Is "Real" Effort More Real?*

E. Glenn Dutcher†
Ohio University

Timothy C. Salmon‡
Southern Methodist University

Krista Jabs Saral§
Webster University Geneva

December 2015

Abstract

In recent years, a growing number of studies have researchers opting to use "real" effort designs for laboratory experiments where subjects complete an actual task to exert effort rather than using what is perhaps a more traditional design of stylized effort where subjects simply choose an effort level from a pre-defined set. The commonly argued reason for real effort is that it makes the results more generalizable and field relevant. Some researchers go further and make a distinction between trivial and useful real effort, i.e. whether the task is only relevant for the experiment or if it leads to tangible production for some purpose outside of the experiment, and claim that the useful effort model is even more likely to be generalizable. We present an experiment designed to test whether these three modes of effort, stylized, trivial, and useful, have any impact on behavior in a public goods setting. We find that all three forms of effort lead to identical decision making and then discuss how these results help to inform us about the use of real effort in laboratory experiments.

JEL Codes: C91, H41 **Key Words:** Real Effort, Stylized Effort, Abstract Effort, Economics Experiments, Public Goods

1 Introduction

There are many different forms of economic experiments in which researchers want experimental subjects to engage in tasks modeled after field tasks that involve expenditures of effort. This includes principal-agent games or other workplace related decision making environments in which a researcher wants to understand how different incentive structures or

*The authors would like to thank Jimmy Walker and Mark Isaac for providing the instructions from their early papers on public goods games which we used as the basis for our instruction scripts as well as Dan Houser for providing data from an earlier paper to which we compare our results.

†Ohio University, Department of Economics, Bentley Annex 3rd Floor, Athens, Ohio 45701. dutcherg@ohio.edu

‡Southern Methodist University, Department of Economics, 3300 Dyer Street, Suite 301 Umphrey Lee Center, Dallas, TX 75275-0496. tsalmon@smu.edu, Phone: 214-768-3547, Fax: 214-768-1821.

§George Herbert Walker School of Business and Technology, Webster University Geneva, Route de Collex 15, CH-1293 Bellevue, Switzerland. Email: kjsaral@webster.ch

environmental features affect effort decisions. It also includes many other standard environments such as coordination games, trust games, and public goods games. This makes the manner in which effort is modeled of key importance to the design of a very wide range of experiments.

Traditionally there are two different ways that one might consider modeling effort. We will refer to the first as "real effort" where subjects perform a task that involves some degree of actual effort and the second as "stylized effort" where a subject chooses some effort level from a given set of alternatives. The stylized effort design is essentially an analog of the structure found in theoretical models involving effort choice. In the stylized effort approach one might allow a subject to choose effort on the range of $[0, 10]$ with higher numbers resulting in a higher cost to the subject than lower numbers. The function defining the cost of any level of effort is therefore induced and under the control of the experimenter. This induced cost function is intended to capture the important aspects of actual physical and mental exertion in which a person chooses how much effort to exert and higher effort levels are assumed to involve higher mental or physical costs.

In a real effort design, the experimenter will have the subjects complete a task such as solving math problems which require actual physical or mental exertion. In these situations, a researcher assumes that there is some cost to the subject of completing the task but the nature of the cost function is unknown and typically not under the control of the experimenter.[1] To the extent that the induced cost function in a stylized effort design captures the cost function in the real effort task, one would expect similar behavior in the two environments. If, however, an induced cost function is not a good match with the cost function in the real effort task, then one should not expect comparable results. Also, if the cost function in the real effort task does not match the one assumed in a theoretical model of behavior, one should also not expect the behavior in the real effort task to match with the theoretical predictions. These points regarding the importance of matching cost functions across contexts are important and yet are commonly ignored in the literature on real effort experiments. They will be a central issue in what we discuss in the current study.

Early experiments that involved subjects engaging in effort were mostly done using the stylized effort design (extensive examples can be found in various literature reviews such as Kuhn and Charness (2011) for labor experiments, Ledyard (1995) and Chaudhuri (2011) for public goods, Devetag and Ortmann (2007) for coordination games, and Johnson and Mislin (2011) for trust games). Recently there has been a substantial shift towards experimenters preferring real effort designs. In attempting to understand the issues behind why one would choose one model or the other, we will first review some of the advantages and disadvantages claimed in the literature for the different ways of modeling effort in the lab.

The stylized effort approach has one clear advantage in its favor: control. By using the stylized effort approach one can establish a very tight connection between the experimental set-up and the underlying theory. For example, in a theoretical model of a principal-agent game, the agent is modeled as choosing an effort level and facing a cost function which translates the effort choice into a monetary equivalent cost. The stylized effort implementation in the lab is identical. This allows an experimenter precise control over the

[1] Gächter, Huang, and Sefton (2015) develop a task that combines real effort with controlled costs in such a way that they claim to achieve the benefits of both design approaches. Their results show that similar results are achieved between this task and stylized effort designs.

cost of a subject's effort which also allows for the ability to manipulate that cost function as needed. Ultimately this allows for the specification and testing of precise hypotheses regarding behavior. This tight connection with the theory also helps to develop a clear understanding for when and perhaps why individual behavior differs from those theoretical predictions. Another advantage of this approach is that the time commitment involved in these experiments is typically less than in real effort experiments which allows the researcher greater flexibility in the topics which can be covered in a typical experiment.

When moving to a real effort design, the cost function is uncontrolled by the experimenter. This loss of control leads to a diminished connection to theory. Consequently, one can often not make precise predictions regarding behavior which makes it more difficult to identify deviations from theoretical predictions and determine their nature. There are, however, claims about advantages from the real effort approach which explain why people use it. A good example of the claimed advantage is found in Kuhn and Charness (2011),"Concerning the objection that the labor task is abstract and artificial, there has been an increasing trend in 'real-effort' experiments...." The claim here is that the increase in the use of real effort designs is due to the fact that they help make the experiments less abstract and artificial. Similarly Van Dijk, Sonnemans, and Van Winden (2001) state that real effort "involves effort, fatigue, boredom, excitement and other affectations not present in the abstract experiments." Their conjecture is that subjects would be willing to work for more hours than if they give the equivalent amount of money to a charity - a conjecture which was supported in Brown, Meer, and Williams (2013). Corgnet, Hernán-González, and Rassenti (2011) further argue that a real effort design helps overcome a problem with laboratory experiments claimed by Falk and Heckman (2009) which is that "There is also a widespread view that the lab produces unrealistic data, which lacks relevance for understanding the 'real world'." This view is crystallized in Gill and Prowse (2011) when the authors state that "The main advantage of using a real effort task over a monetary cost function is the greater external validity of the experiment, which increases with how closely exerting effort in the task replicates the exertion of effort outside of the laboratory." While other studies using real effort designs don't always state this argument so explicitly, they usually make the same claim or a similar one implicitly which is that somehow the real effort specification is a better match with the field. One noticeable failing in most of these studies, though, is the failure to provide a clear argument for why this would be the case or to provide empirical validation for the claim. Both are important points to establish in order to better understand the issue of how to best model effort in an experiment.

As an attempt to establish those points, we can start by trying to understand the logical arguments for why real effort designs might be more externally valid. One argument for why a specific task in an experiment would generalize to the field would be that the cost function of that task shares important characteristics with the field task and these costs cannot be captured in a more controlled setting. To make such a claim one would have to put forward the argument that the cost function of the lab task, e.g. aligning sliders, has the same characteristics of the cost functions in the field tasks such as mail room workers sorting packages, lawyers taking a case to trial or doctors diagnosing patients. This does not appear to be the argument advanced in many papers in the literature possibly because it would be difficult to support. For example, Gneezy, Niederle, and Rustichini (2003) study gender differences in competitive environments where subjects solve mazes. There is

no discussion of what task in the field has a cost function that is well captured by solving mazes. Charness and Villeval (2009) also use a real effort task of completing anagrams to examine how age influences competitive preferences and again it isn't made clear what field competitive environment has the same effort cost function as constructing anagrams. Johnson and Salmon (2016) have subjects solve math problems to represent workplace promotion tournaments but do not provide any argument for why the cost function for solving math problems is a good proxy for that of any specific workplace behavior. While these real effort tasks all involve physical or mental exertion they are also still quite abstract in terms of how they represent the corresponding field situations. These papers are no different than others on this issue and the point is simply that while this seems to be the most obvious way one might support the external validity claim, few seem to use it as their rationale.

An alternative argument is that real effort tasks, by their nature of involving actual mental or physical exertion, are able to trigger certain types of behavior that a stylized design would not be able to. In Ku and Salmon (2012), the authors argue that the reason for choosing real effort over a stylized effort design is due to the belief that subjects might form more of an emotional connection to their effort choices if they represent real effort than if they are simply chosen from a number line. In Fahr and Irlenbusch (2000), subjects crack walnuts to earn their endowment in a trust game because they "wanted the working task to make the workers suffer to a certain extent in order to guarantee that they really felt entitled to the property rights." Again, these are just examples of a common theme found in the literature and while there is certainly intuitive logic to these claims, they are typically just asserted and are not tested.

Some researchers who are concerned about the abstract nature of even these real effort designs go even further and argue that there should be a distinction between what we will refer to as "trivial" versus "useful" real effort. Trivial real effort would be an experiment design in which the effort from the subjects is relevant only for the internal purposes of the experiment. Aligning sliders or solving math problems and mazes would all be examples of trivial effort as these tasks have no relevance outside of the laboratory. To circumvent this issue, other real effort tasks involve having subjects do things like stuff envelopes that will be used for official department business (Carpenter, Liati, and Vickery (2010)) or cracking nuts that are used in a grocery store's holiday cookies (Fahr and Irlenbusch (2000)). This effort is useful since the subjects are engaging in tasks that yield tangible output which is relevant outside of the laboratory. In these experiments, subjects are almost employees and so they could see it as a close analog to an actual job. This useful effort approach to modeling effort in the lab is a very small step from a field experiment and some argue that it is the best way to model effort in the lab. An important caveat is that if one claims that the only way to generalize to the field is to match the tasks precisely then this implicitly involves a claim that envelope stuffing experiments only apply to very similar settings in the field in which people engage in simple repetitive and menial office tasks. Thus this argument can actually be seen as an argument for limits on the generalizability of experiments to the field rather than arguing that experiments based on subjects shelling nuts generalize to a broader set of field situations.

We have demonstrated that while many prior studies advance claims regarding the different approaches to modeling effort in an experiment, the validity of those claims has

not been firmly established. In order to validate these claims, two empirical points must be established:

1. The type of effort in an experiment design has a direct effect on the behavior of subjects.

2. If the behavior in real effort is different than stylized effort, then the behavior generated by the real effort designs does a better job of matching with observed behavior in related field contexts.

The second point is perhaps the more important one but before investigating it, it is necessary to establish the validity of the first. There have been a few prior studies which have begun to address the first point but they have not been able to provide conclusive answers. Bortolotti, Devetag, and Ortmann (2009) examine a weak link game with real effort and find results quite different from many stylized effort experiments suggesting that indeed real and stylized effort yield different behaviors. However, the real effort task they use has a cost function with unknown properties which is unlikely to match the properties of the induced cost function for the stylized effort treatment. Consequently it isn't clear if the differences are due to the type of effort or due to differences in the cost functions. Brüggen and Strobel (2007) examine the issue in the context of gift exchange games and find no difference, but again the cost functions between the two environments are uncontrolled so whether there should be differences is unclear**.** Lezzi, Fleming, and Zizzo (2015) compare three different real effort tasks with an induced effort version and find substantial differences in how people respond to them. This is perhaps similar to results found in Dutcher (2012) where different results were obtained between two different real effort tasks in which one required creative thinking while the other did not. One might conclude that these studies provide conclusive evidence that real effort changes behavior since behavior changes across real effort tasks. That would only be a valid conclusion if the cost of effort functions between these tasks were identical. There seems little reason to believe that they were and so a better interpretation of those results is that they confirm our point made above regarding how different cost functions yield different results.

This now allows us to state very clearly the question that will be the subject of our study: *is there a difference in behavior between real effort and stylized effort experiments which can be attributed purely to the difference in the manner in which effort is modeled?*

We have chosen to investigate this question in the context of public goods games. There are multiple reasons why the public goods game is a useful context in which to test this issue. First, the stylized facts of public goods games with chosen effort have been replicated numerous times in the literature and are well known, see again Ledyard (1995) and Chaudhuri (2011). Thus we have a very strong baseline expectation of what behavior should be like in real effort versions of this game. Second, we will be able to use the context of the public goods environment to maintain careful control of the cost function across contexts, which we have already noted is vital to testing this issue. Third, one reason why real effort may lead to differential behaviors, as noted in prior literature, is that individuals may feel more attached to their earnings from the output of a real effort game and the public goods game provides an excellent vehicle to detect such differences. If individuals feel more entitled to their earnings in a real effort context then it seems likely that they will react more

harshly to group members that are contributing a lesser amount and therefore contributions might be lower in a real effort setting or they might decay faster. Further, the differences between trivial and useful effort may be highlighted here as well as engaging in useful effort could be seen as already providing some sort of public good, which could lead to individuals becoming more or less inclined towards cooperation.

Testing our question requires us to design an experiment in which we will vary the nature of effort provision across treatments between stylized, trivial and useful effort. Doing this requires us to design a public goods game that will have several novel elements compared to standard public goods games and standard real effort experiments. These elements are necessary to ensure that the only thing that varies between treatments is how effort is elicited from the subjects. In the next section we provide a detailed explanation of our experimental design, including an explanation regarding why each element is required.

2 Experiment Design

Our experiment is designed to determine if the nature of effort will have a significant impact on observed behavior. This turns out to be a non-trivial issue to investigate as switching from a standard stylized effort design to a standard real effort design involves changing a number of important aspects of the environment other than the nature of the effort itself, many of which could affect behavior. We will explain these issues as we explain some of the key aspects of the design.

Our design is based on the standard model of a VCM used in many laboratory experiments dating back to Isaac, Walker, and Thomas (1984) and Isaac and Walker (1988). As explained in Ledyard (1995) and Chaudhuri (2011), this basic design has been used to investigate many issues about public good provision and the base results have been replicated many times. Our design will use as its base the exact same incentive structure as these classic VCM designs and our treatments will involve varying how subjects receive the tokens they will be investing in the VCM, while attempting to control all other factors.

The core of all of the treatments possess the same incentive structure. Participants are randomly assigned to a group of 4 and remain matched for the duration of the experiment session. In each period of the session, individual subjects have 10 tokens and can invest these tokens into either an individual private account or a group account. For each token invested in the individual account, the participant earns \$0.20. For each token contributed to the group account, the group earns \$0.40. These group earnings are divided equally between all 4 group members (\$0.10 per group member), leading to a marginal per capital return (MPCR) of 0.5. At the end of the period, participants are provided with feedback that includes a reminder of their contributions to the individual account and group account, the total number of tokens donated by all members of the group to the group account, and a summary of their earnings for the period. This same process was repeated in every period for a total of 10 periods.

In order to address our research question we had to implement these incentives with both real and stylized effort. A typical approach to a real effort version of a public goods game might have players solving math problems or other real effort task and have the capability to

generate earnings either for their group or just for themselves.[2] One immediate difference between these designs and the standard VCM is that the production of the subjects in the real effort versions is unbounded. That would correspond to different subjects having a different number of tokens to invest in a stylized effort design and thus in addition to the change in how effort is modeled, these implementations also introduce heterogeneity in investment capability. The issue of differential ability or differential endowment has been investigated in the context of stylized effort designs (e.g. Cherry, Krol, and Shogren (2005); Buckley and Croson (2006); Reuben and Riedl (2013)) and not surprisingly these alterations to the environment can impact the results. For our purposes, we have to design our experiment to not introduce heterogeneity in this dimension.

Another difference between real and stylized effort experiments is the timing of the decisions. In the standard VCM, a subject must make a single decision about token allocation, i.e. choose how many tokens from 0-10 to contribute to the group account, and periods can go very quickly. In a real effort version, subjects have to spend time on the real effort task producing their tokens or their contributions to the accounts. The timing difference could lead to a person becoming more or perhaps less thoughtful over their contribution choices, which could lead them to either be more or less cooperative. This element must also be eliminated as a difference between treatments.

Real effort tasks may also differ from stylized effort at a cognitive level as engaging in the real effort task will trigger additional cognitive processes that could distract a participant from the underlying incentives of the VCM. While the directional impact of such a distraction is unclear, it seems quite clear that contribution decisions could be impacted. This means that comparisons of real effort to a stylized effort, where subjects only choose contributions and are not also engaging in other cognitive tasks, will have another confound which we seek to eliminate in our design.

We explain how our design dealt with these issues by describing each of the three treatments, beginning with what we will refer to as the Useful Effort (UE) treatment. In the UE treatment, subjects are engaged in a data entry task in which they enter actual financial data from Reuters. This data is an important component of a research project of another faculty member at the university where the experiments were conducted (not a co-author on this project). In the instructions we explain to the subjects very clearly that the data entry task is vital to the research of this faculty member and exhort them to be careful in their work.[3] This was an attempt to have subjects truly see this as useful effort and not some abstract real effort task necessary only for the experiment.

Subjects earned a token by entering in a five-letter fund ticker, a two-letter fund code,

[2] For example, Van Dijk, Sonnemans, and Van Winden (2001) had subjects solve two-variable optimization problems while Cooper and Saral (2013) use GMAT questions. In the first case, subjects were given two problems, task A and task B to work on. Effort on task B was analogous to donations to the individual account, while effort on task A was (in some treatments) similar to the group account. In the second, subjects were asked to donate their answers to GMAT questions to either the individual account or the group account.

[3] We thought carefully about whether and how to error check the entries, but all ways we came up with to do this caused other problems and confounds in the design. The possibility of errors in the data is a problem for the researcher who will be using the data and the capability to enter random nonsense could make the subjects take the task less seriously. Of course this is true of real data entry tasks as well so we decided to settle on not error checking in real time.

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%

Figure 1: Sample of data subjects would enter in the Useful Effort treatment.

and the 1-year, 5-year, and 10-year percentage returns from a sheet printed out with this information. An example of this data and how it is presented to subjects is shown in Figure 1. Each line of data would earn a single token and all subjects were required to enter 10 lines per period so that they would earn 10 tokens per period. They were not able to earn more than 10 tokens per period and they could not advance to a new period without earning all 10. This insured that all subjects had the exact same investment capability in each period. It is also important that they did not earn their tokens and then make their investment decision at the end. Subjects would make their investment choices while generating the tokens using a toggle button on their screen. At the beginning of a period they would have to switch the toggle to either the private or group account and then any tokens they produced would go toward the selected account. They could switch the toggle at any point during a period between the accounts which allowed them to create any split they desired of the tokens between the two accounts. This means that they were essentially choosing which account to work for while producing tokens.

We will refer to our next treatment as the Trivial Effort (TE) treatment and it is conducted identically in all aspects to the previously described UE treatment except that the data subjects enter is presented to them with no context. Subjects in the TE treatment were handed identical data sheets to those in the UE treatment but there was no mention that the data would be used for any external purpose. They were only told that the reason to enter the data is to earn the tokens. To accomplish this, two copies of each data sheet were printed where one went to someone in the TE treatment and one went to someone in the UE treatment.

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 required us to design this treatment to be substantially different from a standard VCM. In this treatment, subjects receive tokens without requiring any effort on their part and so there is no data entry task. In order to ensure that the timing issues were the same between this treatment and the others, subjects would still make their investment choices using the same toggle switch as in the other treatments but instead of earning the tokens through effort, the tokens would arrive at random intervals. The token arrival times were drawn at random from the actual distribution of subject times to complete a data entry line from the UE and TE sessions. The average length of time between tokens was 22 seconds, with a maximum of 73 seconds and a minimum of 8. Subjects would receive a warning that a token was going to be deposited into the selected account 3 seconds before it was to allow them to change the current account to which tokens were accruing if they wanted to.

These design choices were made to ensure the timing and the nature of the decisions were the same across all three treatments. In order to give those in the SE treatment something to do while the tokens appeared that could stimulate some sort of cognitive processes, we allowed them to play Tic-Tac-Toe against a computer opponent for no earnings. We made it very clear that playing this game was not connected to earning tokens and that there were no earnings related to playing. Subjects could however have been just as distracted by playing this task as engaging in the data entry tasks causing them to be distracted from the VCM incentives.[4]

The design of this experiment ensures that the costs of decisions are equivalent between treatments as the cost of working to generate a token to the group account is the earnings/utility given up that could have been generated by working to generate a token to the individual account and vice versa. One might argue that the real effort tokens are more expensive to acquire than the stylized effort tokens but that extra expense has no impact on the decision making margins for contributions as they wash out of the comparison. Further, the main component of cost in an experiment like this is probably the time cost of how long it takes to generate a token and this was common across all treatments. It is in this way that we have ensured costs are common between treatments. What our design does not do is guarantee equivalent utility functions across treatments. As we discussed above, the most compelling argument for why real effort might generate different behavior than stylized effort is the possibility that earning the tokens through real effort could impact the utility functions the subjects have and in this case that would appear through their willingness to be cooperative and such depending on how the tokens are earned. Thus it is through this possible channel that one might expect differences in behavior to emerge.

We conducted 2 sessions of all treatments. Groups of 4 were randomly formed at the start of each session and these groups remained constant throughout the session. All subjects were students at Ohio University and the experiment was programmed using Z-tree software, Fischbacher (2007). Table 1 provides information on the average earnings and the number of subjects who participated in each treatment.

	Useful Effort	Trivial Effort	Stylized Effort
Average Earnings (USD)	\$30.37	\$31.84	\$30.91
Number of Subjects	28	32	28

Table 1: Earnings and number of subjects by treatment.

3 Results

We begin the analysis of the results with a first look at the data in Figure 2 which shows average investment levels into the group account by period for all three treatments. Since our SE treatment is substantially different than a more traditional stylized effort VCM we also include data in the figure from two prior studies which use the more traditional design, Croson (2001) and Houser and Kurzban (2002), and have the same parameterization as

[4] 66% of subjects played at least one game of tic-tac-toe. The average number of games played in a period was 11.

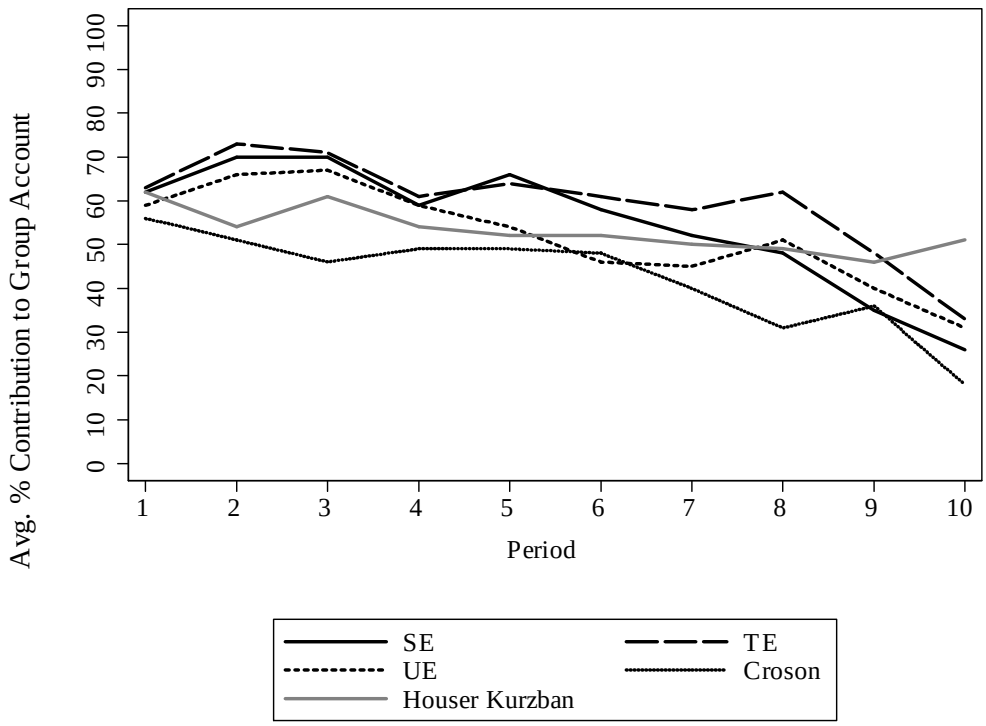

Figure 2: Average the group account by period over all 10 rounds.

our design (MPCR = 0.5, n = 4, 10 periods, partner matching). The figure shows that the results from all three of our treatments are very similar to each other and that all three show the standard pattern observed in other VCM experiments, i.e. contributions starting at a moderately high level and then decaying over time. Despite the design of our experiments being very different, it is also interesting to see that our data fall right in-line with the results observed in the more traditional stylized effort designs from Croson (2001) and Houser and Kurzban (2002). We also provide Table 2 which contains the values from only our data of the average contributions to the group account and their standard deviation for selected rounds.[5] The table provides additional evidence of the similarity between treatments - not only are the average levels of contributions similar between treatments but so too are the standard errors.

Given that contributions are not independent over time, we cannot test any formal hypotheses on these averages as any such test would be incorrectly specified. We instead provide several different regression approaches for dealing with the interactions between group members and over time in Table 3. The regressions are random effects panel regressions where the dependent variable is the amount contributed to the group account. To account for the repeated observations across subjects, errors are clustered at the subject level. Observations are also not entirely independent across group members and so one might also want to cluster at the group level. The data contains only 22 groups which

[5] We provide the same data for all rounds in the appendix.

Treatment	Period 1	Period 5	Period 10	Overall
SE	6.21	6.61	2.64	5.45
	(2.83)	(3.50)	(3.46)	(3.73)
TE	6.25	6.38	3.25	5.92
	(2.71)	(2.93)	(3.38)	(3.41)
UE	5.92	5.86	3.11	5.19
	(3.52)	(3.58)	(3.34)	(3.62)
All Treatments	6.14	6.14	3.01	5.54
	(3.00)	(3.32)	(3.36)	(3.59)

Table 2: Mean and standard deviations of contributions to the group account by treatment.

means that clustering at the group level may not satisfy the asymptotic properties of the estimator. As such, we don't choose this as our default specification but we do provide the results for specifications using the Bias-Reducing Linearization (BRL) procedure (McCaffrey and Bell (2002)) to cluster at the group level which corrects for the small number of clusters. The qualitative results are unchanged by this alternative specification. One might also be concerned about the potential for a censoring problem with the data since subjects could not contribute more than 10 (25% of token allocation decisions to the group were at the maximum of 10) or less than 0 (18% of token allocation decisions to the group were at the minimum of 0, which is the dominant strategy prediction). We therefore also include the results of a Tobit estimator with errors clustered at the subject level in the appendix and again, the qualitative results are unchanged.

Result 1 *There are no statistically significant differences in contributions between treatments.*

Each of the regression specifications include dummy variables for the TE and UE treatments. These coefficients allow us to test whether the average contributions are different between 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 to the group account in TE and UE are different. A post-estimation Wald test yields a $p-$value of 0.24 indicating that those two coefficients are also not significantly different from each other.

We also test whether there are differences over time. Figure 2 indicated that the time paths look similar, but to test this observation formally the second specification in Table 3 includes a time trend and interactions between that time trend and the treatments. This yields the second result.

Result 2 *There are no statistically significant differences between treatments in how contributions adjust across time.*

In this specification, the Period variable is negative and significant indicating that, as in most other similar VCM data, there is a decay in contributions over time in the SE treatment. The interactions between Period and the two other treatments are insignificant

	(1)	(2)	(3)
TE	0.465	-0.175	-0.321
	(0.602)	(0.825)	(0.865)
UE	-0.268	-0.712	-0.497
	(0.692)	(0.973)	(1.018)
Period		-0.423***	
		(0.089)	
Period∗TE		0.116	
		(0.112)	
Period∗UE		0.081	
		(0.117)	
Group_{t-1}			0.522***
			(0.121)
Group_{t-1}∗TE			0.096
			(0.149)
Group_{t-1}∗UE			0.076
			(0.198)
Constant	5.454***	7.781***	2.355***
	(0.474)	(0.692)	(0.638)
Obs (Groups)	880 (88)	880 (88)	792 (88)

Clustered robust standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 3: Random effects panel regressions concerning the investments into the Group account.

and of the same sign indicating that the rate of decay is not different for any of our treatments. Post-estimation Wald tests confirm no difference in the decay (the interaction terms, $p = 0.72$) or the treatments after accounting for the decay (the binary variables for the UE and TE treatments, $p = 0.51$) for the two real-effort treatments.[6]

The last regression specification examines how individuals respond to the average level of investment by their group members from the previous period. Many prior studies conclude that subjects often engage in conditional cooperation in VCMs; they are willing to cooperate if others also cooperate but contributions will decline if they do not see others contributing at their expected level. Arifovic and Ledyard (2012) demonstrate that this behavior can yield the standard decay pattern we observe over time. This specification includes a lag of the average contribution by an individual's group members (Group_{t-1}) to determine if we can see any differences between treatments regarding how individuals react to the contribution levels of others.[7] This leads to our third result.

Result 3 *There are no statistically significant differences between treatments in how contributions adjust to contributions by group members.*

Consistent with conditional cooperation behavior, the significance of the lagged variable indicates that subjects do adjust their contributions based on the contributions of others. The interaction terms, however, are all insignificant which again confirms that there are no differences between treatments in regard to how individuals react to the contributions of others. Post-estimation Walt tests support the lack of a difference between the TE and UE coefficients ($p = 0.86$) or the interaction effect ($p = 0.91$) between the two real-effort treatments.

4 Conclusion

Our goal was to determine if the manner in which effort is simulated in a lab experiment has significant behavioral effects. There are many claims in the literature that making the effort task seem more field-like should enhance the external validity of the results, making them more credible. For that to be true, it should be the case that making the effort more field-like should also have some direct impact on behavior. While the claim regarding external validity has been asserted many times in the literature, there is little in the way of empirical evidence to support it. Our results certainly do not provide such evidence.

We find that all three ways we model effort provision in the lab, stylized, trivial and useful effort, yield identical outcomes. Not only do we find the standard comparative static results common to all VCM experiments across treatments, i.e. initially moderate to high contributions to the public account which decline over time, but the levels of contributions across all three effort designs are indistinguishable from each other. Given these results, it seems difficult to advance a claim that the outcome of an experiment designed on the basis

[6] We do not include the results here, but Result 2 is robust to period by period regressions to determine if we could find differences in any single period. We find no differences between treatments even in period by period regressions.

[7] 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}.

of any of the three effort models should be considered as inherently more or less externally valid or field relevant than any of the others.

We should of course be careful about how our results are interpreted. Our results do not suggest that we should expect any real effort experiment and any stylized effort experiment to yield identical results. There are typically many differences between real and stylized effort designs other than the nature of how effort is modeled which can impact the behavior. Most importantly, the underlying cost function can differ substantially between real and stylized effort designs with the cost function in the real effort design generally unknown and sometimes unknowable. There are also several other potential confounds in comparisons of real effort to stylized effort that our experiment was designed to eliminate, such as the differences in timing of choices and the degree to which subjects were distracted by other tasks. What our results strongly suggest is that should someone observe a difference in behavior between a real effort and stylized effort environment that it is not the nature of the effort itself which is driving the differences but rather it is these others differences, such as differential cost functions, which are driving the behavioral differences.

These results should also be helpful to understand when real effort and stylized effort models might best be used. First, one should not decide between them on the false assumption that one is necessarily more field relevant than the other. A more reasonable basis for selecting between designs has to do with issues concerning the nature of the cost function in the real effort design and the research question being addressed. If there are elements of the cost function for the real effort task which cannot be replicated in an induced cost function yet are important to the issues being tested, then it certainly makes sense to use a real effort design. This might involve situations in which an experimenter is trying to estimate some properties of the cost function or if there are some sort of demographic differences in cost functions of a certain type that are a key interest in the research question. In the latter case, one has to be very careful about the real effort task selected as different ones will have different properties and generate different demographic interactions. The real effort task must be therefore chosen due to it possessing properties very similar to the field situation of interest and it must also be made clear the domains to which those results do and do not apply. On the other hand, should one of these specific issues not be a core element in the research question of the study, there does not seem to be a compelling argument to use a real effort design due to the substantial control lost in doing so.

A valid question regarding our results is the degree to which they might transfer to other games. The trust game might be an important one to consider as it differs in important ways from this public goods model. Games like this add in another dimension that we might refer to as real versus stylized consequences. In a real effort trust game, one might think to implement it by paying subjects different flat wages to engage in a task (e.g. stuffing envelopes) to determine if they respond to high level "gift" wages with higher effort. In this case it may well matter substantially what entity is receiving the benefits of the labor because in addition to the subjects possessing some unknown cost for completing the task they also may receive utility due to some unknown utility function for completing the task on behalf of that entity. One could refer to the situation in which the benefit of the labor accrues to some actual entity as one with real consequences. It would certainly be possible to conduct an experiment in which those consequences are modeled in a stylized manner by inducing utility from the benefits of the labor by paying the subjects money based on the

labor and again, to the extent that the induced utility function matches the homegrown one then similar results should be expected. Which approach one uses here depends on what you want to test. If you want to examine the nature of these unknown utility functions then you would want to use a real consequences design. If, however, you wanted to understand how subjects respond to different versions of those consequences you could conduct the stylized consequences design. While a very different context, an example of both approaches is shown in Isaac, Pevnitskaya, and Salmon (2010) in which the authors investigate charity auctions in which the preferences for the charities are induced in a stylized manner to determine how well the behavior matches with the theoretical predictions and then in other treatments the preferences are induced by having the revenue go to an actual charity to determine if behavior changes in some way based on the homegrown utility functions. The relevant issues here turn out to be very similar to the issues we've investigated above regarding real versus stylized effort and, in our view, the same design principles should translate between contexts.

The results from this study do not provide the answers to all relevant questions regarding where, when, and why someone should use a real or stylized effort design in an experiment but we believe that these results do help to place the relevant questions into proper perspective and help to provide some guidance on important aspects of the question. As with all elements of an experimental design, one must pay careful attention to how the specific design choices one makes affect the results and thought must be given to how reflective those design choices are of the situation to which you wish to apply the results. Our view is that both stylized effort and real effort designs can and should be used with the nature of the underlying research question determining which is preferred. Also, both approaches should be considered to have equal external validity and generalizability though of course careful attention must always be paid to how broadly one attempts to generalize the results of any data driven exercise.

References

Arifovic, J. and J. Ledyard (2012). Individual evolutionary learning, other-regarding preferences and the coluntary contributions mechanism. *Journal of Public Economics 96*, 808–823.

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

Brown, A. L., J. Meer, and J. F. Williams (2013). Why do people volunteer? an experimental analysis of preferences for time donations. Working Paper.

Brüggen, A. and M. Strobel (2007). Real effort versus chosen effort in experiments. *Economics Letters 96*(2), 232–236.

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

Carpenter, J., A. Liati, and B. Vickery (2010). They come to play supply effects in an economic experiment. *Rationality and Society 22*(1), 83–102.

Charness, G. and M. C. Villeval (2009). Cooperation and competition in intergenerational experiments in the field and the laboratory. *The American Economic Review 99*(3), 956–978.

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., S. Krol, and J. F. Shogren (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.

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

Corgnet, B., R. Hernán-González, and S. Rassenti (2011). Real effort, real leisure and real-time supervision: Incentives and peer pressure in virtual organizations. Working Paper.

Croson, R. T. (2001). Feedback in voluntary contribution mechanisms: An experiment in team production. In M. Isaac (Ed.), *Research in Experimental Economics*, Volume 8, pp. 85 – 97. Emerald Group Publishing Limited.

Devetag, G. and A. Ortmann (2007). When and why? a critical survey on coordination failure in the laboratory. *Experimental Economics 10*(3), 331–344.

Dutcher, E. G. (2012). The effects of telecommuting on productivity: An experimental examination. the role of dull and creative tasks. *Journal of Economic Behavior & Organization 84*(1), 355–363.

Fahr, R. and B. Irlenbusch (2000). Fairness as a constraint on trust in reciprocity: Earned property rights in a reciprocal exchange experiment. *Economics Letters 66*(3), 275–282.

Falk, A. and J. J. Heckman (2009). Lab experiments are a major source of knowledge in the social sciences. *Science 326*(5952), 535–538.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox For Readymade Economic Experiments. *Experimental Economics 10*(2), 171–178.

Gill, D. and V. Prowse (2011). A novel computerized real effort task based on sliders. Working Paper.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics 118*(3), 1049–1074.

Gächter, S., L. Huang, and M. Sefton (2015). Combining "real effort" with induced effort costs: The ball-catching task. Working Paper.

Houser, D. and R. Kurzban (2002). Revisiting kindness and confusion in public goods experiments. *American Economics Review 92*(4), 1062–1069.

Isaac, M., S. Pevnitskaya, and T. C. Salmon (2010). Do preferences for charitable giving help auctioneers? *Experimental Economics 13*(1), 14–44.

Isaac, R. M. and J. M. Walker (1988). Group Size Hypotheses of Public Goods Provision: An Experimental Examination. *Quarterly Journal of Economics 103*, 179–199.

Isaac, R. M., J. M. Walker, and S. H. Thomas (1984). Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice 43*, 113–149.

Johnson, D. and T. C. Salmon (2016). Sabotage vs discouragement: Which dominates post promotion tournament behavior? *Southern Economic Journal* –(–), –.

Johnson, N. D. and A. A. Mislin (2011). Trust games: A meta-analysis. *Journal of Economic Psychology 32*, 865–889.

Ku, H. and T. C. Salmon (2012). The incentive effects of inequality: An experimental investigation. *Southern Economic Journal 79*(1), 46–70.

Kuhn, P. and G. Charness (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4A, pp. 229–330. Amsterdam: Noth Holland.

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, pp. 111–194. Princeton, New Jersey: Princeton University Press.

Lezzi, E., P. Fleming, and D. J. Zizzo (2015). Does it matter which effort task you use? a comparison of four effort tasks when agents compete for a prize. Working Paper.

McCaffrey, D. F. and R. M. Bell (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*(2), 169–182.

Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior 77*(1), 122–137.

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

5 Appendix

	(1) BRL1	(2) Tobit1	(3) BRL2	(4) Tobit2	(5) BRL3	(6) Tobit3
TE	0.465	0.690	-0.175	-0.334	-0.908	
	(0.932)	(1.035)	(0.802)	(1.449)	(1.350)	(1.874)
UE	-0.268	-0.427	-0.712	-1.050	-0.690	-1.124
	(1.058)	(1.192)	(0.800)	(1.701)	(1.649)	(2.040)
Period			-0.423***	-0.686***		
			(0.111)	(0.170)		
Period∗TE			0.116	0.190		
			(0.155)	(0.199)		
Period∗UE			0.0807	0.120		
			(0.148)	(0.213)		
Group_{t-1}					0.178***	0.324***
					(0.064)	(0.076)
Group_{t-1}∗TE					0.063	0.064
					(0.069)	(0.097)
Group_{t-1}∗Useful					0.038	0.080
					(0.103)	(0.121)
Constant	5.454***	5.733***	7.781***	9.461***	2.285*	-0.0914
	(0.674)	(0.808)	(0.354)	(1.241)	(1.284)	(1.533)
Obs (Groups)	880 (22)	880 (88)	880 (22)	880 (88)	792 (22)	792 (88)

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

BRLX: Biased-Reduced Linearization clustering at the group level to correct for the small number of clusters.

TobitX: Tobit with bounds [0,10] with standard errors are clustered at the subject level..

Trt	Pd 1	Pd 2	Pd 3	Pd 4	Pd 5	Pd 6	Pd 7	Pd 8	Pd 9	Pd 10	Overall
SE	6.21 (2.83)	6.96 (2.87)	6.96 (3.20)	5.86 (3.95)	6.61 (3.50)	5.75 (3.84)	5.21 (3.90)	4.79 (3.97)	3.54 (3.45)	2.64 (3.47)	5.45 (3.73)
TE	6.25 (2.71)	7.31 (2.79)	7.06 (3.23)	6.13 (3.47)	6.38 (2.93)	6.09 (3.41)	5.78 (3.63)	6.19 (3.19)	4.75 (3.79)	3.25 (3.37)	5.92 (3.41)
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	6.14 (3.00)	6.99 (3.01)	6.91 (3.24)	5.95 (3.62)	6.14 (3.32)	5.52 (3.63)	5.19 (3.64)	5.39 (3.54)	4.14 (3.70)	3.01 (3.36)	5.54 (3.59)

Means and standard deviations of contributions to the Group account by period.