Experimental Economics (2024) 27:1001–1032
https://doi.org/10.1007/s10683-024-09844-y

Is "real" effort more real?

E. Glenn Dutcher[1,2] · **Timothy C. Salmon**[3] · **Krista J. Saral**[1,4]

Received: 22 June 2021 / Revised: 4 September 2024 / Accepted: 6 September 2024 /
Published online: 27 September 2024
© The Author(s) 2024

Abstract

A growing number of studies use "real" effort designs for laboratory experiments where subjects complete an actual task to exert effort rather than using a stylized effort design where subjects simply choose an effort level from a predefined set. The commonly argued reason for real effort is that it makes the results more generalizable and field relevant. We investigate the nature of modeling effort provision by first trying to provide a clear theoretical understanding of the nature of effort costs. We then empirically examine claims about the differences between real effort and stylized effort. A key to our examination is ensuring that we compare the two modes of effort provision keeping effort costs constant, which is a point overlooked in many past examinations. In our data, when controlling for effort costs, we find no differences in behavior between real and stylized effort. Given the importance of effort costs and the lack of a generally accepted way to include them in real effort designs, we provide a simple add-on that any researcher can use with their real effort experiments to incorporate a theoretically appropriate and controlled cost of effort even in a real effort setting. We also discuss ways to better approach modeling effort costs in experiments, whether one is conducting real or stylized designs, to improve inference on research questions.

Keywords Real effort · Stylized effort · Abstract effort · Economics experiments · Public goods · Slider task · Incentives · Coordination games

✉ E. Glenn Dutcher
glenn.dutcher@charlotte.edu

Timothy C. Salmon
tsalmon@smu.edu

Krista J. Saral
ksaral@charlotte.edu

[1] University of North Carolina Charlotte, Charlotte, NC, USA

[2] Ohio University, Athens, OH, USA

[3] Southern Methodist University, Dallas, TX, USA

[4] GATE Lyon Saint-Etienne, Lyon, France

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

JEL Classification C91 · H41

1 Introduction

Many types of economic experiments involve having subjects put forth some form of "effort." While effort has been proxied in different ways, the approaches are generally classified into two categories: stylized (or chosen effort) and real effort. Stylized effort involves subjects choosing a number from a prespecified range to represent their effort level, with higher numbers being more financially costly than lower ones. Real effort designs include a wide range of options in which subjects perform some task as their form of effort, which could be solving mathematics problems, solving mazes or other puzzles, stuffing envelopes, or even in one case shelling walnuts (Fahr and Irlenbusch 2000).[1] The choice of whether to use real or stylized effort is potentially important as it could have substantial implications for the outcome of the experiment. In this paper, we examine the foundations of modeling effort provision in experiments to better understand the important issues in the choice of whether and how to implement stylized or real effort in experiments.

Both approaches to representing effort in the laboratory are attempts to replicate important elements of effort decisions in field settings. Stylized effort is arguably better suited to capture theoretically derived trade-offs present in field effort choices as any trade-off that one believes to be important to a field setting can be specified in a theoretical structure and implemented in a stylized design. There are, of course, potential limitations to stylized designs as they may omit elements of effort provision that are important in the field but not (explicitly) contained in the corresponding theoretical model; for example, contextual issues that are important in field settings may be difficult to model a priori.

The primary argument in favor of the real effort approach is that it involves subjects engaging in actual physical or mental effort to complete a task, as one would do in a field setting. This leads to a belief that real effort better captures trade-offs present in field activity. One can find many claims in the related literature that real effort studies are inherently more generalizable than stylized designs,[2] and while these claims are common, the basis for them is unclear as we can find neither theoretical nor empirical evidence to support them.

In order to empirically examine claims regarding the superiority of real effort designs in replicating a field environment, two things must be demonstrated. First,

[1] See Carpenter and Huet-Vaughn (2019) for a more comprehensive list of different real effort designs.

[2] Examples include this quotation from Gill and Prowse (2011): "The main advantage of using a real effort task over a monetary cost function is the greater external validity of the experiment: exerting actual effort makes the environment more realistic and less sterile, increasing the likelihood that the motivations that drive behavior outside the laboratory carry over to the laboratory." (p.1) This point is reinforced in a recent survey regarding real effort experiments, Charness et al. (2018), in which the authors point out that "The advantage of the real-effort method is that it is closer to the psychology of working. For example, the cost of effort might vary over time: solving mazes might be fun initially but might gradually become less motivating" (p. 75).

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

given comparable real and stylized effort decisions, people make different choices. Second, the choices made using the real effort version are a better match with field behavior. While we are not aware of studies addressing the second, its relevance is conditional on the first being true. We have found several studies trying to address the first issue, but most fail to hold effort costs constant between the two designs. In the stylized designs, there is an explicit monetary cost of choosing higher effort levels. In real effort designs, implicit effort costs are typically "uncontrolled."[3] It is not possible to compare decisions between the two environments if effort costs are not held constant. Doing so requires a better understanding of the nature of effort costs and how they could be implemented/measured in a real effort setting.

To that end, we first present a theoretical characterization of a standard labor supply model to demonstrate what an effort cost is and where it comes from. It is commonly understood by economists that all economic costs are opportunity costs as is taught in Principles courses.[4] While at some level we all know this to be true, an explicit implementation of this insight is often lacking in real effort experiments.

In many real effort experiments, subjects are given the opportunity to engage in some task with their output labeled as "effort" and they are given no alternative activity to engage in; the only thing a subject must give up to generate output is sitting idly. Existing studies show that subjects in experiments would rather experience painful electric shocks than sit idly for a few minutes (Wilson et al., 2014), indicating that the outside option allowed in these designs likely generates negative utility, or a positive cost. While this may seem an odd result, some real effort designs have unintentionally replicated this finding, e.g. Esarey et al. (2012) and Ku and Salmon (2012) accidentally showed that increasing the wages of subjects had no impact on output. Araujo et al. (2016) replicate the finding intentionally as they examine a commonly used real effort mechanism and demonstrate that increasing the piece rate wage 1600%, from \$0.005 per unit of output to \$0.08, has no impact on output. The authors conclude that there is a problem with the specific real effort task they implement but we will argue that the task itself may not actually be the problem. The lack of a wage effect is explainable by the fact that there is effectively no cost of providing effort due to the lack of an alternate activity that generates non-negative utility. Thus, the lowest wage offered compensates the subjects adequately for spending the entire time allotment on the task, and higher wages can induce no greater effort.

Other prior papers have noted some of these issues. For example, in a working paper version of Araujo et al. (2016), Araujo et al. (2015) provide a more thorough bibliography of papers on the slider task and find that when wages are varied *between subjects*, it is common to find that different wages yield equivalent levels of effort. On the other hand, studies that vary the wages *within subject* can find that higher wages induce subjects to engage in more slider tasks. This inconsistency is important and one we will discuss in more detail in our conclusion. Erkal et al.

[3] A notable exception to this is Gächter et al. (2016) which adds an explicit cost to the units of effort in a real effort design with the intent to implement a specific cost of effort in a real effort experiment.

[4] Mankiw (2017), as simply one example of a prominent textbook, sets the following up as the core second principle to economics noting that "The cost of something is what you give up to get it (p. 5)".

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

(2018) and Georg et al. (2019) examine how effort provision changes in a real effort setting under different outside options. Each study has a baseline in which subjects can either perform the task or not. They both compare this to a treatment where the subjects can choose to leave early. Georg et al. (2019) includes other options as well, including one in which subjects can choose to browse the internet. These studies generally show that when subjects possess outside options of various sorts, task performance goes down, demonstrating the potential importance of outside options.

There are also multiple prior studies which show the importance of dealing with effort costs explicitly, as otherwise one runs the risk of a study generating potentially incorrect conclusions. For example, Engel (2010) examines incentive mechanisms in which individuals have to satisfy a production quota to receive a payment. Prior studies had found individuals were willing to work above the required quotas, leading the authors to conclude that individuals possess some intrinsic motivation or moral commitment to work. Yet these prior experiments included no outside option other than sitting idly, so after satisfying a quota, subjects had little else to do but continue working. In contrast, Engel (2010) allowed subjects to switch to an alternate activity once the quota for payment was met or at any other time, and found that in this case, subjects produced their quota and then switched to their alternate activity. These results strongly suggest that earlier studies misidentified the reasons people produced past the quota. Similar to Engel (2010), Erkal et al. (2018) show that participants exert more effort in a contest when prizes increase and an incentivized outside option is used, relative to when subjects' outside option is to sit idly. In DellaVigna et al. (2022), the authors attempt to estimate social preferences using real effort experiments. In their first attempt they use a standard real effort design with no outside option and they find that effort is not responsive to motivation treatments. The conclusion from those results is that social preferences were not substantial. They then conducted a second experiment in which they allow subjects to leave early and find that effort levels respond to treatment conditions allowing them to estimate more substantial levels of social preferences among their subjects. As one understands the real nature of effort costs in these two designs, the reason for the difference in the results becomes quite clear.

Based on our model, we conduct three different experiments to examine different issues that arise in examining real and stylized effort designs. Our first experiment examines a prior claim suggesting that people can better coordinate in real effort than stylized effort designs. That claim was made based on experiments that included explicit effort costs for the stylized design but not the real effort one. In our experiment, we redo the real effort experiment with and without effort costs to demonstrate that it is this difference that drives the ability of subjects to coordinate, not the effort modality. In our second experiment, we propose and test a module that can be added to almost any real effort experiment, allowing an experimenter to induce and control effort costs in a manner consistent with the underlying theory. We then test whether we can add this module on to the experiment in Araujo et al. and generate the expected labor supply effect. We replicate their results and then show that the module works as expected. In our final experiment, we directly test the proposition that one should expect different behavior in real and stylized designs simply because something must be different between the two effort modalities. We

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

design an environment in which the only thing that differs between the real and stylized designs is the effort modality, and we do so in a context where it is reasonable that real effort could trigger something in the preferences of the subjects, leading to different behavior. We find no differences in behavior.

2 Cost of effort

2.1 Theory

To understand the origin of what an effort cost is, we start from a standard version of the neoclassical model of labor supply. This involves an individual deciding how to divide a fixed time budget between multiple activities. For simplicity we will stay with the two classic options of work and leisure and assume these two sources of utility are additively separable. Let $W(y)$ be the utility an individual derives from consumption, which is a direct function of the earnings, y, they receive from time spent working. The earnings an individual receives, y, will be assumed to be a function of the time allocated to the work activity, x, where x is what one usually means when they refer to the effort of a worker. Let ℓ represent the amount of time spent on leisure leading to a utility of $L(\ell)$.

Based on these assumptions, we arrive at the standard labor supply problem in which an individual divides their time budget, T, between labor and leisure, or

$$\begin{aligned} \max_{x,\ell} \; & W(y(x)) + L(\ell) \\ s.t. \; & x + \ell \leq T \\ & x \geq 0, \ell \geq 0 \end{aligned} \tag{1}$$

Given standard non-satiation assumptions, i.e. $W_y > 0$ and $L_\ell > 0$, and the typical assumption that earnings increase with time spent working, $y_x \geq 0$, the inequality constraint is always binding and so the optimality condition for this problem is simply

$$W'y' = L' \tag{2}$$

or the marginal benefit of labor due to the increased consumption possible from more time spent working, $W'y'$, is equal to the marginal cost of effort, L'. This is a standard result used to show that that allocating another unit of time to labor is costly because this requires taking that unit of time away from the leisure activity. The cost of one more unit of labor is therefore equal to the utility decrease from giving up spending that amount of time on leisure.

Most experimental studies that involve effort provision do not explicitly use the standard labor supply model as their foundation, rather they use some variant of a standard principal-agent model. In this specification the agent earns utility from income which is increasing in effort, but the agent also experiences disutility from expending the effort. A simple linearized version of this model would specify the following choice problem

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

$$\max_{e} y(e) - c(e) \tag{3}$$

where $y(\cdot)$ is some function indicating how effort yields income and $c(\cdot)$ measures the cost of effort, normalized into monetary value. The origin of $c(\cdot)$ is generally not specifically addressed in papers using this model as it is simply some function assumed to have positive first and second derivatives. From a theoretical perspective that is enough as it allows one to derive results which hold for a broad range of situations. However, once we try to determine the empirical and experimental representation of $c(\cdot)$, we have to think more deeply about where it comes from.

To understand what is given up by the agent in the principle-agent model, we need to connect it back to the labor supply model. Effort or e in the principal-agent model is typically not denominated in time spent on the effort task but rather it is an abstract notion of effort. In real effort experiments it is usually measured in terms of tasks completed. Both of these are useful simplifications but they are simplifications. In theoretical models, one may prefer to abstract away from specific time frames and production functions. In experiments, measuring effort in completed tasks allows for a convenient definition of payment and cost functions. However, this approach is really a shortcut as a subject in an experiment cannot actually choose an arbitrary number of math problems to complete in 4 min. What they can choose is how much time to spend trying to solve math problems. This time spent is converted to output based on some underlying production function that depends on their capability for solving math problems.

To see the implication of this point, let us redefine the choice problem in the principal agent model to be one of time division. Instead of choosing output, e, we will assume the individual chooses time to devote to the productive task, t, and $e(\cdot)$ becomes a production function which translates time spent on the task to output. With this structure, we can model our agent as choosing how much time, t, to devote towards the labor task out of a total time budget of T. If the agent spends their entire time budget on production, they produce $e(T)$ units of output. If they spend some time $x < T$ on production then the difference $T - t$ indicates the time spent on the leisure activity. Assuming our time constraint holds as an equality, we can define $L(T - t)$. Further, we can define $L(T) = H$ as a constant representing the maximum leisure utility possible from spending no time on the productive task. We can then rewrite $L(T - t) = H - f(t)$ where $f(\cdot)$ measures the decline in leisure utility due to allocating time t to production. Therefore $f(t)$ measures the cost of spending time t on the productive task. With these definitions, we can rewrite the labor supply model, Eq. 1, as

$$\begin{aligned} \max_{t} \; & W(y(e(t))) + H - f(t) \\ & s.t.\, t \leq T \\ & x \geq 0 \end{aligned} \tag{4}$$

It should be clear that Eq. 3 is a simplified version of Eq. 4. The model in Eq. 4 involves choosing time which generates output or effort based on the production function whereas Eq. 3 takes the short cut of assuming the person could choose

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

output directly. The $W(\cdot)$ terms reflects the utility earned from labor and $f(\cdot)$ represents the cost of effort. Seeing this connection between the models is useful because it provides the clear justification for $c(\cdot)$ in Eq. 3. This cost of effort is very clearly the utility lost from spending time working rather than on the next best option.[5]

Understanding the connection between these models is important as it provides key insights on the nature of effort costs.

Insight 1 All effort costs are opportunity costs.

Effort costs represent decreases in utility from not spending the relevant resources on the next best available activity. This point makes it clear why in many prior papers, experiments often find that the output by subjects is not responsive to wage rates. The value of the time spent away from the effort task often generates negative utility so subjects spend their entire time on the productive task. The only way to get subjects not to engage in the productive task, is to find a way to make that task generate even less utility than sitting idly.

Insight 2 The cost of effort is dependent almost entirely on the nature of the outside option, not on the effort task.

The cost of effort is based on utility foregone from not engaging in another task. How much utility is given up depends only on how much resources are diverted from the outside option task to the work task and how much utility those resources could have generated for the person had they been directed to the outside option task. The cost of effort is not directly related to the nature of the work task. Of course the nature of the work task can affect the nature and amount of resources one has to divert from the leisure task and through these mechanisms the work task can affect the way utility is lost from the outside option. Some work tasks may require discrete time increments while others not. Some work tasks could deplete resource budgets in addition to a time budget leading to mental or physical fatigue. In all cases though, the utility lost from any expenditure of time or energy on a productive tasks is dependent on the utility the person could have achieved in their next best option.

This insight is directly contrary to the standard view in the literature which is that the cost of effort derives from the task itself – a point that is summarized in Charness et al. (2018), "Control over the cost-of-effort function, seen as one of the major advantages of the chosen-effort paradigm, has been addressed primarily through qualitative means, for example, by juxtaposing results from 'easy' and 'hard' real-effort tasks." Understanding that effort costs are derived from the outside option is also an important insight in regard to the claimed field relevance of the real effort tasks. In the field, effort is costly because instead of working, workers could be engaging in a broad range of activities, including working for

[5] For example Mas-Colell et al. (1995), chapter 13, p. 438, identifies this term as foregone earnings from home production which is simply another way to note that it is an opportunity cost.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

a different employer, watching a TV show or movie, chatting with friends, online shopping, or any other pursuit. Unless these outside options are included in a real effort experiment, it is quite difficult to claim that the cost of effort reflects effort costs outside the laboratory. Of course including these types of outside options in a laboratory experiment is difficult to do in a practical manner and so other approaches must be taken to incorporate effort costs into the laboratory environment.

There are other elements one can easily add into this model when they are deemed necessary such as the possibility effort may involve expending scarce resources other than or in addition to time. One way of conceptualizing other cost dimensions to the choice framework is to refer to them collectively as the intensity dimension of an individual's work effort. Thus an individual can choose an amount of time to spend working on a task as well as the intensity with which to work on it. Outside the laboratory, this intensity choice is likely to be very important as a manual laborer who works with great intensity for 4 h may exhaust themselves to the point that they diminish how much enjoyment they receive from their leisure time, while a worker who chooses to work with less intensity may still be rested and able to enjoy their leisure time. Inside the laboratory, subjects can certainly vary the intensity but it is not clear that this additional dimension is empirically important. It seems unlikely that working a little harder on a math task in an experiment diminishes the enjoyment of staring at a screen or that the extra focus inside of the experiment leads to substantial fatigue post-experiment. It is likely though that it is these types of costs that are typically referred to as "coginitve costs" in prior literature. While we acknowledge that they can exist, the data in Araujo et al. and similar papers demonstrate that these costs are less than $0.005 per slider, and so it seems reasonable to round them down to zero for typical real effort experiments.

There are also cases in which one might want to add in the possibility that tasks can generate (dis)utility directly which can be done by simply adding a task dependent utility parameter to the function which determines utility earned through working. In some cases, it may also make sense to allow for heterogeneity in this element. An additional important source of heterogeneity could be that some subjects have greater ability in a task than others. This is best achieved by allowing for heterogeneity in the production functions which translate time spent on a task to the number of completed elements. We leave these out now for simplicity but these and other modifications are simple to add in where they are necessary or relevant. We would add a note of caution regarding how ability heterogeneity, in particular, is typically modeled. Many prior studies use an approach based on the notion that differential ability in a task is due to heterogeneity in the cost functions among subjects. Such a shortcut is problematic. While theoretically this simplification will not impact the relevant comparative static predictions, it leads to problems when interpreting the cost function into practical applications. By placing the heterogeneity on the cost function, one suggests that people who are good at math instead derive less utility from outside options. Those that are bad at math therefore derive more utility from their outside options. This does not seem a good way to practically model that some

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

people are faster or slower at solving math problems. In our view, this confound between ability and value for leisure time is best avoided for empirical work.[6]

The insights from this section are important for understanding how to model effort in an experiment. The most important point is that all effort costs are opportunity costs and if effort in an experiment is to be costly, a subject must give up something of quantifiable value to engage in it. Building off of this understanding of effort costs, in the next section, we examine whether a previous claim of real and stylized effort experiments yielding different results was due to the effort modality or just differences in effort costs.

3 Costly coordination design: verifying the importance of effort costs

In this first experiment we aim to understand why some previous investigations of real and stylized effort have shown differences in behavior. The insight from our theoretical investigation suggests that one reason for these differences could be that effort costs between the two designs were not the same. To investigate this point, we examine the experiments in Bortolotti et al. (2009), which suggest that individuals are more willing/able to coordinate when the task is a real effort task rather than stylized effort. We chose this set of experiments due to the fact that its methods are fairly standard in the literature and that the substantive claims made are important to understand. In this paper, the results from a stylized weak link coordination game are compared to a real effort version in which individuals counted coins and were paid the lowest error rate among the members of their group. While the stylized coordination game has costs built directly into it to ensure that effort above the minimum is costly, it is not clear that there were significant effort costs to the coin-counting exercise because this is the only task subjects could engage in. We note though that the authors do include an option on the real effort experiment which would allow subjects to buy extra time to complete their task which could be seen as a clear effort cost. After an initial learning phase, subjects generally seemed to have no need for the extra time meaning it was rarely an actual cost. Whether used or not, there is no indication that these costs were calibrated to be similar to the effort costs in the stylized design and so this difference in effort costs could explain the difference in results rather than a fundamental difference between real and stylized effort.

Similar differences in effort costs using stylized designs were already examined in Van Huyck et al. (1990) (VHBB). In that study, the authors conduct two versions of coordination games, one in which it was costly to contribute above the minimum contribution in the group and another in which it was costless. In the version where

[6] Besley and Ghatak (2005) provide an excellent example of how to bring these types of issues into a theoretical model in a way that translates into empirical analysis. They construct a model of mission oriented workers by proposing that individuals may receive a utility bonus due to the nature of work they do rather than the nature of the job affecting the outside option, i.e. cost of choosing to work, of the worker. This construction has allowed others such as Fehrler and Kosfeld (2014) to estimate the magnitude of this additional utility.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

it was costly to contribute above the minimum of others, coordination failed with groups ending up at the minimum contribution level. When contributing above the minimum of others was not costly, 96% of all subjects chose to coordinate on the highest choice. This is a stark difference and suggestive of the differences found in Bortolotti et al. (2009) between their real and stylized effort designs. Given that there were two elements changed between the treatments in Bortolotti et al. (2009), the effort modality and the effort cost, we want to examine which drove the difference in the ability to coordinate. To that end we conduct an experiment with a real effort coordination game in which we vary whether or not there are effort costs to determine if our results match with what VHBB found in their stylized experiments.

In our real effort version of a coordination game, subjects participate in four-person groups and have the opportunity to complete instances of a task for earnings in one of two between-subjects treatments: Costless Effort and Costly Effort. The task involves the subjects counting the number of 0's in a string of 0's and 1's. Subjects can apply correctly completed tasks toward a team account or (possibly) an individual account to generate earnings. Linking back to the theoretical discussion, tasks directed to the team can be thought of as "effort" towards work output and the effort directed towards the individual account can be viewed as leisure.[7] The team account's earnings follow the same structure as the VHBB weak link game; they receive piece rate earnings based upon the lowest number of tasks directed to the group account by any member of their group. When effort is costly, subjects also earn a piece rate based on how many tasks they complete and direct toward their individual account. When effort is costless the individual account is eliminated from consideration, meaning the subject can only contribute to the team account. This is equivalent to receiving no compensation for tasks directed to the individual account and thus there is no opportunity cost for effort directed towards the group account.[8] We eliminate the option rather than set the compensation to 0 as this better mimics how standard real effort experiments are conducted. Our goal is to test how results compare between a standard real effort design and one with effort costs and these two treatments capture that difference. In all treatments we also include a fixed payment per period.

The costly treatment is mathematically equivalent to VHBB. Let e_i be the number of tasks completed for the group coordination task for which each group member is paid b for the lowest number anyone in the group completes and l_i be the number of tasks directed toward the individual account for which the individual receives c per task then we can represent the payoff function for our costly effort task as

[7] Methodologically, it is important that the tasks to be completed for the individual and group accounts are the same. It makes it easier to measure the effort costs this way as it should take the subjects the same time to complete an instance of this counting task regardless of which account they direct it to. Thus the only difference is the piece rate earnings. If the tasks were different, then we would have to account for the difference in time to complete the two tasks in measuring how much a subject gives up on earnings from one task to complete one unit of the other.

[8] One might object to our labeling this as "costless" effort due to the notion that all effort has to be costly in at least some sense. While at some level that may be true, our label reflects the intentional lack of induced effort costs.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

$$\pi_i(e_i, e_{-i}) = f + b * \min(e_i, e_{-i}) + c * l_i \quad (5)$$

We implement a standard version of a coordination game with costly effort by setting $f = 60$, $b = 20$ and $c = 10$. We allow unbounded output, however for ease of exposition and comparison to VHBB, if the total number of tasks were capped at 7, as in VHBB, the payoffs in the costly treatment would mimic those in Table 1.[9] For the costless effort case, we set $c = 0$ and obtain the matrix in Table 2.

In the costly effort case, we see the standard coordination game structure in which group members choosing the same number is an equilibrium with the Pareto dominant case at the maximum contribution. Coordination may be difficult of course because contributing above the team minimum is costly. In our real effort version, coordination may be particularly difficult because there is not a common upper bound to coordinate on since subjects will be heterogeneous in their ability. In the costless effort case, there is no cost of contributing effort above the team minimum and the Pareto dominant solution is still the maximal contribution to the group account.

These payoff specifications have clear implications. In the costly effort treatment, completing more tasks for the group account than the minimum of others is costly because the individual could have completed them for their own account and made \$0.10 each rather than \$0. In the costless effort case, completing more tasks for the group account than others has no opportunity cost other than the utility foregone by not sitting idly. As noted above, that translates into a negligible or even negative opportunity cost.

On the productivity range of 0–7, Tables 1 and 2 accurately reflect the relative incentives. The same pattern continues to higher levels of productivity. They engaged in this game for 10 periods with the same four-person group and feedback was given after each period.[10]

This Costly Coordination (CC) experiment and another experiment, the Effort Cost Module (ECM), to be explained later, were conducted in common sessions. An initial round of data collection took place in 2019 and a replication check was conducted in 2022.[11] Because both experiments comprised two treatments, each session had subjects participate in one treatment from each experiment where the treatment

[9] VHBB had integer effort choices ranging from 1 to 7. We also include the option of 0 in our examples to the subjects, so we have presented this version.

[10] Examining Tables 1 and 2 shows that there are two differences between the treatments. The first is the intended one of the treatments differing as to whether contributing above the team minimum was costly or not. The second difference is in the returns to coordination. These returns are higher in the costless effort version. This difference occurs due to the elimination of earnings from the outside option in the costless effort case. VHBB gets around this problem by imposing a rule that essentially makes effort costly up to the group minimum but not above. This rule would have been awkward to implement and explain in our real effort setting yielding the unfortunate difference in our two treatments. This element is a potential confound of our results which we will examine in the next section.

[11] For the main analysis in the paper, we pool both sets of data. We also ran specifications of the main regression tables with the initial set of data only and the replication set separately to demonstrate that both data sets lead to the same substantive conclusions. These specifications are available upon request, or can be replicated from the online replication package.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

order was varied between sessions. In 2019, the ECM experiment always preceded the CC experiment, while in 2022, the order was reversed to correct for any possible ordering effects from the first set of data. Out of the nine sessions conducted, four took place at Ohio University in 2019, two were conducted at Southern Methodist University (SMU) in 2022, and the rest at Ohio University (OU) in 2022.[12] The experiment was programmed using Z-tree software, Fischbacher (2007), and lasted 60–90 min yielding an average payment of $31.12, including a $5 show-up fee for the OU sessions, $10 for SMU sessions and the earnings from both experiments. In the CC experiment, there were a total of 92 subjects; 40 in the costless treatment and 52 in the costly treatment.[13]

3.1 Costly coordination results

The question to be addressed is whether providing an opportunity cost for contributing to a group account in a weak link coordination game has the same impact on coordination in a real effort study as in the stylized effort experiments in VHBB. Figure 1 shows the average contribution of tokens to the team account in all 10 periods separated by treatment with the 95% confidence interval around the average indicated for each bar. It also contains the total production, which for the Costless Effort treatment is exactly the same as the contributions to the team account, while for the Costly Effort treatment, it includes the contributions to the private account. The basic results are readily apparent. The total team contributions are very high in the Costless Effort treatment, with an average of 13.27, while they are much lower in the Costly Effort Treatment, with an average 4.80. On the other hand, total production appears to be approximately the same between both treatments, with averages of 13.27 and 12.88, respectively. This means subjects completed essentially the same total instances of the task in both treatments but in the Costly Effort case, many of these instances are completed for the individual account even though completing them for the team account would be payoff dominant should all team members choose to do the same.

Table 3 provides the statistical analysis to support the visual results. We provide regressions examining how team contributions, the minimum of the team contributions, and total contributions vary by treatment. We also provide specifications examining whether there is a time trend. These regressions are all random effects panel regressions with the standard errors clustered at the team level. In the case of the minimum of team contributions, the observations are at the team level, whereas for team contributions and total production, the observations are at the individual level.

[12] Only two sessions were conducted at SMU as data collection was disrupted due to COVID-related factors requiring data collection to be relocated back to Ohio University.

[13] The initial wave had 24 subjects in the costless and 24 in the costly. The follow-up had 16 in the costless and 28 in the costly.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Table 1 Coordination game with effort cost

		Minimum of the team's contribution							
		7	6	5	4	3	2	1	0
Own	7	200	180	160	140	120	100	80	60
Contribution	6		190	170	150	130	110	90	70
	5			180	160	140	120	100	80
	4				170	150	130	110	90
	3					160	140	120	100
	2						150	130	110
	1							140	120
	0								130

Table 2 Coordination game without effort cost

		Minimum of the Team's Contribution							
		7	6	5	4	3	2	1	0
Own	7	200	180	160	140	120	100	80	60
Contribution	6		180	160	140	120	100	80	60
	5			160	140	120	100	80	60
	4				140	120	100	80	60
	3					120	100	80	60
	2						100	80	60
	1							80	60
	0								60

Result 1 In a real effort design, costless effort yields high coordination, while costly effort leads to a breakdown of coordination.

As predicted, we find that for team contributions and the minimum of team contributions, the treatment effect of Costly Effort is large and highly significant. We also find a time trend where the individual contributions to the team and consequently the minimum contributions to the team are declining in the Costly Effort treatment, yet they are constant or rising in the Costless Effort treatment. Further, we find no difference in the base treatment effect or in the time trend when we look at total production meaning that these effects can be eliminated as potential confounds. Thus, we find that real effort coordination games demonstrate the same properties as in the VHBB stylized games; coordination largely fails when effort is costly yet "succeeds" when effort is not costly.[14] These results suggest that the results found

[14] It is possible to argue that subjects were better able to coordinate in the costless effort case due to the higher marginal returns from that coordination. This is not the best interpretation of the results. If we examine the coefficients on the time variables in Table 3, we see a very strong negative time trend of individual contributions to the team account in the Costly Effort treatment but not the Costless Effort treatment. In the first, this is likely from a slow breakdown in willingess to contribute at a high level based on seeing others contributing less. The lack of a time trend in the other treatment is likely due to

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

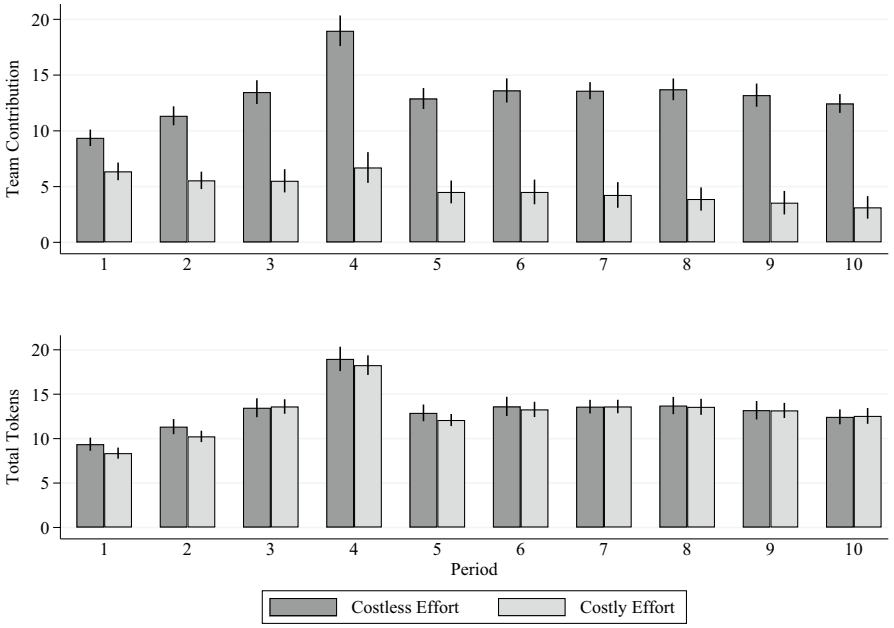


Fig. 1 Paired bar graphs by period of average contributions to the group and total production for both costless and costly effort coordination games. Error bars represent the 95% confidence interval around the average

Table 3 Random effects panel regressions on team and total production

	Team Contributions		Min of Team Contributions		Total Production	
	(1)	(2)	(3)	(4)	(5)	(6)
Costly Effort	−8.466***	−7.480***	−7.303***	−6.529***	−0.387	−0.696
	(0.882)	(0.817)	(0.921)	(0.904)	(0.612)	(0.632)
Last Half		0.105		0.640***		0.105
		(0.149)		(0.214)		(0.149)
Last Half * Costly Effort		−1.970***		−1.548***		0.618***
		(0.590)		(0.588)		(0.200)
Constant	13.27***	13.22***	10.08***	9.760***	13.27***	13.22***
	(0.468)	(0.487)	(0.592)	(0.627)	(0.468)	(0.487)
Obs (Groups)	920 (23)	920 (23)	230 (23)	230 (23)	920 (23)	920 (23)

Standard errors clustered at team in parentheses, ***$p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Footnote 14 (continued)
the fact that subjects do not mind having done more than others given that it is not costly to do so and so do not pull back productivity over time. It is this difference in dynamics which appears the best explanation for the overall differences in contributions to the team.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

in the prior study of behavioral differences between real and stylized effort were observed mainly because the cost of effort differed between those two modalities, not because stylized effort inhibits coordination while real effort makes coordination easier.

4 Effort cost module experiment: incorporating effort costs into real effort design

The CC results make it clear that the proper specification of an outside option is important for real effort experiments. That leads to the question of how one could bring effort costs into a standard real effort design. In this section, we propose and test the effectiveness of an effort cost-inducing module that could be added to any real effort experiment to allow effort costs to be included and controlled for. The fundamental innovation is adding an outside option that implements an effort cost function as specified in the standard model. We add this outside option to the experiment design of Araujo et al. (2016) and attempt to determine if we can recover the theoretically predicted wage effect once a cost of effort is included in the experiment.

Several prior papers add outside options, but most of these prior implementations have drawbacks that keep them from being universally applicable. A commonly thought of and occasionally used outside option is to allow subjects to browse the internet (e.g., Corgnet et al. 2015b).[15] While these approaches have relevance to many external situations, implementation in the laboratory is problematic for a couple of reasons. First, the time structure of most designs have production periods that only last a few minutes. Subjects switching between internet browsing and the experiment involves relatively substantial switching costs and subjects may not find 30 s of internet browsing valuable, especially using an unfamiliar browser lacking their normal bookmarks. This is despite the fact that office workers may choose to spend hours online rather than engaging in their primary work activity. The other important drawback of methods like this is that the experimenter has neither knowledge of nor control over the value of these activities to their subjects. This means that for some subjects giving up 30 s of time browsing the internet is very costly but not for others. This idiosyncratic difference which may have little relevance to any treatment condition could drive treatment differences.

Prior papers also provide outside options with fixed or linear utility structures (e.g., Johnson and Salmon 2016; Erkal et al. 2018). While these designs can be effective in some cases, those payoff structures are insufficient to guarantee an interior optimum for the subjects. That is, if the "effort" activity pays back at some piece rate and the leisure option pays back at some other rate, the effort activity could still dominate the alternative through the entire production period. While this

[15] Kessler and Norton (2016) and Corgnet et al. (2015a) are other papers which allow subjects an outside option of internet browsing. Additional examples of leisure options include reading magazines (Charness et al., 2014) and watching funny videos (Kamei and Markussen, 2022). Kamei and Markussen (2022) added additional control by paying a small per minute wage to subjects engaged in watching videos.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

demonstrates a revealed preference relationship that working is revealed preferred to this alternative option, it can be difficult to observe treatment effects in some cases due to the boundary solutions. Erkal et al. (2018) provide a design with a piece rate outside option where the main effort task is a competitive tournament and find that effort responds to the prize value but competitive settings like this often generate different behavioral responses than piece rate settings. While that design worked there, the results from Johnson and Salmon (2016) demonstrate that it may not be as effective if the main task has a piece rate structure.

An outside option with a payoff structure that is non-linear, as assumed in the standard theoretical models of effort provision, is needed so that it is possible to expect interior optima for effort provision.[16] The design should also fix the cost of effort to be the same between all subjects or induce heterogeneity in a known manner. Our design satisfies these needs. In the Araujo et al. experiment, subjects engaged in a standard slider alignment task based on Gill and Prowse (2011) in which they are paid a piece rate wage for each slider aligned. In our version, subjects have 3 min to align as many sliders as they wish, facing either a \$0.01/slider or \$0.04/slider wage rate. The wages are varied between, not within subject. After subjects complete an initial period of this task with no outside option, we introduce the option to engage in an alternative to the slider task. The alternative activities we provide are playing Tic-Tac-Toe (TTT) against a mildly challenging computer algorithm and solving word search matrices where the subjects find words embedded into matrices of letters. The nature of these tasks is not important; they merely need to be active and perhaps mildly amusing for the subjects. The key is in how these alternative activities are incentivized.

When an outside option is available, subjects can spend time aligning sliders or switch to a screen with the outside option tasks. Their earnings are based on how many sliders they complete (piece-rate wage) and on the total time spent on the outside option screen. To make the experiment easier, they begin the production period facing the slider screen and can switch to the outside option screen whenever they like, but the switch is only allowed once - after switching away from aligning sliders, they cannot switch back. This is not a necessary design element but chosen to allow for a cleaner design.[17] If a subject chooses to spend the entire 3 min on the outside option screen, they earn a fixed amount which is set at \$1.19. This fixed payment intentionally does not depend on how many TTT games they win or how many words they find in matrices. Any amount of time subjects choose to spend aligning sliders before switching over to the outside option screen decreases this amount. The total cost of any amount of time spent aligning sliders is $0.006t^{2.3}$ where t represents seconds spent aligning sliders. This generates a convex time cost function, as standard models assume. Note that this structure matches the specification in Eq. 4. To

[16] One could of course also find a way to make the pain of the effort task to increase in such a way to generate an interior solution if desired.

[17] For an experiment with longer production periods allowing frequent switching would be reasonable and easy to implement. For our simple experiment here, it seemed an unnecessary complication to explain to subjects.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

make the implications of this cost function easier to understand, we do not present the mathematical version of this function rather, we alert them on their screen to how much their earnings will decrease for the next 5 s they spend aligning sliders. We chose that time increment because in the Araujo et al. data, subjects on average aligned one slider every 5 s. The subjects also complete one initial period of slider alignment without the outside option. We use this initial period performance to calculate how many sliders they align on average in 5 s and provide this information to them during the production periods with the outside option. This makes it easier for a subject to determine the point at which they believe aligning sliders is no longer worthwhile for them. Figure 2 shows the total cost function and the marginal cost of effort for 5-s intervals. This construction allows for a way to predict when a subject will stop working and choose the outside option if they are sensitive to incentives in a real effort task. At the \$0.01/slider wage rate, an individual who completes on average 1 slider per 5 s will find that time spent aligning sliders is more valuable than their outside option for 40 s. After that, the marginal cost of foregoing the outside option dominates their earning ability in the slider alignment task. At the \$0.04/slider wage rate, aligning sliders is more lucrative than the alternative up to 110 s. If an individual is faster or slower at aligning sliders, then their optimal time to spend aligning sliders will shift accordingly but it should still be the case that the switchover point should move up with a wage increase.

In total, 92 subjects participated in this experiment with 44 in the Low PR treatment and 48 in the High PR treatment.[18] Other experimental procedures and related information were already described above.[19]

4.1 Effort cost module experiment results

The issue we wish to examine is whether the introduction of an outside option allows us to observe a wage effect and whether this effect is predictable using a standard model. Figure 3 provides a visual representation of the data. It shows the average number of sliders aligned per period with the 95% confidence interval around that mean indicated for each bar. Recall that in period 1, we do not allow the outside option but do in periods 2-4. The immediate observation from this figure is that subjects aligned more or less the same number of sliders in period 1 in both treatments while in periods 2-4 they align many more sliders in the high piece rate treatment than the low piece rate treatment. We also provide lines indicating the number of sliders that on average would be optimal for subjects to complete based on their speed of slider alignment in period 1. We note that our subjects exceed this prediction in all periods. This is in part due to a learning effect as subjects were able

[18] In the initial wave, 24 subjects were in each treatment, while in the second wave 24 were in the High PR and 20 were in the Low PR. We note again that all data is pooled in the analysis presented in the paper. A comparison between the two waves of data was conducted, demonstrating no differences in the conclusions between data sets. Once again, these specifications are available upon request, or can be replicated from the online replication package.

[19] Instructions are available in the replication archive at https://doi.org/10.3886/E208628V1.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

to complete sliders more quickly in periods 2–4 than in period 1. Given that, it is reasonable that they should align more sliders in periods 2–4 than indicated by those lines.[20]

The first formal statistical result we need to test from this data is to examine the output in period 1 to determine if our data replicates the Araujo et al finding of no differences in productivity due to the piece rate wage differential when the outside option is not present. We do find that the mean production in the Low PR treatment is 29.3 for period 1 while it is 34.5 for the High PR treatment. This is suggestive of a possible small difference but a t-test or Wilcoxon test both show a lack of a significant difference in these distributions (p-values of 0.12 and 0.11, respectively).

To examine the treatment effect in periods 2–4, Table 4 provides three different regression approaches to examining the treatment effect and we provide two specifications for each regression type. All regressions include only data from periods 2–4 as this simplifies the inference but we include each player's period 1 productivity as a control variable to take into account any heterogeneity in ability. Columns 1 and 2 present OLS random effects regressions with the standard errors clustered at the subject level. Columns 3 and 4 provide random effect Tobit regressions to correct for the fact that a number of observations are piled up at 0, especially in the Low PR treatment. Columns 5 and 6 provide standard Tobit regressions with the standard errors clustered at the individual level. For each approach, we provide one specification to examine the base treatment effect and then a second to examine the data for any time trends as well as whether the time trend depends on the treatment. These regressions support our next result.

Result 2 Output levels are higher in the high piece rate condition than in the low piece rate condition when an outside option is provided.

The results are consistent across all specifications. We find that the treatment variable for the High PR treatment is large and highly significant indicating that a wage effect is clearly present in the data for periods 2–4. We also find some evidence of a time trend as subjects in the low piece-rate treatment appear to reduce output over time and subjects in the high piece-rate treatment increase output over time which magnifies the treatment effect. Period 1 productivity is highly correlated with later period performance, as would be expected.

These results allow us to refine the conclusions drawn in Araujo et al. regarding the reliability of the slider task. In that paper, the authors suggest the lack of a wage effect in their experiments is because the slider task may not be an effective or useful task to use in laboratory experiments. Our analysis shows that there is nothing wrong with the slider task itself, but there is a potential concern in how it is typically implemented. The lack of a viable outside option creates the lack of a

[20] We could have instead demonstrated the wage impact by showing that subjects spend more time aligning sliders with a higher wage as they spend 101 s on average aligning sliders in the High PR treatment but only 60 seconds in the Low PR treatment. As the two measures are almost perfectly collinear, we chose to use sliders completed as the more typical metric.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

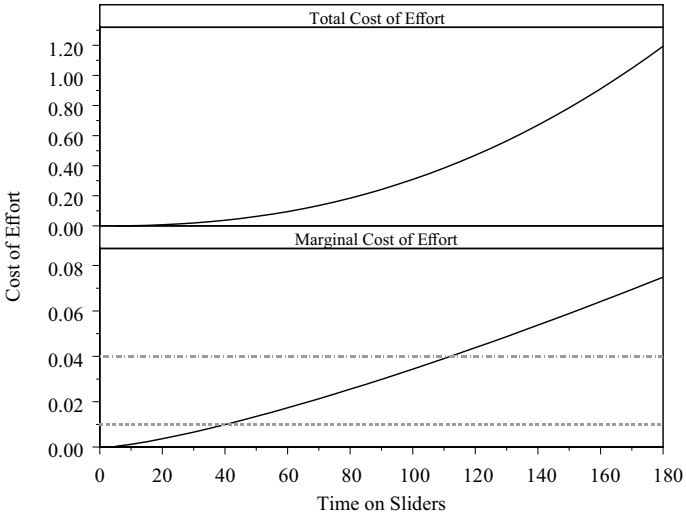


Fig. 2 Total cost of effort and then marginal cost for 5-s intervals

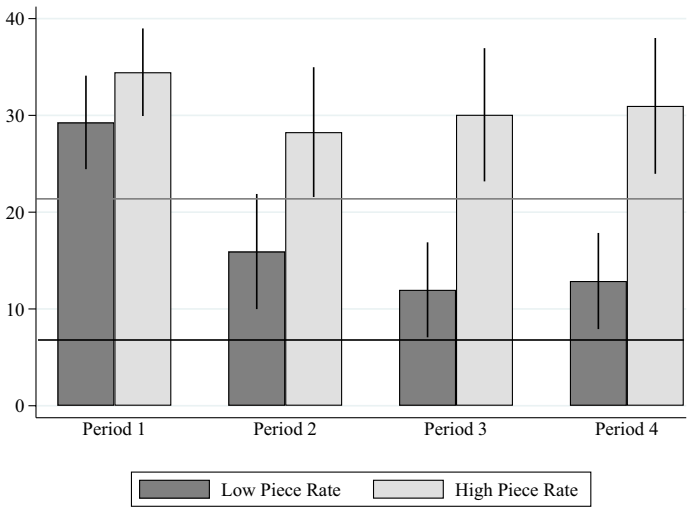


Fig. 3 Average number of sliders aligned by treatment in the main production periods. Light gray line is the predicted number of sliders based on period 1 speed for the High Piece Rate treatment. Black line is prediction for the Low Piece Rate treatment. Error bars represent the 95% confidence interval around the average

wage effect. Here we show that a simple augmentation to the experiment design which incorporates an outside option leads to the expected wage effect. The same will be true of many other tasks used in real effort experiments, as the general issues we examine here should not be expected to relate only to the slider task.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Table 4 Random effects panel regressions of sliders completed

	RE with CSE		Panel Tobit		Tobit with CSE	
	(1)	(2)	(3)	(4)	(5)	(6)
High PR	11.59***	7.758**	16.43***	12.04***	16.07***	11.88***
	(2.861)	(3.292)	(3.492)	(4.044)	(3.441)	(3.885)
Period 3		−3.955		−5.137*		−5.281
		(2.831)		(2.743)		(4.124)
Period 4		−3.045		−4.069		−3.570
		(1.999)		(2.742)		(2.478)
Period 3 X High PR		5.746*		6.809*		6.888
		(3.184)		(3.603)		(4.402)
Period 4 X High PR		5.754**		6.379*		5.679*
		(2.894)		(3.607)		(3.399)
Period 1 Total	0.883***	0.883***	1.016***	1.016***	1.007***	1.006***
	(0.0972)	(0.0979)	(0.109)	(0.109)	(0.106)	(0.106)
Constant	−12.26***	−9.930***	−22.27***	−19.17***	−21.59***	−18.61***
	(2.917)	(3.383)	(4.249)	(4.507)	(4.155)	(4.592)
Obs (Subjects)	276 (92)	276 (92)	276 (92)	276 (92)	276 (92)	276 (92)

Cols (1), (2), (5), and (6) have std. errors clustered at the individual level. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

Note that we only include data from periods 2–4

5 Real effort VCM: is real effort just different?

The first two experiments demonstrate the importance of effort costs in real effort experiments and provide ways to induce them. In this third experiment, the Real Effort VCM (REVCM), our focus shifts to testing a claim often made in the literature, which is that real effort is somehow just different than stylized effort, and it should be expected to yield different results. Our goal in this section is to test the two approaches to modeling effort controlling as many differences as possible between real and stylized designs to determine if other elements can still drive a difference. We do this in a setting in which there is an a priori reason why one might expect the two forms of effort to lead to different results, which we hope gives real effort a clear chance to generate different behavior than stylized.

We implement this test in the context of a standard public goods experiment matching the basic design stemming from Isaac et al. (1984) and Isaac and Walker (1988). One benefit of using this design is that, as explained in Ledyard (1995) and Chaudhuri (2011), the base results have been replicated many times, allowing us ample comparisons with prior work. We examine real and stylized versions of this design and all treatments possess the same incentive structure. Participants are randomly assigned to a group of four and remain matched for the duration of the experiment. In each period, individual subjects accumulate tokens over time either by their own effort or by a random arrival process and can choose to direct the tokens towards a group account or an individual account. For each token invested in the

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

individual account, the participant earns \$0.20. For each token invested in the group account, the group earns \$0.40 with group earnings divided equally among 4 group members. This leads to a marginal per capita return (MPCR) of 0.5 and sets up the standard incentive conflict known as a social dilemma. At the end of each period, participants are provided with feedback that includes a reminder of their contributions, the total number of tokens contributed to the group by all members of the group, and a summary of their earnings for the period. This same process repeats for ten periods.[21]

An important aspect of our design is the need to keep the maximum possible effort common across all subjects and common across effort modes. In many real effort designs, including our first two experiments, maximum effort or productivity is unbounded and therefore subject specific. If we allowed this in our VCM design, it would induce heterogeneity in endowments in the real effort treatments while the stylized effort treatment would exclude this feature. The literature utilizing stylized designs already shows that heterogeneity will affect behavior (e.g. Cherry et al. (2005); Buckley and Croson (2006); Reuben and Riedl (2013)) and so this would be a substantial confound in comparing behavior between treatments. Therefore we fix the token budget at 10 regardless of the effort type.

Another potential source of behavioral differences between real and stylized effort designs is the amount of time involved in decision making. In the standard VCM, a subject makes a single decision about token allocation and periods can go very quickly. In the real effort version, subjects spend time on the real effort task producing tokens and contributions to the accounts. The timing difference could, for instance, lead to a person becoming either more or less thoughtful over their contribution choices which could lead them to be either more or less cooperative. This element must also be eliminated as a difference between treatments.

Real effort tasks may also differ from stylized effort due to potential differences in cognitive load. Engaging in the real effort task could distract participants from the VCM task triggering different behavior. While the directional impact of such a distraction is unclear, it seems quite clear that contribution decisions could be impacted. Our design must therefore equalize the cognitive load between treatments to the extent possible.

One further element we examine is the difference between Useful Effort and what we term Trivial Effort. In the former, the effort is on a task that seems useful and could benefit someone, e.g. stuffing envelopes with mailers for an academic department, or shelling walnuts for a bakery. In Trivial Effort, the effort is clearly just an artifact of the experiment and is conducted for no other purpose, e.g. aligning sliders. There are some claims in the literature that for real effort designs to be effective the effort involved should be plausibly useful to someone, see for example Carpenter and Huet-Vaughn (2019). This is an interesting claim for which we have not seen evidence, and so we have designed our treatments to test this as well. In the Useful Effort (UE) treatment, subjects engaged in a data entry task in which they enter

[21] There are other papers that have real effort in VCM's, for example, Van Dijk et al. (2001) had subjects solve two-variable optimization problems while Cooper and Saral (2013) use GMAT questions.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

actual financial data into a database. In the instructions, we explain to the subjects very clearly that the data entry task is to create a database to be used for research purposes by another faculty member at Ohio University (not a co-author on this project) and that it is vital to this faculty member's research. We further exhort them to be careful in their work so that the final database will be free from errors. This was an attempt to have subjects truly see this as useful effort and not an abstract task necessary only for the experiment.[22]

Subjects earn a token by entering a line of data from a sheet provided to them. An example of this data is shown in Fig. 4. Each data line entered earns a single token, and all subjects must enter exactly ten lines per period. While earning the tokens, subjects choose how to allocate their productive time between the private and group accounts. They do this by using a toggle button on their screen. At the beginning of a period, they must switch the toggle to either the group or private account. After they make an initial choice, they begin producing tokens, and any tokens they produce go to the selected account. They can switch the toggle as many times as they like and at any time, with tokens accruing to whichever account is active when they submit a line of data. Subjects can therefore choose how much of their time/effort to devote to working for the public or individual account.

The Trivial Effort (TE) treatment is conducted identically in all aspects to the UE treatment except that the data subjects enter is presented to them with no context. Subjects in the TE treatment are handed sheets with identical data as those in the UE treatment but there was no mention that the data would be used for any external purpose. They are only told that the reason to enter the data is to earn the tokens. To accomplish this, two copies of each data sheet were made where one went to someone in the UE treatment with the contextual elements and one went to someone in the TE treatment with the contextual elements removed.[23] The data generated by the TE treatment was discarded while the data generated by the UE treatment was turned into a database and given to the researcher who needed it.

Designing a treatment to represent Stylized Effort (SE), which has the properties of a standard VCM but that differs from the previous two treatments only by how the tokens are earned, requires the design to be somewhat different from a standard VCM. In this treatment, subjects receive tokens but do not engage in the data entry task. To ensure that the timing issues are the same between this treatment and the others, subjects make their investment choices using the same toggle switch as in the other treatments but instead of earning the tokens through data entry, the tokens arrive at random intervals. This means that they are making a stylized effort choice as they choose how to allocate effort/tokens towards each account, but no actual effort is required from them. The token arrival times are drawn randomly from the actual distribution of subject times to complete a data entry line from the UE and

[22] After the experiment was concluded we did collect the entered data into a database and hand it off to the other faculty member. We did not follow-up to verify how or if it was ever used.

[23] The titles of the type of data used in UE were completely stripped; the UE titles of ticker, type, and 1, 5, and 10 yr returns were changed to code 1, code 2, 1, 2, 3 percents, respectively in the TE treatment. So the data to input were identical, but no framing was given in TE.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

TE sessions. The average length of time between tokens is 22 s, with a maximum of 73 s and a minimum of 8. Before the token appears, subjects receive a warning that a token will be deposited into the selected account in 3 s, allowing them time to change the current account to which tokens are accruing. In order to provide a similar cognitive load to the data entry task, subjects are able to play Tic-Tac-Toe against a computer opponent for no earnings while the tokens are arriving. It is made clear that playing this game is not connected to earning tokens and that there are no earnings related to playing.[24]

The underlying incentives for all treatments are a precise analog of the labor supply model shown in Eq. 1. Each individual has a fixed budget of tokens (time) to split between two utility-providing alternatives. In our context, an individual chooses how much effort to supply to the group account (labor), with the rest allocated to the private account (leisure). The value of this structure is that the cost of each token/unit of effort contributed to the group account is the foregone private account earning. This is true for all treatments.

This REVCM experiment is intentionally designed to provide a channel through which the mode of effort could affect decisions. The base incentives make it a dominant strategy for all effort to be allocated to private production. Still, we know from many prior public goods studies that this does not usually occur. People generally engage in conditional cooperation in which they contribute to the public good so long as others do, but we also generally observe that people's willingness to contribute declines over time. There is also literature demonstrating that subjects often feel a greater sense of entitlement to earned money than to money received at random, see for example Hoffman et al. (1994). A plausible hypothesis is that compared to someone gifted with tokens, a person who must work for their tokens might experience greater disutility from finding that they have contributed more to the group account than other group members. If valid, then the real effort treatments would lead to different behavior than the stylized effort treatment as subjects may be less willing to risk cooperation with others, or their contributions might decline faster if they are disappointed with the contributions of their group members. The need to allow for such a channel is why the public goods framework was chosen. There are certainly other such channels through which real effort could differ from stylized effort, and we make no claims here of comprehensive testing.

The design may seem similar to one that would test whether earned endowments in public good games lead to different behavior than unearned endowments. This has been previously investigated in the public goods framework (e.g. Cherry et al. 2005; Clark 2002) with the general finding that earned endowments do not change contributions to the public good relative to unearned endowments. These designs generally have subjects engage in tasks to earn tokens prior to the public good phase and so token earning and contribution decisions are separated. In our case, the token production phase is integrated into the token contribution decisions. There are certainly overlaps between the two designs but it is also plausible that behavioral

[24] 66% of subjects played at least one game of Tic-Tac-Toe. The average number of games played in a period was 11.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Fund Name	Ticker	NAV as of 10/5/2015	Total Net Assets	Load Adjusted Returns			
				1 Yr Return	5 Yr Return	10 Yr Return	Since Inception
Advisory Rsrch Glbl Val	ADVWX	11.47	$14,700,000	-7.28%	9.11%	N/A	9.38%
AllianBer Gl Value:A	ABAGX	N/A	N/A	0.90%	6.26%	2.35%	2.59%
AllianBer Gl Value:Adv	ABGYX	N/A	N/A	5.61%	7.49%	3.09%	3.33%
AllianBer Gl Value:B	ABBGX	N/A	N/A	0.70%	6.36%	2.02%	2.29%
AllianBer Gl Value:C	ABCGX	N/A	N/A	3.67%	6.42%	2.06%	2.32%

Fig. 4 Sample of data subjects would enter in the Useful Effort treatment

differences will emerge due to a tighter connection between effort to earn each token and where the individual contributes it.

We conducted two sessions of all treatments in a first wave of data collection in 2015 and followed up with two more sessions of the Trivial Effort and Stylized effort treatments as a replication check.[25] All subjects were students at Ohio University and the experiment was programmed using Z-tree software, Fischbacher (2007). 52 subjects participated in the TE and SE treatments, while 28 participated in UE. Including the show-up fee, subjects earned about $35.50 for an experiment that lasted a little less than 90 min. Instructions are available in the replication archive at https://doi.org/10.3886/E208628V1..

5.1 Real effort VCM results

Figure 5 shows average contribution levels into the group account by period for all three treatments; each bar in the figure also has the 95% confidence interval indicated. Since the SE treatment is different from previous stylized effort VCM designs, we also include data from two prior studies in line graphs, Croson (2001) and Houser and Kurzban (2002), which use the more traditional design and also have the same parameterization as our design in regards to group size and MPCR.[26] The figure shows that the results from all three treatments are very similar to each other and the behavior matches qualitatively with what was observed in the prior studies using the traditional stylized design. Specifically, contributions to the public good start at around 60% of the token budget and steadily decline over time. This similarity suggests that some of the confounds we worried about in the design of our SE treatment

[25] By the time we conducted the second set of experiments, the website used to gather the original financial data was longer in existence. So, for the second set of experiments, new sheets were generated using Microsoft Excel's stock data type feature. We used this feature to generate the name and returns for mutual funds used in the first set of experiments and formatted the sheets to look similar to what was used in the first set of experiments. The second set of experiments only included Trivial Effort and Stylized Effort treatments because the database construction story for an external researcher would no longer have been valid. Also given that the initial data showed no difference between the UE and TE environments, we did not see a strong need to create a new UE design. As before, we pool all data in the results presented here and a replication check can be reproduced in the replication package to show no substantive differences.

[26] These papers do use different token budgets from ours. They allow 25 and 50 per period, respectively. We have collapsed these down to match our scale of 10 tokens and so the best interpretation of the comparison between our data and theirs is thinking of the Figure as showing the percentage of the token budget allocated to the group account in each period.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

may not have been empirically substantial, but that is, of course, only knowable once the data have been gathered. Importantly, were claims true that real effort designs will generate fundamentally different behavior we would have expected much lower contributions in the TE and UE treatments compared to the SE treatment or perhaps a much more dramatic decline in contributions. We instead observe the same basic pattern observed in any standard VCM previously conducted. This suggests that whatever differences the two effort modalities might yield, the substantive conclusions from the analysis remain the same.

Table 5 summarizes the average contributions to the group account and their standard deviation by period for our three treatments. Initial tests of these aggregate statistics find no significant differences between any treatments ($p > 0.36$ for all pairwise comparisons using Wilcoxon tests on the average contribution to the group account for each group over all 10 periods and $p > 0.34$ for t-tests of same). If we conduct these tests on group averages by period, again all differences remain insignificant ($p > 0.16$ for all pairwise comparisons using Wilcoxon tests and $p > 0.18$ for all t-tests).[27]

These tests do not correct for the panel structure of the data, nor do they allow us to investigate the conditional nature of the decisions. To account for this, Table 6 reports results from random effects regressions where the dependent variable is the amount an individual contributed to the group account, with errors clustered at the subject level.[28]

Result 3 There are no statistically significant differences in contributions between stylized effort, trivial real effort, and useful real effort treatments.

Each regression specification includes dummy variables for the TE and UE treatments. The first specification includes only these dummy variables and a constant, which provides a clean test for differences between the overall contribution levels. Neither coefficient is significant which indicates that the contributions to the group account in TE and UE are not significantly different from SE. Since the two coefficients are opposite signs, it could be the case that the average level of contributions

[27] For the $t-$tests on overall contributions by groups we find that assuming a power level of 80%, for the observed effect size to be significant at the 5% level between the SE and TE treatments we would need 222 groups; between the TE and UE treatments we would need 182 groups; and between the SE and UE treatments we would need 4568 groups. These sample sizes are beyond feasible sample sizes for most experiments. It is certainly the case that our results cannot guarantee that no true effect exists, but they should be enough to suggest that if any effect exists it is quite small. We can also calculate the effect size that our study was powered to identify. Given the standard errors and sample sizes in the combined data set and a power level of 80%, were the differences in average effort of the magnitude 1.15 tokens between SE and TE or 1.32 between SE and UE, we could have detected a significant difference. A difference of 1.32 would have been significant between the TE and UE treatments. Any level differences less than 1 token seem economically quite insignificant suggesting that our experiment is sufficiently powered to detect differences of the magnitude necessary to investigate the relevant question.

[28] In the replication package, we provide Tobit regressions to correct for censoring and specifications and also use sandwich regressions (McCaffrey and Bell 2002; Tipton and Pustejovsky 2015; Pustejovsky and Tipton 2018) to cluster at the group level despite the small number of groups. Qualitative results are unchanged.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

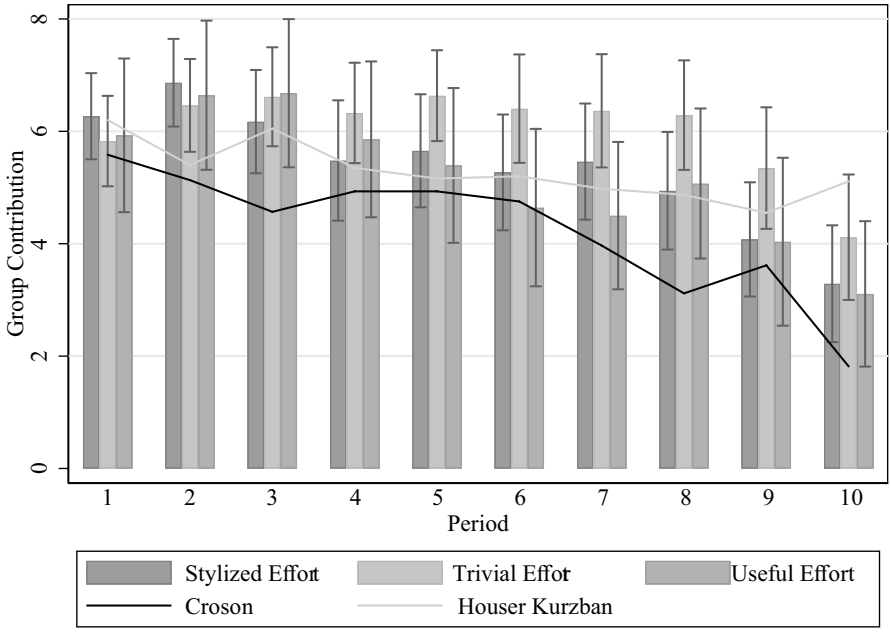


Fig. 5 Average contribution to the group account by period over all 10 periods. Error bars represent the 95% confidence interval around the average

Table 5 Mean and standard deviations of contributions to the group account by treatment (Trt) and period (Pd)

Trt	Pd 1	Pd 2	Pd 3	Pd 4	Pd 5	Pd 6	Pd 7	Pd 8	Pd 9	Pd 10	Overall
SE	6.27 (2.75)	6.87 (2.80)	6.17 (3.30)	5.48 (3.85)	5.65 (3.61)	5.27 (3.70)	5.46 (3.71)	4.94 (3.76)	4.08 (3.65)	3.29 (3.73)	5.35 (3.61)
TE	5.83 (2.89)	6.46 (2.97)	6.62 (3.16)	6.33 (3.21)	6.63 (2.90)	6.40 (3.47)	6.37 (3.62)	6.29 (3.50)	5.35 (3.89)	4.12 (4.01)	6.04 (3.43)
UE	5.92 (3.52)	6.64 (3.42)	6.68 (3.40)	5.86 (3.58)	5.39 (3.55)	4.64 (3.61)	4.50 (3.38)	5.07 (3.44)	4.04 (3.85)	3.11 (3.34)	5.19 (3.62)
All	6.02 (2.75)	6.66 (2.99)	6.45 (3.25)	5.89 (3.54)	5.98 (3.35)	5.58 (3.63)	5.61 (3.65)	5.50 (3.62)	4.57 (3.81)	3.58 (3.76)	5.59 (3.56)

to the group account in TE and UE are different. A post-estimation Wald test yields $p = 0.15$ indicating that those two coefficients are also not significantly different from each other.

Given the decay over time we observe in Fig. 5 we also examine the degree to which that decay is common across treatments. To do that we add a regression specification which includes a linear time trend along with interactions with the two treatments. Given that one possible reason for that decay is how people respond to prior contribution levels among their group, we also include a specification with a

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Table 6 Random effects panel regressions on contributions to the group account

	(1)	(2)	(3)
TE	0.690	−0.242	−0.639
	(0.453)	(0.591)	(0.665)
UE	−0.162	−0.0438	−0.782
	(0.606)	(0.824)	(0.905)
Period		−0.321***	
		(0.0650)	
Period*TE		0.170*	
		(0.0939)	
Period*UE		−0.0216	
		(0.0995)	
$Group_{t-1}$			0.468***
			(0.0823)
$Group_{t-1}$*TE			0.182*
			(0.107)
$Group_{t-1}$*UE			0.132
			(0.176)
Constant	5.348***	7.113***	2.638***
	(0.337)	(0.464)	(0.446)
Obs (Clusters)	1320 (132)	1320 (132)	1188 (132)

Clustered robust standard errors in parentheses

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

lag of average group contributions again interacted with the treatments.[29] In general, our results remain unchanged in that the main treatment variables are insignificant indicating that overall, contribution levels do not change between treatments. We do get two marginally significant interaction terms with the TE treatment. These are very small effects and seem to be false positives since we do not observe a similar result for the UE treatment. They also go in the opposite direction of the predicted effect. The predicted effect would be that when people earn their tokens they feel more entitled to them and less likely to contribute if others are not contributing. The opposite possibility is indicated by these marginally significant results and it is difficult to justify. Also, we already noted the period by period tests showing that in every period we find a lack of significant differences in the contribution levels between treatments. This makes it clear that these marginally significant effects are not resulting in substantial differences between the contribution levels.

[29] Because of the near perfect (negative) correlation between the lagged term and Period ($p < 0.0001$), the regression from column two cannot include both Period and $Group_{t-1}$.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

6 Conclusion

It is commonly argued that an experiment design based on a stylized effort task does not generalize to field settings as readily as a real effort experiment. There are task-specific arguments made (e.g., intrinsic motivation related to a specific task), but the most generalizable argument lobbied against stylized effort is that effort cost in a real effort experiment better represents effort costs outside the laboratory. Such a claim indicates that the nature of the task changes the inherent cost associated with completing it. We demonstrate that this claim is simply unfounded.

Our first experiment showed that real and stylized effort do not yield different behavior in the context of coordination games as claimed in Bortolotti et al. (2009). Our results suggest that a better interpretation of those results is that behavior differed between real and stylized treatments because effort costs differed markedly between the two treatments and not because of the effort modalities themselves. Our second experiment provides a demonstration of how to include effort costs in a standard real effort experiment that better matches theory. This was motivated by Araujo et al. (2016), which found no wage effect when implementing the slider task in its usual implementation (Gill and Prowse, 2019). We find that once effort costs are implemented in a real effort experiment, the expected standard wage effects emerge that were lacking from prior studies, which did not include relevant outside options. There are prior studies that have also introduced outside options in their design (Corgnet et al., 2015b; Engel, 2010); however, those studies do not measure the value of the outside option and thus cannot determine more precise theoretical predictions. Our last experiment utilizes the social dilemmas present in many VCM games to determine if there is some other intrinsic reason that suggests behavior in a real effort design to be different from a stylized effort design. We chose this design because of the consistent behavioral deviations observed in traditional VCM experiments, which we believe lends itself to identifying if the modality of the experiment is tied to the behavioral effects or if these behaviors generalize. We construct tightly connected real and stylized versions of an experiment and find that subjects generally make the same decisions regardless of whether their choices are based on real or stylized effort.

While our last experiment showed no differences between real and stylized effort, there can be other environments where differences might exist, and future research can work to identify those environments. Also, none of our results suggest that either real or stylized effort should always be preferred in any experiment. For example, if demographic-based diversity in ability is important to a research question, a real effort design would certainly be indicated to estimate the degree of naturally occurring diversity and exploit it in a way that a stylized design cannot. See, for instance, Coffman et al. (2021) for a relevant example. What our results do suggest is that for a broad range of questions, there may well be no gain from implementing a real effort design given the substantial costs and potential confounds of these designs.

For cases in which it makes sense to use a real effort design, our second experiment gives researchers a useful tool to implement an actual effort cost in a real

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

effort design to ensure that the choice environment matches the theoretically relevant one. The key element of the methodology involves a non-linear payment to subjects based on how much time they spend engaging in an outside option, which constitutes an opportunity cost of time spent on the productive activity. The non-linearity adds convexity to the cost so that interior solutions can be obtained. Simply including an outside option that is available in the external environment may not be sufficient to effectively test typical labor supply research questions. For instance, in short production periods, switching costs tied to externally relevant activities are likely too high and may still lead to boundary solutions. Likewise, while allowing subjects to leave an experiment early works well to incorporate field-relevant outside options, it is not appropriate if one needs more than one period of production or more than one interaction. Similarly, the utility gained from an externally relevant outside option may vary for the subjects, and this variance may differ from what is observed in the external setting, leading to questionable external validity and a confound in the design. Our design element, where time away from the real effort task is increasingly costly, removes the constraints of using any particular outside option and allows a researcher to use any outside option (e.g., catching a ball, surfing the internet, perusing a magazine or working on puzzles) appropriate for their specific design and vary that cost in any way they deem relevant.

Araujo et al. (2015) and Gill and Prowse (2019) discuss within versus between-subjects design, and Gill and Prowse (2019) suggests that the inconsistency in findings for between and within-subject designs is due to the increased statistical power for within-subjects designs. We show that the reason for the lack of a difference is not typically large standard errors but small differences between treatments. An alternative explanation for the difference is that the effect found within subjects is a simple demand effect. When one changes a wage in an experiment, subjects could see that as a signal that they should change something about their behavior. A between-subjects design eliminates the demand effect. Thus, our results, and those in the prior literature, support that the within-subjects design experiments likely suffer from an internal validity problem. Given that and the consistency with the theoretical analysis we provided, we would argue that conducting such tests within subjects does not fix the actual incentive problem from conducting real effort studies with no effort cost.

The overall point from this analysis is that the choice of real versus stylized effort in an experiment design is both simpler and more complicated than is often considered in prior work. It is simpler in the sense that if implemented properly, for most research questions, there should be little expectation of a difference in the results between real and stylized effort-based experiments. Our results suggest that the broad claims in the literature on these differences are not well-founded. Our second experiment also shows that properly modeling effort and costs is more complex than often appreciated, and more care needs to be taken in the design of real effort experiments to ensure that there is an actual cost for supplying effort. Without actual effort costs in a real effort design, researchers risk generating inappropriate conclusions to research questions.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Acknowledgements The authors would like to thank Jimmy Walker and Mark Isaac for providing the instructions from their early papers on public goods games, and Lise Vesterlund for providing the instruction scripts, software programs, and data we used in the design of one of our treatments. We would also like to thank participants at the ESA World Meetings in Vancouver, ESA North American meetings in Dallas and the Texas Experimental Conference at Rice for many useful comments and suggestions. All funding for this research was provided by our home universities. The replication material for the study is available at https://doi.org/10.3886/E208628V1

Funding Open access funding provided by the Carolinas Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., & Wilson, A. J. (2015). The effect of incentives on real effort: Evidence from the slider task. CESIfo Working Paper No. 5372.

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., & Wilson, A. J. (2016). The slider task: An example of restricted inference on incentive effects. *Journal of the Economic Science Association, 2*(1), 1–12.

Besley, T., & Ghatak, M. (2005). Competition and incentives with motivated agents. *American Economic Review, 95*(3), 616–636.

Bortolotti, S., Devetag, G., & Ortmann, A. (2009). Exploring the effects of real effort in a weak-link experiment. Working Paper.

Buckley, E., & Croson, R. (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics, 90*(4), 935–955.

Carpenter, J., & Huet-Vaughn, E. (2019). Real effort tasks. In A. Schram & A. Ule (Eds.), *Handbook of research methods and applications in experimental economics, Chapter 19* (pp. 368–383). Cheltenham: Elgar.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization, 149*, 74–87.

Charness, G., Masclet, D., & Villeval, M. C. (2014). The dark side of competition for status. *Management Science, 60*(1), 38–55.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics, 14*(1), 47–83.

Cherry, T. L., Krol, S., & Shogren, J. F. (2005). The impact of endowment heterogeneity and origin on public good vontributions: Evidence from the lab. *Journal of Economic Behavior & Organization, 57*(3), 357–365.

Clark, J. (2002). House money effects in public good experiments. *Experimental Economics, 5*(3), 223–231.

Coffman, K. B., Exley, C. L., & Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science, 67*(6), 3551–3569.

Cooper, D. J., & Saral, K. J. (2013). Entrepreneurship and team participation: An experimental study. *European Economic Review, 59*, 126–140.

Corgnet, B., Gómez-Miñambres, J., & Hernán-Gonzalez, R. (2015). Goal setting and monetary incentives: When large stakes are not enough. *Management Science, 61*(12), 2926–2944.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Corgnet, B., Hernán-González, R., & Schniter, E. (2015). Why real leisure really matters: Incentive effects on real effort in the laboratory. *Experimental Economics, 18*(2), 284–301.
Croson, R. T. (2001). Feedback in voluntary contribution mechanisms: An experiment in team production. In M. Isaac (Ed.), *Research in experimental economics* (Vol. 8, pp. 85–97). Bingley: Emerald Group Publishing Limited.
DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2022). Estimating social preferences and gift exchange at work. *American Economic Review, 112*(3), 1038–1074.
Engel, R. (2010). Why work when you can shirk? Worker productivity in an experimental setting. *Journal of Applied Business and Economics, 11*(2), 104–119.
Erkal, N., Gangadharan, L., & Koh, B. H. (2018). Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review, 101*, 528–545.
Esarey, J., Salmon, T. C., & Barrilleaux, C. (2012). What motivates political preferences? Self-interest, ideology, and fairness in a laboratory democracy. *Economic Inquiry, 50*(3), 604–624.
Fahr, R., & Irlenbusch, B. (2000). Fairness as a constraint on trust in reciprocity: Earned property rights in a reciprocal exchange experiment. *Economics Letters, 66*(3), 275–282.
Fehrler, S., & Kosfeld, M. (2014). Pro-social missions and worker motivation: An experimental study. *Journal of Economic Behavior and Organization, 100*, 99–110.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for readymade economic experiments. *Experimental Economics, 10*(2), 171–178.
Georg, S. J., Kube, S., & Radbruch, J. (2019). The effectiveness of incentive schemes in the presence of implicit effort costs. *Management Science, 65*(9), 3949–4450.
Gill, D., & Prowse, V. (2011). A novel computerized real effort task based on sliders. Working Paper.
Gill, D., & Prowse, V. (2019). Measuring costly effort using the slider task. *Journal of Behavioral and Experimental Finance, 21*, 1–9.
Gächter, S., Huang, L., & Sefton, M. (2016). Combining "real effort'' with induced effort costs: The ball-catching task. *Experimental Economics, 19*, 687–712.
Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior, 7*(3), 346–380.
Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economics Review, 92*(4), 1062–1069.
Isaac, R. M., & Walker, J. M. (1988). Group size hypotheses of public goods provision: An experimental examination. *Quarterly Journal of Economics, 103*, 179–199.
Isaac, R. M., Walker, J. M., & Thomas, S. H. (1984). Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice, 43*, 113–149.
Johnson, D., & Salmon, T. C. (2016). Sabotage vs discouragement: Which dominates post promotion tournament behavior? *Southern Economic Journal, 82*(3), 673–696.
Kamei, K., & Markussen, T. (2022). Free riding and workplace democracy: Heterogeneous task preferences and sorting. *Management Science*. https://doi.org/10.1287/mnsc.2022.4556
Kessler, J. B., & Norton, M. I. (2016). Tax aversion in labor supply. *Journal of Economic Behavior & Organization, 124*, 15–28.
Ku, H., & Salmon, T. C. (2012). The incentive effects of inequality: An experimental investigation. *Southern Economic Journal, 79*(1), 46–70.
Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton, NJ: Princeton University Press.
Mankiw, N. G. (2017). *Principles of microeconomics* (8th ed.). Cengage: Boston.
Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York: Oxford University Press.
McCaffrey, D. F., & Bell, R. M. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology, 28*(2), 169–182.
Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683.
Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior, 77*(1), 122–137.
Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press

Van Dijk, F., Sonnemans, J., & Van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review, 45*(2), 187–214.

Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review, 80*(1), 234–248.

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L., & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science, 345*(6192), 75–77.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1007/s10683-024-09844-y Published online by Cambridge University Press