cost than slipping off a ledge on the edge of a cliff with a thousand foot chasm beneath it.

Regarding the second issue, the possibility that mappings between memorial difficulty and physical difficulty may suffice without positing an abstract, amodal representation of difficulty per se, we cannot rule out this possibility for the data we have. Conceivably there may be values of memorial difficulty and values for physical difficulty with some mathematically well-defined mapping between the two, with no intervening representations. On the other hand, neural network modeling has shown that neural networks capable of reasonably complex learning must have an intermediate hidden layer

as well as an input layer and an output layer. In our case, the input layer could be for memory difficulty and the output layer could be for physical difficulty; we have no way of distinguishing between these possibilities and have no reason to try. The important point is that the intermediate hidden layer would be task difficulty. Given that intermediate hidden layers are well known to be essential for successful neural modeling of attention, perception, and psychophysics, it is hardly surprising that extensive evidence exists for an abstract, amodal common code for perception and action (Prinz, 1990; Prinz & Hommel, 2002). We therefore think that difficulty is also represented in some abstract, amodal common code probably represented in a hidden intermediate layer of the relevant neural substrate.

Regarding the third issue, the value of pursuing numerical values for research in this area, we have been moved by recent arguments from Yarkoni and Westfall (2017), who have suggested that models in psychological science should be able to predict new numerical values, much as physics and other sciences have long done. The numerical prediction we made here was primitive by the standards of physics, for all we could predict was that the value of S would be larger than in the predecessor study. That prediction was supported, suggesting we were on the right track. In a future study, we might next ask a more subtle question such as this: Over a range of narrow gap sizes in a within-subject design, with 6, 7, or 8 memory items per choice and a fixed-width wider gap, how will S vary with the size of the narrow gap? Will S be a linear function of the narrow gap size or a logarithmic function? Science progresses by answering questions of this sort.

Regarding the fourth issue, concerning the promise of our approach with special reference to the use of the 2-alternative forced choice procedure, we would like to point out that the procedure we have used here has proven, time and again, to yield lovely, interpretable data (e.g., Rosenbaum et al., 2013). We have sufficient faith in the 2-alternative forced choice procedure to recommend it to others interested in assessing the perception of task difficulty, both in basic research where it can add to the understanding of multi-modal experience and be useful in applied contexts. For example, in clinical settings, the method can be used to show how patients view the difficulty of performing a task. If hemiparetic patients judge the difficulty of moving an affected arm as being comparable to the difficulty of memorizing *five* digits soon after stroke but as being comparable to memorizing *two* digits later on, that outcome can provide a quantitative index of the change in the judged difficulty of the arm-movement task. If it is clear that the memory abilities remain the same, the measured change can be used to gauge recovery and design future treatments.

We turn finally to the fifth issue, the limitations of the present study. In the current work, we investigated a small range of memorization demands (6, 7, and 8 digits) and only two levels of navigation demands. Based on the common code hypothesis, these two demands should be lawfully comparable in other ranges as well. That said, we make no claim how the relationship would change outside the ranges used here – for example, whether the perceived difficulty of the same navigation challenge would be similar to what we measured here if we used 2, 3, and 4-digit lists. This topic needs more investigation.

The common code hypothesis also predicts that other aspects of a task, like energy expenditure, time, utility, and consequence of mistakes, should be convertible to the perceived difficulty and hence be systematically comparable. Given that different demands have different levels of evaluability (Dunn et al., 2017), further experiments are needed to better understand how different demands are compared. We did not explore all of these potential contributors to perceived difficulty. For example, we did not track possible differences in speed-accuracy tradeoffs.

Lastly, measuring each participant's digit span could help reveal the impact of navigation on memory performance. Measuring each participant's digit span could also be used to equate the memorization challenge across participants and thereby have more control over the demands of the memorization tasks. Pursuing this question, like the others raised above, should help advance understanding in this area of study.

Acknowledgments

Assistance with data collection was provided by Ryan Kim, Shania Hunsinger, Nicole Habib, Jessica Michel, Yenyen Tran and Samreet Atwal. We thank Gideon Caplovitz and three anonymous reviewers for their useful comments in the review process.

Declarations

Funding: Not applicable.

Conflicts of interest/Competing interests

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

Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Consent to participate.

Informed consent was obtained from all individual participants included in the study.

Consent for publication.

Not applicable

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The code used in this experiment are available on reasonable request.

Authors' contributions: All the authors contributed to data collection, data analysis and preparing the manuscript.

Open Practices Statement

None of the experiments was preregistered, and the data or materials for the experiments reported here are available on reasonable request.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*, 1–10.

Allan, L. G. (1979). The perception of time. Perception & Psychophysics, 26, 340–354.

André, N., Audiffren, M., & Baumeister, R. F. (2019). An integrative model of effortful control. *Frontiers in Systems Neuroscience, 13*, 79.

Apps, M., Grima, L., Manohar, S., & Husain, M. (2015). The role of cognitive effort in subjective reward devaluation and risky decision-making. *Scientific Reports, 5*, 16880. https://doi.org/10.1038/srep16880

Apps, M., & Ramnani, N. (2014). The anterior cingulate gyrus signals the net value of others' rewards. *Journal of Neuroscience, 34*, 6190–6200.

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology, 87*, 25–34.

Burgess, P. R., & Jones, Larry, F. (1997). Perceptions of effort and heaviness during fatigue and during the size-weight illusion. *Somatosensory & Motor Research, 14*, 189–202. https://doi.org/10.1080/08990229771051

Cos, I. (2017). Perceived effort for motor control and decision-making. *PLoS Biology*, 15, e2002885.

Craig, A. D. (2013). An interoceptive neuroanatomical perspective on feelings, energy, and effort. *Behavioral and Brain Sciences, 36*, 685.

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: Roles of perceived error-likelihood and time demands. *Psychological Research, 83*, 1033–1056. https://doi.org/10.1007/s00426-017-0943-x

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

Dunn, T. L., Lutes, D. J. C., & Risko, E. F. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 1372–1387. https://doi.org/10.1037/xhp0000236

Fechner, G. T., Howes, D. H., & Boring, E. G. (1966). *Elements of psychophysics (Vol. 1)*. Holt, Rinehart and Winston New York.

Feghhi, I., & Rosenbaum, D. A. (2019). Judging the subjective difficulty of different kinds of tasks. *Journal of Experimental Psychology: Human Perception and Performance, 45*, 983–994. https://doi.org/10.1037/xhp0000653

Feghhi, I., & Rosenbaum, D. A. (2020). Effort avoidance is not simply error avoidance. *Psychological Research*. https://doi.org/10.1007/s00426-020-01331-2 {still not available in print form as of June 1, 2021}

Fisher, J., & Steele, J. (2014). Questioning the resistance/aerobic training dichotomy: A commentary on physiological adaptations determined by effort rather than exercise modality. *Journal of Human Kinetics, 44*, 137–142.

Franchak, J. M. (2017). Exploratory behaviors and recalibration: What processes are shared between functionally similar affordances? *Attention, Perception, & Psychophysics, 79*, 1816–1829.

Franchak, J. M. (2020). Calibration of perception fails to transfer between functionally similar affordances. *Quarterly Journal of Experimental Psychology, 73*, 1311–1325.

Franchak, J. M., van der Zalm, D. J., & Adolph, K. E. (2010). Learning by doing: Action performance facilitates affordance perception. *Vision Research, 50*, 2758–2765.

Glover S & Dixon P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*, 791-806.

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

Halperin, I., & Emanuel, A. (2020). Rating of perceived effort: Methodological concerns and future directions. *Sports Medicine*, 1–9.

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

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences, 12*, 273–280.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*, 665–682. https://doi.org/10.1037/a0020198

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*, 661–679. https://doi.org/10.1017/S0140525X12003196

Labinger, E., Monson, J. R., & Franchak, J. M. (2018). Effectiveness of adults' spontaneous exploration while perceiving affordances for squeezing through doorways. *PloS One, 12*, e0209298.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. Journal of *Experimental Psychology: Human Perception and Performance, 12*, 517.

Miltner, W. H., Lemke, U., Weiss, T., Holroyd, C., Scheffers, M. K., & Coles, M. G. (2003). Implementation of error-processing in the human anterior cingulate cortex: A source analysis of the magnetic equivalent of the error-related negativity. *Biological Psychology, 64*, 157–166.

Montero, B. G. (2016). *Thought in action: Expertise and the conscious mind*. Oxford University Press.

Morel, P., Ulbrich, P., & Gail, A. (2017). What makes a reach movement effortful? Physical effort discounting supports common minimization principles in decision

making and motor control. *PLOS Biology, 15*, e2001323. https://doi.org/10.1371/journal.pbio.2001323

Mulert, C., Menzinger, E., Leicht, G., Pogarell, O., & Hegerl, U. (2005). Evidence for a close relationship between conscious effort and anterior cingulate cortex activity. *International Journal of Psychophysiology, 56*, 65–80.

Pageaux, B. (2014). The psychobiological model of endurance performance: An effort-based decision-making theory to explain self-paced endurance performance. *Sports Medicine, 44*, 1319.

Pitts, B., Riggs, S. L., & Sarter, N. (2016). Crossmodal Matching: A Critical but Neglected Step in Multimodal Research. *IEEE Transactions on Human-Machine Systems, 46*, 445–450. https://doi.org/10.1109/THMS.2015.2501420

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

Prins, N., & Kingdom, F. A. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes toolbox. *Frontiers in Psychology, 9*, 1250. https://doi.org/10.3389/fpsyg.2018.01250

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann, & W. Prinz (Eds.), *Relationships between perception and action (pp. 167-201)*. Berlin: Springer.

Prinz, W. & Hommel, B. (Eds.). (2002). *Common mechanisms in perception and action: Attention and performance*, Volume XIX. Oxford, UK: Oxford University Press.

Rosenbaum, D. A. (2012). The tiger on your tail: Choosing between temporally extended behaviors. *Psychological Science, 23*, 855–860. https://doi.org/10.1177/0956797612440459

Rosenbaum, D. A., & Bui, B. V. (2019). Does task sustainability provide a unified measure of subjective task difficulty? *Psychonomic Bulletin & Review*, 26, 1980–1987.

Rosenbaum, D. A., Chapman, K. M., Coelho, C. J., Gong, L., & Studenka, B. E. (2013). Choosing actions. *Frontiers in Psychology*, Volume 4, Article 273, doi:10.3389/fpsyg.2013.00273

Rosenbaum, D. A., & Gregory, R. (2002). Development of a method for measuring movement-related effort. *Experimental Brain Research, 142*, 365–373. https://doi.org/10.1007/s00221-001-0925-4

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79*, 217–240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience, 40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Song, J., Kim, S., & Bong, M. (2019). The more interest, the less effort cost perception and effort avoidance. *Frontiers in Psychology, 10*, 2146.

Steele, J. (2020). What is (perception of) effort? Objective and subjective effort during task performance [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kbyhm

van Iersel, M. B., Ribbers, H., Munneke, M., Borm, G. F., & Rikkert, M. G. O. (2007). The effect of cognitive dual tasks on balance during walking in physically fit elderly people. *Archives of Physical Medicine and Rehabilitation, 88*, 187–191. https://doi.org/10.1016/j.apmr.2006.10.031

Walton, M. E., Bannerman, D. M., Alterescu, K., & Rushworth, M. F. (2003). Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. *Journal of Neuroscience, 23*, 6475–6479.

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L., & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science, 345*, 75–77.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100–1122.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley Press.

CHAPTER 6 –TASK DIFFICULTY AND TASK SCHEDULING

Although little is known about how people schedule their tasks, task scheduling has received a lot of attention in computer science and robotics. Computers need to have a scheduler system to assign resources to different tasks optimally. Advancements in computer science would have been impossible without solving this problem. On the other hand, psychologists have overlooked this brain's vital ability without harming their effort to understand other crucial functions of the brain. Per the current *status quo*, it is worth having a general idea about task scheduling in computer science.

In computer science, schedulers may optimize four variables: 1) Maximizing the total amount of work completed, 2) Minimizing the wait time, 3) Minimizing the response time, and 4) Scaling allocated resources to the priority of each work (Liu & Layland, 1973). To optimize these variables, several scheduling disciplines have been proposed (e.g., Feitelson, 2015; Silberschatz, et al., 2012). Some of these disciplines are not relevant to how humans schedule their tasks. The main reason is that task-switching is not as costly for computers as for humans. So, schedulers with a high rate of switching could be designed for computers but not for humans. For example, with a round-robin schedule, the system cycles through the to-be-performed tasks (queue) with a fixed amount of time (called time slices) allocated to each task. It is an effective and easy-to-implement scheduler for a computer, but it is not how humans schedule their tasks, at least short-term tasks. For instance, if you have to answer five emails, it is so unlikely that you spend, say, 1 min on each email and cycle through them.

Nevertheless, some concepts related to scheduling in computer science might resemble, or be helpful in understanding, how we schedule tasks. For example, our brain may need to solve the scheduling problem on multiple levels, similar to a computer. Our brain should have a high-level (long-term) scheduler to decide which task is in line with long-term goals and prevent unrelated tasks from breaching the queue. Similarly, a long-term scheduler in a computer needs to authorize or deny different tasks from using computational resources. Another similarity between these high-level scheduling processes in computers and our brains is the frequency of using them. In both cases, despite their importance, they are needed infrequently.

The focus of this work is on the scheduler that is needed more frequently, the short-term scheduler. Assuming that the system has already decided which tasks are more critical, it should choose, at the short-term level, the order of performing the to-be-performed tasks. A straightforward algorithm is a first in, first out (FIFO) – also known as first come, first served (FCFS) - algorithm (Tanenbaum & Bos, 2015). It may seem not to be very effective in many situations, but it is proven useful in some cases. Is there any evidence that humans might also have FIFO schedulers?

Pre-crastination (Rosenbaum et al., 2014), doing something early at the expense of extra effort, could be the outcome of a FIFO scheduler. Rosenbaum and his colleagues asked participants to walk through an alley and pick one of the buckets located to the left and right side and carry the bucket to the alley's end. Their original plan was to vary the weight and to-be-carried-distance of the left and right bucket and investigate how participants equate these two different costs. To their surprise, most participants picked

the bucket with higher to-be-carried-distance even when both buckets had equal weights. What makes this study relevant to the FIFO scheduler is that from the perspective of a participant who was walking through the alley, the bucket with a higher to-be-carried-distance was the bucket that "came first" in the path, so it is more likely to be the bucket that was "served first". Their paradigm defers from the scheduler problem significantly, though; only one task was needed to be done.

Fournier et al. (2018) made a chance in the pre-crastination paradigm, which made it a scheduler problem paradigm. They asked participants to fetch two transparent buckets where each of them was located at different distances from the home position and had different numbers of ping-pong balls inside them. In a sense, participants had two options: 1) pass the close bucket, grab the far one, carry it to the close bucket position, grab the close bucket, and carry both of the buckets to the home position; 2) grab the close bucket, carry it to the far bucket position, grab the far bucket, and carry both of the buckets to the home position. In line with the previous research in pre-crastination, most participants picked the close bucket on the way to getting the far bucket (did something early) when they could have passed it and take care of it on the way back (even at the expense of extra effort). As you can see, this paradigm could be seen as a scheduler problem. If they used a FIFO scheduler, they should have chosen the less optimal solution and picked the close bucket first. That was what they did. So, it seems that this simple and sometimes not optimal solution for ordering different tasks might be, in fact, a solution that we use in some conditions. It has been shown that when the consequence of serving the first things first is detrimental, people are less likely to pre-crastinate

(Rosenbaum & Sauerberger, 2019). So, there might be some trade-off between the cost of using some more sophisticated schedulers and the cost of wasting resources by using the most straightforward scheduler.

More related to the current study, how do we decide what to do now when several tasks unfold simultaneously? For example, if you need to do the dishes and answer an email, what would you do first? Some computer algorithms can provide some insight here as well. We might use a method called the earliest deadline first (EDF) algorithm in computer science (Short, 2011). As its name implies, EDF prioritizes tasks based on their deadline. In our example, if the email is urgent, you would do it first, but if you need to clean the dishes to serve food to your kids that are hungry and need to go to school in 10 minutes, you will do dishes first. EDF would be useless if both tasks have the same deadline. In the current experiment, we asked participants to imagine the situation they need to do both tasks, and none of the tasks had any specific deadline. In these conditions, the shortest job first (SJF) algorithm would be handy.

SJF needs advanced knowledge (or estimation) about the required time to do each task in the queue. This method is the optimum method for tasks that simultaneously come to the queue and do not have a predefined priority (Arpaci-Dusseau & Arpaci-Dusseau, 2018). What makes this algorithm interesting for cognitive psychologists is the evidence suggesting that humans also have an SJF scheduler. The first piece of evidence is documented by a revered ancient Indian Sanskrit philologist, Dakṣiputra Pāṇini, whose dates are probably somewhere between the seventh to fourth B.C.E. He observed that when ordering two words seems to be arbitrary, people are more comfortable with

uttering the short word first. For example, William Hanna and Joseph Barbera could have named their masterpiece Jerry and Tom, but they preferred Tom and Jerry because it rolls off the tongue. In other words, it is easier to utter the shorter words first. This principle is known as Pāṇini's law.

More recently, Miller, Ulrich, and Rolke (2009) found that when the time between presenting two stimuli (SOA) in a psychological refractory period (PRP) experiment is small enough to process both stimuli simultaneously, participants opt to do the task with shorter reaction time (RT) first (see also Leonhard et al., 2011, Fernandez et al., 2011). They showed that by doing the task with shorter RT first, one could decrease the total RT and hence decrease the total performance time. Although they considered RT rather than performance time, this result is well-connected to Pāṇini's law because other things being equal, the shorter a task, the faster the RT. Another way to connect their results to Pāṇini's law is to propose a mechanism that explains the law. By doing short tasks first, one can minimize total RT and hence total performance time. This aligns with what Beaty et al. (2020) proposed in explaining Pāṇini's law.

Beaty et al. (2020) suggested that the overarching approach in planning speech and also motor behavior is to use incremental planning - executing some parts of a task while planning other parts (Lashley 1951, Rosenbaum et al., 2007). According to them, incremental planning would be most beneficial if people do easy tasks first. So, it would not be surprising to observe the easy first principle in a variety of conditions. They showed that in the same way that more accessible sequences are likely to happen early in

a phrase (Levelt, 2008), music improvisers use more straightforward melodies early in their performance.

Here we want to see whether the easy-first principle can hold for other tasks or not. Our particular interest is to test this principle in both physical tasks and mental tasks. We do so by providing participants with two options and ask them about their perceived difficulty/ease (which is easier?) and also ask them about the order in which they prefer doing both of them (which is first?). If people prefer doing easy tasks first, then the choices in "which is easier?" and "which is first?" conditions should be similar.

Experiment 1

In this experiment, we wanted to investigate the easy-first principles for physical tasks. We used 2-alternative forced-choice and asked participants to choose the easier option of the two and the order in which they prefer doing both tasks. The two options were two levels of a physical task. The task was to move a bucket back and forth to the two sides of a mat. The two levels of the bucket moving task were created by manipulating the weight of the bucket. In one case, the bucket was empty/light, and in the other case, the bucket was weighted/heavy. Comparing the pattern of choices in "which is easier?" and "which is first?" conditions was thought to provide insight into the easy-first principle.

Method

Participants

Thirty undergraduate students (17 female and 13 male) from the University of California, Riverside, participated in this experiment for course credit. The participants ranged in age from 19 years to 23 years, with an average of 19.22 years and a standard deviation of 1.12 years. All participants signed an informed consent form before the experiment. That number let us exceed the value of n = 500 observations recommended for evaluation of logistic regression models (Cohen, Cohen, West, & Aiken, 2013; Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997). With 30 subjects, the number of choices per participant was 50, so there were 1500 observations in the "which is easier?" condition and 1500 observations in the "which is first?" condition.

Materials and Procedure

After signing the consent form, participants were asked to practice doing both the light-bucket task and heavy-bucket (1.5 kg) task 6 times. A random half of participants did the light-bucket first, and the other half did the heavy-bucket first. The purpose of the exposure practice trials was to ensure that all participants have a clear idea about the difficulty of these tasks. In both conditions, at the start of each trial, a beach bucket was located in the middle of an 84 cm by 53 cm mat. The mat was situated in the center of a 132 cm by 92 cm table. The height of the table was 81 cm. Participants were instructed to move the bucket back and forth to the two sides of the mat and touch the table with the

bucket. They were asked to do the task at a leisurely pace. A random half of participants performed the light bucket first. The other half performed the heavy bucket first.

After the exposure phase, participants were asked to answer two sets of 25 questions about the difficulty of doing the light-bucket N times (5, 10, 20, 40, and 80 times) vs. the difficulty of doing the heavy-bucket N times (5, 10, 20, 40, and 80 times). They were also asked to answer two sets of 25 questions about the order in which they prefer doing both of the tasks. So, the total number of question that each participant answered was 100. A random half of participants answered the "which is easier?" questions first and the other half answered the "which is first?" questions first. The order of questions and answers in each block of questions was randomized. Questions were presented one at a time using a MATLAB code.

Results

To analyze choices, we fitted a logistic curve to the probability of choosing the light-bucket as a function relative N, N for the light-bucket relative to the N for the heavy-bucket. Relative N was defined as:

$$\frac{N_{light}}{N_{light} + N_{heavy}}$$

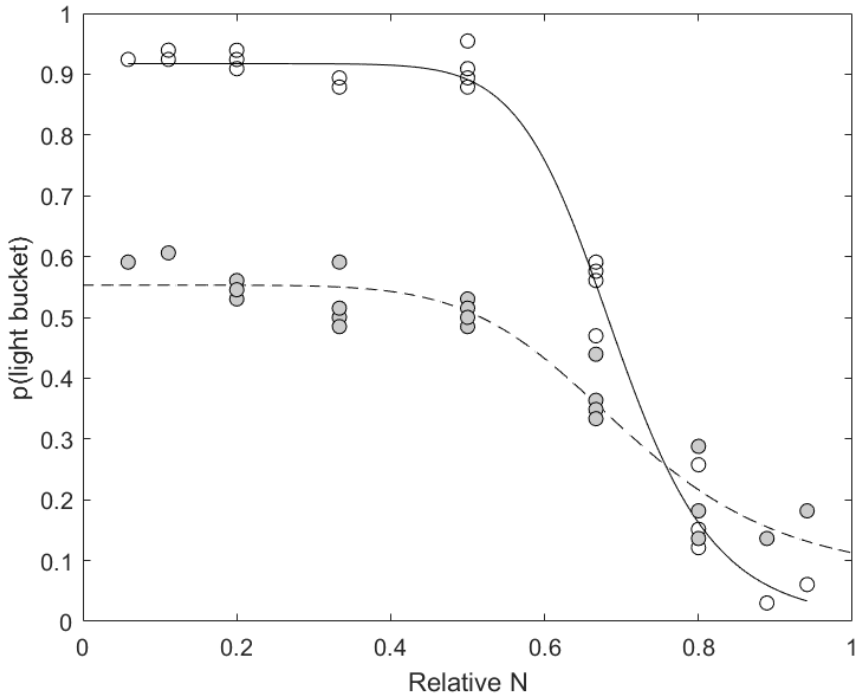
To fit the curve, we used L4P (Cardillo G. (2012) Four parameters logistic regression - There and back again, https://it.mathworks.com/matlabcentral/fileexchange/38122), which fits data points with a four points logistic regression in MATLAB. L4P exploit MATLAB's Curve Fitting functionality to fit

$$F(x) = \frac{Min + (Min - Max)}{1 + (\frac{x}{C})^k}$$

to the data. Where Min is the minimum range of the curve, Max is the maximum range of the curve, C is the mid-point (inflection point), x is the data, and *k* is the steepness of the curve. L4P provides the best estimates of all the four free parameters and provides a 95% confidence interval for each of them. We used the 95% CI to compare each of the four free parameters in "which is easier?" and "which is first?" conditions. The method proposed by Cumming (2009) was used to compare two means with known confidence intervals around the means.

There was a robust correlation between *p*(light bucket) in "what is easier?" and "what is first?" conditions, $r(24) = .97$, $p < .001$. Regardless of the strong correlation, results revealed a noticeable difference between the choice pattern in the "what is easier?" and "what is first?" conditions (Figure 6.1). While in both cases, the logistic curves have a positive slope (10.78, 95% CI = [7.4, 14.1], and 6.71 95% CI = [2.39, 11.04] for "what is easier?" and "what is first?" conditions, respectively) choices for the "what is first?" condition were restricted in a narrower range. As you can see in Table 6.1, probability of choosing the light bucket, *p*(light bucket), ranged between Min = .00, CI 95% CI = [-.08, .09] and Max = .91, 95% CI = [.89, .94] for the "which is easier?" condition but it ranges between Min = .07, 95% CI = [-.11, .25] and Max = .55, 95% CI = [.52, .58] for the "which is first?" condition. Based on the guidelines provided by Cumming (2009), the only free parameter that is reliably different between the two conditions is the Max (See Figure 6.1).

Figure 6. 1 *p*(light bucket) as a function of relative N when the alternative option was the heavy bucket task.



Note: Empty circles show the average *p*(light bucket) across all participants in each of the 25 conditions of the "which is easier?" condition. Gray circles show the average *p*(light bucket) across all participants in each of the 25 conditions of the "which is first?" condition. The graph also shows the fitted curves. The solid line shows the fitted curve for the "which is easier?" condition, and the dashed line shows the fitted curve for the "which is first?" condition.

Table 6. 1. Best estimates of logistic curve's free parameters in all 3 experiments presented in this chapter.

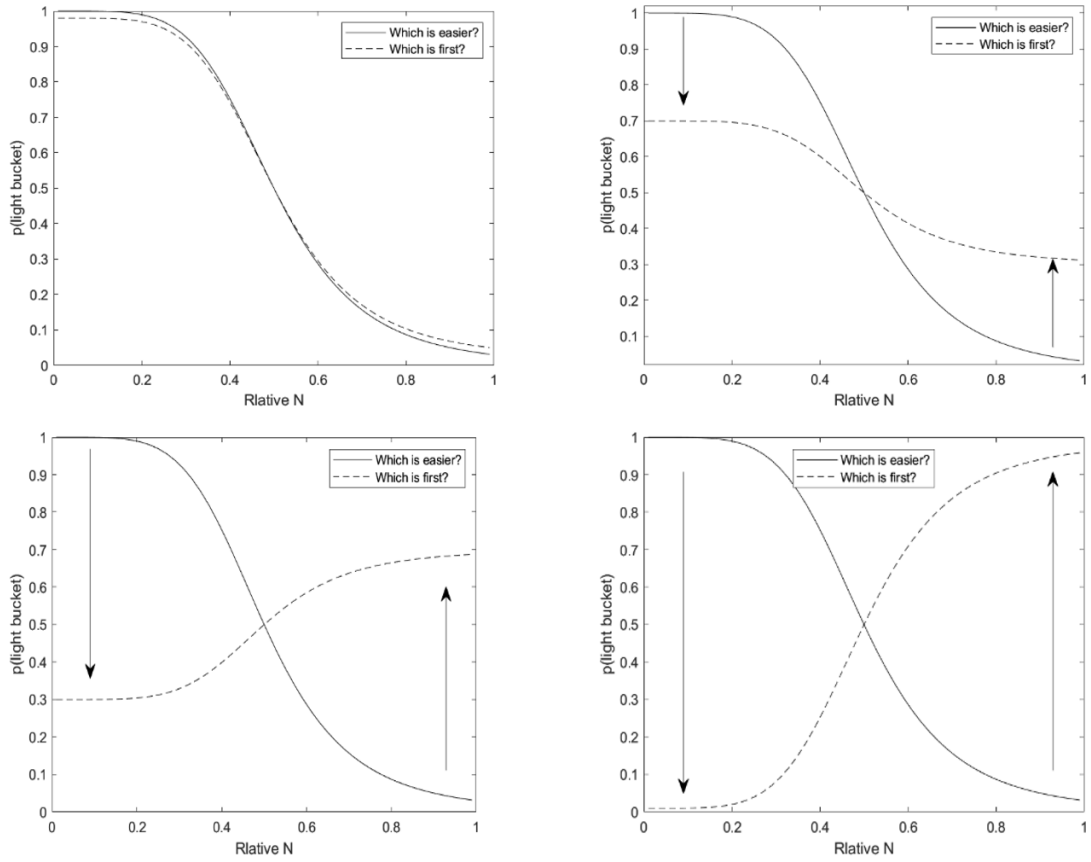
		Min	Max	C	k
Empty bucket vs. heavy bucket	Which is easier?	.00 (-.08, .09)	.91 (.89, .94)	.69 (.67, 0.71)	10.78 (7.4, 14.1)
	Which is first?	0.07 (-.11, .25)	.55 (.52, .58)	0.70 (.60, .81)	6.71 (2.39, 11.04)
2-term vs. 6-term	Which is easier?	0.13 (.03, .24)	0.98 (.96, 1.01)	0.72 (.60, .74)	13.1 (8.65, 17.55)
	Which is first?	0.19 (.06, .31)	0.95 (.92, .97)	0.72 (.68, .75)	11.20 (6.58, 15.80,)
Empty bucket vs. 2-term	Which is easier?	.00 (-.15, .08)	0.89 (.85, .92)	0.53 (.51, .54)	6.40 (5.25, 7.56)
	Which is first?	.00 (-0.10, 0.10)	0.83 (.79, .86)	0.52 (.47, .56)	4.12 (2.89, 5.359)
Empty bucket vs. 6-term	Which is easier?	0.04 (.05, .14)	0.93 (.90, .96)	0.64 (.61, .67)	6.81 (5.11, 8.50)
	Which is first?	.00 (-.16, .16)	0.94 (.91, .97)	0.66 (.60, .71)	5.12 (3.70, 6.54,)
Heavy bucket vs. 2-term	Which is easier?	.00 (-.06, .09)	0.89 (.83, .96)	0.37 (.34, .39)	3.59 (2.87, 4.31,)
	Which is first?	0.01 (-.06, .09)	0.84 (.77, .92)	0.36 (.32, .39)	4.16 (2.51, 5.81)
Heavy bucket vs. 6-term	Which is easier?	.00 (-.11, .11)	0.91 (.85, .97)	0.49 (.45, .53)	5.20 (2.99, 7.40,)
	Which is first?	0.02 (-.05, .09)	0.91 (.87, .96)	0.47 (.44, .50)	4.76 (3.40, 6.12,)

Note: The values in parentheses show 95% confidence interval around the best estimate.

Discussion

Experiment 1 was conducted to investigate how people schedule physical tasks. We asked participants to get exposed to two levels of a physical task: moving an empty bucket or a weighted bucket back and forth to the two sides of a mat. After performing the exposure trials and having a clear sense of the difficulty of each of them, we asked them to answer two series of questions. In one series of questions, we asked them to compare the ease/difficulty of doing each of the tasks N times, where N could be 5, 10, 20, 40, or 80. In another series of questions, we asked them to assume that they need to do both tasks and then decide which of the two tasks they want to do first. If all of them chose to do easy task first, the two conditions should have the same pattern (Figure 6.2 top left panel). If a majority of them preferred doing easy task first (some prefer doing hard task first), still a positive correlation is expected, but the range of *p*(light bucket) should be restricted (Figure 6.2 top right panel). If a majority of them preferred doing hard task first, a negative correlation is expected (Figure 6.2 bottom left panel). Lastly, if all of them preferred doing hard task first, the pattern of choice should flip completely (Figure bottom right panel).

Figure 6. 2. Changing the choice patterns based on the proportion of easy-first choices. By decreasing the proportion of easy-first choices, the deviation between "which is easier?" and "which is first?" choice patterns should increase.



Note: The top left panel shows the all-easy-first model. The top right panel shows the majority-easy-first model. The bottom left panel shows the majority-hard-first model. And, the bottom right panel shows the all-hard-first model.

Given the positive correlation between the two conditions and restricted range of *p*(light bucket) in the "which is first?" condition compared to the "which is easier?" condition, the result of this experiment is clearly in line with the model that suggested a mixture of easy-first and hard-first preferences (the majority-easy-first model). Potts and Rosenbaum (2021) also found that in scheduling two physical tasks, a majority (~65%) of participants preferred doing easy tasks first and the rest (~35%) preferred doing the hard

tasks first. Their results also point to a higher chance of doing the easy task first when both tasks are cognitive. Experiment 2 is designed to replicate and extend this result.

Experiment 2

Experiment 1 shows the inclination toward doing easy tasks first. Based on Potts and Rosenbaum's (2021) results, when both to-be-scheduled tasks are cognitive, one should expect an even stronger tendency to do easy tasks first. To test this prediction, here, we used a similar paradigm as the previous experiment but used a task that was more cognitively demanding than physically demanding.

Method

Participants

Thirty undergraduate students (19 female and 11 male) from the University of California, Riverside, participated in this experiment for course credit. The participants ranged in age from 18 years to 22 years, with an average of 19.22 years and a standard deviation of 1.12 years. All participants signed an informed consent form before the experiment. The current sample size was similar to the sample size of the previous experiment.

Materials and Procedure

After signing the consent form, participants were asked to solve 12 math problems as the exposure trials. All math problems required adding and/or subtracting

single digits. For half of the problems, the number of addends was two (e.g., 4 – 9 = ?; 5 + 1 = ?). For the other half, the number of addends was six (e.g., 2 + 3 – 9 + 7 – 5 + 4 = ?; 8 – 1 + 5 – 4 + 6 – 8 = ?). A random half of 2-digit problems were addition problems, and the other problems were subtraction problems. In the 6-digit condition, we altered addition and subtraction operations for successive operations. A random half of 6-digit problems started with the addition, and the other half started with subtraction. A random half of participants solved 2-digit problems first, and the other half began with 6-digit problems.

A MATLAB program was written to present one question at a time, collect participants' responses, give them "correct/incorrect" feedback, and then present questions pertaining to task scheduling and task difficulty (as in Experiment 1). Just like Experiment 1, "which is easier?" (task difficulty) questions and "which is first?" (task scheduling) questions were generated by varying the N associated with 2-term and 6-term problems. Given that N could be 5, 10, 20, 40, and 80, an example of a task difficulty question is: Which is easier? A) 40 of 2-term problems, or B) 5 of 6-term problems. The task scheduling counterpart of this question is: Which would you rather do first? A) 40 of 2-term problems, or B) 5 of 6-term problems. The MATLAB code recorded the time and accuracy of responses in the exposure trials.

Results

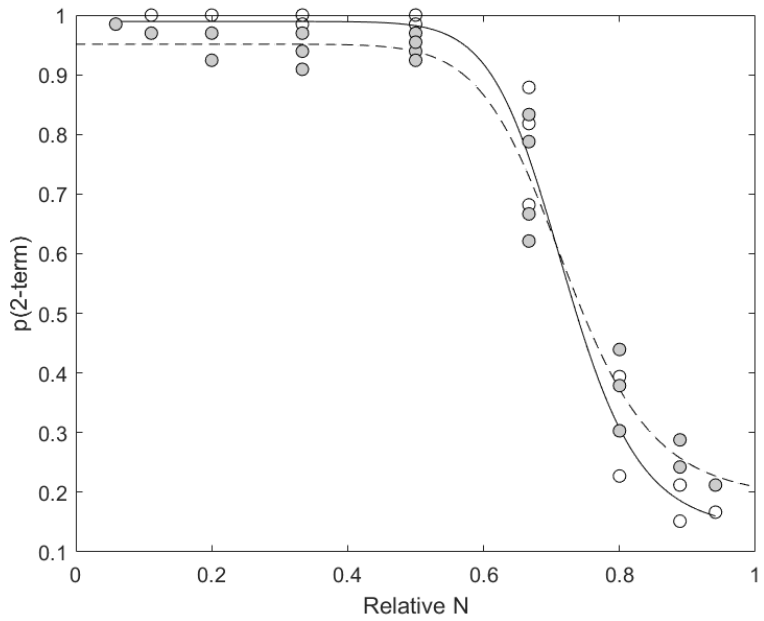
Performance times and error rates of the exposure trials were analyzed using t-test. Results showed that 6-term math problems took more time to solve (M = 4.22), than

2-term math problems (M = 1.10), t = 23.12, p < .001. Also, 6-term math problems were more prone to error (M = .34) than 2-term math problems (M = .09), t = 46, p < .001.

Choice data were analyzed the same way that it was analyzed in Experiment 1. The same models were also developed to characterize choices.

There was a very strong correlation between p(light bucket) in "what is easier?" and "what is first?" conditions, $r(24)$ = .99, p < .001. Figure 6.3 shows that in addition to having a high correlation, the logistic fits to the two choice patterns are pretty similar. In fact, using the method introduced Cumming and Geoff (2005) and developed by Cumming (2009), Min, k, and C of the two fits are not reliably different, and the Max for "which is easier?" condition, Max = 0.98, 95% CI = [0.96, 1.01] is just marginally higher than the Max for the "which is first?" condition, Max = 0.95, 95%CI = [0.92, 0.97]. According to them, 56% overlap between two marginal errors is approximately indicating p = .05. The overlap between marginal errors for Max is 50% which shows a marginal difference.

Figure 6. 3. *p*(2-term) as a function of relative N when the alternative option was the 6-term task.



Note: Empty circles show the average *p*(2-term) across all participants in each of the 25 conditions of the "which is easier?" condition. Gray circles show the average *p*(2-term) across all participants in each of the 25 conditions of the "which is first?" condition. The graph also shows the fitted curves. The solid line shows the fitted curve for the "which is easier?" condition, and the dashed line shows the fitted curve for the "which is first?" condition.

Discussion

This experiment aimed to investigate the effect of perceived task difficulty on task scheduling of two cognitive tasks. After getting exposed to 2 levels of solving math problems, participants were asked to indicate which of the two tasks is easier to do N times (5, 10, 20, 40, and 80) and similarly, if they need to do both of the tasks N times which they would rather do first. As expected, there was a remarkable similarity between "which is easier?" and "which is first?" choice patterns. It shows that participants preferred doing easy cognitive tasks first, even more so than the preference to do easy

physical tasks first. The next obvious question is how do people schedule doing a physical and a mental task?

Experiment 3

In line with Potts and Rosenbaum's (2020) results, we found a higher inclination toward doing easy tasks first when both of the to-be-ordered tasks were cognitive (Experiment 2) compared to the condition that both were physical (Experiment 1). Here we want to examine how task difficulty affects scheduling tasks when one of the to-be-ordered tasks is physical and the other is mental.

Method

Participants

Sixty undergraduate students (36 female and 24 male) from the University of California, Riverside, participated in this experiment for course credit. The participants ranged in age from 18 years to 24 years, with an average of 19.7 years and a standard deviation of 1.72 years. All participants signed an informed consent form before the experiment. Participants answered all possible combinations of the two levels of the physical and mental tasks once. So, to have the same number of observations per condition as the previous experiments in which each participant answered each question twice, we doubled the number of participants here.

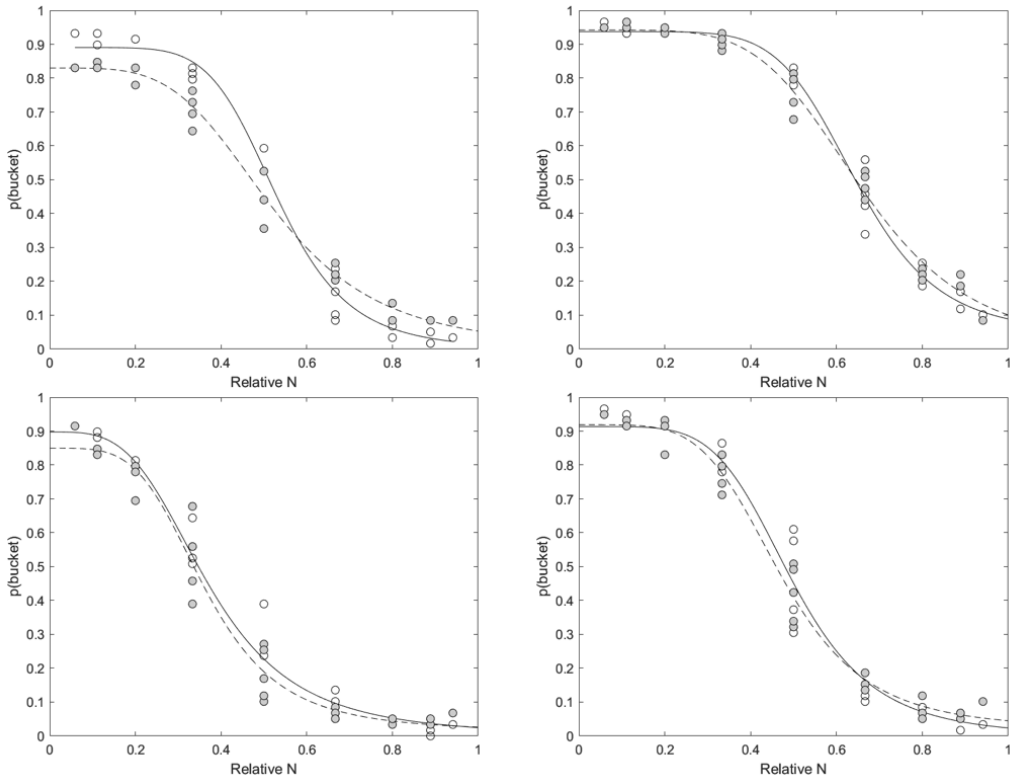
Materials and Procedure

In this experiment, the two levels of the physical task (light-bucket and heavy-bucket) we used in Experiment 1 were crossed with the two levels of the cognitive task (2-term and 6-term) we used in Experiment 2. Each pair of physical and mental tasks were presented in a separate block. There was a total number of 8 blocks (four blocks pertaining to the "which is easier?" questions and four pertaining to the "which is first?" questions). A random half of the participants did all the four blocks of the "which is easier?" questions before doing the four blocks of the "which is first?" questions. At the start of each block, there was an exposure phase. Like Experiments 1 and 2, the exposure phase consisted of doing each of the tasks six times. The order of presenting the tasks in the exposure phase was random. After getting exposed to the two tasks related to a given block, participants answered 25 questions about the ease ("which is easier?") or the preferred order ("which is first?") of the two tasks, which were similar to the questions that were asked in Experiments 1 and 2.

Results

Figure 6.4 shows the similarity between the choice patterns for the "which is easier?" questions and the "which is first?" questions in all four combinations of the two levels of the physical task and the two levels of the mental task.

Figure 6. 4. *p*(bucket) in four conditions. (1) *p*(empty bucket) as function of relative N when the alternative option was the 2-term task (top left panel) and (2) when the alternative option was the 6-term task (top right panel). Also (3) *p*(heavy bucket) when the alternative option was the 2-term task and when the alternative option was (4) when the alternative option was the 6-term task.



Note: Empty circles show the average *p*(bucket) across all participants in each of the 25 conditions of the "which is easier?" condition. Gray circles show the average *p*(bucket) across all participants in each of the 25 conditions of the "which is first?" condition. The graph also shows the fitted curves. The solid line shows the fitted curve for the "which is easier?" condition, and the dashed line shows the fitted curve for the "which is first?" condition.

Statistical analyses confirm the similarity between choice patterns in the "which is easier?" and the "which is first?" conditions. Using the guidelines provided by Cumming (2009), there is no difference between any four free parameters of logistic curves of any of the four panels of Figure 6.4.

Discussion

The aim of this experiment was to see how perceived difficulty affects the preferred order of performing a physical task and a mental task. We used the same tasks as Experiment 1 and 2, but instead of using tasks of the same modality (physical in Experiment 1 and mental in Experiment 2), we crossed the modalities. Given that we had two levels of physical demands (lifting light bucket and heavy bucket) in Experiment 1 and two levels of mental demands (solving 2-term and 6-term math problems) in Experiment 2, in this experiment, we had four possible pairs of physical and mental tasks. Like the previous experiments in this study, the effect of perceived difficulty on task scheduling was investigated by looking at choice patterns in the following conditions: A) comparing the ease of doing the physical and the mental task, B) selecting the preferred order in doing both tasks. As expected, participants preferred doing easy tasks first.

General Discussion

This study was designed to test the easy-first principle in scheduling tasks. We did so by comparing choice patterns in the "which is easier?" and the "which is first?" conditions. In both of these conditions, participants had to choose between doing a physical (or a mental) task N times (5, 10, 20, 40, or 80) and another physical (or mental) task N times (5, 10, 20, 40, or 80). For the "which is easier?" condition, we asked them to compare the ease/difficulty of the options and pick the one that is easier. For the "which is easier?" condition, we asked them to assume that they need to do both of the tasks and then pick the option that they would rather do first. In Experiment 1, both options were

physical. The task associated with one option was moving an empty bucket back and forth to the two sides of a mat. The task associated with the other option was moving a weighted buck and forth to the two sides of a mat. In Experiment 2, we used the same procedure but used two levels of a cognitive task: solving 2-term math problems and 6-term math problems. In Experiment 3, we crossed the two levels of the bucket task with the two levels of the math task and asked the participants to compare the ease of doing a bucket task N times to the ease of doing a math task N times ("which is easier?" condition). As in the previous experiments, we also asked them to select their preferred order for doing both of the tasks ("which is first?" condition).

The easy-first principle of ordering tasks has been documented in ordering words when the reverse ordering seems to be as legitimate as easy-first ordering (Panini's Law). For example, naming a company Jerry and Ben's will not violate any grammatical rules, but putting the short (easy) section first, Ben and Jerry's, makes it sound more natural. The easy-first principle has been shown in more prolonged and sophisticated speech production behaviors as well (Levelt, 2008). Such that phrases usually start with easier sequences. In the same way, jazz musicians use the easy-first principle in their improvisations. They begin with straightforward arrangements early in their performance (Beaty et al., 2020).

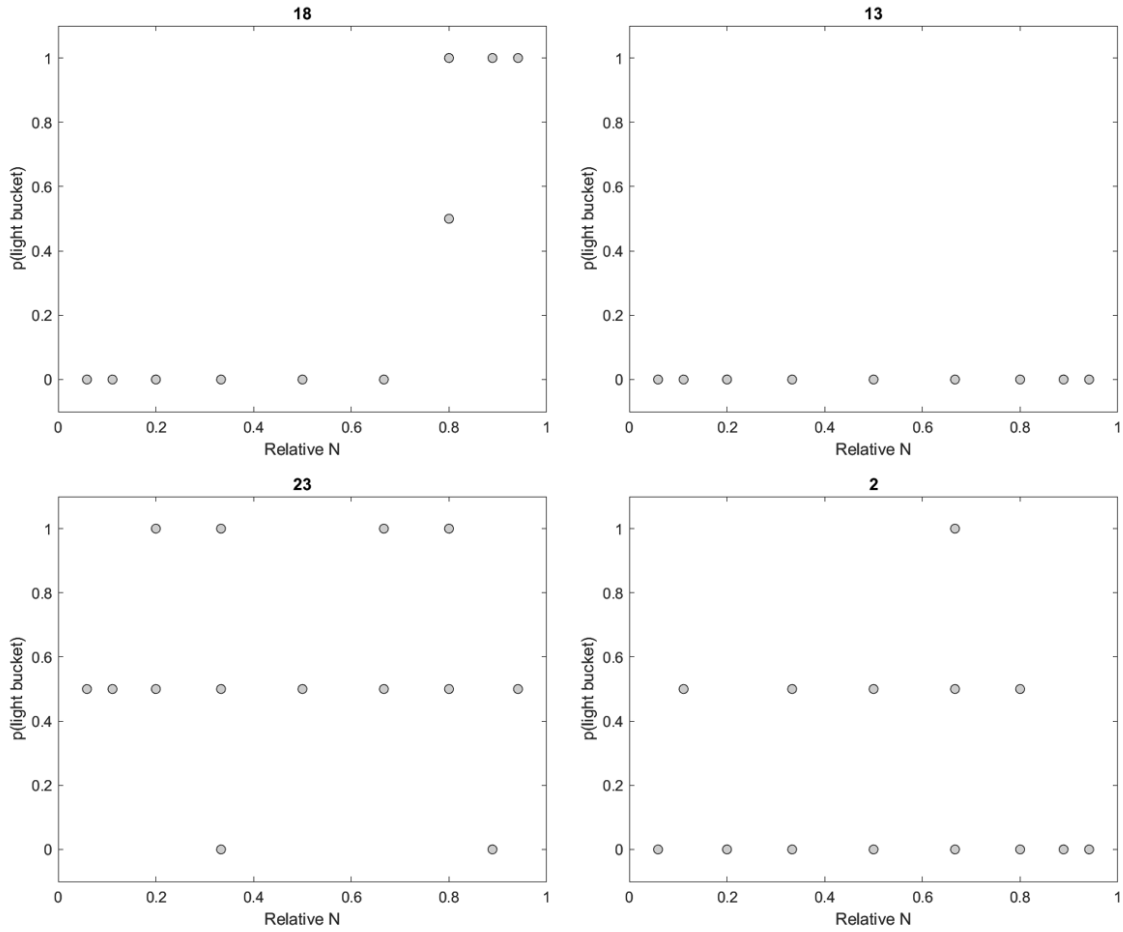
The easy-first principle resembles the short job first (SJF) algorithm for a scheduler in a computer. In computer science, it is well-known that if the tasks in the queue have the same priority and the same deadline, SJF provides the optimum solution for ordering the tasks (Arpaci-Dusseau & Arpaci-Dusseau, 2018). Similarly, in cognitive

science, Ruiz Fernández et al., (2012) showed that doing short tasks first can reduce the overall performance time. Potts and Rosenbaum's (2020) results are also in line with the easy-first principle. Nevertheless, inspecting their results more carefully suggests that the easy-first principle should be taken with a grain of salt.

Potts and Rosenbaum (2021) provided a group of participants with two tasks and asked: Assume you need to do both tasks; which would you rather do first? While some of the tasks are clearly easier than their alternative in a given question, for example, walking at a leisurely pace for .5 minutes vs. running as quickly as possible for 4.5 minutes, a considerable proportion of their participants opt to do the hard task first. In the example given, 65% of the participants preferred doing the easy task first. It is taken to be evidence for the easy-first model, but it also means that 35% of participants preferred doing the hard task first. Why should around one-third of participants prefer the hard-first approach? More importantly, the same pattern could emerge with a drastically different proportion of easy-first and hard-first approaches. The easy-first choices could be as low as only 30% choices. It is possible if one also considers a random scheduler or a scheduler that is independent of ease/difficulty. If 30% of cases were driven by easy-first scheduling and the rest (70%) were driven by random scheduling, the observed percentages will be 65% easy and 35% hard first. It is the case because 50% of the 70% random choices (50% × 70% = 35% of the total) were the easy option. So, 30% might have chosen the easy option based on an easy-first scheduler, and 35% have chosen the same based on a random scheduler.

In the current experiment, instead of looking at one number, we looked at the pattern of choices in the "which is easier?" and the "which is first?" conditions. The similarity between the two patterns in Experiment 2 and Experiment 3 is taken to provide strong support for the easy-first principle. In Experiment 1, though, there is a striking difference between the two patterns. The range of *p*(light bucket) was more restricted in which "which is first?" condition compare to the "which is easier?" condition. This pattern could happen because a portion of participants had chosen the hard task to do first. It could also be the case that a portion of participants did not use the information about ease of the physical task to inform their decisions about task scheduling. Knowing that the other experiments reported in this work suggest that the difficulty of the two tasks is something that people consider for scheduling their task, and, to the best of our knowledge, no one has reported otherwise, this possibility seems to be unlikely. Still, looking at each participant's choice data separately, one cannot exclude this possibility. While some participants clearly preferred doing hard task first (top left panel in Figure 6.5), there were also others who preferred doing the empty bucket regardless of the number of times it should be performed (independent of the difficulty level) (top right panel in Figure 6.5), and participants whose choices did not clearly fall into a specific category (panel in Figure 6.5).

Figure 6. 5. Four individual participant's data of *p*(light bucket) in the "which is first?" condition of Experiment 1.



Note: Numbers on top of each plot show the participant's number.

This result leads us to two remaining remarks of this paper. First, to the best of our knowledge, having a hard-first, or a lighter-first, or a random-choice scheduler has not been documented yet. The current study suggests that, unlike scheduling cognitive tasks, which seems to be mainly through an easy-first scheduler, for scheduling two physical tasks, different people might use different strategies. To clarify the point, it might help explain how we came up with the idea of testing the easy-first principle for a

physical task. The idea emerged from an observation that the first author had. Looking at how his neighbors carried their belongings to a moving truck, he observed that they carried big, heavy stuff first. It stroked the author as a contradictory piece of evidence for the easy-first principle. Then we tested this anecdotal observation in the lab. The results clearly show that the easy-first principle is not as universal for physical tasks as it is for cognitive tasks.

Second, in several places in the manuscript, we used easy-first and short-first interchangeably. It entails equating time and difficulty. One can safely do that if the system under investigation is a computer. Equating time and difficulty is debatable if we talk about humans, though. On the one hand, there are accounts and evidence suggesting that time, or a variation of time like subjective time, could be the determinant of difficulty (Gray et al., 2006; Potts et al., 2018; Rosenbaum & Bui, 2019). On the other hand, there are accounts and evidence suggesting that time could not be the sole determinant of difficulty (e.g., Kool et al., 2010). So, should we call it easy-first or short-first then? Given that ease/difficulty could be seen as an abstract code pertinent to any task and encompasses time (Feghhi and Rosenbaum, 2021), we think easy-first is a more valid name.

References

Arpaci-Dusseau, R. H., & Arpaci-Dusseau, A. C. (2018). *Operating systems: Three easy pieces*. Arpaci-Dusseau Books LLC.

Beaty, R., Frieler, K., Norgaard, M., Merseal, H. M., MacDonald, M., & Weiss, D. (2020). Spontaneous Melodic Productions of Expert Musicians Contain Sequencing Biases Seen in Language Production. *PsyArXiv*. https://doi.org/10.31234/osf.io/qdh32

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Cumming, G. (2009). Inference by eye: reading the overlap of independent confidence intervals. *Statistics in Medicine*, *28*, 205-220.

Feghhi, I., & Rosenbaum, D. A. (2021). Effort avoidance is not simply error avoidance. *Psychological Research*, 85, 1462-1472.

Feghhi, I., & Rosenbaum, D. A. (2021, accepted). Towards A Common Code For Difficulty: Navigating A Narrow Gap Is Like Memorizing An Extra Digit. *Attention, Perception, and Psychophysics*.

Feitelson, D. G. (2015). *Workload modeling for computer systems performance evaluation*. Cambridge University Press.

Fournier, L. R., Coder, E., Kogan, C., Raghunath, N., Taddese, E., & Rosenbaum, D. A. (2019). Which task will we choose first? Precrastination and cognitive load in task ordering. *Attention, Perception, & Psychophysics*, 81, 489-503.

Haber, R. N. (1966). Nature of the effect of set on perception. *Psychological Review, 73*, 335-351. doi:http://dx.doi.org/10.1037/h0023442

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16, 965-980.

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

Lashley, K. l., (1951). The problem of serial order in behavior. In *Cerebral mechanisms in behavior*. Bobbs-Merrill Oxford, United Kingdom.

Leonhard, T., Fernández, S. R., Ulrich, R., & Miller, J. (2011). Dual-task processing when task 1 is hard and task 2 is easy: Reversed central processing order? Journal of *Experimental Psychology: Human Perception and Performance*, 37, 115–136. https://doi.org/10.1037/a0019238

Levelt, W. J. (2008). *An introduction to the theory of formal languages and automata*. John Benjamins Publishing.

Liu, C. L., & Layland, J. W. (1973). Scheduling algorithms for multiprogramming in a hard-realtime environment. *Journal of the ACM (JACM)*, *doi:10.1145/321738.321743, 20*, 46-61.

Miller, J., Ulrich, R., & Rolke, B. (2009). On the optimality of serial and parallel processing in the psychological refractory period paradigm: Effects of the distribution of stimulus onset asynchronies. *Cognitive psychology*, 58, 273-310.

Naharudin, M. N., & Yusof, A. (2013). Fatigue index and fatigue rate during an anaerobic performance under hypohydrations. *PLoS One*, *8*, e77290.

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & Van Der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, 26, 525-554.

Rosenbaum, D. A., Gong, L., & Potts, C. A. (2014). Pre-crastination: Hastening subgoal completion at the expense of extra physical effort. *Psychological Science*, 25, 1487-1496.

Rosenbaum, D. A., & Bui, B. V. (2019). Does task sustainability provide a unified measure of subjective task difficulty? *Psychonomic Bulletin & Review*, 26, 1980-1987.

Rosenbaum, D. A. & Sauerberger, K. S. (2019). End-state comfort meets pre-crastination. Psychological Research, 1-11. DOI 10.1007/s00426-018-01142-6

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

Potts, C. A., & Rosenbaum, D. A. (2021). Does attention solve the "apples-and-oranges" problems of judging task difficulty and task order? *Psychological Research,* 1-8.

Fernández, S. R., Leonhard, T., Rolke, B., & Ulrich, R. (2011). Processing two tasks with varying task order: Central stage duration influences central processing order. Acta psychologica, 137(1), 10-17.

Short, M. (2011). Improved schedulability analysis of implicit deadline tasks under limited preemption EDF scheduling. In *ETFA2011*. IEEE.

Silberschatz, A., Galvin, P. B., & Gagne, G. (2012). *Operating system concepts, 9th edition*, Wiley Publishing.

Willem J.M. Levelt (2008) The architecture of normal spoken language use. In *Linguistic Disorders and Pathologies: An International Handbook*, pages 1–15. De Gruyter Mouton, 2008.

Willem J.M. Levelt, Ardi Roelofs, and Antje S. Meyer. (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–75.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & cognition*, *26*, 716-730.

Tanenbaum, A. S., & Bos, H. (2015). *Modern operating systems*. Pearson.