Strategy selection under equal accuracy: Dissociating time, effort, and engagement

Taylor Mathis (a02443118@usu.edu)
Department of Psychology, Utah State University
2810 Old Main Hill, Logan, UT 84341 USA

Zachary A. Niese (zachary-adolph.niese@uni-tuebingen.de)
Department of Psychology, University of Tübingen
4 Schleistrasse, Tübingen, 72076 Germany

Emily R. Weichart (emily.weichart@usu.edu)
Department of Psychology, Utah State University
2810 Old Main Hill, Logan, UT 84341 USA

Abstract

Humans primarily seek to maximize accuracy during task-based decision-making, but prefer to minimize time and effort as well. As time and effort demands are often coupled, the current work examines the relative influences of these secondary minimization goals on decision strategy selection. University students (N=144; ages 18-37) learned two strategies for completing a classification task involving fictional microorganisms: one that demanded moderately effortful integration over learned feature associations, and a minimal-effort alternative that required participants to wait for an answer to be provided. An adaptive experimental design ensured that both strategies yielded equivalent long-run accuracy within-subject, and wait times for the latter were manipulated to be longer or equal to deliberation times in the former. Participants overwhelmingly preferred the more effortful, time-saving strategy. Surprisingly, this preference persisted even when the strategies were matched for both accuracy and time, suggesting the perceived benefit of engagement outweighed the costs of effort.

Keywords: concepts and categories; decision making; problem solving; statistical learning

Introduction

Decision makers seek to maximize accuracy or reward, but also appear to have internal limits on the time and effort they are willing to invest in an individual choice. The costs of time and effort, however, are often tightly coupled, leaving open questions about whether decision makers are primarily motivated to minimize one or the other, or if both draw upon a single limited metabolic resource (Shenhav et al., 2017).

In perceptual decisions, for example, requirements to resolve conflict or uncertainty necessitate longer periods of *evidence accumulation* before a confidently-correct response can be made (Smith & Ratcliff, 2004; Ditterich, 2006). Computational work has demonstrated that evidence accumulation itself is costly, such that humans are willing to make decisions on the basis of less information as deliberation time wears on (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Palestro, Weichart, Sederberg, & Turner, 2018). Other work has shown that humans avoid cognitively effortful tasks (Hull, 1943), even if applying less effort does not result in reciprocal minimization of overall task time (Kool, McGuire, Rosen, & Botvinick, 2010; Westbrook, Kester, & Braver, 2013). Similar tradeoffs have emerged in learning tasks as well, such that humans seek to maximize accuracy while minimizing expenditure of attention across sources of information (Weichart, Galdo, Sloutsky, & Turner, 2022). Given that attention-driven mechanisms for sampling and integrating information require both time *and* effort (Braunlich & Love, 2022), it is unclear which underlying cost functions drive apparent reductions in attention across learning. To address this gap, the current work seeks to dissociate time and effort minimization goals by requiring participants to repeatedly choose between strategies that differ in deliberation time and subjective effort (Figure 1).

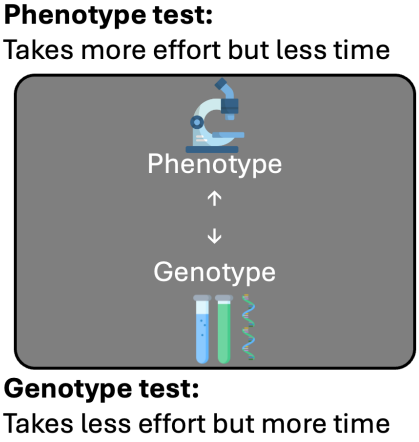

Figure 1: Test selection screen. Example of the interface where participants selected whether to use the "Phenotype" or "Genotype" Test to classify an organism on each trial. Through prior experience, participants learned that the Phenotype Test was more effortful, whereas the Genotype Test required longer deliberation times.

Participants were told they would be helping a scientist classify microorganisms into two species groups. They learned two strategies for achieving this: A) the *Phenotype Test* required viewing an organism and effortfully integrating across learned feature-to-category associations; B) the *Genotype Test* did not involve viewing an organism at all, but rather waiting for an answer to be explicitly provided after an amount of time that was equal to or longer than was required for the Phenotype Test. Correct/Incorrect feedback was provided in both cases, and an adaptive design was used so that the tests would yield equivalent long-run accuracy within-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

subject. Participants then chose which strategy they wanted to use on each trial. We hypothesized that systematic biases toward one strategy or the other would indicate whether decision makers prioritize time or effort minimization in the absence of accuracy differences between strategies. We suggest that this line of inquiry is particularly relevant to the domain of higher education as students increasingly rely upon Artificial Intelligence (AI) tools to reduce time and/or effort when completing academic tasks. To foster genuine learning in the age of AI, educators first need to understand the motivational factors that explain student decisions to cognitively engage or offload tasks to external tools.

Method

Participants

All research was approved by the Institutional Review Board at Utah State University (USU; Protocol #13870). Participants (N=169) were recruited from USU's student population in Logan, Utah, and were compensated with partial course credit. The study was conducted in-person and all participants verbally self-reported eligibility based on the following inclusion criteria: age 18+, English language fluency, normal or corrected-to-normal vision, and no history of photosensitive epilepsy. Data from twenty-five (25, 14.8%) participants were excluded from analysis for failing a learning check (see description of *Phase 1*). The reported sample consists of 144 participants (98 females, 45 males, 1 unreported gender; age range 18-37 years, mean=19.76, SD=2.45).

Apparatus and Stimuli

Each participant was seated in an individual testing room in view of an experimenter. Stimuli were presented on an EIZO FlexScan EV2451 monitor using PsychoPy 1.85.0 (Peirce, 2009). Responses were made using a standard keyboard and mouse. Before beginning the experiment, participants provided informed consent and completed 10 experimenter-guided practice trials where they learned to map features to categories via trial-and-error. The practice set was intended to convey the structure of the main task, but stimuli used different features to avoid carryover.

Stimuli in the main task were fictional microorganisms that were presented as if viewed through a microscope (Figure 2A) (Weichart & Darby, 2026). Each stimulus consisted of a common background image with five spatial locations where organelles could appear. One of two possible organelles could appear in each location. The background image and organelles subtended 18.6° and 1.6° of visual angle, respectively, when participants were seated 61 cm from the monitor. Organelles that could occur in the same location across trials were designed to be visually similar, whereas organelles at different locations were visually distinct. To create each stimulus, organelles were drawn from two un-presented species prototypes ("Dax" and "Zim"), which each consisted of a distinct set of five organelles. "True" feature-to-category associations were randomized between subjects.

Procedure

Participants were randomly assigned to one of three manipulation groups. For all groups, the Phenotype Test was designed to be more effortful than the Genotype Test (Effort Level: 0). Organisms presented to Groups 1 and 2 during the Phenotype Test contained 3/5 organelles from a common Dax or Zim prototype (Effort Level: 2), whereas organisms presented to Group 3 contained 4/5 organelles from a common prototype (Effort Level: 1). Genotype Test wait durations between trial start time and species label presentation were set to be equal to Phenotype stimulus viewing times for Group 1, but were 5 seconds (s) longer for Groups 2 and 3. A summary of the between subjects design is provided in the top row of Table 1.

Phase 1: Feature-to-Category Associations Before being introduced to the two organism classification strategies, participants learned which organelles were typical of each species. Organelles were presented individually, and participants pressed the spacebar when they were ready to judge whether the organelle was typical of a Dax or Zim. The stimulus was replaced by text corresponding to the two options, and participants used the left and right arrow keys to make a selection. Response mappings were counterbalanced across subjects. Feedback appeared for 1.2 s after the response: "correct" (green text) or "incorrect" (red text). Each unique organelle was presented 8 times in a semi-randomized order, resulting in 80 trials. After learning feature-to-category associations via feedback, participants completed an additional 30 trials (3 presentations of each organelle) without feedback. Participants were excluded from analysis if accuracy was not significantly above chance as determined by binomial tests ($H_0 = 0.5$, $\alpha = 0.05$).

Phase 2: Practicing the Phenotype and Genotype Tests Participants were told that they would be practicing two tests for classifying organisms. They first completed 20 trials of the Phenotype Test. Organisms consisted of 5 organelles, and participants were instructed to choose the species associated with *most* of the organelles. Category structures are shown in the lower panel of Figure 2A. The trial progression was similar to Phase 1, and correct/incorrect feedback was provided. The only difference was that each stimulus consisted of five previously-seen, concurrently-presented organelles instead of one. An adaptive experimental design was used, such that each subject's accuracy and response times (RTs) across the latter 10 trials of the Phenotype Test affected the progression of subsequent exposure to the Genotype Test.

When being introduced to the Genotype Test, participants were told that they would wait for a result before selecting a species. Ten trials were completed. Each began with an image of test tubes alongside instructions to "Please wait" (Figure 2B). The rightmost test tube was animated so that it appeared to "fill" over time like a progress bar. Wait times were randomly sampled from each participant's own distribution of RTs from trials 11-20 of the Phenotype Test, then increased

Table 1: Manipulation checks of accuracy and deliberation times in Phase 2, and subjective effort.

Measure	Test	Group 1	Group 2	Group 3
Design	Phenotype	Time: t; Eff: L2	Time: t; Eff: L2	Time: t; Eff: L1
	Genotype	Time: t; Eff: L0	Time: t+5; Eff: L0	Time: t+5; Eff: L0
Accuracy	Phenotype	0.78 (0.20)	0.78 (0.18)	0.91 (0.12)
	Genotype	0.80 (0.23)	0.77 (0.23)	0.91 (0.13)
	Δ (P−G)	$t = -0.94$, $p = .23$	$t = 0.68$, $p = .50$	$t = -0.32$, $p = .74$
Time (s)	Phenotype	6.43 (2.92)	6.52 (2.53)	5.14 (2.84)
	Genotype	6.44 (2.81)	11.60 (2.51)	10.23 (2.97)
	Δ (P−G)	$t = 0.22$, $p = .82$	$t = -85.37$, $p < .001$***	$t = -129.68$, $p < .001$***
Effort	Phenotype	3.77 (0.65)	3.39 (0.86)	3.06 (0.82)
	Genotype	1.73 (0.88)	1.95 (1.02)	1.73 (1.13)
	Δ (P−G)	$t = 17.17$, $p < .001$***	$t = 7.51$, $p < .001$***	$t = 6.25$, $p < .001$***

Note. t refers to the participant's distribution of time spent studying organisms in Trials 11-20 of Phenotype Test. Genotype Test wait times were based on this distribution. L0, L1, and L2 refer to levels of our effort manipulation, as presented in Figure 2. Cells for Accuracy/Time/Effort report *M* (SD). t and p values indicate results of paired-samples t-tests. Δ denotes Phenotype minus Genotype.

by 0 s (Group 1) or 5 s (Groups 2 and 3). After the wait time ended, the text "Test complete." and "Result: [DAX/ ZIM /Inconclusive]" appeared. When species labels were provided, they corresponded to the correct response 100% of the time. "Inconclusive" results meant that participants would have to guess a species (expected accuracy: 50%). For every n incorrect responses during trials 11-20 of the Phenotype Test, there were $2n$ (up to 10) trials with "Inconclusive" results during the Genotype Test. This was intended to equate expected accuracy between the two tests on average. As in the Phenotype Test, participants pressed the spacebar when they were ready to make a species judgment. Text corresponding to the two species options was presented side-by-side, participants used the left and right arrow keys to indicate "DAX" or "ZIM," and feedback was presented for 1.2 s.

Phase 3: Test Selection Participants were instructed to choose which test they wanted to use to classify each of the remaining organisms. Forty trials were completed. Each trial began with a test selection screen (Figure 1), and participants used the up and down arrow keys to choose between the Phenotype and Genotype Tests. Response mappings were counterbalanced between participants. Trial progressions then proceeded identically to the practice sets, depending on whether the Phenotype or Genotype Test was selected. Wait times and proportions of "Inconclusive" results in the Genotype Test were fixed based on Phase 2 Phenotype Test performance and were not adaptively updated according to ongoing performance in Phase 3.

Post-Experiment Subjective Effort Ratings At the end of the experiment, participants were presented with questions in the form: "How would you rate the level of effort required for you to complete the [GENOTYPE/ PHENOTYPE] test (1=no effort; 5=maximum effort)?" Participants responded by using the mouse to click 1 of 5 radio buttons.

Results

After excluding participants who did not learn the feature-to-category associations beyond a chance-level criterion, there were 48 participants in Group 1, 44 in Group 2, and 52 in Group 3. Among these participants, Phase 1 feature-to-category association accuracy was reasonably high during both initial feedback-based learning (80 trials; M=0.86, SD=0.09) and the learning check without feedback (30 trials; M=0.89, SD=0.11).

Manipulation Checks Table 1 shows accuracy and deliberation time comparisons between tests in Phase 2. Comparisons between post-experiment effort ratings are shown as well. Per our adaptive design, there were no significant differences in accuracy between the Phenotype and Genotype Tests as determined by paired-samples t-tests (all $ps > 0.05$). Also as intended, deliberation times were equal between tests for Group 1 ($t[47] = 0.22$; $p = 0.82$) and the Genotype Test required longer deliberation times than the Phenotype Test for Groups 2 and 3 (Group 2: $t[43] = -85.37$; $p < 0.001$***; Group 3: $t[51] = -129.68$; $p < 0.001$***). All three groups rated the Phenotype Test as requiring more effort than the Genotype Test (all $ps < 0.001$***).

Effortful Phenotype Test Unilaterally Preferred If participants prioritized minimizing effort, we would expect the complexity of the Phenotype Test to impact how frequently it was selected (Phenotype preference: Groups 1 and 2 < Group 3). If participants prioritized saving time, we would expect the relative wait time of the Genotype Test to impact how frequently Phenotype Test was selected (Phenotype preference: Groups 2 and 3 > Group 1). An initial ANOVA and pairwise tests (outcome measure: proportion of Phenotype selections

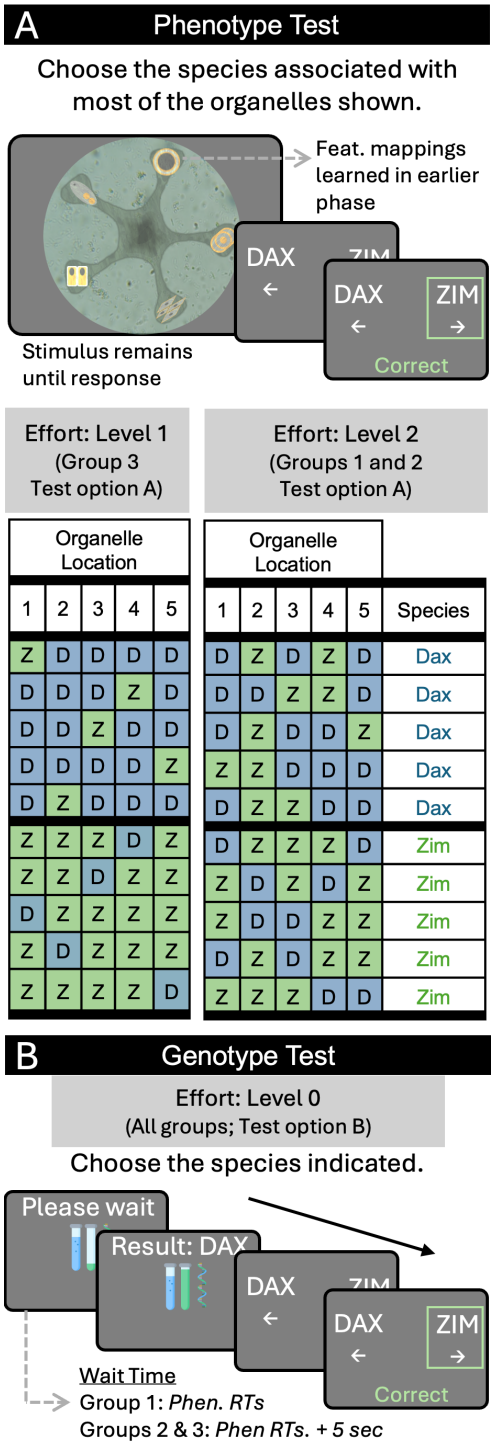

Figure 2: Progression of test trials. D/Z=features associated with Dax/Zim species labels in Phase 1; RTs=response times. **(A)** Top: In Phase 1, 10 organelle-to-species associations were learned. In the Phenotype Test, organisms were classified based on the majority of organelles. Bottom: Category structures. Between subjects, either 4/5 or 3/5 organelles were associated with a common species. **(B)** In the Genotype Test, species labels were provided after a wait. The test could provide the correct answer, or could be "inconclusive." Between subjects, wait times either matched the amount of time taken to study stimuli in the Phenotype Test or were 5 seconds longer.

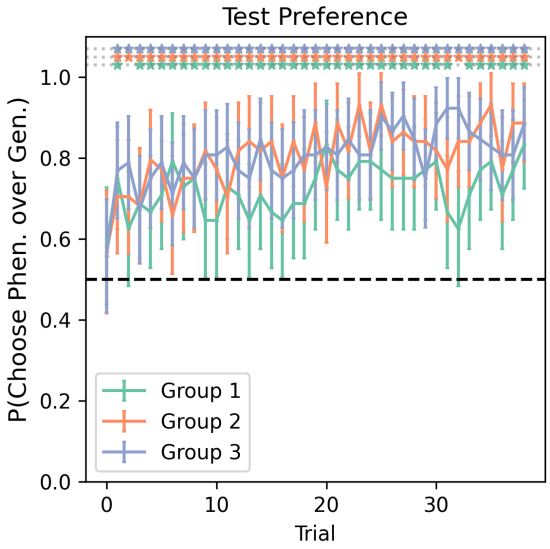

Figure 3: By-trial preference for Phenotype Test over Genotype Test in Phase 3. One-sample t-tests were used to compare mean choice proportions to 0.5 (equal preference) within each group. P values were corrected for false discovery rate. Error bars show 95% confidence intervals. Stars indicate significant deviation from 0.5 at $\alpha = 0.05$. All groups preferred Phenotype Test throughout Phase 3.

in Phase 3) trended toward the latter account ($F(2) = 3.154$, $p = 0.047$, $\eta^2 p = 0.048$; 1 vs. 2: $g = -0.452$, $p = 0.076$, 1 vs. 3: $g = -0.457$, $p = 0.063$, 2 vs. 3: $g = 0.008$, $p = 1.000$).

Figure 3 shows trial-level proportions of participants in each group that selected the Phenotype Test during Phase 3. Group-level means were compared to equal test preference (0.5) using one-sample t tests, and resulting p values were adjusted for false discovery rate (FDR) (Benjamini & Hochberg, 1995). All groups showed a robust preference for the Phenotype Test over the Genotype Test across trials. This pattern would be expected for Groups 2 and 3 if participants prioritized minimizing time over effort under conditions of equal accuracy. In these groups, the Phenotype test was perceived as more effortful but participants could respond approximately 5 s faster than they were permitted during the Genotype Test. Surprisingly, however, Group 1 overwhelmingly preferred the Phenotype Test as well even though accuracy *and* deliberation time were matched between tests and Phenotype Test was perceived as *more* effortful.

One potential explanation is that although effort may be considered costly in some cases (Kool et al., 2010), another dimension such as "engagement" may have been more salient when choosing a test option here. Participants may have intrinsically valued the active, self-directed nature of the Phenotype Test over the easier-but-passive nature of the Genotype Test. To briefly examine this hypothesis, an additional sample of 27 participants completed the experiment under conditions where the tests were equalized for accuracy and time during training (3 excluded for learning check failure; final sample: N=24, 17 females, 7 males; ages: 18-22, M=19.58,

Table 2: Mixed-effects logistic regression coefficients for predicting probability of Phenotype Test selection in Phase 3

Predictor	Estimate	Std. Error	z	p	Sig.
Intercept	0.904	0.213	4.236	$< .001$	***
Trial	0.062	0.056	1.099	.272	
Δ Accuracy	0.261	0.069	3.779	$< .001$	***
Δ Time	-0.102	0.096	-1.058	.290	
Δ Effort	0.083	0.106	0.786	.432	
Δ Accuracy × Trial	0.037	0.041	0.912	.362	
Δ Time × Trial	-0.157	0.038	-4.142	$< .001$	***

Note. The outcome variable was test selection in Phase 3; 0=Genotype, 1=Phenotype. Δ refers to Phenotype minus Genotype. Accuracy and Time predictors are cumulative test-level means across trials. All predictors were scaled by standard deviation. By-subject intercepts were included as random effects.

SD=1.35). These participants were asked to rate their enjoyment of each test in addition to the effort required. Participants reported enjoying the Phenotype test (1=no enjoyment, 5=maximum enjoyment; M=3.63; SD=0.70) more than the Genotype test (M=2.13, SD=1.30), and this difference was statistically significant as determined by a paired-samples t-test (t(23)=4.41, $p < 0.001$***). Importantly, they enjoyed the Phenotype test despite finding it to be more effortful than the Genotype test, consistent with results from Groups 1-3 (Phenotype: M=3.08, SD=0.91; Genotype: 2.03, SD=1.10; t(23)=3.61, $p < 0.01$**). This suggests that choice policies in our main experiment may have been driven by factors other than a tradeoff between time and effort, such as by the subjective value of engagement. This point will be revisited in the *Discussion*.

Accuracy Prioritized First; Saving Time Predicts Choice Later We used a logistic mixed-effects regression model to examine how participants' accumulated experience with each test impacted their selection across trials. The model was implemented in R version 4.5.1 (R Core Team, 2025) using the package *lme4* version 1.1.37 (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was test choice (0=Genotype, 1=Phenotype). Fixed effects included between-test differences in cumulative mean accuracy and deliberation time, their interactions with trial number, and differences in subjective effort ratings. Random intercepts were included for each participant. Predictors were scaled by their standard deviation instead of z-scoring to retain Trial 1 as a baseline. The model was fit using maximum likelihood with a Laplace approximation to the marginal likelihood, and significance of fixed effects was assessed using Wald z tests. The model converged without warnings. Figure 4 illustrates general changes in each predictor across trials, and Table 2 shows fixed effect estimates from the model predicting test selection. Predictors are presented in the format "ΔX" to indicate Phenotype minus Genotype differences.

We observe a strong initial bias toward selecting the Phenotype Test ($\beta = 0.904$, $z = 4.236$, $p < 0.001$***), corresponding to a predicted selection probability of 71% if tests were matched for accuracy, deliberation time, and effort. We suggest that this subsumes the previously-discussed, unanticipated influence of task engagement on selection.

We additionally observe that a one-SD increase in ΔAccuracy was predicted to increase the odds of choosing the Phenotype test immediately after test exposure by 29.8%. The tests, however, were matched in accuracy during Phase 2 (Table 1). We suggest that although tests did not differ in accuracy across subjects on average, idiosyncratic within-subject differences may have guided test selection at the outset of Phase 3. To examine this possibility, we sorted participants by their test choice on Trial 1 of Phase 3, and examined their relative performances on the Phenotype and Genotype Tests during Phase 2. Participants who showed an initial bias toward the Genotype Test during Phase 3 (N = 46) had performed slightly better during Genotype practice than Phenotype practice on average (Δ[P-G]: M = -0.011, SD = 0.14), whereas participants who were biased toward the Phenotype Test (N = 98) showed the opposite pattern (Δ[P-G]: M = 0.002, SD = 0.11). This difference was not statistically significant as determined by an independent samples t-test ($t(76.18) = 0.55$, $p = 0.58$), but may nevertheless explain the predictive value of ΔAccuracy shown in Table 2. Consistent with the finding that cumulative accuracy differences did not meaningfully deviate from zero during Phase 3 (leftmost panel of Figure 4), the interaction between ΔAccuracy and trial did not significantly predict test selection.

Despite large differences in Genotype deliberation times between groups (0 s vs. 5 s), ΔTime was not a significant predictor of test choice at the start of Phase 3 (β =-0.102, z=-1.058, p=0.290). The interaction ΔTime and trial number, however, was a significant predictor ($\beta = -0.157$, $z = -4.142$, $p < 0.001$***). This suggests that as the task progressed, participants increasingly selected the test that could be completed faster. Indeed, the middle panel of Figure 4 shows a gradual decrease in ΔTime, consistent with faster Phenotype deliberation times due to practice while Genotype deliberation times remained fixed.

Differences in subjective effort did not significantly predict test choice after controlling for ΔAccuracy and ΔTime ($\beta = 0.083$, $z = 0.786$, $p = 0.432$). Participants selected the Phenotype Test despite perceiving it as more effortful, in a manner that carried through to significantly higher subjective effort ratings on the post-experiment questionnaire (rightmost panel of Figure 4).

Discussion

We designed our experiment to examine whether time or effort minimization goals take priority when controlling for accuracy. Our findings suggest that while participants initially seek to maximize accuracy, they increasingly preferred the faster, more effortful strategy as the task went on (Table 2). Surprisingly, the preference for the effortful Phenotype Test persisted even when it did not offer a time-saving advantage over the Genotype Test (Figure 3).

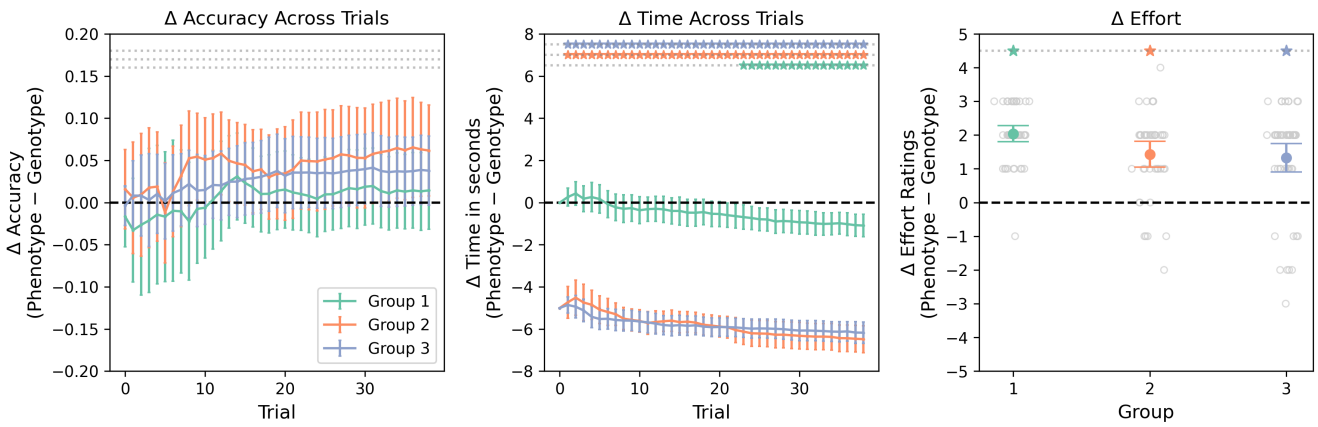

Figure 4: Accuracy, deliberation time, and effort differences between tests in Phase 3. Δ refers to Phenotype minus Genotype. Panels show the results of one-sample t-tests comparing mean differences to 0, indicating no difference between tests. Error bars are 95% confidence intervals. Stars indicate significant deviation from 0 at $\alpha = 0.05$. P-values were corrected for false discovery rate. **Left Panel.** Trial-level differences in cumulative accuracy. Tests did not significantly differ in accuracy for any group. **Middle Panel.** Trial-level differences in deliberation time. Phenotype Test was always significantly faster than Genotype Test for Groups 2 and 3. For Group 1, Phenotype Test became faster than Genotype Test in latter trials. **Right Panel.** Difference in post-experiment subjective effort ratings. All groups rated Phenotype Test as significantly more effortful than Genotype Test.

The latter result seemingly contradicts many findings suggesting that decision makers seek to reduce cognitive effort in a resource-rational manner that mirrors reduction of physical effort (Navon & Gopher, 1979; Kagel, Battalio, & Green, 1995; Kool et al., 2010; McGuire & Botvinick, 2010; Westbrook et al., 2013). Cognitive effort has been proposed to be costly insofar as it increases metabolic demand in the brain and is subject to diminishing accuracy returns due to mental capacity constraints (Shenhav et al., 2017). Other work, however, has suggested that effort adds subjective value to reward (Inzlicht, Shenhav, & Olivola, 2018). Effortful performance has been shown to amplify neurophysiological signals related to feedback-related reward positivity, which are muted or absent following non-effortful performance (Lallement et al., 2013; Ma, Meng, Wang, & Shen, 2014; Wang, Zheng, & Meng, 2017). Longstanding literature on phenomena like need for cognition (Cacioppo, Petty, Feinstein, & Jarvis, 1996; Westbrook et al., 2013), flow (Csikszentmihalyi, 2014), and regulatory fit (Freitas & Higgins, 2002; Niese, Libby, & Pfent, 2021) describe how observing alignment between effort and goal-related outcomes increases motivation. The robust preference for the effortful Phenotype Test that we observed here may therefore be attributed to the motivational value of active engagement. Indeed, a follow-up participant sample reported enjoying the Phenotype Test, despite finding it more effortful than the Genotype Test.

We note that operationalizing "effort" is not straightforward, which may partially account from the apparent conflict between our results and other studies that have demonstrated preferences toward effort minimization. For example, several findings of effort minimization have come from *Demand Selection Tasks*, whereby effort demand is increased via the addition of set-shifting requirements (Botvinick & Rosen, 2009; Kool et al., 2010; Dunn, Lutes, & Risko, 2016). When the surface qualities of two alternative tasks and time requirements are matched, participants overwhelmingly avoid the option that requires set-shifting (i.e. AAABBB design preferred over ABABAB). Our experiment, in contrast, operationalized effort demand as a feature integration requirement relative to simply waiting for a result. This design choice was inspired by a real-world dilemma that today's students continuously face: should I effortfully engage with assignments, or offload to ChatGPT? To reconcile diverging results between implementations, we suggest that effortful engagement may be avoided when it is experienced as inefficient relative to an alternative, but preferred when it is perceived as directly supporting achievement of a goal.

In line with pedagogy emphasizing the benefits of *desirable difficulties* on learning (Bjork & Bjork, 2020), our findings may offer a bit of hope to educators as they navigate a new AI-laden landscape. Students may effortfully engage if 1) they can observe a direct impact of effort on performance; and 2) effortful engagement does not take more time than offloading to external tools.

Acknowledgments

This work was funded by Research Catalyst grant A62271 awarded to Emily Weichart by the Office of Research at USU. The authors gratefully acknowledge Ethan Bardsley, Isabelle Lloyd, Mylee Strong, Jared Vance, and Talmage Walters for assisting with data collection.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bjork, R., & Bjork, E. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. doi: 10.1016/j.jarmac.2020.09.003

Botvinick, M., & Rosen, Z. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research PRPF*, *73*, 835–842. doi: 10.1007/s00426-008-0197-8

Braunlich, K., & Love, B. (2022). Bidirectional influences of information sampling and concept learning. *Psychological Review*, *129*(2), 213–234.

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. doi: 10.1037/0033-2909.119.2.197

Csikszentmihalyi, M. (2014). *Flow and the foundations of positive psychology*. Springer Dordrecht. doi: 10.1007/978-94-017-9088-8

Ditterich, J. (2006). Evidence for time-variant decision making. *European Journal of Neuroscience*, *24*(12), 3628–3641. doi: 10.1111/j.1460-9568.2006.05221.x

Drugowitsch, J., Moreno-Bote, R., Churchland, A., Shadlen, M., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628. doi: 10.1523/JNEUROSCI.4010-11.2012

Dunn, T., Lutes, D., & Risko, E. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1372–1387. doi: 10.1037/xhp0000236

Freitas, A., & Higgins, E. (2002). Enjoying goal-directed action: The role of regulatory fit. *Psychological Science*, *13*(1), 1–6. doi: 10.1111/1467-9280.00401

Hull, C. (1943). *Principles of behavior: An introduction to behavior theory*. Appleton-Century.

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

Kagel, J., Battalio, R., & Green, L. (1995). *Economic choice theory: An experimental analysis of animal behavior*. Melbourne, Australia: Cambridge University Press.

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

Lallement, J., Kuss, K., Trautner, P., Weber, B., Falk, A., & Fliessbach, K. (2013). Effort increases sensitivity to reward and loss magnitude in the human brain. *Social Cognitive and Affective Neuroscience*, *9*(3), 342–349. doi: 10.1093/scan/nss147

Ma, Q., Meng, L., Wang, L., & Shen, Q. (2014). I endeavor to make it: Effort increases valuation of subsequent monetary reward. *Behavioural Brain Research*, *261*(3), 1–7. doi: 10.1016/j.bbr.2013.11.045

McGuire, J., & Botvinick, M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences*, *107*(17), 7922–7926. doi: 10.1073/pnas.0910662107

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*(3), 214–255. doi: 10.1037/0033-295X.86.3.214

Niese, Z., Libby, L., & Pfent, A. (2021). When the going gets tough, the committed get going: Preexisting goal commitment determines the consequences of experiencing regulatory nonfit. *Journal of Personality and Social Psychology*, *121*(3), 447–473. doi: 10.1037/pspa0000271

Palestro, J., Weichart, E., Sederberg, P., & Turner, B. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychonomic Bulletin & Review*, *25*, 1225–1248. doi: 10.3758/s13423-018-1479-9

Peirce, J. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*.

R Core Team. (2025). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

Smith, P., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. doi: 10.1016/j.tins.2004.01.006

Wang, L., Zheng, J., & Meng, L. (2017). Effort provides its own reward: Endeavors reinforce subjective expectation and evaluation of task performance. *Experimental Brain Research*, *235*, 1107–1118. doi: 10.1007/s00221-017-4873-z

Weichart, E., & Darby, K. (2026). Overcoming the costs of selective attention: Resolving rule-uncertainty supports acquisition of generalizable memory traces. *Cognition*, *272*(1), 106505. doi: 10.1016/j.cognition.2026.106505

Weichart, E., Galdo, M., Sloutsky, V., & Turner, B. (2022). As within, so without; as above, so below: Common mechanisms can support between- and within-trial category learning dynamics. *Psychological Review*, *129*(5), 1104–1143. doi: 10.1037/rev0000381

Westbrook, A., Kester, D., & Braver, T. (2013). What is the subjective cost of cognitive effort? load, trait, and aging effects revealed by economic preference. *PloS One*, *8*(7), e68210. doi: 10.1371/journal.pone.0068210