13. Conducting real effort experiments

David Hagmann and Luxi Shen

Economists are naturally interested in studying labor supply decisions: How many hours people choose to work, how intensely they work during those hours, and how they respond to incentives shape everything from business productivity to the revenue raised by income taxes. Given the importance of these questions to economics, researchers have long sought ways to study labor supply empirically. A particular challenge lies in designing experiments that can measure work in a controlled laboratory setting. While formal economic models abstract away from the physical and mental costs of labor, experimental economists must consider these features in their research design. This chapter introduces some of the widely used experimental tasks that put participants to work and highlights some of the features and trade-offs that need to be considered in their use.

The evolution of methods in the study of effort provision reflects a broader tension in experimental research: the trade-off between internal and external validity. Early laboratory studies relied on so-called stated effort tasks. In this paradigm, participants observe a production function that relies on a numeric input labeled "effort." Participants enter a value into a textbox reflecting how much effort they want to exert, incurring an economic cost for each unit of effort and some payoff based on the output. This presents an optimization problem that requires some mathematical skills and cognitive effort but nothing otherwise laborious (see Bull et al., 1987 for an early example). The advantage of such tasks is the clearly defined effort costs (the rate at which effort translates into outputs) that could be fixed across participants or manipulated as a treatment. The disadvantage, however, is a lack of realism: inputting a value into a textbox that reflects a monetary cost taken out of the experimental earnings has little in common with the work people normally do.

More recent experiments ask participants to complete tasks that require manual or cognitive effort, hence the name "real effort" tasks. These tasks generally do not require any special skills, such that anyone who is willing to put in the work can complete them. The aim of this chapter is to provide a practical guide for choosing from the large set of available tasks and for implementing them in an experimental context. We do not attempt to provide an exhaustive review of real effort tasks (see Charness et al., 2018, for a recent review). Moreover, we focus on tasks that can be implemented in the laboratory, omitting the highly realistic but context-specific tasks that can be used in the field (e.g., the speed at which pear packers pack pears, Chang and Gross, 2014, or how fast and how long bike messengers work to deliver packages, Fehr and Goette, 2007). Instead, we seek to illustrate different trade-offs that experimentalists make and how tasks can be adapted to answer illustrative questions of interest. With these tools, we hope that readers can pick and adapt an existing task for their research needs.

Some tasks we discuss blur the line between the laboratory and the field. This distinction notably highlights the key tension in real effort tasks: an experimenter desires realism to capture the economic and psychological factors determining people's willingness to exert effort.

However, the more realistic the task is, the less control the experimenter has over its features (such as how difficult – or "costly" – it is for participants) and the more challenging it is to implement. This fundamental trade-off between external and internal validity is not unique to real effort tasks and applies to experiments more broadly, but it may be especially important in the context of labor supply, given that a willingness to "work" is generally the topic of inquiry – and this willingness may be sensitive to any number of task features. The question then is whether results obtained from stylized experimental tasks conducted in the laboratory really are informative of people's labor supply decisions in the real world.

What makes a good real effort task? First, it should pose a similar challenge to all participants, so differences in output are best explained by experimental variations rather than individual differences. For example, asking American participants to translate German text into English would be a poor task because it would be easy for those fluent in both languages but impossible for those who are not, regardless of how they are incentivized to complete the task. Second, the number of tasks completed should be easily countable and comparable to provide an unambiguous outcome measure. This requires both clear task completion criteria and consistent difficulty across tasks and participants. Consider proofreading: one grammatical error might be obvious while another is subtle, making error detection an unreliable measure of effort. Finding six errors versus five may not represent 20% more work if the sixth error was particularly challenging to spot. Additionally, any scoring system would need to account for false positives where participants incorrectly flag text as errors, introducing additional complexity. Third, task difficulty should allow for meaningful variation in completion rates. Tasks that are too easy lead to ceiling effects, where participants hit an upper limit, while tasks that are too difficult create floor effects, where few tasks are completed at all. The ideal difficulty level allows participants' effort choices to meaningfully translate into different output levels. Fourth, participants should not be able to devise strategies that make a task easier (we will illustrate this in the matrix search task), unless that is explicitly of interest to the researchers.

A fifth feature is that the task should not be inherently enjoyable. As Dickinson (1999, footnote 20) notes: "A task that is enjoyable to the subjects would blur the line between labor and leisure and make interpretation of the results much more difficult. Subject comments on finishing the experiment assured me that I had succeeded in finding a 'no-leisure' task for them to perform." We similarly assure readers that the real effort tasks presented in this chapter pose no risk of being enjoyable to the participants. However, we note here (and return to this point again at the end of the chapter) that this may be an underappreciated but consequential deviation from real labor supply decisions. Work provides both meaning and enjoyment for many people (Cassar and Meier, 2018). The present methods, by design, explicitly omit this dimension.

The remainder of this chapter proceeds as follows. We begin by introducing seven example real effort tasks and describing their defining features and the challenges involved in using them. The first two are intended to take place in person and have high external (or face) validity. Participants are asked to engage in laborious tasks (stuffing envelopes for a mailing campaign and using a foot pump to inflate balloons), which resemble the kind of repetitive, manual work that characterizes some occupations. Notably, these tasks are effortful not only for the participants but also for the experimenters. The remaining five examples represent more abstract tasks. Although they are less realistic, they nonetheless capture the drudgery of a monotonous task (perhaps amplified by the transparent understanding that, unlike the

previous tasks, they serve no purpose but to test the participants' willingness to engage in the task). Their undeniable advantage, however, is that they are much more easily implemented by the experimenter and can even be conducted remotely over the internet.

After illustrating the seven real effort tasks in detail, we present some suggestions for when to use which one. Although this is likely driven by the question of interest, and an experimenter may have good reason for using a particular task (including tasks we do not discuss here), we believe each has features that give it an advantage for certain types of inquiries. For example, opportunities for "cheating" can be introduced (and detected) in some, which may be desirable for some research questions and not so for others. We then discuss incentivization and the decision of whether to introduce a time limit for completing the task. For researchers desiring to conduct experiments over the internet, we discuss the current policies of two widely used platforms (Amazon Mechanical Turk and Prolific) and how they are relevant to real effort tasks. Finally, we provide some closing reflections on the future of real effort tasks.

13.1 ILLUSTRATIONS OF REAL EFFORT TASKS

Example 1: Stuffing Envelopes

The "gold standard" for conducting a real effort task is to give people a realistic assignment, much like they may encounter in the real world, but that features the desirable properties discussed earlier: comparable difficulty across participants, no opportunities for simplifying strategies, an easily countable outcome with meaningful variation, and no inherent enjoyment. We illustrate this using an experiment by Hedegaard and Tyran (2018), which blurs the line between a laboratory and a field experiment. As we will see, this offers high external validity but comes at the expense of limited sample size and substantial effort by the experimenters.

In the experiment, secondary school students from Denmark helped with a mailing campaign in exchange for payment. Each participant was placed at a workstation by themselves, so that how much they worked was not influenced by how much others who happened to be engaged in the task at the same time were working. They were given 90 minutes to fold letters and place them into envelopes. The task was made more difficult in that they had to match each letter to a type of mailing and, in some cases, include a gift in the mailing. Moreover, they were required to look up additional information in a binder and sort the envelopes into different bins (see the online appendix of the paper for a series of photos showing this setup). While some might have worked as fast as they could to maximize the number of envelopes they stuffed, others may have procrastinated by using their phones. When the time was up, they were paid a fixed payment of approximately 50 Euro cents for each correctly processed mailing. Participants stuffed 107 envelopes on average.

Since this experiment was part of a real mailing campaign and participants did not realize they were part of a study, the task is very high on realism. It is also apparent what the challenges are: experimenters can only give instructions to one participant at a time, each participant requires their own private space (which may be difficult in smaller experimental laboratories), and determining the error rate for mailings requires additional effort on the part of the researcher. Thus, the time cost of conducting this experiment is extremely high. Moreover, the financial costs of conducting such an involved design are substantial, and the

sample size is thus inherently limited. The study includes 169 participants, whose payments for this task alone exceed 9,000 Euros. For comparison, participants in some of the online tasks we present later earn less than one Euro for their labor. Thus, it is likely not feasible with this design to have sufficiently large sample sizes to conduct between-subjects experiments (much less experiments with multiple treatments).

There is one more feature of this task that makes it particularly interesting: it can be completed by each participant on their own, as we have discussed so far, or by multiple participants working together. Indeed, the central research question of Hedegaard and Tyran (2018) is the choice of partner to work with on this task. After completing the first round, students were invited to return and informed of the first names of two potential partners, as well as how many envelopes they had stuffed in the first round. The authors find that students are willing to pay a small (but not a large) price to work with someone of their own ethnic group, as revealed through the first name. In addition to partner choice, this task may be adaptable to examine group dynamics, how participants split up tasks (i.e., if there are returns to specialization that they can capitalize on), the kind of conversations workers engage in (if any) during repetitive tasks, and similar questions for which group dynamics are relevant.

Example 2: Inflating Balloons

This second example similarly presents a laborious in-person task, but one that requires less time and allows for higher intensity output. Participants (students at Hong Kong universities) were brought into a laboratory and tasked with inflating balloons (Shen and Hirshman, 2023). They received individual instructions from a research assistant and engaged in a practice round prior to the start of the main experiment to ensure that they would be familiar with the process and the tools. Specifically, they had access to a foot-operated pump, which they would press down in order to inflate balloons. Each participant completed this task independently, but the research assistant remained in the room. Proper inflation required 50 presses of the pump, and they were asked to count out loud so that the research assistant could keep track of their progress. Participants received a large box of balloons so they would not run out of materials and had 15 minutes to inflate as many of them as they wanted to. The instructions also advised them that they would not have enough time to inflate all the balloons in the box and that they could take breaks freely if they wanted.

The number of properly inflated balloons defined the participants' effort. Each balloon was compensated with a piece-rate payment of HK$0.40 (approximately US$0.05). At the end of the period, the research assistant reported the total count to the participant, along with the corresponding payment. In the initial round, the payment was framed as HK$4 for every 10 balloons. The experimental treatments involved a wage change in a subsequent round, as well as whether the wage change was framed as a smaller/larger payment for the same number of balloons or the same payment for a greater/smaller number of balloons.

To recruit participants, the authors advertised a job opportunity on a mailing list used by multiple local universities and were able to recruit 132 participants. The task maintained some sense of realism, as participants believed they were there to complete a real job for payment, rather than participate in an experiment. Moreover, they experienced a task that required them to exert physical effort and overcome fatigue. Like the envelope-stuffing task, the balloon-inflation task presented similar logistical and sample size challenges. The authors resolve this

with a larger online experiment relying on a different design and replicate their findings from the laboratory.

Computerized Tasks

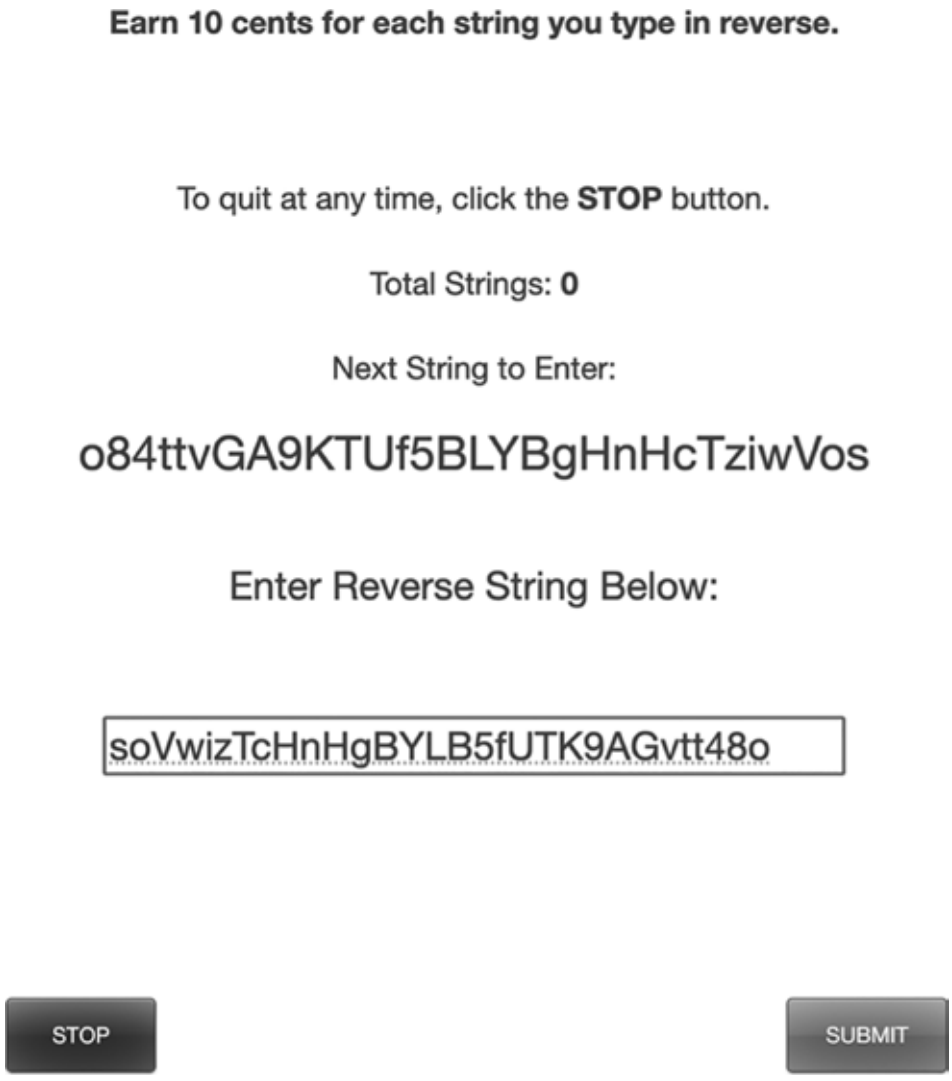
The two example tasks discussed so far clearly measure "effort" and capture realistic features of work, such as the boredom of a prolonged repetitive task (stuffing envelopes for 90 minutes) or the fatigue from high-intensity exertion of physical force (using a foot pump to inflate balloons). They also require substantial effort on the part of the researchers (and research assistants), who coordinate sequential data collection and manage the scheduling and flow of participants. In many cases, such in-person experiments are not feasible, creating the need for tasks that participants can complete in parallel or even over the internet.

As a result, most real effort tasks in experiments are more abstract, but can be implemented on computers without the direct involvement of a researcher or research assistant. While this comes at the cost of being more stylized, it does offer considerable advantages. First, effort data can easily be collected via existing experimental platforms, such as Qualtrics, z-Tree, oTree, or Empirica – or with custom-built web services, which can now be developed with the help of AI tools even by those with no programming experience. This also allows data to be collected remotely over the internet, using participant platforms like Amazon Mechanical Turk or Prolific. Second, participants' work tasks can automatically be checked for accuracy. This also means participants can get feedback while they are completing the task, rather than having tasks invalidated at the end of the session, if this is desired. Third, performance can be tracked without direct involvement by the experimenter and while preserving the participants' anonymity. This allows for large sample sizes that would not be feasible in person and may also reduce the pressure or anxiety that can arise when participants are directly observed. Fourth, and perhaps most important for a reader of this chapter, the tasks can easily be shared and reproduced by other experimenters. This makes it easier to replicate and extend existing research without having to design a novel task. To facilitate this process, we provide template Qualtrics files for each task described in this section via the Open Science Framework (https://osf.io/a9dgf). The repository also includes JavaScript code to showcase the tasks in a web browser and, for the Matrix and Counting tasks, code to generate additional stimuli. Note that tools like Claude Code and OpenAI Codex can quickly implement real effort tasks in HTML and JavaScript for inclusion in a survey (e.g., via Qualtrics), or develop and deploy stand-alone React applications. The capabilities of these tools are evolving quickly and are beyond the scope of this chapter.

Example 3: String Reversal Task

The string reversal task developed by Huck et al. (2015) exemplifies abstract, computerized tasks. Rather than aiming for realism, it offers participants a laborious assignment that transparently benefits no one. Instead, it offers a simple way of mapping effort into a payment. Figure 13.1 shows a version of this task taken from Hagmann et al. (2026).

In this task, participants see a randomly generated alphanumeric string (here of length 30) and are asked to enter it in reverse. After doing so, they click the "SUBMIT" button, and the string is validated. If it is reversed correctly, the "Total Strings" counter increments by one,



Note: In the *string reversal* task, participants type randomly generated alphanumeric strings in reverse in exchange for a piece-rate payment.

Figure 13.1 The string reversal task

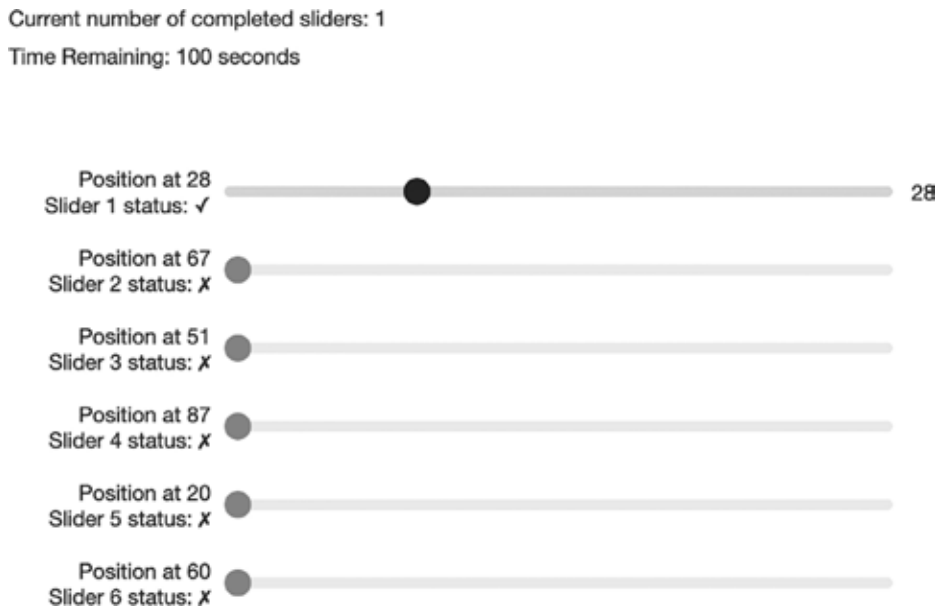
and a new random string is generated. If the participant makes an error, they are informed that their submission is incorrect and can revise their string (which remains in the textbox and is not deleted). Participants can try as often as they want to, and they do not face a time limit (we return to this design decision later in this chapter). Once they are done and do not wish to complete further tasks, they can click the "STOP" button to proceed to the remainder of the survey.

While this task is easily explainable to participants, researchers must be mindful to avoid ambiguous characters and guard against cheating. Specifically, uppercase 'O' and zero '0' and uppercase 'I' and lowercase 'l' are difficult to distinguish, and participants may quit early if they erroneously believe that the task is not working properly. Moreover, there are online tools available that can reverse strings. If copying and pasting text from the page is not disabled, the task becomes trivial to complete. With this in mind, participants appear to exert substantial effort on this task. In Hagmann et al. (2026), 400 participants recruited from Amazon Mechanical Turk reversed 16 strings on average, and 20% of participants reversed all 50 strings (the limit was not announced in advance).

The task can easily be adjusted depending on the research hypothesis. For example, difficulty can be manipulated by adjusting the length of the strings to be reversed. Moreover, the task could be completed with or without a time limit and with or without the option of quitting early. We return to some of these variations later in the chapter.

Example 4: Slider Task

The slider task introduced by Gill and Prowse (2012) relies on hand-eye coordination. Participants see a list of vertically aligned sliders, where the slider's position is coded to an integer response from 0 to 100, inclusive. To the left of each slider is a number corresponding to the target position to which the slider must be set (see Figure 13.2). Participants then use



Note: In the *slider task*, participants position a slider in a specific location by clicking and dragging their mouse. A checkmark indicates that a slider is positioned correctly.

Figure 13.2 The slider task

their mouse to drag each slider to the specified position and, when it is positioned correctly, receive a visual cue (a checkmark). The task becomes difficult because there is a time limit that is insufficient to complete all sliders, allowing for a measure of "effort" corresponding to the number of correctly positioned sliders.

An experimenter wishing to implement this task does not need to worry about cheating via copy-and-paste, as in the string reversal task, or people giving up early because of frustration with the validation check. However, for the task to be difficult, it is imperative to disable adjustment of the slider using the keyboard. Moreover, this task is less suited than others for implementation without a time limit: because the effort per slider is low, participants might position a large number of sliders without some constraint.

One note of caution: the slider task does not appear well-suited for testing responsiveness to financial incentives. Whether participants were paid half a cent for each correctly positioned slider or eight cents, their average performance was virtually unchanged (approximately 27 sliders completed in 2 minutes, Araujo et al., 2016). This may suggest that hand-eye coordination is not responsive to incentives, or it may be that people are already intrinsically motivated to do well. Alternatively, incentives might need to be much larger to motivate additional effort. In any case, for a range of plausible incentives, participants do not seem to be able to position sliders faster or more accurately. We return to the question of incentives in a later section.

Example 5: Button-Pressing Task

If the previous computerized tasks seem like they offer some variety or the opportunity of a "challenge" for participants that may provide some intrinsic reward for doing well, the button-pressing task offers the kind of repetitive task that quickly leads to boredom and fatigue (in our experience). DellaVigna and Pope (2018) asked participants to alternate between pressing the "a" and "b" keys, earning one point for each successful alternation. Figure 13.3 presents an adaptation of the task that includes a visual cue for which button needs to be pressed.

The authors assigned participants to one of 18 different incentive treatments to see how many pairs of button presses they performed within ten minutes. They observed a range from 1,521 points (when the number of button presses does not affect payment) to 2,188 points when participants received an 80-cent bonus for obtaining a score of at least 2,000 points. A number of additional treatments examined incentives informed by behavioral economic principles, such as loss aversion or time discounting.

To get a sense of participants' responsiveness to incentives on this task, it is informative to look at the subset of treatments in which they earned a bonus for every 100 points scored. There was a substantial increase from paying nothing to 1 cent per 100 points (1,521 to 2,029), but a much smaller increase for a payment of 4 cents (2,132) and 10 cents (2,175). This experiment illustrates the advantage of using a task that can be conducted over the internet. The authors recruited approximately 10,000 participants via Amazon Mechanical Turk, resulting in more than 550 participants per incentive treatment. A puzzling observation, however, is that even in the absence of an incentive to do the task, people still performed more than 1,500 pairs of a-b button presses. Moreover, a 10-fold increase in the piece-rate incentive increased the effort provided by only 7%.

Note: In the *a-b button-pressing task*, participants have 10 minutes to alternate between pressing the "a" and "b" keys on their keyboard. Each pair of key presses increases the score by one.

Figure 13.3 The button-pressing task

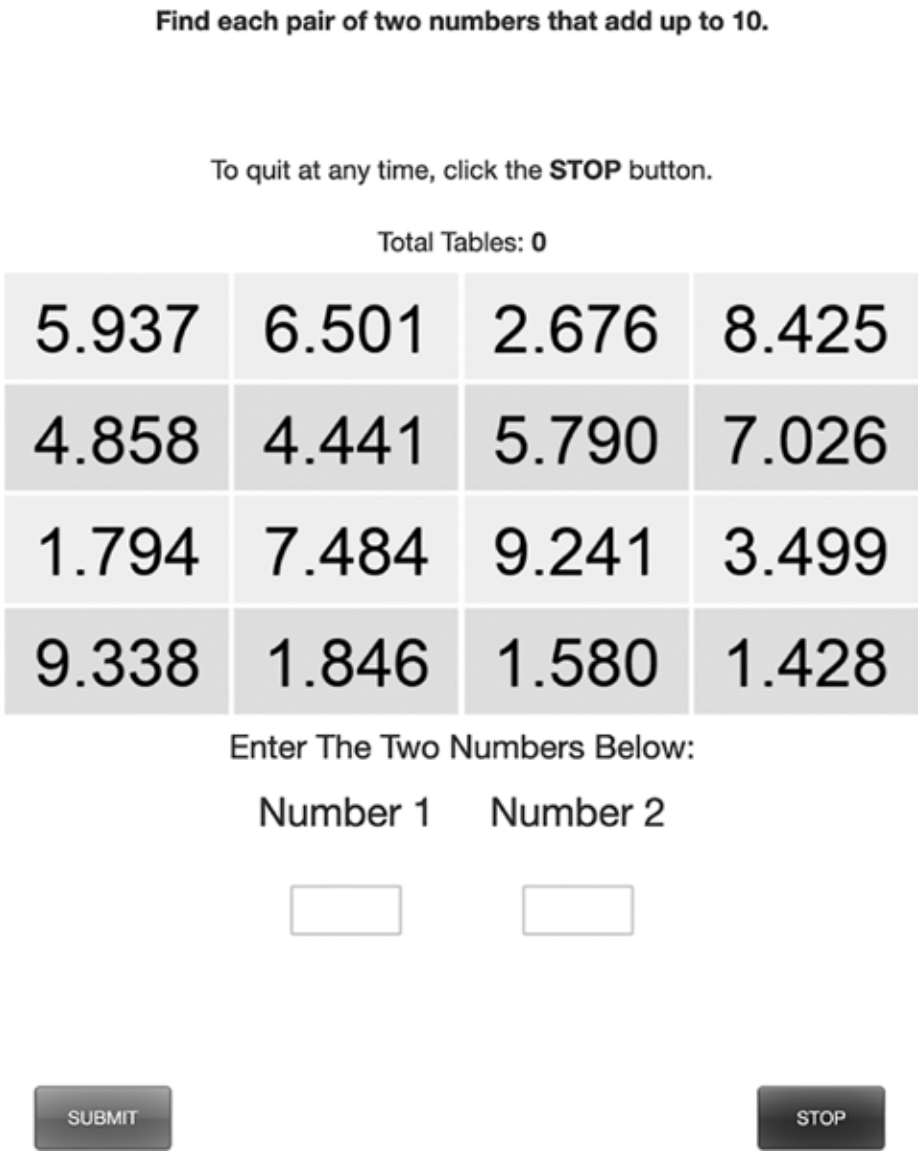
Example 6: Matrix Search

The *matrix search task* developed by Mazar et al. (2008) differs from the previous tasks in that it demands cognitive attention rather than physical endurance or precision. Each task consists of a grid of decimal numbers (see Figure 13.4 for an example). There is a unique pair of numbers adding up to exactly ten, and participants need to enter those two numbers into two textboxes. The simple additive operation required ensures that every participant can solve the tasks.

However, unlike the prior examples where participants cannot make use of "strategies" to simplify the tasks, here, a good search strategy can be rewarding. Specifically, someone moving through the grid methodically, e.g., checking the first and last digits of the first value against all others before moving on to the second number, may identify pairs faster than someone who engages in random comparisons. This may be a desired feature for some research questions, or an unwanted source of noise in others. Similar to physical fatigue from the previously described computerized tasks, participants anecdotally report getting tired from looking at number grids. Thus, the number of correctly solved grids appears to be a meaningful measure of the cognitive effort it takes to remain focused.

Unlike the previous three computerized tasks, the matrix task also lends itself to non-computerized experiments. The grids can be printed on paper and distributed to participants outside of a laboratory. Participants then report the two numbers for the table, and a research assistant can quickly determine whether the grid was solved correctly.

The task's difficulty can be manipulated in one of two ways. First, the size of the grid can be changed, and the task becomes more difficult as the number of cells increases. Second, one can add additional decimal digits to the numbers that need to be compared. Notably, the latter



Note: The *matrix search task* consists of a grid of decimal numbers. Participants have to find the unique pair of numbers that adds up to 10. In this example, the correct answer is 6.501 and 3.499.

Figure 13.4 The matrix search

does not necessarily make the task more difficult. A participant checking the first and last digits, rather than trying to add up each pair of numbers, may not require much additional effort with additional digits. Of course, this means that manipulating the number of decimals could

help identify participants who use such a search strategy and who would not be slowed down by this change. This could, in turn, be used as a proxy for other forms of strategic reasoning or sophistication. To our knowledge, performance on real effort tasks has not previously been used to identify types of participants.

Caution is warranted when using this task in online data collection. In our testing, we found that Claude Haiku 4.5 can correctly identify the solution pair from a screenshot with no further instructions in less than five seconds. As a result, participants may have a difficult-to-detect way of cheating. We believe this task is therefore best used for experiments conducted in the laboratory, where access to generative artificial intelligence can be prevented.

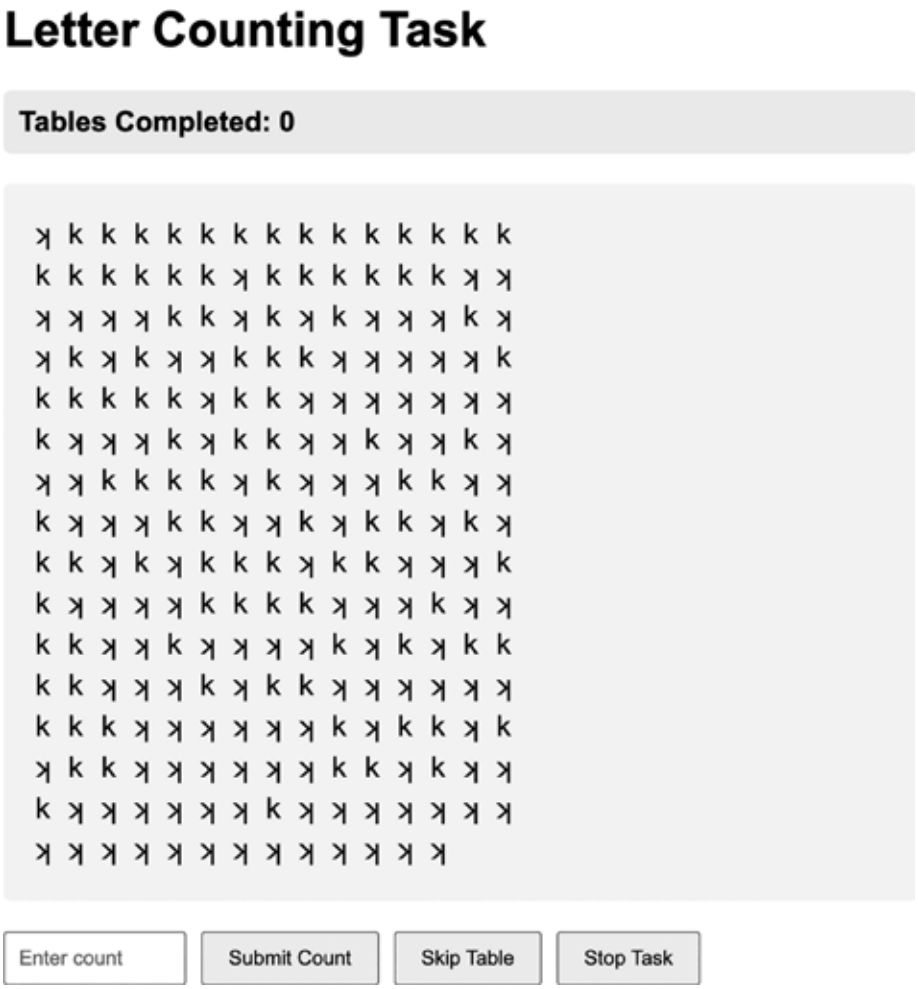
Example 7: Counting Tasks

For our last example, we present one version of a counting task developed by Saccardo and Permut (unpublished). Participants see a large grid consisting of between 235 and 245 letters "k" or upside-down rotated "k's" (see Figure 13.5). There are closely related versions, such as counting the occurrences of "0's" in a large grid consisting of 150 0's and 1's (Abeler et al., 2011).

Similar to the matrix task, this grid can be generated in advance and printed on paper for an in-person experiment, or shown as an image in a Qualtrics survey. However, the counting task differs in two notable dimensions from the matrix task. First, there is no optimal search strategy here. Participants unavoidably have to count the number of instances they see in the grid since they are not rewarded for guessing and are unlikely to arrive at the correct number by chance. Second, no addition is required, perhaps reducing cognitive fatigue relative to the matrix task. Conversely, participants may experience frustration if their initial submission is wrong and they must start counting again.

We picked the upside-down "k" counting task because of an interesting variation specific to the computerized version, and at the core of the study conducted by Saccardo and Permut. The grid and the letters were included as text on the survey page rather than embedded as an image. While we would normally recommend using an image to avoid "cheating" via a browser's search function, this study was designed precisely to test whether participants would make use of disallowed shortcuts. Specifically, participants were instructed to manually count the "k's" in the table and not use any automatic means of counting. Since the grid was text, however, it would be tempting for participants to use the search function in the browser (e.g., CTRL + F or COMMAND + F on Windows and Mac computers, respectively) to save themselves the effort and frustration of counting and arrive at an immediate count.

To test if participants engaged in such cheating, the authors embedded additional letters "k" in a hidden table that would not display to participants who faithfully counted the instances in the table. They were, however, included in the count generated by the browser search function. Thus, there were three potential responses: participants could report an accurate count, make a mistake and report an erroneous count (which would lead to an error message), or precisely report the count resulting from using the disallowed search shortcut (in which case the survey also proceeded to the next task). Thus, the researchers could unobtrusively measure how often participants "cheated." Adding sufficiently (but plausibly) many hidden "k's" ensures that the "cheating" count is distinct from a count arising from an honest mistake.



Note: In the *counting task*, participants report how many instances of a right-facing letter "k" they observe in the large grid. The task includes a random number of hidden "k's" embedded in the HTML code. These are counted when participants use the search function in their browser, which they were instructed not to do.

Figure 13.5 The counting task

We note that here, too, there is a concern that generative artificial intelligence can provide the correct answer. Importantly for the detection of cheating via the disallowed search function, the count from the screenshot does not include the hidden "k's" that would indicate cheating. Thus, the response would look as if someone faithfully counted the letters. If this task is used, it may be necessary to preregister a precise exclusion rule for participants who submit unusually quickly.

13.2 CREATING PERFORMANCE DATA WITH REAL EFFORT TASKS

In some cases, an experimenter may not be directly interested in participants' effort provision. For example, Hagmann et al. (2026) recruited people on Amazon Mechanical Turk to complete the string reversal task (as presented in Example 3). The survey did not include an experimental manipulation: participants received an identical fixed payment for each string they reversed and faced no time constraints. The purpose was to generate real performance data, determined by the number of strings successfully reversed before deciding to stop (or completing all 50 tasks).

The participants who completed the string reversal task then became "job candidates" for other participants in a series of main experiments. Those new participants observed demographic information (along with other information that differed by treatment) and selected those they thought completed more tasks. The lack of a time constraint made the prediction task more challenging: participants had to form opinions not only about who could reverse strings more quickly, but also who would be willing to do so for longer.

When experimenters want to avoid stereotypes about task performance, the arbitrary real effort tasks described here may be particularly suitable. Participants are unlikely to believe any particular social group to be better at typing strings in reverse or alternating between button presses. Conversely, performance tasks that rely on knowledge or ability, rather than effort, have been used to examine the role of stereotypes (Coffman et al., 2021).

Are results obtained in real effort experiments sensitive to the chosen task? We are unaware of experiments designed to compare performance or treatment effects across different tasks. However, in the next section, we report previously unpublished data that establish the robustness of a result in three real effort tasks: the string reversal and matrix tasks discussed in this chapter, and a Greek transcription task that we will introduce briefly.

13.3 RESPONSIVENESS TO INCENTIVES AND TIME CONSTRAINTS

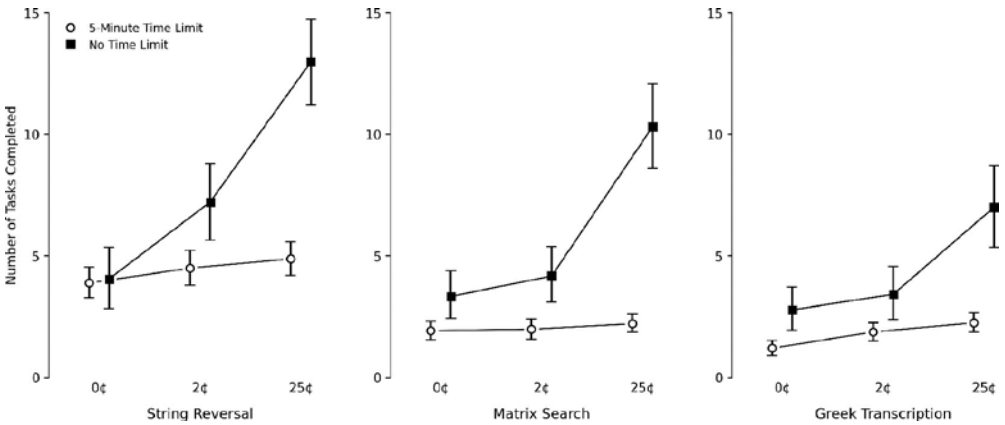
Having picked an appropriate task, the next question is how to incentivize participants. Many real effort experiments explicitly seek to test how people respond to various financial or behavioral incentives. We have previously mentioned Araujo et al. (2016) in the context of the slider task, finding no significant difference in performance when increasing the per-slider incentive from 0.5 cents to 8 cents. Similarly, we noted that the incentive effects of DellaVigna and Pope (2018), in which people alternated between pressing "a" and "b" on their keyboards, were limited (conditional on paying something). Of course, the two experiments are not directly comparable: the former asked students in the laboratory to position sliders over multiple two-minute rounds, whereas the latter asked Amazon Mechanical Turk participants to press buttons during a single ten-minute round. Nonetheless, a plausible inference is that participants respond to incentives to some degree but sustain their "maximum effort" for at least 10 minutes even with low payments.

Here, we report unpublished data from an experiment in which participants faced three different experimental tasks, three different incentive levels, and the presence or absence of

a time constraint. A total of 1,788 participants from Amazon Mechanical Turk completed a study in which they were randomly assigned to different real effort tasks: String Reversal (Example 3), Matrix Search (Example 6), and a Greek Transcription task that involved transcribing blurry Greek letters by clicking on corresponding buttons (see Augenblick et al., 2015). For each task participants completed, they were either paid no additional incentive, two cents, or 25 cents, in addition to a 20-cent fixed payment. Moreover, half the participants were given at most five minutes to work on the task (with a countdown showing the remaining time), whereas the other half could work as long as they wanted (up to a maximum of 25 tasks, which we did not announce in advance). In all cases, they could click on a "STOP" button to immediately advance to the demographic questionnaire and complete the survey. They would be paid their earnings, and clicking stop early (including before completing a single task) would not affect their fixed payment. To avoid issues with selection, where participants might drop out after learning about the low bonus payment, we included participants who dropped out of the study after the screen that explained their task and bonus payment. These participants (n = 215) returned the task to the platform and were not counted toward the completed-submission target but were included in the analysis, yielding an analytic sample of 2,003.[1]

These tasks differed in their difficulty. Looking at the treatments with a five-minute time constraint, the most productive workers, across all wage conditions, reversed 15 strings, found 9 pairs of numbers that added up to ten, and transcribed 7 strings of Greek letters. Average performance was considerably lower: 4.42 strings, 2.04 pairs of numbers, and 1.76 Greek strings. Of course, the intent of this experiment was to examine responsiveness to incentives as a function of time constraints.

As is evident from Figure 13.6, which depicts the average number of tasks completed across the experimental treatments, we see much greater responsiveness to incentives when



Note: Mean number of tasks completed in three real effort tasks (String Reversal, Matrix Search, Greek Transcription), by piece-rate wage (0¢, 2¢, or 25¢ per task) and time condition (5-minute limit vs. no limit). Error bars are 95% bootstrap percentile confidence intervals (10,000 resamples).

Figure 13.6 Response to incentives

participants are not constrained by a time limit. In the string reversal task, for example, effort with a time limit ranged from 3.88 with no piece-rate incentive to 4.88 with a 25-cent-per-reversed-string payment ($t(221) = 2.10,\ p = .037$). In the absence of a time constraint, participants still completed 4.05 tasks when receiving no piece-rate payment, and 13.0 tasks with a 25-cent payment ($t(221) = 8.02,\ p < .001$).

In Table 13.1, we show OLS regressions for each of the tasks, both with and without a time constraint. For each regression, the baseline condition is the 2-cent low payment treatment. With a five-minute time limit (the first three columns), participants complete the same number of tasks regardless of whether they are paid two cents or 25 cents. In the Greek task, performance is slightly lower in the absence of an incentive; in the other two tasks, a small incentive does not increase effort relative to no incentive. Turning our attention to the treatments without time constraints (columns 4–6), participants complete significantly more tasks with a high incentive than with a low incentive. For two of the three tasks, however, effort does not differ between no incentive and the low incentive; only in the string reversal task does the small incentive lead to higher performance.

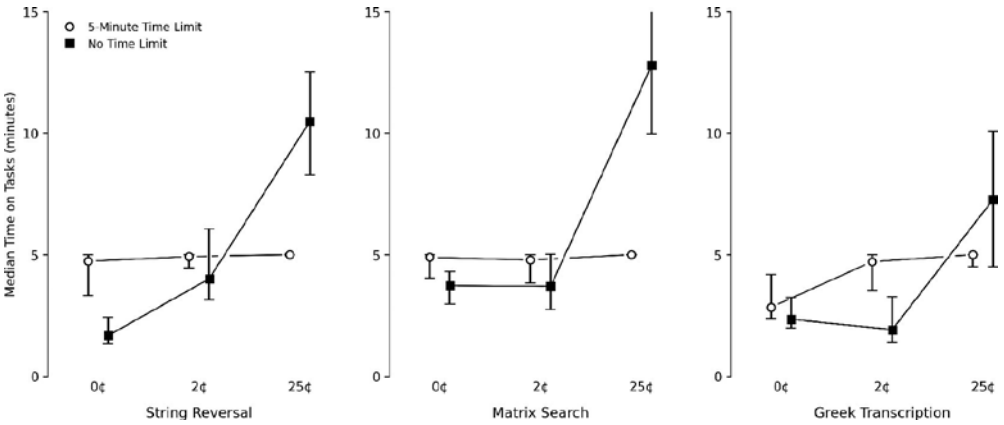
We next turn our focus to *productivity*, which we define as the number of tasks completed per minute on the task. We collapse the observations across the time constraint dimension and exclude participants who did not advance past the instruction screen or who completed no tasks. Participants in the low incentive treatment complete on average 1.24 string reversal tasks, 0.76 matrix search tasks, and 0.77 Greek transcription tasks per minute. Productivity does not differ across incentive levels, with the absolute value of coefficients ranging from 0.00 to 0.06. That is, participants do not seem to be able to complete the tasks any faster when they are offered higher pay.

This suggests that the effects of a higher incentive come not from the intensity with which people exert effort, but from how long they are willing to spend on the task. Recall that participants in all treatments could click a "STOP" button to advance to the demographic questions and conclude the survey. In Figure 13.7, we plot the median amount of time participants chose to spend working on the task. Looking at the treatment with a five-minute time constraint, we

Table 13.1 OLS regressions for the number of tasks completed

	5-Minute Time Limit **String Reversal**	5-Minute Time Limit **Matrix Search**	5-Minute Time Limit **Greek Transcription**	No Time Limit **String Reversal**	No Time Limit **Matrix Search**	No Time Limit **Greek Transcription**
No Incentive	-0.607	-0.063	-0.665**	-3.157**	-0.836	-0.640
	(0.491)	(0.282)	(0.256)	(1.128)	(0.973)	(0.903)
High Incentive	0.392	0.232	0.386	5.771***	6.137***	3.579***
	(0.492)	(0.282)	(0.263)	(1.126)	(0.973)	(0.920)
Constant	4.491***	1.982***	1.868***	7.211***	4.170***	3.421***
	(0.348)	(0.199)	(0.180)	(0.801)	(0.691)	(0.638)
N	333	335	329	332	340	334

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Responsiveness to receiving a high incentive (25 cents per completed task) or no incentive, compared to a low incentive (2 cents) baseline.



Note: Median time spent on each task, by piece-rate wage and time condition. Error bars are 95% bootstrap percentile confidence intervals (10,000 resamples).

Figure 13.7 Time spent on tasks

observe that the median amount of time spent is about five minutes in all but one treatment (the Greek transcription task with no incentive). That is, most participants appear to spend all available time working on these tasks, regardless of their compensation.

Next, we examine the treatments without a time constraint. Participants across all tasks spent more time working for a high incentive than for a low incentive. The low incentive, however, only led to additional effort for the string reversal task, relative to no incentive. Perhaps most interesting is the comparison between the treatments with and without a time constraint and low or no piece-rate incentives. Here, participants chose to stop sooner if they did not have a time constraint than if they had at most five minutes. Note that this does not hold for the *average* number of minutes on the task, where the treatments without a time constraint have outliers who spend more than 20 minutes even without a piece-rate incentive.

In Table 13.2, we specifically look at the marginal impact of a high incentive (vs. a low incentive) on effort across all three tasks. The left column shows an OLS regression for the number of tasks completed, with the low incentive string reversal task with a time constraint as the baseline treatment. The "High Incentive" coefficient shows that participants did not reverse more strings in the presence of a time limit when they were paid 25 cents vs. 2 cents per completed task. The same result holds for the other two tasks, as reflected by the non-significant interactions between tasks and high incentives. For the treatments without a time constraint, we observe that all interactions between the tasks and high incentives are significant. That is, participants completed more tasks (exerted more effort) when incentives were high. In the right column, the dependent variable is the amount of time (in minutes) spent working on the task. This measure replicates the findings from the number of tasks completed: participants spend more time on the task when incentives are high and they have no time constraint, but not when they are limited to working for at most five minutes.

It appears that people are willing to sustain some effort and complete a non-zero number of tasks even when they are not being paid at all. Larger incentives do not increase the intensity

Table 13.2 The marginal impact of a high incentive vs. a low incentive on effort across all three tasks

	Tasks Completed	Minutes on Task
High Incentive	0.392	0.446
	(0.831)	(0.924)
Matrix (timed)	-2.509**	-0.111
	(0.829)	(0.945)
Greek (timed)	-2.622**	-0.194
	(0.826)	(0.933)
Greek (untimed)	-1.070	2.062*
	(0.826)	(0.937)
Matrix (untimed)	-0.321	2.908**
	(0.829)	(0.945)
Reverse (untimed)	2.720**	4.147***
	(0.835)	(0.937)
High Incentive × Matrix (timed)	-0.160	0.255
	(1.171)	(1.338)
High Incentive × Greek (timed)	-0.005	0.109
	(1.183)	(1.337)
High Incentive × Greek (untimed)	3.187**	5.798***
	(1.177)	(1.346)
High Incentive × Matrix (untimed)	5.745***	8.749***
	(1.169)	(1.334)
High Incentive × Reverse (untimed)	5.379***	3.369*
	(1.175)	(1.318)
Constant	4.491***	3.613***
	(0.589)	(0.655)
N	1328	1218

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Across all three tasks, participants who had a time constraint were not responsive to an increase in the piece-rate wage from 2 cents to 25 cents. Participants without a time constraint completed more tasks when payment was higher, across all three tasks. The time on task analysis has fewer observations because only those who completed this stage are included.

of effort, as measured by the number of tasks completed per minute. However, higher payments motivate people to work longer (consistent with field evidence from Fehr and Goette, 2007). These results suggest that experimenters may need to be careful about imposing time limits when the focus of the study is people's responsiveness to incentives.

13.4 CHOOSING A TASK AND AN ONLINE PLATFORM

Given the many tasks we have introduced here, a natural question is how to choose from among them. The first question is whether the experiment can be conducted in person or will take place online. If the experiment is *in person*, another key question is whether participants will complete the task individually or in the same room as others. If participants are recruited *sequentially*, we believe the balloon pumping task (Example 2) balances realism and feasibility well. Notably, however, this remains a very labor-intensive task and depends on access to research assistants who can monitor progress. If participants can complete the task *in the same room as others*, we believe a paper version of the matrix task (Example 6) is a good solution. Cheating concerns do not apply in this version, and a research assistant can easily verify the solutions at the end of the task. While the task is more arbitrary, it captures many dimensions of "effort" that require overcoming mind-wandering and remaining engaged in a task that may not be inherently interesting.

If the experiment is conducted *online*, we believe that the string reversal task (Example 3) offers a good balance of requiring effort while remaining easy to implement and without opportunities for cheating (as long as copy-paste is disabled). Moreover, our experiment found a strong responsiveness to the magnitude of the piece-rate incentive in the absence of a time constraint. The task's difficulty can also be easily manipulated by changing the length of the string. The slider task (Example 4) may work well when participants experience pressure to act quickly and when responding to different incentive amounts is not central to the research. If including people on mobile devices is important, the matrix task (Example 6) may be the best option despite the possibility of cheating. Transcribing long strings and precisely positioning sliders are poorly suited to mobile devices. Alternatively, the button-pressing task could likely be adapted to a touch-screen interface, with participants asked to alternately tap buttons on their mobile device. Finally, experiments that require an unobtrusive measure of cheating may find the counting task (Example 7) most appropriate. Even after completing the study, participants may remain unaware that they have been "caught" and thus fail to adjust their behavior when facing similar tasks.

Many researchers want to conduct their experiments via online labor platforms like *Amazon Mechanical Turk* (*MTurk*) or *Prolific*. The key advantage is that they offer sample sizes that exceed what can be obtained with in-person laboratory experiments (e.g., the 10,000 participants recruited by DellaVigna and Pope, 2018). Such large sample sizes permit well-powered tests of multiple treatments at a feasible cost. Thus, even a certain attrition rate or additional noise due to non-compliance may be more than offset by the larger sample size afforded relative to in-person experiments. Moreover, online participants represent a more diverse population than undergraduate students (Buhrmester et al., 2011), not to mention that the latter may not be the most conscientious research participants either.

We begin with some general advice for conducting real effort tasks on online platforms, particularly because we anticipate that new entrants may emerge and platform policies are constantly evolving. First, researchers may need to anticipate some dropout due to technical problems. In the laboratory, experimenters have full control over the software used for the task and thus can exhaustively test the task. When people participate remotely, however, any number of issues with outdated browsers or extensions may interfere with the experimental task (e.g., some privacy-focused extensions block JavaScript that many tasks rely on). In the

event of inevitable technical issues, we recommend making the full payment and being lenient with approval decisions.

Platform policies regarding minimum payments may also present challenges. At present, the policies of *Prolific* make it less suited for real effort experiments than *MTurk*. First, *Prolific* imposes a minimum participant payment that does not consider potential bonus payments. Thus, recruiting participants requires a (relatively) large fixed payment, which could reduce the marginal incentive for completing tasks during the experiment. *MTurk,* in contrast, requires no minimum payment, allowing for a low show-up fee combined with a (relatively) high bonus incentive.

For experimenters considering *untimed* effort tasks, *Prolific* poses a second challenge. Each study needs to be advertised for a specific time (with a fixed payment appropriate for this amount of time). Tasks automatically expire if participants do not submit them within a "reasonable" time that is based on the originally estimated study length. Thus, if participants could spend anywhere from one minute to one hour on a task, the study would have to be announced for one hour and provide a large base payment. As a result, the marginal incentive for working longer may be very small, and participants may be motivated to terminate the study immediately. Notably, however, we are not aware of any studies explicitly testing the effect of the magnitude of a fixed show-up fee on willingness to provide effort for a piece-rate wage.

13.5 CONCLUDING THOUGHTS

We introduced this chapter with features of a good experimental task, and a cautionary note by Dickinson (1999): "A task that is enjoyable to the subjects would blur the line between labor and leisure and make interpretation of the results much more difficult." This grim view of labor supply permeates the experimental literature on real effort tasks, and indeed all the examples we have listed here – as is readily apparent to anyone who has tested these tasks themselves. Of course, removing any sense of intrinsic motivation is sensible for many research questions.

However, we should not lose sight of what we are trying to study: people's decision to put effort into real jobs. As we noted, people derive meaning and enjoyment from their work (Cassar and Meier, 2018). Ironically, the same economists who may enthusiastically work late into the night on their research – deriving immense satisfaction from doing so – deliberately create unenjoyable tasks to study their participants' choice of how much to work. Indeed, work is also an important venue for social connections, not merely an endless repetition of disconnected tasks. While the real effort tasks presented here have *deliberately* removed enjoyment, we may have lost something that is essential to effort provision rather than an unwanted confound. What features of a task make it meaningful and rewarding, and how do we capture the social features of work in the laboratory?

We reported perhaps surprising findings that participants worked on precisely these boring tasks for five minutes, even when they were not paid for completing them and could stop at any time. While they did not respond to financial incentives when facing a time constraint, they were quick to quit without this constraint. It may be that people inherently enjoy a challenge, which was present with a time constraint but not in its absence. Even when the task itself is not fun, *doing well* on the task may be enjoyable.

Finally, the real effort tasks presented here largely emulate manual labor, and the criteria we use to evaluate them reflect that focus. What might a real effort task designed to capture modern knowledge work look like? While generative artificial intelligence has been featured in this chapter as a problem for some experimental tasks, workers in real jobs are increasingly relying on large language models to automate some of their work. Future tasks may deliberately build in opportunities for participants to leverage outside tools, allowing for creative ways that lessen the effort required by a task. Rather than a problem that introduces noise, this may open up new avenues of inquiry.

NOTE

1. The study was preregistered on AsPredicted.org (https://aspredicted.org/w689-88md.pdf) and conducted by Hagmann and Rees-Jones. Data and analysis code are available via OSF: https://osf.io/ay7tx/

REFERENCES

Abeler, J., Falk, A., Goette, L., & Huffman, D. 2011. "Reference Points and Effort Provision." *American Economic Review*, 101 (2), 470–492. https://doi.org/10.1257/aer.101.2.470

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R. et al. 2016. "The Slider Task: An Example of Restricted Inference on Incentive Effects." *Journal of the Economic Science Association*, 2 (1), 1–12. https://doi.org/10.1007/s40881-016-0025-7

Augenblick, N., Niederle, M., & Sprenger, C. 2015. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks." *The Quarterly Journal of Economics*, 130 (3), 1067–1115. https://doi.org/10.1093/qje/qjv020

Buhrmester, M., Kwang, T., & Gosling, S. D. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science*, 6 (1), 3–5. https://doi.org/10.1177/1745691610393980

Bull, C., Schotter, A., & Weigelt, K. 1987. "Tournaments and Piece Rates: An Experimental Study." *Journal of Political Economy*, 95 (1), 1–33. https://doi.org/10.1086/261439

Cassar, L., & Meier, S. 2018. "Nonmonetary Incentives and the Implications of Work as a Source of Meaning." *Journal of Economic Perspectives,* 32 (3), 215–238. https://doi.org/10.1257/jep.32.3.215

Chang, T., & Gross, T. 2014. "How Many Pears Would a Pear Packer Pack if a Pear Packer Could Pack Pears at Quasi-Exogenously Varying Piece Rates?" *Journal of Economic Behavior & Organization*, 99, 1–17. https://doi.org/10.1016/j.jebo.2013.11.001

Charness, G., Gneezy, U., & Henderson, A. 2018. "Experimental Methods: Measuring Effort in Economics Experiments." *Journal of Economic Behavior & Organization*, 149, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Coffman, K. B., Exley, C. L., & Niederle, M. 2021. "The Role of Beliefs in Driving Gender Discrimination." *Management Science*, 67 (6), 3551–3569. https://doi.org/10.1287/mnsc.2020.3660

DellaVigna, S., & Pope, D. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies*, 85 (2), 1029–1069. https://doi.org/10.1093/restud/rdx033

Dickinson, D. L. 1999. "An Experimental Examination of Labor Supply and Work Intensities." *Journal of Labor Economics*, 17 (4), 638–670. https://doi.org/10.1086/209934

Fehr, E., & Goette, L. 2007. "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review*, 97 (1), 298–317. https://doi.org/10.1257/aer.97.1.298

Gill, D., & Prowse, V. 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review*, 102 (1), 469–503. https://doi.org/10.1257/aer.102.1.469

Hagmann, D., Sajons, G., & Tinsley, C. 2026. "Base Rate Neglect as a Source of Inaccurate Statistical Discrimination." *Management Science*, 72 (6), 4974–4999. https://doi.org/10.1287/mnsc.2023.00603

Hedegaard, M., & Tyran, J.-R. 2018. "The Price of Prejudice." *American Economic Journal: Applied Economics*, 10 (1), 40–63. https://doi.org/10.1257/app.20150241

Huck, S., Szech, N., & Wenner, L. M. 2015. "More Effort with Less Pay: On Information Avoidance, Belief Design and Performance." CESifo Working Paper No. 5542, Center for Economic Studies and ifo Institute (CESifo), Munich. https://www.econstor.eu/handle/10419/123172

Mazar, N., Amir, O., & Ariely, D. 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research*, 45 (6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Saccardo, S., & Permut, S. "Working for Good: Incentives, Effort and Cheating." Unpublished manuscript.

Shen, L., & Hirshman, S. D. 2023. "As Wages Increase, Do People Work More or Less? A Wage Frame Effect." *Management Science*, 69 (8), 4721–4732. https://doi.org/10.1287/mnsc.2022.4591