Psychonomic Bulletin & Review (2019) 26:1980–1987
https://doi.org/10.3758/s13423-019-01631-8

BRIEF REPORT

Does task sustainability provide a unified measure of subjective task difficulty?

David A. Rosenbaum[1] · Bill V. Bui[1]

Published online: 25 June 2019

Abstract

What accounts for the subjective difficulty of a task? It is easy to suggest ad hoc measures, such as how many individuals can do the task, how long it takes them to do it, how likely they are to complete it, how much attention it requires, and so on. But having such ad hoc measures may miss the point that it is possible to judge the relative difficulty of different kinds of tasks, suggesting that there may be a common basis for judging task difficulty. If there is such a common basis, it might be used to compare the difficulties of different kinds of task. We tested two hypotheses about what the common basis might be. One was that time serves this role. This hypothesis was attractive because time is amodal and previous studies have provided support for the hypothesis that time might be an index of task difficulty. The other hypothesis was new. According to the new hypothesis, the subjective difficulty of a task corresponds to its estimated sustainability. We obtained results consistent with the time hypothesis, but our data were less supportive of the sustainability hypothesis.

Keywords task switching or executive control · timing · subjective difficulty · action

Suppose you are in a library and need information. You know there is a very clearly written book on the topic of interest several flights up, but the elevator is not working, and you would have to walk up many stairs to get the book. You also know there is another, less clearly written book on a shelf on the other side of the room you are in. With a few steps, you could get that closer book, which is less clearly written. Which book would you get? To reach the decision, you would need to compare the difficulty of walking with the difficulty of reading, but those two costs are, or seem to be, different kinds. One cost is, roughly speaking, "physical." The other is, roughly speaking, "mental." How do we make such decisions?

This problem was broached by Potts, Pastel, and Rosenbaum (2018), who asked university students to choose between a physical task (picking up and carrying a bucket) and a more cognitive task (counting). Potts et al. hypothesized that participants might make the choice based on task-completion times. These authors reasoned that, all else being equal, a task that takes less time would be judged easier than a task that takes more time. The results were consistent with this hypothesis and fit with the notion, promoted earlier by Gray, Sims, Fu, and Schoelles (2006), that time is a critical feature of task difficulty.

Even so, time may not be useful in all circumstances. Consider this anecdote. One of the authors of this article is an avid amateur chamber music player who loves to play string quartets for hours on end—including playing pieces that are very technically challenging for any player (e.g., late Beethoven quartets). But while he finds it easy to play challenging music for hours and hours, he finds it extremely difficult to engage in seemingly simple tasks like counting from 1 to 10 over and over for more than a minute. Counting is tedious. Playing string quartets is not.

Anecdotes are not dispositive, of course. Still, the anecdote poses a problem for the widely held view that tasks are more difficult if they require more attention. As tasks are learned, they require less attention (e.g., Fitts, 1964; Logan, 1988; Schneider & Shiffrin, 1977). People also avoid tasks requiring greater cognitive demands, all else being equal (Droll & Hayhoe, 2007; Dunn, Lutes, & Risko, 2016; Fournier et al., 2018; Kool, McGuire, Rosen, & Botvnick, 2010). Given these observations, if more attention-demanding, longer lasting tasks (e.g., playing string quartets) are judged easier than less

✉ David A. Rosenbaum
david.rosenbaum@ucr.edu

[1] Department of Psychology, University of California, Riverside, CA 92521, USA

attention-demanding, shorter lasting tasks (e.g., counting), that outcome could signal the need for a new measure of subjective task difficulty.

We pursued a new measure here. The measure we pursued was based on a new hypothesis—namely, the longer a task is deemed sustainable, the greater its subjective ease. To the best of our knowledge, this is a new hypothesis for psychology, neuroscience, and related fields.

Several comments will help explain our thinking. First, we focused on how a long a task is thought or expected to be sustainable because expectation is, or is here regarded as, a subjective measure, which subjective difficulty is as well, of course. We were interested in relating one subjective measure to another. The question of what the relation is between *estimated* and *actual* sustainability is an interesting one, but not one we explored.

Second, the sustainability hypothesis makes sense from the point of view that, as individuals carry out tasks, they get fatigued (Ackerman, 2011). Watching radar screens for hours on end leads to vigilance decrements, for example (Pattyn, Neyt, Henderickx, & Soetens, 2008).

Third and finally, as shown in Fig. 1, when one must decide which of two tasks is easier, if the tasks seem equally hard or easy when done once (the leftmost points in the figure), tiny differences between them may grow as the tasks repeat (points farther to the right in the figure). Hence, a possible heuristic for choosing between tasks is to consider their sustainability after multiple repetitions.

Method

We conducted a follow-up to Potts, Pastel, and Rosenbaum (2018), focusing on eight tasks. Four were simply to count from 1 up to 8, 12, 16, or 20, doing so aloud and at a comfortable pace. The other four tasks involved walking down a 16-foot-long alley, picking up a bucket to one side of the alley, and carrying the bucket to a stool at the end of the alley (see Fig. 2). One bucket task was picking up an empty bucket from a stool that was *near* (.15 m) the alley. Another was picking up the same empty bucket on a stool *far* from (.75 m) the alley. The two other tasks were the same, except that the buckets were loaded with 7 pounds of pebbles.

The eight tasks were completed by each participant once at the start of the experiment so participants could get familiar with the tasks and so we could record the times to complete them based on off-line video analysis. We needed the times to test the hypothesis that choices between the tasks might be based on their times, as found by Potts et al. (2018).

After the familiarization, participants in one group chose between each of the bucket tasks and each of the counting tasks and then made sustainability judgments about the tasks. Participants in the other group did these things in the opposite order. The assignment of participants to groups was random except for the constraint that an equal number (18) of participants needed to be in each group.

In the choice task, the choice was simply to perform whichever task seemed easier—the possible counting task or the

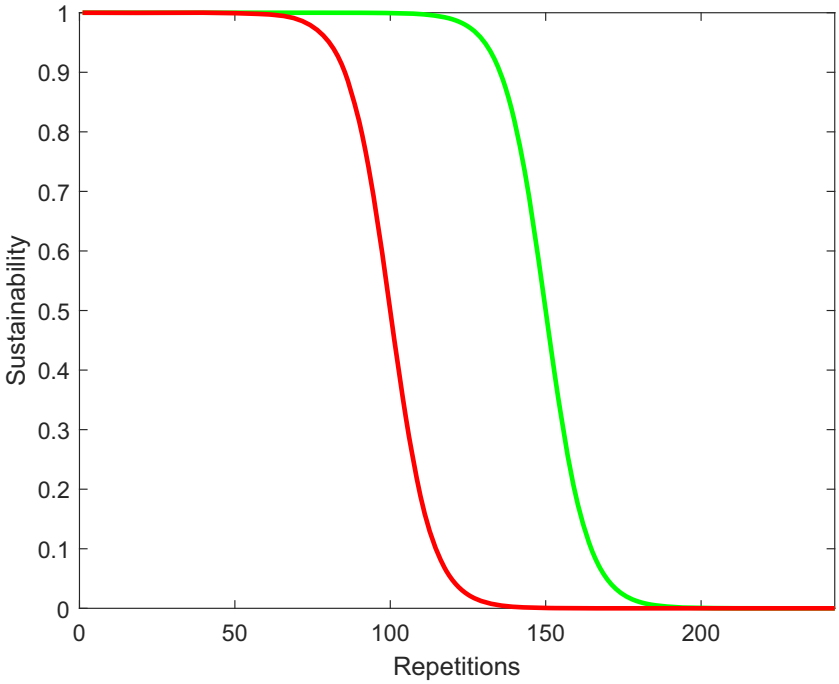

Fig. 1 Theoretical curves showing sustainability as a function of repetitions for a task with low sustainability (left, red curve) and high sustainability (right, green curve). (Color figure online)

Fig. 2 Setup for the experiment. Each individually tested participant stood at the blue line looking out at the six stools. If the participant chose to do the bucket task, he or she walked down the alley, picked up the bucket, which stood near or far from the alley; the far case is shown here. The bucket was either empty or loaded with 7 pounds of pebbles. Its load status was announced before each trial. If the participant chose to pick up the bucket, she or he walked straight ahead, leaned over and reached for the bucket, and carried it to the remote stool on the same side as the bucket. If the bucket was on the left, it was to be lifted with the left hand and carried to the far left bucket; otherwise, it was to be lifted with the right hand and carried to the far right bucket. The choice in each trial was either doing the bucket task or counting up from 1 to 8, 12, 16, or 20, as indicated on a computer laptop that stood on another stool by the participant's side at the blue line. The laptop is not shown here. The side of the laptop was fixed per participant, as was the side of the bucket (opposite the laptop side). The experiment was conducted in an outdoor arcade on the University of California, Riverside, campus. (Color figure online)

bucket task. The counting tasks were communicated via a laptop that stood to one side of the participant at his or her starting point at the foot of the alley (see the blue line in the photo in Fig. 2). For a random half of the subjects, the laptop stood to the left; for the other participants, the laptop stood to the right. If the laptop was to the left, the *bucket* stood on a stool on the right, and vice versa if the laptop stood on a stool on the right. The reason for varying the side of the bucket and laptop was to control for possible side bias. (The results indicated that there was none.) All four stools stood midway along the length of the alley. Figure 2 shows a typical setup. The experimenter always told the participant whether the bucket was loaded or empty. These terms were meaningful since participant had carried the bucket with and without a load (and from the near and far locations) in the familiarization phase.

In addition to the choice task, participants gave sustainability judgments. These were obtained by asking participants whether they thought they could do each of the eight possible tasks 3, 9, 27, 81, or 243 times. The rationale for using these numbers was that, according to Fechner's law, subjective magnitude scales with the logarithm of stimulus intensity (see, for example, Gescheider, 1997). We assumed that a comparable relation would apply to subjective difficulty, so there would be a roughly constant rise in subjective difficulty with each of these successively larger numbers, which are 3^1, 3^2, 3^3, 3^4, and 3^5, respectively. Our participants (36 university students who participated for course credit) were told that

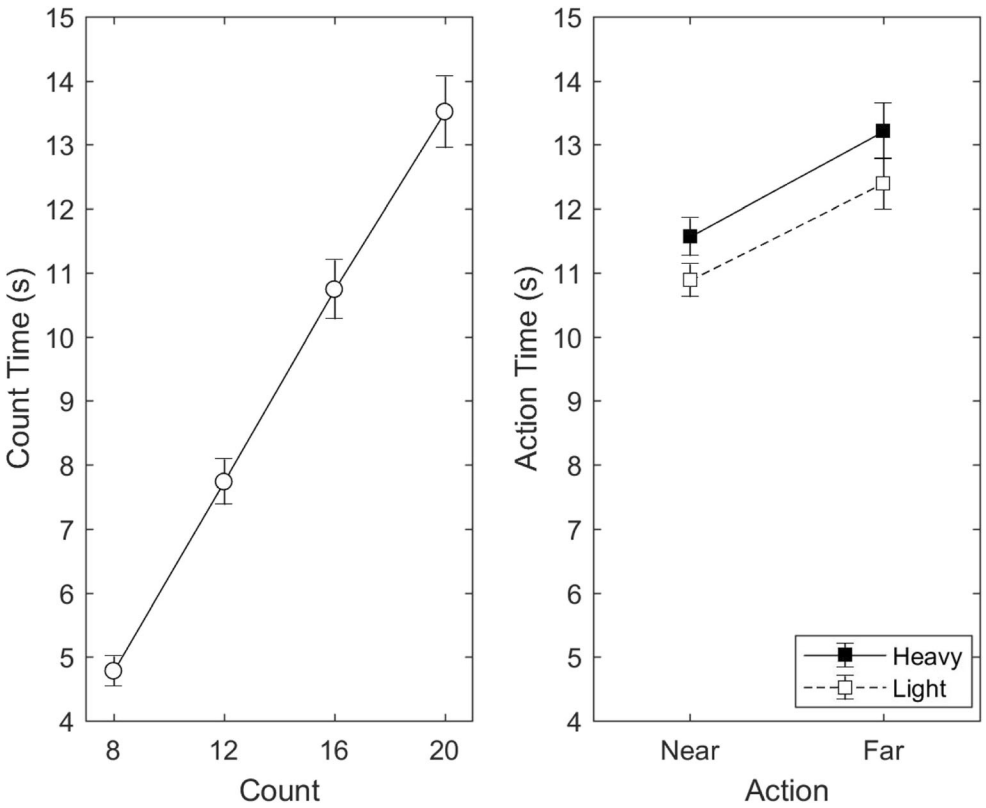

Fig. 3 Mean (±1 *SE*) count times and action times

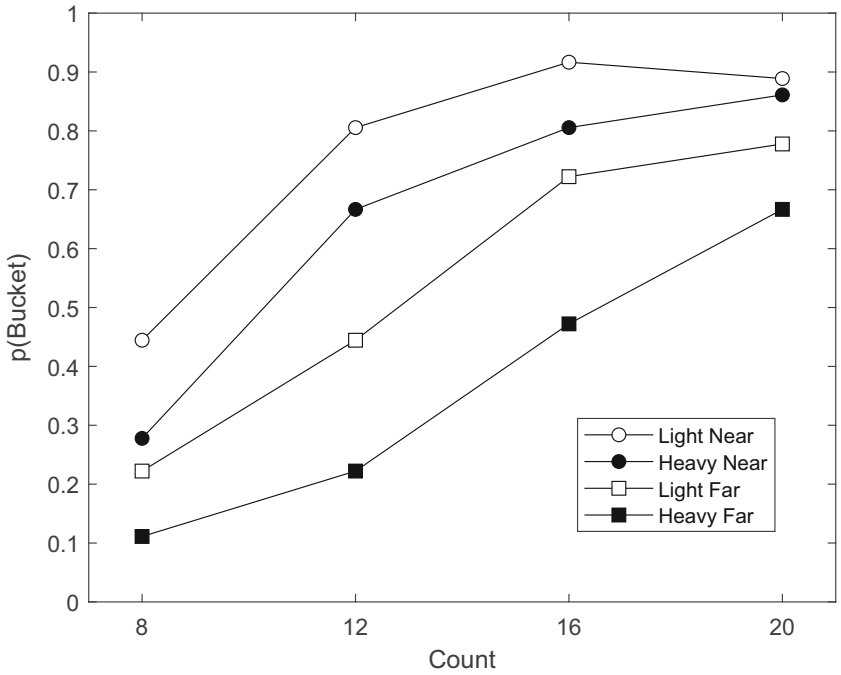

Fig. 4 Probability of choosing the bucket when the bucket was either light and near, heavy and near, light and far, or heavy and far, and when the alternative task was counting to 8, 12, 16, or 20

the values of 3, 9, 27, 81, or 243 corresponded to 3^k, where k could be 1, 2, 3, 4, or 5. Participants were told that the numbers were meant to provide an overall sense of how many times the named task would be performed and that the question was being asked hypothetically. The question for each number of named times was simply whether the participant thought she or he could repeat the task that number of times. The participants simply said "yes" or "no" in each case. The 20 possible task combinations (near light bucket, far light bucket, near heavy bucket, far heavy bucket × the 5 values

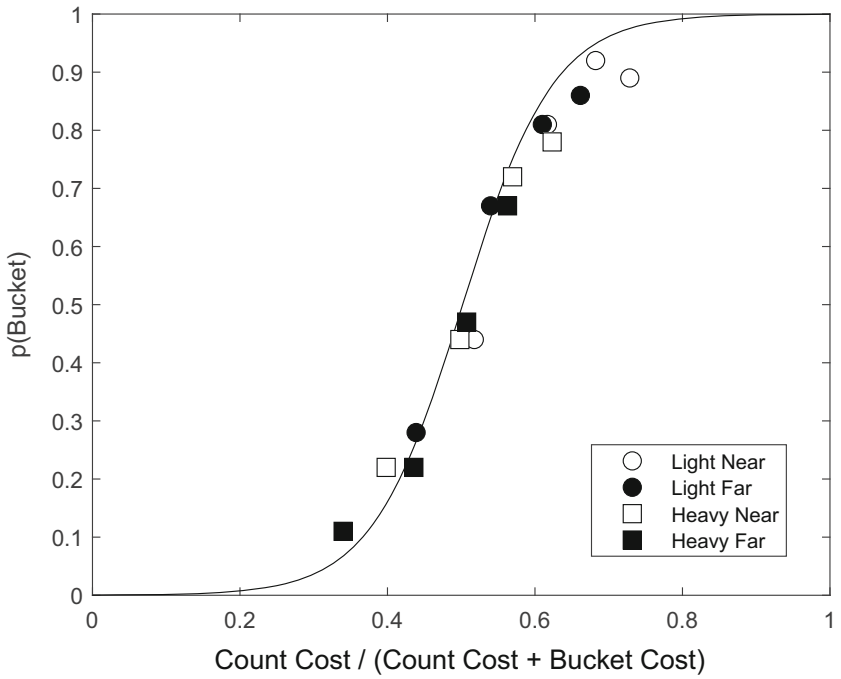

Fig. 5 Probability of choosing the bucket rather than counting, as a function of a single cost, the cost of counting from 1 up to 8, 12, 16, or 20, divided by that cost plus the cost of picking up and carrying the bucket when it was light and near, light and far, heavy and near, or heavy and far. The fitted logistic curve minimized the sum of squared deviations between the observed and predicted values (R^2 = .95). See text for more details

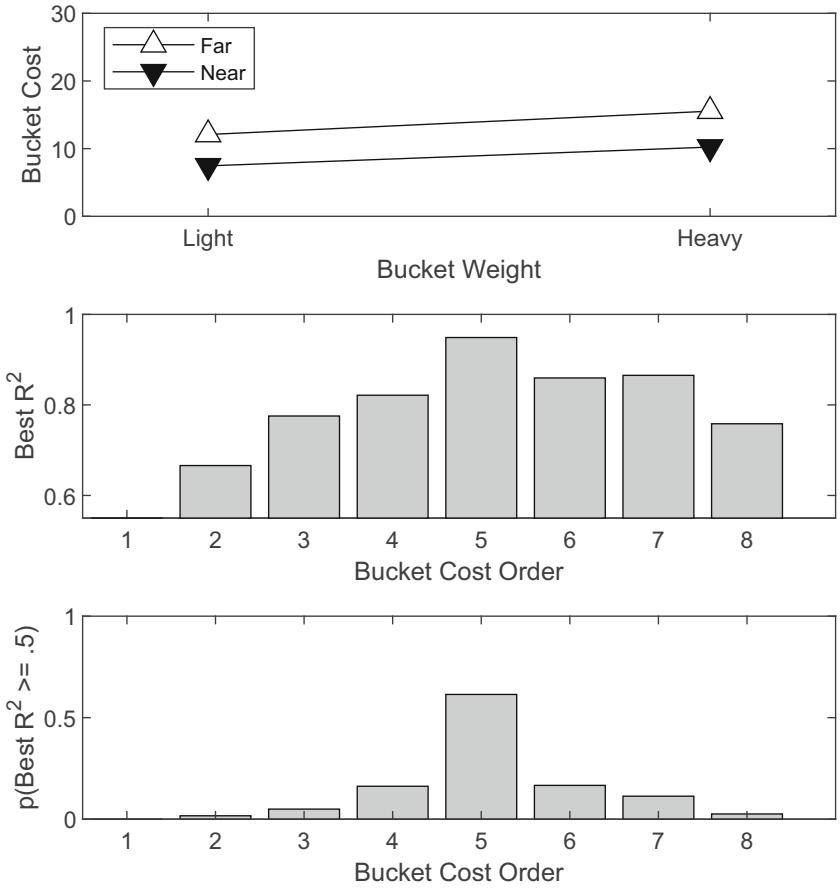

Fig. 6 More model-fitting results. Top panel: Bucket costs for the best fitting model. Middle panel: Best R^2 values for the eight possible orderings of bucket costs for the light–near (LN), heavy–near (HN), light–far (LF), and heavy–far (HF) conditions: (1) NL ≥ NH ≥ FL ≥ FH; (2) NL ≥ NH ≥ FL ≤ FH; (3) NL ≥ NH ≤ FL ≥ FH; (4) NL ≥ NH ≤ FL ≤ FH; (5) NL ≤ NH ≤ FL ≤ FH; (6) NL ≤ NH ≤ FL ≥ FH; (7) NL ≤ NH ≥ FL ≤ FH; (8) NL ≤ NH ≥ FL ≥ FH. Bottom panel: Probability of getting a best R^2 value equal to or exceeding .5 for each of the eight possible bucket-cost orders

of k) were administered in a random order per subject. The task and number of times were displayed on the computer screen, and the experimenter read the task and number of times out loud to the participants, who stood at the blue start line, as in the actual performance situation, and answered "yes" or "no," whereupon the experimenter input the answer to the computer.

Participants who gave the sustainability judgments first were not told that they would have to perform the tasks in the choice context afterward, and participants who performed the tasks in the choice context first were not told that would have to give sustainability judgments afterward. Participants in both groups were told that there were no right or wrong answers. The experiment was approved by the University of California, Riverside, Institutional Review Board.

Results and discussion

Figure 3 shows the times to perform the tasks. The times grew with the count value, were longer for far reaches than for near reaches, and were longer for the weighted bucket than for the unweighted bucket. Similar results were reported by Potts et al. (2018).

Figure 4 shows the probability, p (Bucket), of choosing the bucket rather than counting when participants decided which task to do. As seen in Fig. 4, p (Bucket) increased with the count, and p (Bucket) was larger for near reaches than for far reaches and was larger for the empty bucket than the loaded bucket. These effects did not depend on whether the choice task (bucket versus counting) was undertaken before or after the sustainability task, and neither were the sustainability judgment affected by whether the sustainability decisions were made before or after the choice task, as confirmed with t tests for each condition.

To evaluate the p(Bucket) data, we applied Cochran's Q to the 16 p(Bucket) proportions to test the null hypothesis that the proportions were the same. Using this nonparametric test (Siegel & Castellan, 1988), where the response variable can only take on two possible values, we could reject the null hypothesis, $Q(15) = 193.96$, $p < .001$. The present p(Bucket) data closely resemble those observed by Potts et al. (2018).

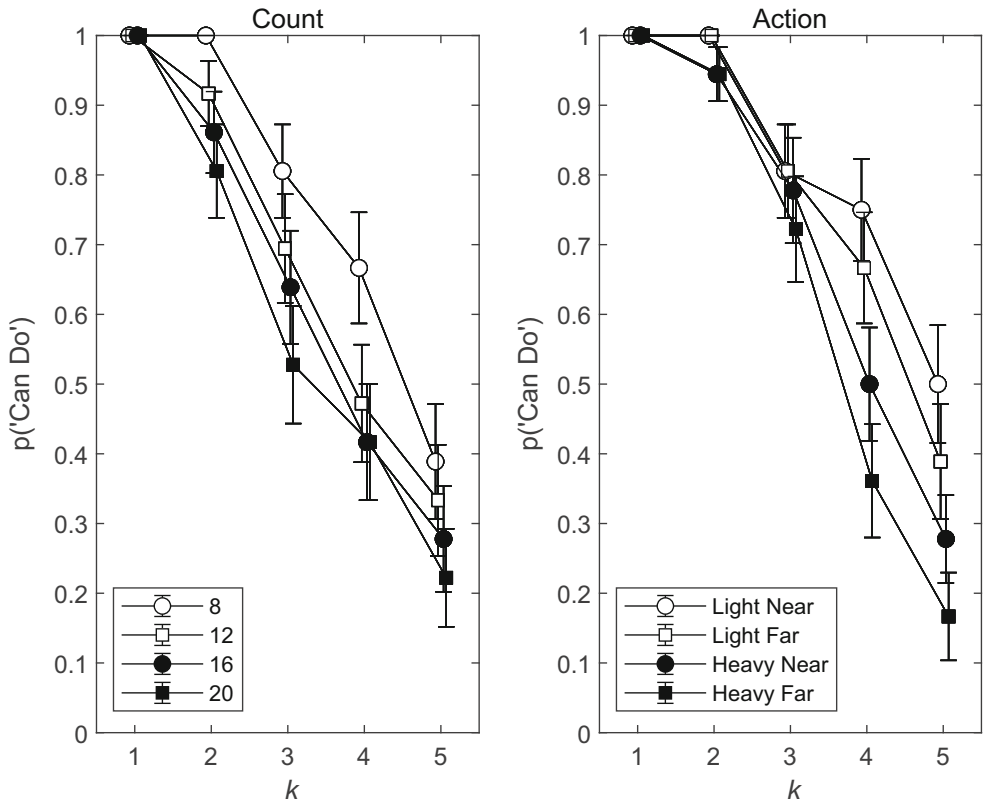

Fig. 7 Mean (±1 *SE*) probability, p('Can Do'), of participants saying they could count to 8, 12, 16, or 20 3^k times (left panel), or carry out the action of transporting the light near bucket, light far bucket, heavy near bucket, or heavy far bucket 3^k times (right panel). The numbers of times participants were actually asked about were 3, 9, 27, 81, and 243, and the basis for those numbers was explained to the participants (3 raised to the first, second, third, fourth, and fifth power). The values on the abscissa are the possible values of *k*

To further evaluate the *p*(Bucket) data, we associated the *p*(Bucket) values with a single theoretical curve. The expected values on the curve came from a logistic (S-shaped) function whose inputs were treated as costs for each of the 16 tasks, expressed as Luce ratios. The Luce ratio for choosing one task (or stimulus), A, rather than another task (or stimulus), B, is simply the probability, *p*(A), of choosing A divided by the sum of the probabilities, *p*(A) + *p*(B), where *p*(A) + *p*(B) = 1 (Luce, 1959; Pleskac, 2015). We assumed that 1 minus the probability of choosing each count value indexed the relative subjective cost of counting up to 8, 12, 16, or 20, divided by the sum of that cost plus the cost of performing whichever bucket task was possible. We also assumed that the costs for counting had reference values of 8, 12, 16, and 20 (arbitrary units). Because we were interested in *relative* costs, we noted that the costs for the bucket tasks could be expressed in the same arbitrary units, so the ratio became dimensionless once the units in the numerator and denominator canceled out.[1]

We wrote a computer program (in MATLAB) to randomly assign possible costs to each of the four bucket tasks. The possible costs were randomly chosen real numbers in the range [0, 28]. That range extended ([−8, 8]) beyond the limits of the count costs ([8, 20]) and was found, in trial runs, to go beyond the range of bucket costs that permitted R^2 values exceeding .10. We generated 1,000 random cost combinations for the four bucket tasks, and then, for each of those four-cost vectors, generated 16 associated Luce ratios for the full set of count and bucket-task combinations. Next, we generated expected values for a logistic curve whose inputs were all of the 16 Luce ratios and whose governing values for the logistic curve's shape and position were randomly chosen 1,000 times for each set of Luce ratios. The program we wrote kept track of the proportion of variance explained, R^2, between the predicted and obtained values of *p*(Bucket) over the 1,000 simulations per Luce ratio set.

The best outcome (R^2 = .95), is shown in Fig. 5. The associated bucket costs were 7.45, 10.23, 12.09, and 15.54, for the light–near (LN), heavy–near (HN), light–far (LF), and heavy–far (HF) bucket tasks, respectively. These four values are shown in Fig. 6 (top panel). The middle panel of Fig. 6 shows the best R^2 values for the eight possible orders of bucket costs. The fifth of the possible orders, LN < HN < LF < HF, was the

[1] A similar approach was taken by Feghhi and Rosenbaum (2019) in a study of choices between tasks involving physical navigation and memorization. These authors found that costs of navigation could be expressed in terms of number of to-be-memorized digits.

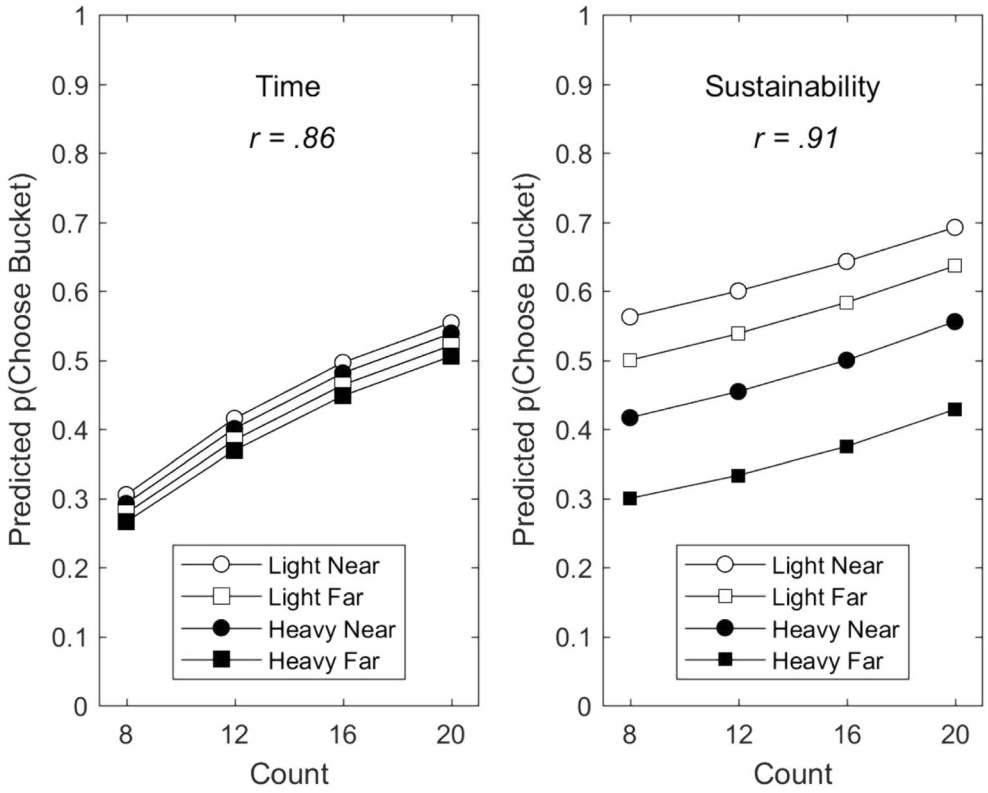

Fig. 8 Predicted probabilities of choosing the bucket based on Luce ratios of task times (left panel) and sustainability values for the largest number (3^5) of repetitions

best of all, as indicated above. The bottom panel of Fig. 6 shows the probability of getting a best R^2 value equal to or exceeding .5 for each cost order. The largest of these probabilities was for the fifth cost order, named above, where the probability was 0.61. The next largest probability was associated with the sixth cost order, NL ≤ NH ≤ FL ≥ FH, where the probability was .16. Based on these results, we can say that, with respect to the criteria just given, the *p*(Bucket) data were about .61/.16 = 3.81 more likely to have come from the fifth cost order than the sixth cost order. The likelihood of the fifth cost order relative to the others was even greater.

Figure 7 shows the sustainability results. The probability, *p*(Can_Do), of "yes" responses decreased with *k*. Moreover, the rate of decrease differed for different tasks. As seen in Fig. 7, *p*(Can_Do) for counting decreased at a higher rate for higher count values than for lower count values. For k = 2, 3, 4, and 5, the top-down ordering of *p*(Can_Do) values was 8 > 12 > 16 > 20. Turning next to the *p*(Can_Do) for the bucket tasks, these results were also orderly. For k = 2, 3, 4, and 5, *p*(Can_Do) was larger for the light (empty) bucket than for the heavy (loaded) bucket. There was a difference between *p*(Can_Do) for near buckets and for far buckets such that *p*(Can_Do) was larger for near buckets than for far buckets, but this difference did not emerge until k equaled 4 or 5.

Figure 8 shows how the observed *p*(Bucket) values were related to the predicted values based on the task times (left panel) and sustainability judgments (right panel), using Luce ratios again. For the time-based predictions, the Luce-ratio numerators were the times for each bucket task, and the denominators were those same times plus the time for the alternative possible counting tasks. The correlation between the obtained and predicted values of *p*(Bucket) based on the times was reasonably good (r = .86), consistent with the expectation that participants would care about time, or that time would serve as a proxy for one or more other variables that participants relied on to make their choices. This outcome accords with the findings of Potts et al. (2018) and the emphasis on time by Gray, Sims, Fu, and Schoelles (2006).

The right side of Fig. 8 shows how the observed *p*(Bucket) values were related to the predicted values based on the sustainability judgments, again using Luce ratios. The numerator was the observed value of *p*(Can_Do) for the largest number (3^5) repetitions of a given bucket task and the denominator was that same value plus the value of *p*(Can_Do) for the 3^5 repetitions of the associated counting task. We focused on the case of 3^5 repetitions (the largest number of repetitions) because that was where the *p*(Can_Do) distinctions were greatest. We reasoned that if there were a problem with the fit of the sustainability model to the choice data, it would be most evident for the 3^5 case.

The correlation between the predicted and obtained values for the sustainability model was somewhat higher than for the

time-based model (r = .91 compared with r = .86), but this difference did not approach so-called statistical significance. Even if it had reached so-called significance, an aspect of the sustainability model results was troubling. Whereas the time model made the correct prediction about the ordering of curves for the LN, HN, LF, and HF conditions, the sustainability model did not. When participants gave sustainability judgments, they apparently believed that bucket weight would more strongly impact sustainability than would reaching distance. By contrast, when the same participants made task choices, they apparently believed, or acted as if they believed, reaching distance was more important than bucket weight.

Final remarks

Where does this leave us? First, the fact that the time model made the correct predictions does not imply that participants used time per se. For example, as argued by Potts et al. (2018), estimates of times might be a better predictor than actual times, and times (actual or estimated) might be a proxy for some other control variable. Second and more importantly, the misordering of the predicted levels of p(Bucket) based on the sustainability model is damning to this model. The misordering must be taken seriously given the model-fitting described above, where we showed that the p(Bucket) data were most likely to have come from LN < HN < LF < HF and not some other order. The order corresponding to the sustainability model was the seventh order, NL ≤ NH ≥ FL ≤ FH, and the likelihood that the p(Bucket) data came from that order was .11, in which case the likelihood of order 5 relative to order 7 was .61/.11 = 5.55. Although it is possible that further work would show that the sustainability model does a better job predicting p(Bucket) data, the most straightforward conclusion for now is that the sustainability model is wrong. Time remains the most promising index of subjective difficulty.

Author note This work was supported by a University of California, Riverside, Committee on Research Grant to the first author. Much of the work done by the second author was completed when he was an undergraduate at UCR. The authors are indebted to Stephen Goldinger and two anonymous reviewers for helpful suggestions on the manuscript.

References

Ackerman, P. L. (Ed.). (2011). Cognitive fatigue: Multidisciplinary perspectives on current research and future applications. Washington, DC: American Psychological Association.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1352–1365. doi:https://doi.org/10.1037/0096-1523.33.6.1352

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,* 1372–1388. doi:https://doi.org/10.1037/xhp0000236

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

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), Categories of human learning (pp. 243–285). New York, NY: Academic Press.

Fournier, L. R., Coder, E., Kogan, C., Raghunath, N., Taddese, E., & Rosenbaum, D. A. (2018). Which task will we choose first? Precrastination and cognitive load in task ordering. *Attention, Perception & Performance.* Advance online publication. doi: https://doi.org/10.3758/s13414-018-1633-5

Gescheider, G. A. (1997). Psychophysics: The fundamentals. Mahwah, NJ: Erlbaum.

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.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665-682. doi:https://doi.org/10.1037/a0020198

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York, NY: Wiley.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? *Physiology & Behavior*, *93*, 369–378.

Pleskac, T. J. (2015). Decision and choice: Luce's choice axiom. In J. D. Wright (Ed.), International encyclopedia of the social & behavioral sciences (Vol. 5, pp. 895–900). New York, NY: Elsevier.

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

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York, NY: McGraw-Hill.