REAL EFFORT TASKS

Jeffrey Carpenter[1] and Emiliano Huet-Vaughn[2]

Abstract

We review the use of real effort tasks in economic experiments. To this point, the paradigm has been mostly used to model principal-agent relationships in the laboratory, the focus of our review. We first discuss the rationales for choosing between real and chosen effort when designing an experiment. To facilitate this discussion, we present a taxonomy of the common tasks that people have used, discuss some issues to keep in mind when implementing a real effort task and then survey the limited literature that compares the two methods. We end by offering a few recommendations on topics that could use additional investigation.

Keywords: real effort, chosen effort, principal-agent, incentive, intrinsic motivation, experiment.

[1] Department of Economics, Middlebury College. jpc@middlebury.edu.
[2] California Policy Lab, UCLA. ehuetvaughn@ucla.edu.

1. INTRODUCTION

The canonical principal-agent setting used to model the incentives embedded in various employment relationships like fixed and variable pay, bonuses and tournaments and even co-ops and team production has become a workhorse in the economic theory of contracts, personnel economics and labor supply, more generally. When using experiments to test various aspects of this theory, two frameworks have dominated: "chosen effort" and "real effort".

In the typical chosen effort experiment, preferences are induced, as discussed in Smith (1976) and Smith (1982). What this means is that, in addition to providing monetary incentives, often based on the amount of effort chosen, all the costs associated with providing effort are also explicitly monetary, instead of being paid in "blood, toil, tears and sweat." In most cases, these experiments can be presented in tabular form. The participant weighs the explicit monetary benefits and costs of choosing an effort level between some minimum allowed and a maximum. Examples of this design include the canonical gift exchange experiment of Fehr et al. (1993), the influential tournament experiments in Nalbantian and Schotter (1997) and the more recent team production experiment discussed in Carpenter and Dolifka (2017). What is key is that in the chosen effort paradigm, no actual effort is expended to achieve an outcome, and, instead "effort" levels are chosen from a menu.

By contrast, in real effort experiments, participants must actually work, be it manual, mental, or both to some degree, to achieve outcomes. Thus, unlike in chosen effort experiments, participants actually experience exerting effort. Though the earliest real effort experiment we know of is Wyatt (1934), an illustrative early example of the method is Swenson (1988) who set out to test the Laffer-curve conjecture that to maximize tax revenue, one must account for the labor supply response of workers. Participants were paid a one cent piece rate to repeatedly press the "!" key and then the "enter" key – the real effort task. Their final payoff was their net of tax earnings minus their unobserved effort cost. Another important feature of this experiment, something that we will return to below, is that instead of working, the participants in Swenson's experiment could either flip through current editions of popular magazines, try their luck at a video game version of "Concentration" or play a trivia card game. In other words, there was some opportunity cost to working.

In the remainder of this chapter, we will discuss the rationales for choosing between real and chosen effort when designing a principal-agent experiment. To facilitate this discussion, we present a taxonomy of the common tasks that people have used, discuss some pointers to keep in mind when implementing a real effort task and then survey the limited literature that compares the two methods. We end by offering a few recommendations on topics that could use additional investigation.

2. THE PROS AND CONS OF REAL EFFORT

What are some of the standard arguments used by experimenters when deciding on chosen versus real effort? In this section, we discuss some of the most common rationales offered for using real effort and a few things to think twice about.

Considering the "pros" of implementing a real effort task, perhaps the most often cited benefit is external validity. However, this term is often misused and the true benefit, in psychological terminology, is "realism," in this case, the poorly termed sub-classification, "mundane realism," to be exact Aronson and Carlsmith (1968). Yes, using a real effort task may facilitate running the same experiment with different subject populations or in different contexts wherein participants may be less familiar with the tabular representation of data or math, more generally, but extending the external validity of a result in this way is not unique to real effort experiments. The same could be done in carefully constructed chosen effort protocols. Instead, real effort enhances the mundane realism of the experiment because performing an actual work task is considerably more similar to what participants do outside the laboratory in the context in which we are interested – the workplace.

A second perceived benefit of using a real effort task, one that is in the same vein as realism, has to do with important elements of the workplace that are often assumed away in the chosen effort environment. As van Dijk et al. (2001), one of the earliest real effort experiments in this literature, point out, in real effort experiments it seems more easy and natural to incorporate the social dimension of work, the fact that people spend time interacting with others. Further, work "involves effort, fatigue, boredom, excitement and other affections not present in abstract experiments."

The concern is that the induced cost of effort function in a chosen effort design may not accurately capture these other phenomena that are important determinants of work. A proponent of chosen effort would likely respond that this objection can be overcome by selecting the right cost of effort function to aggregate the different phenomenon incentivizing work. But, to put it a bit boldly for rhetorical purposes, you might simply be kidding yourself as an experimenter to think that you have selected the right cost of effort function and actually induced preferences. Surveying the literature suggests there are many examples of factors that have not been induced affecting decisions and outcomes. The obvious example is that of social preferences. This very large literature at the heart of behavioral economics is founded on experiments like Gueth et al. (1982), Kahneman et al. (1986) and Berg et al. (1995) that show that the induced preferences do not predict choices. In fact, this literature is vast and robust precisely because it is hard to induce preferences completely. This potential disconnect between induced preferences and actual ones, makes it difficult to know what to make of chosen effort results that conform with this or that theoretical prediction, as one has to be careful when extrapolating from the behavior participants exhibit while fulfilling an induced role to what participants would naturally do while inhabiting the role as themselves.

More broadly, the debate between chosen and real effort often has the same flavor as the dichotomy surrounding reduced form and structural analysis. Assigning a specific utility function to participants allows for the structural estimation of treatment effects but, as is often said of this form of analysis, the assumed utility function may be unrepresentative of the actual preferences that govern real-world work. Likewise, a common utility function ignores the important heterogeneity of preferences that might be at the core of explaining the phenomenon under scrutiny.

Lastly, though rarely mentioned explicitly, there is also some sense expressed in existing work that real effort experiments are just more engaging and if the experimenter worries about maintaining control, engaged participants are clearly better. While this may be true “on average” it is not hard to find counterexamples. For instance, would participants be more engaged by playing something resembling a role-playing board game or typing “!” and “enter” over and over again?

Taking all this into account, many researchers have begun to see real effort as the preferred way forward; however, there are some shortcomings of the method or, if not shortcomings, things

of which researchers should be mindful. Most importantly, one has to be careful to not lose control of the experiment when implementing a real effort task. If done correctly, however, there should be little or no tradeoff between realism and internal validity.

The perception of a validity tradeoff arises for the following reason. Assuming preferences can be induced completely (despite the argument made in the previous paragraphs), chosen effort experiments are attractive to many experimenters because chosen effort allows one to control the important first-order elements of the principal-agent problem (e.g., payments and the cost of effort) and make irrelevant the unobserved factors such as ability that may affect performance on a real effort task. In brief, chosen effort experiments seem "cleaner" and more "powerful" in a modest sample than in real effort experiments where the cost of effort function is not assigned and underlying parameters for ability and other possibly confounding factors can vary by worker. Actually, however, all this really means is that you should err on the side of a larger sample with a real effort experiment to increase power and, most importantly, be very careful to randomize participants to the treatments (Note: randomization is discussed further in Chapter II.2).

Consider, for example, experiments in which real effort is used to make participants feel more entitled to their endowments (e.g., Thaler and Johnson 1990; Hoffman et al. 1994; or Carpenter et al. 2014). Rather than just being given the endowment, they have to earn it. Participants will, undoubtedly have heterogeneous unobserved abilities for whatever task is used and as a result they will also have endowments that vary due to this factor. Therefore, any analysis is likely to suffer from omitted variable bias. If the experimenter is diligent about randomization, however, unobserved ability will be orthogonal to the treatments and, while it might affect endowments, ability will not affect the assessment of the treatment effects. Another strategy to mitigate the effect of unobserved ability, as used in Ball et al. (2001), is to obfuscate the link between real effort and treatment assignment. Although it may seem like doing well on a trivia quiz, for example, will put you in the "high status" treatment, status is actually assigned randomly (e.g., treatment assignment might be a function of observed effort and a large random component). The obvious problem with this alternative is that you have to be careful to not skate too closely to the thin ice of deception.

Another thing that practitioners worry about is that participants may be "intrinsically motivated" to work hard in real effort experiments (though sometimes allowing for intrinsic

motivation in the work task is the point). As detailed in Deci and Ryan (1985), participants may be driven by an inherent interest or enjoyment in the task itself, yet another unobserved factor. If participants all work hard, regardless on the incentives, because they are intrinsically motivated or feel some duty to do so or work hard to please the experimenter (i.e., experimenter demand effects discussed in Chapter V.5), then treatment effects will be underestimated. Again, however, diligent randomization should mitigate this problem too as it will leave intrinsic motivation orthogonal to the treatment arms. Unless intrinsic motivation interacts with the treatments (as in the "control aversion study of Falk and Kosfeld (2006) for example), people may all work hard in some tasks and less hard in others but this baseline level of effort should be the same across treatments. Given the weight of this concern among practitioners (e.g., DellaVigna and Pope 2017 or Erkal et al. 2018), we return to the question of sufficient output variation when tasks are intrinsically motivating in Section 3.

A final, related, concern to that of intrinsically motivated participants is that in some cases without knowing the parameters of participants' actual utility functions, it may be hard to calibrate the incentives of a real effort experiment – setting piece rates, for example is often a "crap shoot." Should you pay them one cent per keystroke, ten cents or ten dollars? Imagine that although participants have heterogeneous costs of effort in a given task, the functions are all relatively flat (despite being increasing and convex). Without knowing this, it could easily be the case that piece rates or other marginal incentives are set too high or too low and everyone either works as hard as possible or as little as possible. Unless the experimenter can identify the incentive "sweet spot" treatment effects will be artificially negligible by design. A similar argument is made in Araujo et al. (2016) and we return to this point in Section 4.

3. A TAXONOMY OF REAL EFFORT TASKS

Say you've decided to go with a real effort design. The first thing to consider is the question of which task to implement. In this section, we aim to provide an overview of the common real effort tasks employed in the existing literature and to discuss some of the relevant considerations for a researcher when choosing among these alternatives. In approaching our review, we focused on seminal works using real effort designs in economics and provided greater attention to works of the last two decades (for an earlier survey on related topics that include a fair number of real effort

experiments, see Camerer and Hogarth (1999). The review benefited greatly from compilations of papers by Christina Gravert via her recent query to the Economic Science Association discussion list and Juan Andrade-Vera, our diligent research assistant. While there are most certainly tasks and studies missing from our survey of the literature, we think the 92 real effort papers reviewed give readers a more or less representative sense of the existing options before them.

Table 1 groups the real effort tasks employed in the surveyed papers into common categories. They run the gamut from rather mindless, motor tasks (counting, moving a slider on a computer screen, typing and data entry, envelope stuffing) to the more cognitively challenging (arithmetic, decoding, maze-solving). Considering arithmetic, a representative example is the study of Niederle and Vesterlund (2007) in which participants earn money by correctly summing five two-digit numbers. Summing numbers is extremely common, though some studies opt for multiplication instead (e.g., Blumkin et al. 2012). Overall, arithmetic accounts for 21% of the papers we considered.

Another popular category, comprising 15% of the sample is clerical tasks. One benefit of these tasks is that they tend to be more representative of what entry-level clerical staff would do for a real employer. In these experiments, participants typically stuff envelopes, sort things or do data entry. A more concrete example of this type of task comes from Linardi and McConnell (2011) who asked participants to do internet searches for educational resources and input the results into a database that would be used by tutors of homeless children.

The next category is what we call computer tasks. In 13% of the studies we reviewed, experimental subjects were asked to do some computerized task designed specifically for the experiment. They clicked on a box as it moved across the screen, they "caught" balls as they fell down the screen and they dragged a ball around the computer screen. However, by far the most utilized computer task is the "slider task" created in Gill and Prowse (2012). Participants in the slider task are confronted with a computer screen full of 48 sliders and they are asked to move the indicator on each slider to exactly the middle of the line.

Table 1: A typography of common real effort tasks (1997-2016)

Task	Canonical example	Frequency in our sample (overall)	Frequency (1997-2012)	Frequency (2013-2016)	Is production typically useful?	Is production intrinsically interesting?
Arithmetic	Niederle and Vesterlund 2007	19 (21%)	23%	20%	No	No
Clerical	Linardi and McConnell 2011	14 (15%)	15%	15%	Yes	Yes
Computer	Gill and Prowse 2012	12 (13%)	7%	18%	No	No
Counting	Abeler et al. 2011	10 (11%)	3%	18%	No	No
Decoding	Sillamaa 1999a	10 (11%)	8%	14%	No	No
Puzzle	Charness and Villeval 2009	18 (19%)	31%	10%	No	Yes
Typing	Greiner et al. 2011	6 (7%)	12%	2%	No	No
Other	Fahr and Irlenbusch 2000	3 (3%)	3%	3%	No	Depends

Note: List of the 92 studies included in the table: Abeler et al. 2011; James Alm 2012; Ariely 2008; Augenblick et al. 2015; Azar 2015; Barr et al. 2016; Bartling et al. 2009; Belot and Schröder 2013; Berger and Pope 2011; Bhui 2016; Blumkin et al. 2012; Bruggen and Strobel 2007; Cadsby et al. 2013; Calsamiglia et al. 2013; Carpenter and Gong 2016; Carpenter et al. 2010; Cason et al. 2010; Charness and Villeval 2009; Charness et al. 2013; Charness et al. 2016; Chaudhry and Klinowski 2016; Corgnet 2012; Corgnet et al. 2011; Corgnet et al. 2015a; Corgnet et al. 2015b; Dasgupta and Mani 2015; Dasgupta et al. 2015; DellaVigna et al. 2016; Dickinson 1999; Dickinson and Villeval 2012; Dohmen and Falk 2011; Douoguih 2011; Dutcher 2012; Dutcher et al. 2016; Ellis et al. 2016; Eriksson et al. 2009; Erkal et al. 2011; Fahr and Irlenbusch 2000; Falk and Ichino 2006; Fan and Gómez-Miñambres 2016; Fehr 2016; Gaechter et al. 2016; Gerhards and Gravert 2015; Gill and Prowse 2012; Gneezy and Rustichini 2000; Gneezy et al. 2003; Goldstein and Hogarth 1997; Greiner et al. 2011; Gupta et al. 2013; Hargreaves Heap et al. 2016; Healy and Pate 2011; Hennig-Schmidt et al. 2010; Heyman and Ariely 2004; Hogarth and Villeval 2014; Huang and Murad

2016; Imas 2014; Ivanova-Stenzel and Kübler 2011; Jones and Linardi 2014; Kessler and Norton 2016; Kidd et al. 2013; Koch and Nafziger 2016; Konow 2000; Kraut et al. 2011; Kuhn and Villeval 2013; Lefgren et al. 2016; Linardi and McConnell 2011; Niederle and Vesterlund 2007; Noussair and Stoop 2014; Petrie and Segal 2015; Pikulina et al. 2014; Pikulina et al. 2016; Ravid et al. 2016; Rosaz et al. 2016; Rubin et al. 2016; Rutstrom and Williams 2000; Shurchkov 2012; Sillamaa 1999a; Sillamaa 1999b; Takahashi et al. 2016; van Dijk et al. 2001; Weber and Schram Forthcoming; Wozniak et al. 2014.

Experimental participants are also asked to simply count things, often it is the number of 1s and 0s in a table of numbers as in Abeler et al. (2011). They might also count the number of letters instead. Overall, 11% of the studies we consider implement a counting task. Similar to counting, another 11% of the studies are ones in which the researchers have participants decode numbers into letters (or letters into numbers). An early example of this is Sillamaa (1999a) who asked subjects to take five two-digit numbers and use a paper decoding sheet to translate the numbers into a pattern of letters.

Another widespread task category is puzzles. Puzzles are second in use only to arithmetic problems. In these experiments, participants disentangle difficult mazes, solve Sudoku matrices, Kanji puzzles or the Tower of Hanoi, they try to forecast by discovering an underlying linear relationship or they do crossword puzzles. In an early example of solving puzzles, the participants in Charness and Villeval (2009) solved anagrams.

Simply typing is not utilized very often. Here the task varies from typing the same paragraph over and over to the mind-numbing task of just typing the same keys (e.g. “a” and “b”) over and over again. In (Greiner et al. 2011), for example, the subjects copied three-digit decimal numbers into an input mask on a computer screen.

Our residual category “Other” captures a grab-bag of interesting tasks that range from cracking walnuts (Fahr and Irlenbusch 2000) to simply waiting a predetermined amount of time (Noussair and Stoop 2014) or squeezing a hand dynamometer (Imas 2014).

Looking at the third and fourth columns of Table 1, gives us a sense of how the use of real effort tasks has evolved over time. In column (4) we see the distribution of task categories in the

earlier half of our sample (those studies that occurred in 2012 or earlier). In this time period, arithmetic tasks and puzzles dominated the research landscape and only a few researchers were using counting, decoding or computer tasks. In the past five years or so, however, the distribution of task choices has shifted. Now we see far fewer puzzles, less typing but more computer tasks (the slider task, in particular), counting and decoding. Clerical and arithmetic tasks appear to be used with constant regularity. It is not clear what is behind the shift, though much of it is due to the current popularity of the slider task, a trend that we and others (e.g., Araujo et al. 2016) feel needs a bit more scrutiny given the observed elasticity of performance with respect to monetary incentives, the point we made at the end of Section 2.

Within this type-space, there are some categories of real effort tasks that exemplify what has been termed "useless" or "trivial" real effort. In these experiments, the subjects' work at the task has no obvious productive use outside of its function in the study. Perhaps the most popular current example of such a task is moving sliders, though we offer our assessment of the usefulness of the typical task in each category in the sixth column of Table 1. Of course, we can bicker about the usefulness of the output of this task or another, but our assessment indicates that the vast majority of tasks employed in the literature are trivial. For instance, while many of the data entry and typing experiments follow Swenson (1988) in compensating workers for clearly useless repetitions of strings of text or letters/characters, we note that it does not take a great deal of extra effort to convert the data entry task into one with a putatively useful purpose, as with the inputting of bibliographic data into a database for researcher, library, or firm needs (see Hennig-Schmidt et al. 2010; Tonin and Vlassopoulos 2014 or Huet-Vaughn 2015), extending the underlying theme that real effort tasks are interesting only when they extend the realism of the experiment. As an extension of this theme, with clerical tasks, the useful real effort framing is the natural default (e.g., Konow 2000; Falk and Ichino 2006; Carpenter et al. 2010; Carpenter and Gong 2016 or DellaVigna et al. 2016).

Beyond the dimension of useful vs. useless real effort, the experiments summarized in Table 1 also differ in the degree to which the task inspires intrinsic motivation in the worker or not. While this is, of course, person-dependent, in general, it is probably the case that a counting task, for instance, inspires very little intrinsic motivation while maze solving, on average, does. Researchers attempting to study the comparative effects of pecuniary and non-pecuniary

incentives, including the crowding out phenomenon (Frey and Jegen 2001), should be particularly mindful of the implications of the real effort task selected in this respect.

This question of intrinsic motivation invites consideration of the effort cost function more generally, as even tasks which may inspire intrinsic motivation (those that can be thought of as having a non-negative net effort cost) will likely still vary in the intrinsic reward for the first and the hundredth unit produced (even those who may enjoy arithmetic tasks presumably become fatigued after enough computation). Whether worker effort cost will be flat, convex, or non-negative will depend on the real effort task selected, in addition to the relevant range of production allowed (or alternatively, the time constraint imposed) and the degree to which outside on-the-job leisure is allowed (see Section 4).

While thinking about the likely cost of effort function for a given real effort task, it is also important to consider how it will be shaped by performance ceilings where increased incentives will be ineffectual (for instance, in an arithmetic task) and the degree to which training requirements are essential (for instance, with uncommon computerized tasks).

4. CHOOSING AND IMPLEMENTING A TASK

Given the wide variety of tasks presented in the previous section, it is clear that there are many options for the experimenter looking to implement a real effort task. Based on which criteria has this choice been made (or should it be made)?

Before we get to the task selection criteria, it is important to point out that it is often the case that little to no rationale is given for this choice. As a matter of course, we urge authors to explicitly discuss in the "Methods" section of their papers why they are using the selected task and not some other task. Further, we feel that it is important for the stated rationale to be more detailed than a list of references that use the same task (unless, of course, the study is a replication).

Surveying the literature, it is clear that the most common selection criterion seems to be whether performance on the task does or does not depend on the characteristics of the participants. In the majority of these papers, experimenters typically state that they chose tasks that were perceived to be monotonous, to have little intrinsic value and to be commonplace or accessible to

all workers (e.g., Hogarth and Villeval 2014). The argument for these choices seems reasonable on first blush – if ability and intrinsic motivation are an unobserved nuisance, pick a task that is likely to be unaffected by either. On top of ability and tastes, there is sometimes thought given by authors to avoiding correlation with other demographic characteristics (e.g. gender or IQ). It quickly becomes obvious what the problem is with this sort of rationale, however. If you want a task that does not interact with demographics and it is not clear, *ex ante*, which demographics will matter most, then it is not long before you run out of tasks. As important, these assessments of the tasks are mostly arbitrary – with few exceptions, nobody knows with any confidence whether participant characteristics matter, *a priori*.

Instead of cycling down this rabbit hole, it is important to point out that randomization may attenuate this issue to a great extent too (Randomization is discussed in detail in Chapter II.2). If participants are properly randomized (and perhaps stratified on characteristics such as gender that have robustly been shown to affect performance in some tasks), the characteristics may affect performance, but, if balance is achieved, they will not bias the treatment effects. The thing to remember, however, is that you will be estimating a treatment effect that is averaged across all the different subgroups of participants. That is, without gathering the demographics, you cannot test for heterogeneous treatment effects.

An alternative to guessing which tasks are and are not correlated with participant demographics is to remember why the real effort paradigm is supposed to be valuable and to strive to employ a more realistic task. Like Falk and Ichino (2006), Gneezy and List (2006) and Kube et al. (2012), for example, a solid starting point is clerical work, especially since the experimental participants are often a convenience sample of students who may do this sort of work at their jobs on campus. One exception to this simple rule may come when it is important to separate the pure effect of effort from ability, in which case, it might make sense for the researcher to trade control against realism and pick a task like the slider. That said, there are a number of clerical task for which the ability component is likely to be low (e.g., stuffing envelopes).

From a practical point of view, the choice of task should also be based on somewhat mechanical considerations. When you think about it, because we do not know the specific cost of effort function for the various tasks, it is hard to set the compensation parameters to make sure the marginal benefits and costs of working on the task intersect, an important problem mentioned

above (at the end of Section 2). As a sort of "fixed point theorem" for real effort tasks, to insure an intersection, you may try, for a given piece rate in a limited timed work period, to implement a task in which the initial marginal cost of effort is low but grows relatively quickly. Obviously, this is easier said than done, but it shouldn't be too hard to argue that for some tasks (e.g., alphabetizing files which is easy when the sorted pile consists of two or three but becomes a bit harder as the pile grows with every next sorted addition) the first unit or two does not take much effort but to add subsequent units is increasingly onerous, especially compared to other tasks, like keystrokes for which, over a short period of time the marginal cost must be relatively constant. Of course, increasing the time allowed for the task will serve a similar function as eventually even initially low and flat marginal cost tasks, like typing, become increasingly onerous.

Once the task has been chosen, it is important to realize that there are still plenty of choices to be made concerning the protocol. Starting with the dependent variable (often proxied by output), without variation in effort across participants, one cannot measure treatment effects. This seems obvious but it is not uncommon to run an untried task only to find that all your participants are equally productive, regardless of the treatment. This can be because there are thresholds or other nonlinearities in the production function or the issue may result because of the problem of incentive calibration, of which we just spoke. Along with making sure your procedures are sound, this is an important reason for piloting your experiment.

Along with a pilot, it might also make sense to run a baseline treatment with no incentives just to get a sense for the importance of intrinsic motivation or the general category of experimenter demand effects. Exactly how hard are participants willing to work on the task when there is nothing (financially) at stake? With this knowledge, your incentive treatment effects can be estimated more accurately. In the same vein, you can imagine that people learn and get better at many tasks and therefore, especially in within-subject designs, it may be valuable to run a baseline or "practice" period long enough for participants to get the hang of the task. As important, in within-subject settings, it will be important to "block" to make sure that the treatment order is not perfectly correlated with any effect of learning on performance.

Something we cannot stress enough, especially because it is nearly costless to implement, is to run either a pre- or a post-experimental questionnaire with your participants. One concern with doing this has always been that your survey (if pre-) may prime your participants or (if post-

) your survey responses will be confounded by the fact that participants will try to be consistent between what they did and how they justify doing it. This may all be true, in general, but in many cases, it is unlikely that standard demographics will prime participants and we should be able to agree that there is little chance that experimental behavior will change one's demographics. Given the weight we have put on randomization, collecting demographics is crucial so that you can assess whether it has worked. Are the treatments balanced on the observables that you collected? A balance table should be the first table presented in any study. Another trick that has been used recently (e.g., Carpenter and Gong 2016 or Huet-Vaughn et al. 2017) is to run your survey a considerable amount of time before the experiment so that it is unlikely that respondents even remember what they said in the survey on the day of the experiment. If this is done, you will also have more confidence in the exogeneity of the other survey responses (i.e., non-demographics) you gather.

In the introduction, we pointed out that participants in Swenson's (1988) experiment could either work for pay or do something else, in that case read magazines or do brain-teasers. It turns out that only a minority of real effort experiments have this feature so, in most cases, participants can either work on the assigned task or sit and do nothing. In other words, the opportunity cost of working is zero. This lack of an alternative activity may inflate output numbers and bias treatment effects, though by how much economists are only beginning to examine (e.g., Corgnet et al. 2015c or Koch and Nafziger 2016). Of course, the counter-point here is that there are plenty of jobs out there in which you need to show up and don't have an alternative. In this case, instead of offering another task, the experimenter might let participants leave if they do not wish to work and switch focus to the extensive margin, something one might argue is more consistent with studies using naturally occurring data as described in Heckman (1993). As always, there are no panacea and the obvious problem with this design choice is that you could imagine large peer effects being triggered by the first participant to leave. Despite this, clever work-arounds have been implemented in Linardi and McConnell (2011) and Weber and Schram (2016), for example.

When push comes to shove, the primary sticking point with principal-agent experiments seems to be the cost of effort. In chosen effort experiments it is induced but may not be representative or capture all the aspects of the choice that participants in the role find compelling. In real effort experiments, we simply don't know what the cost of effort function looks like and so

it is hard to formulate precise hypotheses based in theory and, at a minimum, calibration can be problematic. Is there some way to take only the best aspects of the two paradigms to create a hybrid that solves this issue? This is precisely the point of the study by Gaechter et al. (2016) who introduce the "ball-catching task" in which computerized balls fall down the screen on either the left or the right side and the participant needs to capture them to earn money. The "catch" however, is that the experimenters charge the participants a cost for clicking to switch the side of the screen. Here the relationship between the cost of switching and the number of switching clicks induces a specific cost of effort function. This seems like a promising way forward but we look forward to a version of this hybrid paradigm that offers a bit more realism (again, the point of real effort experiments).

5. DOES THE PARADIGM CHOICE MATTER?

The correct way to answer this question is for experimenters to run both chosen effort and real effort as parts of a single study. However, we run into the cost of effort problem again – without knowing this cost function in the real effort context, it can be tricky to calibrate the incentives so that one is confident of setting up an "apples to apples" comparison. Though not plentiful, there are a few studies that attempt just this.

The first comparison that we could find is the study by Bruggen and Strobel (2007). In a within-subjects gift exchange experiment, Bruggen and Strobel have participants play both a chosen effort game similar in design and incentives to Fehr et al. (1993) and a version of the game in which participants do arithmetic. In the real effort version, participants record baseline scores of the arithmetic task a week before the experiment and then their incentivized scores during the experiment are compared to the baseline to see what fraction of their "ability" they decided to give based on the wage they were offered. The results are not straightforward, nor is the analysis that was conducted. In a regression model that is not fully saturated and treats the two observations per participant as independent, there is some evidence (at the 10% level) that workers are more sensitive to the wage offered when the effort is real (i.e., they are more reciprocal) but in a second regression in which the dependent variable is the difference in output to account for the lack of independence, the results are not as strong. The researchers conclude that gift exchange is equally strong in the two paradigms.

To test whether social comparisons affect worker effort choices, Charness et al. (2016) create a wage delegation experiment wherein firms vary the amount of information workers have about each other's wages and whether the boss or the workers pick those wages. For our purposes, what is interesting is that the authors compare the results in two different settings, one of which was chosen effort and the other real effort – arithmetic. Qualitatively, the authors conclude that the results are similar in the two paradigms. Specifically, delegating the wage choice has a positive effect on effort (not unlike Mellizo et al. 2017) in both paradigms.

In something closer to a team production setting, Dutcher et al. (2016) allow participants to work entering financial data either for the team or for themselves and compare this to a similar game in which they are given tokens over the course of time that they can then allocate to the two accounts. The primary result is that contributions to the team look similar between settings both overall and across time.

There are also a few other less closely related studies. For instance, Lezzi et al. (2015) compare four tasks in a tournament setting: a chosen effort contest and three real effort contests (a slider task, an arithmetic task and a counting task). However, despite ostensibly focusing on the generalization across tasks, performance across the tasks is never compared. In a weak-link experiment, Bortolotti et al. (2009) run a real effort (money counting) version of a production-framed weakest link game by Brandts and Cooper (2006) and note that coordination in the real effort task is much higher than what is typically observed in other chosen effort experiments. Lastly, Pikulina et al. (2014) examine the link between overconfidence and effort using both a chosen effort game and a real effort task. They find that the effect of confidence on effort and investment is similar across tasks.

Based on the studies that attempt to directly compare real and chosen effort, the evidence suggests that there is not much difference in the qualitative results. Behavioral patterns tend to be similar, though there is much less evidence on treatment effect sizes.

6. CONCLUDING THOUGHTS

After studying a broad sample of experiments and forcing ourselves to pause and think a bit more about experimental design and our own experience, what is our final assessment of the real effort

paradigm? As the reader can observe, there are two important themes to our review that summarize our assessment: (i) realism and (ii) randomization.

The stated purpose of real effort experiments is to extend the (mundane) realism of experiments in the hope that doing so will insure that the results we observe will be more similar to what participants do (or would do) in the workplace. This seems like a very sensible goal to us. The problem, perhaps, is that the majority of the studies we have surveyed do not press particularly hard on the realism margin. In fact, the trend seems to be to develop and implement more and more arcane computerized tasks, a trend we feel is a bit contrary to the goal of realism, especially considering much of the resulting work is "useless". In Table 1, only a relatively stable 15% of experiments utilize clerical work, the real effort tasks that we argue extend realism unambiguously.

As we have seen, the methodological discussion of chosen versus real effort tends to circle back to the cost of effort and other potential unobserved determinants of effort and output. Our assessment of this debate is that, yes, there are many potential correlates of effort, including ability, gender and even personality and when effort is real, one can add the cost of effort to the list. That point conceded, let's remember the basics of experimental design – aspects of the design that you cannot explicitly control should be randomized. When push comes to shove, isn't the real difference between chosen and real effort experiments that, while there are N unobserved factors in the chosen effort paradigm, there are N+1 in the real effort paradigm (that is if you believe you can adequately induce the cost of effort). If yes, then the solution is the same: in both cases you should make sure to randomize participants to the treatments and run a survey to assess whether you have done so adequately.

Given our assessment of the state of the real effort experimental paradigm, what do we consider to be some fruitful areas for future research? First, we were only able to identify a small handful of experiments that test whether behavior differs depending on whether the effort is real or hypothetical. As mentioned in Section 5, this is a more complicated task than one might first think because one needs to take steps to ensure that the two effort treatments are comparable. Does the shape of the cost of effort function from the chosen effort experiment look anything like the one in the real effort comparison? One thing to consider is that some real tasks (as in Gaechter et al. 2016) do follow an identifiable production process which imposes a cost of effort function, a function that can also be imposed on participants in a chosen effort treatment.

Building on the first recommendation, we also think it would be beneficial to design new tasks that are more plausibly realistic. These tasks should obviously incorporate some aspect of real work but our gut tells us that they should also be realistic in that the output is useful. As important, the tasks should be designed so that the effort cost has the correct theoretical properties (i.e., increasing and convex). The example mentioned above – sorting files – even if done on the computer, is an integral part of many jobs, it can easily be made useful and it has the property that the next file added to a pile (or list) is harder to place than the one before it because the list of comparisons grows with each additional unit of output.

Lastly, another reason for implementing a survey along with your experiment is to assess which demographics actually affect effort and output in a task. Implicit in the current literature are a lot of claims about what demographics affect performance on the tasks that are regularly implemented, but there is little evidence to back up these claims. Because it is easy to gather basic demographics after your experiment, the demographic correlates of task performance should be regularly reported as a contribution to the public good.

REFERENCES

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference Points and Effort Provision. American Economic Association 101, 470-492.

Araujo, F., Carbone, E., Conell-Price, L., Dunietz, M., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S., Wilson, A., 2016. The Slider Task: An Example of Restricted Inference on Incentive Effects. Journal of Economic Science Association 2, 1-12.

Ariely, D., 2008. Predictably Irrational. Harper Collins, New York.

Aronson, E., Carlsmith, J.M., 1968. Experimentation in social psychology, in: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology. Addison-Wesley, Reading, MA.

Augenblick, N., Niederle, M., Sprenger, C., 2015. Working over time: dunamic inconsistency in real effort tasks. Quarterly Journal of Economics 130, 1067-1115.

Azar, O., 2015. Does Relative Thinking Exist in Mixed Compensation Schemes? , 1-42.

Ball, S., Eckel, C., Grossman, P., Zame, W., 2001. Status in Markets. Quarterly Journal of Economics 155, 161-181.

Barr, A., Miller, L., Ubeda, P., 2016. Moral Consequences of Becoming Unemployment. Proceedings of the National Academy of Sciences 113, 4676-4681.

Bartling, B., Fehr, E., Marechal, M.A., Schunk, D., 2009. Egalitarianism and competitiveness. American Economic Review 99, 93-98.

Belot, M., Schröder, M., 2013. Sloppy Work, Lies and Theft: A Novel Experimental Design to Study Counterproductive Behaviour. Journal of Economic Behavior and Organization 93, 233-238.

Berg, J., Dickaut, J., McCabe, K., 1995. Trust, Reciprocity and Social History. Games and Economic Behavior 10, 122-142.

Berger, J., Pope, D., 2011. Can losing lead to winning. Management Science 57, 817-827.

Bhui, R., 2016. Falling Behind: Time and Expectations-Based Reference Dependence. 1-52.

Blumkin, T., Ruffle, B., Ganun, Y., 2012. Are Income and Consumption Taxes Ever Really Equivalent? Evidence From a Real-Effort Experiment with Real Goods. European Economic Review 56, 1200-1219.

Bortolotti, S., Devetag, G., Ortmann, A., 2009. Exploring the effects of real effort in a weak-link experiment, in: paper, w. (Ed.).

Brandts, J., Cooper, D., 2006. A change would do you good... an experimental study of how to overcome coordination failure in organizations. American Economic Review 96, 669-693.

Bruggen, A., Strobel, M., 2007. Real effort versus chosen effort in experiments. Economics Letters 96, 232-236.

Cadsby, C.B., Servátka, M., Song, F., 2013. How Competitive are Female Professionals? A Tale of Identity Conflict. Journal of Economic Behavior and Organization 92, 284-303.

Calsamiglia, C., Franke, J., Rey-Biel, P., 2013. The incentive effects of affirmative action in a real-effort tournament. JOurnal of Public Economics 98, 15-31.

Camerer, C., Hogarth, R., 1999. The Effects of Financial Incentives in Experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty 19, 7-42.

Carpenter, J., Dolifka, D., 2017. Exploitation Aversion: When financial incentives fail to motivate agents. Journal of Economic Psychology 61, 213-224.

Carpenter, J., Gong, E., 2016. Motivating agents: How much does the mission matter? Journal of Labor Economics 34, 211-236.

Carpenter, J., Holmes, J., Matthews, P.H., 2014. An Introduction to "Bucket Auctions" for Charity. Games and Economic Behavior 88, 260-276.

Carpenter, J., Matthews, P., Schirm, J., 2010. Tournaments and Office Politics: Evidence from a Real Effort Experiment. American Economic Review 100, 504-517.

Cason, T., Masters, W., Sheremeta, R., 2010. Entry Into Winner-take-all and Proportional-prize Contests: An Experimental Study. Journal of Public Economics 94, 604-611.

Charness, G., Cobo-Reyes, R., Lacomba, J., Lagos, F., Perez, J.M., 2016. Social comparisons in wage delegation: experiental evidence. Experimental Economics 19, 433-459.

Charness, G., Masclet, D., Villeval, M.-C., 2013. The dark side of competition for status. Management Science 60, 38-55.

Charness, G., Villeval, M.-C., 2009. Cooperation and competition in intergenerational experiments in the field and the laboratory. American Economic Review 99, 956-978.

Chaudhry, S., Klinowski, D., 2016. Enhancing Autonomy to Motivate Effort: An Experiment on the Delegation of Contract Choice. 1-12.

Corgnet, B., 2012. Peer evaluations and team performance: when friends do worse than strangers. Economic Inquiry 50, 171-181.

Corgnet, B., Hernan-Gonzalez, R., Rassenti, S., 2011. Real effort, real leisure and real-time supervision: Incentives and peer pressure in virtual organizations, Chapman University, Economic Science Institute Working Papers.

Corgnet, B., Hernán-González, R., Rassenti, S., 2015a. Firing Threats: Incentive Effects and Impression Management. Games and Economic Behavior 91, 97-113.

Corgnet, B., Hernán-González, R., Schniter, E., 2015b. Why Real Leisure Really Matters: Incentive Effects on Real Effort in the Laboratory. Experimental Economics 18, 284-301.

Corgnet, B., Hernan-Gonzalez, R., Schnitter, E., 2015c. Why real leisure rally matteers: incentive effects on real effort in the laboratory. Experimental Economics 18, 284-301.

Dasgupta, U., Gangadharan, L., Maitra, P., Mani, S., Subramanian, S., 2015. Choosing to be Trained: Do Behavioral Traits Matter? Journal of Economic Behavior and Organization 110, 145-159.

Dasgupta, U., Mani, S., 2015. Only Mine or All Ours: Do Stronger Entitlements Affect Altruistic Choices in the Household. World Development 67, 363-375.

Deci, E., Ryan, R., 1985. Intrinsic motivation and self-determination in human behavior. Plenum, New York.

DellaVigna, S., List, J., Malmendier, U., Rao, G., 2016. Estimating Social Preferences and Gift Exchange at Work, NBER Working Paper 22043.

DellaVigna, S., Pope, D., 2017. What Motivates Effort? Evidence and Expert Forecasts. Review of Economic Studies Forthcoming.

Dickinson, D., 1999. An Experimental Examination of Labor Supply and Work Intensities. Journal of Labor Economics 17, 638-670.

Dickinson, D., Villeval, M.C., 2012. Job Allocation Rules and Sorting Efficiency: Experimental Outcomes in a Peter Principle Environment. Southern Economic Journal 78, 842-859.

Dohmen, T., Falk, A., 2011. Performance pay and multidimensional sorting: productivity, preferences and gender. American Economic Review 101, 556-590.

Douoguih, K., 2011. Essays in experimental economics with implications for economic development.

Dutcher, E.G., Salmon, T., Saral, K., 2016. Is "real" effort more real?, Southern Methodist University Working Paper.

Dutcher, G., 2012. The Effects of Telecommuting on Productivity: An Experimental Examination. The Role of Dull and Creative Tasks. Journal of Economic Behavior & Organization 84, 355-363.

Ellis, S., Fooks, J., Messer, K., Miller, M., 2016. The Effects of Climate Change Information on Charitable Giving for Water Quality Protection: A Field Experiment. Agricultural and Resource Economics Review 45, 319-337.

Eriksson, T., Poulsen, A., Villeval, M.C., 2009. Feedback and Incentives: Experimental Evidence. Labour Economics 16, 679-688.

Erkal, N., Gangadharan, L., Koh, B.H., 2018. Monetary and Non-Monetary Incentives in Real-Effort Tournaments. European Economic Review Forthcoming.

Erkal, N., Gangadharan, L., Nikiforakis, N., 2011. Relative Earnings and Giving in a Real-Effort Experiment. American Economic Association 101, 3330-3348.

Fahr, R., Irlenbusch, B., 2000. Fariness as a Constraint on trust and reciprocity: earned property rights in a reciprocal exchange experiment. Economics Letters 66, 275-282.

Falk, A., Ichino, A., 2006. Clean evidence on peer pressure. Journal of Labor Economics 24, 39-57.

Falk, A., Kosfeld, M., 2006. The Hidden Cost of Control. American Economic Review 96, 1611-1630.

Fan, J., Gómez-Miñambres, J., 2016. Non-binding Goals in Teams: A Real-Effort Coordination Experiments. 1-29.

Fehr, D., 2016. Is Increasing Inequality Harmful? Experimental Evidence. 1-21.

Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does Fairness Prevent Market Clearing? An Experimental Investigation. Quarterly Journal of Economics 108, 437-459.

Frey, B., Jegen, R., 2001. Motivation Crowding Theory. Journal of Economic Surveys 15, 589-611.

Gaechter, S., Huang, L., Sefton, M., 2016. Combining "real effort" with induced effort costs: the ball-catching task. Experimental Economics 19, 687-712.

Gerhards, L., Gravert, C., 2015. Grit Trumps Talent? An Experimental Approach. University of Gothenbrug Working Paper.

Gill, D., Prowse, V., 2012. A structural analysis of disappointment aversion in a real effort competition. American Economic Review 102, 469-503.

Gneezy, U., List, J., 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. Econometrica 74, 1365-1384.

Gneezy, U., Niederle, M., Rustichini, A., 2003. Performance in Competitive Environments: Gender Differences. Quarterly Journal of Economics 118, 1049-1074.

Gneezy, U., Rustichini, A., 2000. Pay enough or don't pay at all. Quarterly Journal of Economics 115, 791-810.

Goldstein, W., Hogarth, R., 1997. Research in judgement and decision making: current, connections, and controversies. Cambridge University Press, New York.

Greiner, B., Ockenfels, A., Werner, P., 2011. Wage Transparency and Performance: A Real-Effort Experiment. Economic Letters 111, 236-238.

Gueth, W., Schmittberger, R., Schwarze, B., 1982. An Experimental Analysis of Ultimatum Bargaining. Journal of Economic Behavior and Organization 3, 367-388.

Gupta, N.D., Poulsen, A., Villeval, M.C., 2013. Gender Matching and Competitiveness: Experimental Evidence. Economic Inquiry 51, 816-835.

Hargreaves Heap, S., Ramalingam, A., Rojo Arjona, D., 2016. Social Information 'Nudges': An Experiment with Multiple Group References. 1-36.

Healy, A., Pate, J., 2011. Can teams help to close the gender competition gap? Economic Journal 121, 1192-1204.

Heckman, J., 1993. What Has Been Learned About Labor Supply in the Past Twenty Years? American Economic Review 83, 116-121.

Hennig-Schmidt, H., Rockenbach, B., Sadrieh, A., 2010. In search of workers' real effort reciprocity - A field and a laboratory experiment. Journal of the European Economic Association 8, 817-837.

Heyman, J., Ariely, D., 2004. Effort for payment: A tale of two markets. Psychological Science 15, 787-793.

Hoffman, E., McCabe, K., Shachat, J., Smith, V., 1994. Preferences, Property Rights, and Anonymity in Bargaining Games. Games and Economic Behavior 7, 346-380.

Hogarth, R., Villeval, M.-C., 2014. Ambiguous Incentives and the Persistence of Effort: Experimental Evidence. Journal of Economic Behavior and Organization 100, 1-19.

Huang, L., Murad, Z., 2016. Impact of Relative Performance Feedback on Beliefs, Preferences and Performance Across Dissimilar Tasks. 1-45.

Huet-Vaughn, E., 2015. Do Social Comparisons Motivate Workers? A Field Experiment on Relative Earnings and Labor Supply, Middlebury College Working Paper.

Huet-Vaughn, E., Robbett, A., Spitzer, M., 2017. A Taste for Taxes: Minimizing Distortions Using Political Preferences.

Imas, A., 2014. Working for the "warm glow": On the benefits and limits of prosocial incentives. Journal of Public Economics 114, 14-18.

Ivanova-Stenzel, R., Kübler, D., 2011. Gender Differences in Team Work and Team Competition. Journal of Economic Psychology 32, 797-808.

James Alm, T.L.C., Michael Jones, Michael McKee, 2012. Social Programs as Positive Inducements for Tax Participation. Journal of Economic Behavior and Organization 84, 85-96.

Jones, D., Linardi, S., 2014. Wallflowers: Experimental Evidence of an Aversion to Standing Out. Management Science 60, 1757-1771.

Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness and the Assumptions of Economics. Journal of Business 59, s285-s300.

Kessler, J., Norton, M., 2016. Tax Aversion in Labor Supply. Journal of Economic Behavior and Organization 124, 15-28.

Kidd, M., Nicholas, A., Rai, B., 2013. Tournament Outcomes and Proscial Behaviour. Journal of Economic Psychology 39, 387-401.

Koch, A., Nafziger, J., 2016. Gift exchange, control, and cyberloafing: A real-effort experiment. Journal of Economic Behavior & Organization 131, 409-426.

Konow, J., 2000. Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. American Economic Review 90, 1072-1091.

Kraut, R., Sunder, S., Telang, R., Morris, J., 2011. Pricing Electronic Mail to Solve the Problem of Spam. Human-Computer Interaction 20, 195-223.

Kube, S., Marechal, M.A., Puppe, C., 2012. The Currency of Reciprocity: Gift Exchange in the Workplace. American Economic Review 102, 1644-1662.

Kuhn, P., Villeval, M.C., 2013. Are Women More Attracted to Co-operation than Men? The Economic Journal 125, 115-140.

Lefgren, L., Sims, D., Stoddard, O., 2016. The Other 1%: Class Leavening, Contamination and Voting for Redistribution. 1-29.

Lezzi, E., Fleming, P., Zizzo, D., 2015. Does it matter which effort task you use? A comparison of four effort tasks when agents compete for a prize, working paper.

Linardi, S., McConnell, M., 2011. No excuses for good behavior: volunteering and the social environment. Journal of Public Economics 95, 445-454.

Markey, A., Chin, A., Vanepps, E., Loewenstein, G., 2014. Identifying a Reliable Boredome Induction. Perceptual & Motor Skills: Perception 119, 237-253.

Mellizo, P., Carpenter, J., Matthews, P., 2017. Ceding control: an experimental analysis of participatory management. Journal of the Economic Science Association 3, 62-74.

Nalbantian, H., Schotter, A., 1997. Productivity Under Group Incentives: an experimental study. American Economic Review 87, 314-341.

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics 122, 1067-1101.

Noussair, C., Stoop, J., 2014. Time as a Medium of Reward in Three Social Preference Experiments.

Petrie, R., Segal, C., 2015. Gender Differences in Competitiveness: The Role of Prizes. 1-32.

Pikulina, E., Renneboog, L., Tobler, P., 2014. Overconfidence, effort and investment, working paper.

Pikulina, E., Renneboog, L., Tobler, P., 2016. Do Confident Individuals Generally Work Harder? , 1-22.

Ravid, O., Malul, M., Zultan, R.i., 2016. The Effect of Economic Cycles on Job Satisfaction in a Two-Sector Economy. 1-17.

Rey-Biel, P., Sheremeta, R., Uler, N., 2016. When Income Depends on Performance and Luck: The Effects of Culture and Information on Giving.

Rosaz, J., Slonim, R., Villeval, M.C., 2016. Quitting and Peer Effects at Work. Labour Economics 39, 55-67.

Rubin, J., Samek, A., Sheremeta, R., 2016. Incentivizing Quantity and Quality of Output: An Experimental Investigation of the Quantity-Quality Trade-off. 1-53.

Rutstrom, E., Williams, M., 2000. Entitlements and fairness: an experimental study of distributive preferences. Journal of Economic Behavior & Organization 43, 75-80.

Shurchkov, O., 2012. Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints. Journal of the European Economic Association 10, 1189-1213.

Sillamaa, M.-A., 1999a. How work effort responds to wage taxation: A non-linear versus a linear experiment. Journal of Economic Behavior & Organization 39, 219-233.

Sillamaa, M.-A., 1999b. How work effort responds to wage taxation: An experimental test of a zero top marginal tax rate. Journal of Public Economics 73, 125-134.

Smith, V., 1976. Experimental economics: induced value theory. American Economic Review 66, 274-279.

Smith, V., 1982. Microeconomic Systems as an Experimental Science. American Economic Review 72, 923-955.

Swenson, C., 1988. Taxpayer behavior in response to taxation. Journal of Accounting and Public Policy 7, 1-28.

Takahashi, H., Shen, J., Ogawa, K., 2016. An Experimental Examination of Compensation Schemes and Level of Effort in Differentiated Tasks. Journal of Behavioral and Experimental Economics 61, 12-19.

Thaler, R., Johnson, E., 1990. Gambling wiht the house money and trying to break even: the effects of prior outcomes on risky choice. Management Science 36, 643-660.

Tonin, M., Vlassopoulos, M., 2014. An experimental investigation of intrinsic motivations for giving. Theory and Decision 76, 47-67.

van Dijk, F., Sonnemans, J., van Winden, F., 2001. Incentive Systems in a Real Effort Experiment. European Economic Review 45, 187-214.

Weber, M., Schram, A., 2016. The non-equivalence of labour market taxes: a real-effort experiment. The Economic Journal forthcoming, 1-29.

Weber, M., Schram, A., Forthcoming. The Non-Equivalence of Labour Market Taxes: A Real-Effort Experiment. The Economic Journal, 1-43.

Wozniak, D., Harbaugh, W., Mayr, U., 2014. The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. Journal of Labor Economics 32, 161-198.

Wyatt, S., 1934. Incentives in repetitive work; a practical experiment in a factory. H. M. Stationery Office, London.