Journal of Behavioral and Experimental Economics 103 (2023) 101972

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Don't pay the highly motivated too much[☆]

Zack Dorner[a,*], Emily Lancsar[b]

[a] Waikato Management School, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand
[b] Department of Health Services Research & Policy, Australian National University, 62 Mills Road, Acton ACT 2601, Australia

ARTICLE INFO

Keywords:
Incentives
Intrinsic motivation
Lab-in-the-field
Crowding out
Real effort task

ABSTRACT

One remaining puzzle in the literature on intrinsic motivation and extrinsic incentives is the conflict between two oft-cited rules: "pay enough or don't pay at all" and "pay, but not too much". There is some evidence to suggest differing impact of incentives, depending on level of initial motivation. Our lab-in-the-field experiment, on a heterogeneous sample of the general public, allows us to test our intuitive prediction: that crowding out will be less (more) likely when motivation is low (high) to begin with. We use a real effort task and a within- and between-subject design to measure motivation before, during and after an incentive is applied. On average, we find support for "pay, but not too much", as a low power monetary incentive is as effective as a high power monetary incentive when applied. However, high power monetary incentives strongly crowd out motivation for the high types. Monetary incentives (low or high) do not crowd out motivation in low motivation types. A charity incentive does not increase performance in the real effort task, but also has no crowding out effect. We show the importance of accounting for initial intrinsic motivation, and that the highly motivated are more susceptible to crowding out.

1. Introduction

Seminal and highly cited works such as Deci (1971) and Titmuss (1970) demonstrate the potential for extrinsic incentives to undermine intrinsic motivation. Crowding out of intrinsic motivation can be observed either when an individual's effort is reduced during the application of an extrinsic incentive, or reduced below initial effort after the removal of a temporary incentive. The voluminous literature that has developed since those studies have found seemingly contradictory results on the interaction between size of incentives and motivation. With regard to monetary incentives, the first pattern that has been observed, comparing incentives when they are applied, is a V shape. In this case, effort with no incentive is the top left point of the V; a low power monetary incentive worse than no incentive (bottom of the V) and a high power monetary incentive is at least as good as no incentive (top right of the V), leading to the rule "pay enough or don't pay at all" (Gneezy & Rustichini, 2000). Next is an inverse V, whereby a high incentive is worse than a low one, which suggests the rule "pay, but not too much" (Pokorny, 2008). Of course, a third possibility is a monotonic response where effort increases with the incentive level (DellaVigna & Pope, 2018). Reconciling these findings is important not only for theory, but also for public policy. Examples where extrinsic incentives have been used or considered include payments for test scores in education, or monetary incentives for exercise, to attend disease screenings or quit smoking (Gneezy et al., 2011; Promberger & Marteau, 2013).

An important dimension to the puzzle that has received less attention in the literature is this: for an individual with a high level of intrinsic motivation for a task, there is little space to crowd in their motivation, but lots of space to crowd out their motivation. The opposite holds for someone with low motivation for a task. While some papers have found differences in responses to incentives between boring versus interesting tasks (Giusti & Dopeso-Fernández, 2020; Takahashi et al., 2016), less attention has been paid to heterogeneity of motivation

[☆] Lata Gangadharan in particular is acknowledged for the extensive supervision given throughout this project. A thank you also for the supervision provided by Anke Leroux and Paul Raschky. Acknowledgements are also made to those who provided comments during seminars at the Department of Economics at Monash University and at the 29th Ph.D. Conference in Economics and Business. In particular, for providing us with detailed comments and discussion, we thank Tim Cason, Philip Grossman, Kenan Kalayci, Elias Khalil, Lana Friesen and Andreas Lange. Finally, we are grateful to Kate Ennis for providing research support. Errors and omission are the sole responsibility of the authors. Funding for this project is from the Australian Research Council (ARC), Australia, through a Discovery Early Career Researcher Award (DECRA) (Grant ID DE140101260). This research was supported by an Australian Government Research Training Program (RTP) Scholarship, Australia. The project has been approved by the Monash University Human Research Ethics Committee; MUHREC#: CF16/618 - 2016000300.

[*] Corresponding author.
E-mail address: zack.dorner@waikato.ac.nz (Z. Dorner).

https://doi.org/10.1016/j.socec.2022.101972
Received 3 November 2021; Received in revised form 6 December 2022; Accepted 22 December 2022
Available online 5 January 2023

between individuals for a given task.[1] Based on our earlier point, we can expect to find greater levels of crowding out with high motivation types, than with low. Thus, incentives may be more likely to show monotonic improvements in effort for low types than high types. For high types, crowding out may even outweigh the effect of the incentive, causing a lowering of effort either when an incentive is applied at a low level (a V shape), or indeed as an incentive is increased in size (the inverse V).

Indeed, one key variable missing from many studies on extrinsic incentives is a measure of intrinsic motivation for the task, and a measure of crowding out from incentives. The most relevant lab studies for our purposes tend to compare effort over a range of incentives between treatment groups, in a single observation for each group (DellaVigna & Pope, 2018; Heyman & Ariely, 2004; Kajackaite & Werner, 2015; Pokorny, 2008; Takahashi et al., 2016). In these studies, there is no baseline observation of intrinsic motivation for the task. Theory demonstrates that an extrinsic incentive can increase effort by more than it undermines intrinsic motivation (Bénabou & Tirole, 2003). Therefore, in a purely between subject study, we may not be able to fully understand whether one incentive undermines intrinsic motivation more than another. Furthermore, as with any task, we should expect heterogeneity in initial level of intrinsic motivation. Hence, it is important to be able to measure how intrinsic motivation changes with incentives, and account for any heterogeneity within a subject pool.

The aim of this paper is to better understand the impact of type and size of extrinsic incentives on heterogeneous initial intrinsic motivation. In particular, we aim to reconcile existing conflicting results in the literature on the effect of extrinsic incentives. Our lab experiment utilises a real effort task over three rounds. The first round measures initial intrinsic motivation for the task, meaning we can test for heterogeneous responses to the incentives. In round 2, we apply an extrinsic incentive, depending on between subject treatment group. This allows us to measure the effectiveness of each type of incentive for low versus high motivation subjects, in terms of increasing performance. However, performance in round 2 consists of both intrinsic and extrinsic motivation. Given intrinsic motivation is likely to have changed after the incentive is applied, we remove incentives for round 3 in order to measure intrinsic motivation again. Thus, the difference-in-differences between rounds 2 and 1 shows the effectiveness of each incentive at increasing effort for those with low and high initial motivation; the difference-in-differences between rounds 3 and 1 shows the change in intrinsic motivation due to the extrinsic incentives; and the difference-in-differences between rounds 3 and 2 shows the performance level due to the extrinsic incentive.

The incentives we test include low and high power monetary incentives. Additionally, we test a charity (non-monetary) incentive, as this type of incentive has been shown to be less likely to crowd out motivation, but there is a gap in the literature on whether this type of incentive has differing effects across a heterogeneous population (DellaVigna & Pope, 2018; Heyman & Ariely, 2004; Imas, 2014). Our lab-in-the-field study utilises a heterogeneous adult population to provide sufficient variation in intrinsic motivation, while still affording us the control provided by a lab.[2]

As far as we are aware, this is the first lab study on intrinsic motivation and extrinsic incentives to test both (1) the within-subject effect of applying and taking away extrinsic incentives (by comparison across rounds) and (2) the between-subject effect of different types and magnitudes extrinsic incentives on crowding in or out. We find low power and high power monetary incentives are as effective as each other on average and for low motivation individuals. After incentives are removed (for round 3) we can observe that the high power incentive strongly crowds out intrinsic motivation in the high motivation group. Hence, we caution against paying high motivation individuals too high a piece rate. The charity incentive has no impact when applied or evidence of crowding out. Overall, our results suggest some of the conflicting findings in the literature may be explained by the distribution of baseline intrinsic motivation for a given task.

The remainder of this paper is organised as follows. The next section defines key terms and presents the background literature in more detail. In Section 3 we outline the theoretical framework, experimental procedures and design, hypotheses and our analytical approach. The results are presented in Section 4 and discussed in Section 5 to conclude.

2. Background

We define intrinsic motivation in accordance with the major theoretical perspectives, which is an individual performing a task for its own sake (Bénabou & Tirole, 2003; Ryan et al., 2021). This definition is useful as it encapsulates the dimension of heterogeneity we are interested in. Specifically, the individual's desire to undertake a task, which we assume will vary at an individual level while holding the task constant. Additionally, we assume "for its own sake" manifests as the level of effort observed in the absence of an extrinsic incentive.

To define extrinsic incentives, we follow Cerasoli et al. (2014, p. 981): "anything provided by an external agent contingent on performance of particular standards of behaviour(s)". Hence, we use piece rate incentives to test extrinsic incentives. Another important distinction to make is type of incentive. We use both monetary and non-monetary incentives. Monetary incentives can have a crowding out effect relative to non-monetary incentives of the same value (Heyman & Ariely, 2004).

One of the main theories explaining crowding out comes from self-determination theory (SDT) (Ryan et al., 2021). It is grounded in the idea that people have innate psychological needs in terms of competence, autonomy and relatedness. Intrinsic motivation to undertake an activity comes particularly from the need to feel competent and autonomous (Deci & Ryan, 2000). Extrinsic incentives in the form of tangible rewards can undermine an individual's feelings of competence by reinforcing the message that they require rewards to be motivated, rather than allowing them to feel they can be motivated themselves. Hence, once the rewards are withdrawn, intrinsic motivation has been crowded out rather than maintained or crowded in Deci and Ryan (1980). These ideas were developed as part of cognitive evaluation theory, which is considered a sub-theory within SDT (Ryan et al., 2021). We argue this theory is compatible with our main hypothesis, that those with higher intrinsic motivation could be more susceptible to crowding out from monetary incentives, compared to those with low motivation, who have less potential for crowding out to begin with.

Another theoretical perspective, based more in the economics literature, has a stronger focus on how an extrinsic incentive can affect task perception. Crucially, such a theoretical perspective suggests that the individual undertaking the task has incomplete information about the nature of the task. An extrinsic monetary incentive can thus be seen as bad news about the nature of the task, in that it suggests that the principal who sets the task has information that it is costly to undertake (Bénabou & Tirole, 2003). In a similar vein, individuals may be intrinsically motivated to undertake a task that has pro-social benefits, to improve their image (social image or self-image). An extrinsic monetary incentive can undermine this motivation by suggesting ulterior motives; the individual is no longer motivated by the pro-social benefits, but partly or mostly by the private monetary reward they receive. This is termed an "overjustification effect", explained in this instance by the additional noise in understanding a signal about an individual's motive (by an external party or introspectively) when they

[1] We canvas some examples in background section, but there are few papers that look at size of incentive in relation to low versus high levels of intrinsic motivation.

[2] While the study was undertaken in a standard university lab, our non-standard subject pool qualifies the study as a lab-in-the-field, or an artefactual field experiment (Harrison & List, 2004; Viceisza, 2016).

undertake a pro-social activity associated with a private award (Bénabou & Tirole, 2006). Again, our main hypothesis is compatible with these two theories in that higher initial intrinsic motivation is more susceptible to crowding out from extrinsic incentives.

All of these theories point to the importance of context in terms of underlying intrinsic motivation, and likely response to incentives. Our study relates to other lab studies, where some form of task like a puzzle or a real effort task is undertaken in the lab or lab-in-the-field (Deci, 1971; DellaVigna & Pope, 2018; Pokorny, 2008). Lab studies have revealed many important insights for contexts such as education and work (Deci & Ryan, 1980; Gneezy et al., 2011; Ryan & Deci, 2020); though caution of course must be used when applying the results to a specific context. Given these points, we thus focus the rest of this literature review mainly on lab or lab-in-the-field studies.

Monetary incentives have two important dimensions to consider: applying a monetary incentive compared with no incentive; and size of monetary incentive. Both these dimensions are at the heart of a V or inverse V shape, with the pointy end of the V demonstrating the response to a monetary incentive compared with no incentive. Both Gneezy and Rustichini (2000) and Heyman and Ariely (2004) explain their V shaped finding by noting the crowding out potential of money itself. They argue that the move from putting in effort voluntarily, to being paid, undermines intrinsic motivation, regardless of the size of the monetary incentive. Hence, a larger monetary incentive will compensate individuals enough to see them increase their effort, as the crowding out is somewhat constant across levels of the monetary incentive. Gneezy and Rustichini (2000) recommend to “pay enough or don’t pay at all”. The treatments in these studies are varied only at the between subject level, and do not account for heterogeneity of motivation.

A further examination of the Gneezy and Rustichini (2000) data by Rydval and Ortmann (2004) demonstrates that between subject heterogeneity within treatment groups is higher than treatment differences, and that there are likely differential effects of the incentives across the distribution. It appears the very low payment treatment impacts the lowest performers more, which is counter to our hypothesis that crowding out will be greater among high performers. However, we argue our task relies less on ability compared to the IQ test employed by Gneezy and Rustichini (2000). Furthermore, with only a between subject comparison of incentives, their data cannot provide the fuller picture that our within- and between-subject design can.

Pokorny’s (2008) finding of an inverse V pattern shows that crowding out can be dependent on the size of the monetary incentive too. Hence, they recommend “pay — but do not pay too much”. They explain their results with reference dependent preferences, arguing student subjects try to achieve just their expected earnings from a lab experiment. This finding is consistent with the minimum threshold treatment in Kajackaite and Werner (2015). A third finding in the literature of a monotonic pattern of incentive size and increases in effort (DellaVigna & Pope, 2018) suggests either no crowding out with monetary incentives, or any crowding out is outweighed by the power of the incentive. However, we note that small monetary incentives have small or no statistical effect in the study by DellaVigna and Pope (2018), meaning “pay enough or don’t pay at all” may still be a reasonable rule to apply in this case. Again, all of these papers rely on between subject treatments only.

Additionally, studies have looked at a range of non-monetary extrinsic incentives, such as charity payments, which are often found to have significant power for increasing effort. This power holds even compared with monetary incentives, particularly smaller monetary incentives (DellaVigna & Pope, 2018; Heyman & Ariely, 2004; Imas, 2014). These incentives can be hypothesised to not crowd out intrinsic motivation, hence their effectiveness. It remains to be seen whether a lack of crowding out from charity incentives applies to the same extent across a heterogeneous population.

The importance of underlying intrinsic motivation for the given task has been widely acknowledged, but less has been done to look at incentive size and how this interacts with low or high levels of intrinsic motivation. Takahashi et al. (2016) compares what they assert is a boring task, and an interesting task. They find a monotonic response to incentives for the boring task, and an inverse V (a la Pokorny, 2008) for the interesting task. Thus, they find higher monetary incentives can crowd out motivation more when subjects are more motivated on average. Giusti and Dopeso-Fernández (2020) also find performance-based incentives are more effective for a boring task, compared with an interesting one. Both of these studies are thus consistent with our hypothesis that crowding out will be more likely when intrinsic motivation is high.

In terms of heterogeneous motivation for the same task, Taylor (2020) looks at low versus high scorers on the CRT cognitive test. They find the low scorers are responsive to a monetary incentive, whereas the high scorers are not, suggesting the low scorers have low scores due to a mixture of low ability and low motivation. They do not test incentives of different sizes, however. Pascual-Ezama et al. (2013) also separates subjects by motivation for a puzzle task and finds differences depending on motivation levels. However, these differences are in response to monitoring regimes rather than incentive sizes, which they did not investigate. Finally, Dessí and Rustichini (2015) experimentally manipulates intrinsic motivation, finding in the high motivation treatment incentives are not effective, and in the low motivation treatment they are, again consistent with our hypothesis. We contribute to this literature by looking within the same task, with varying levels of incentive size and type, and measuring baseline intrinsic motivation before incentives are applied.

In terms of empirical evidence of drivers of intrinsic motivation itself, Segal (2012) provides evidence that links intrinsic motivation to personality traits. More widely, research in neuroscience shows the important role of subjective feelings in underlying intrinsic motivation (eg. Lee & Reeve, 2017). Thus, there is a suggestion of a more transient component of intrinsic motivation. Indeed, an immediate change in neurological state can be seen when extrinsic incentives act to crowd out intrinsic motivation (Ma et al., 2014). Other studies in the education and labour literatures show how effort is related to non-cognitive abilities, which are defined as “personality traits, goals, character, motivations, and preferences that are valued in the labour market, in school, and in many other domains” (Kautz et al., 2014). For example, Cubel et al. (2016) find neurotic subjects are less productive and conscientious subjects are more. Gneezy et al. (2019) find evidence of higher motivation among Chinese students than US, given a stronger response to incentivisation among US students. There are mixed findings in the literature however, such as Staněk and Krčál (2019) who find patience does not explain unincentivised test scores (which are more influenced by intrinsic motivation), but does correlate with incentivised test scores. Overall, there is a large body of evidence to suggest these non-cognitive abilities in general, including intrinsic motivation, are important for labour, educational and other lifetime outcomes (Balart et al., 2018; Gneezy et al., 2019; Hitt et al., 2016; Kautz et al., 2014; Segal, 2012).

Our contribution is in testing how extrinsic incentives of differing size and type interact with underlying intrinsic motivation. We do so using a heterogeneous population undertaking the same task, so that we can better understand whether heterogeneity in underlying intrinsic motivation can lead to differing levels of crowding out (or in) of intrinsic motivation in response to extrinsic incentives. Testing two levels of monetary incentive is important to unpack whether size of monetary incentive matters to crowding out across the heterogeneous group. Testing a non-monetary incentive is important to understand how low and high motivation groups respond to a non-monetary extrinsic incentive.

Table 1
Experimental activities timeline.

Between subject treatment group	Practice round	Round 1	Round 2	Round 3
Control group	Effort task explained and practice round given.	Effort task with no incentives.	Effort task with no incentives.	Effort task with no incentives.
Extrinsic incentive groups (three separate groups, each with a different type of incentive)			Effort task with extrinsic incentive, type depending on treatment group.	
Task time limit	2 min	5 min	5 min	5 min

3. Method

3.1. Theory

Here we consider a very simple descriptive model so that we are able to inform our experimental design, generate our hypotheses and interpret our results. Let us assume effort level is directly observable, and is the result of two components; intrinsic motivation and extrinsic motivation. Let us also assume that when both intrinsic and extrinsic motivation contribute to effort, we cannot observe how much each component contributes. This simple model has two main implications. First, when we observe effort level with no extrinsic incentive, we are observing effort as a result of intrinsic motivation only.

The model's second implication relates to observing crowding out of intrinsic motivation due to an extrinsic incentive. We can only be certain we are observing crowding out when the reduction in effort from crowding out is larger than the increase in effort from the extrinsic motivation provided by the incentive. If we observe an increase in effort with an incentive compared with no incentive, we cannot determine whether the increase in effort is due to the extrinsic motivation being larger than the crowding out of intrinsic motivation, or whether intrinsic motivation is unchanged or even crowded in (increased). However, we can remove the extrinsic incentive and observe intrinsic motivation again, given a no-incentive observation will give us effort from intrinsic motivation only.

Hence, we have the rationale for our three round within- and between-subject experimental design: initial intrinsic motivation is observed in round 1; effort due to both intrinsic and extrinsic motivation is observed in round 2 when the incentive is applied; and change in intrinsic motivation is observed in round 3, when the incentive is removed. This full design is shown in Table 1 in the experimental design section below. The difference between round 2 and round 1 shows the how the combination of intrinsic and extrinsic motivation compares with just initial intrinsic motivation. Round 3 versus round 1 shows how post-incentive intrinsic motivation compares with initial intrinsic motivation. A comparison between rounds 3 and 2 shows us the size of the extrinsic motivation if we assume removing the incentive does not significantly change intrinsic motivation. This assumption implies that it is only the application of the incentive that changes intrinsic motivation. We return to this point in our discussion at the end of this section. Overall, this design also allows us to measure difference-in-differences between treatment groups. Finally, our baseline measure of intrinsic motivation in round 1 means we can compare changes between subjects with differing levels of underlying intrinsic motivation, and observe any heterogeneous effects of extrinsic incentives.

We now discuss how each incentive may impact individuals with low versus high initial intrinsic motivation for a task, in light of the literature outlined in the previous section. We start by considering the monetary incentives and an individual with low underlying intrinsic motivation for a task (low type). Their intrinsic motivation is unlikely to be significantly crowded out, due to its low starting point. Additionally, monetary incentives are unlikely to crowd in their intrinsic motivation either. This assertion follows from both SDT and Bénabou and Tirole (2003). Monetary incentives either undermine/confirm already low feelings of autonomy and competency (as per SDT) or reinforce the initial belief about the high cost of undertaking the task (as per Bénabou & Tirole, 2003). Thus, we predict there will be little change in already low intrinsic motivation due to any of the extrinsic incentives. Therefore, effort will be highest under the high power incentive, when applied in round 2, and lowest with no incentive.

For those with high initial intrinsic motivation for a task (high types), there is considerable scope for potential crowding out. As outlined in the previous section, there is mixed evidence in the literature as to whether the size of the monetary incentive impact the level of crowding out or whether all monetary incentives have a similar impact on crowding out regardless of size. This latter relationship would suggest that effort level when the low monetary incentive is applied could be lower than no incentive, as the crowding out of intrinsic motivation will be higher than the extrinsic motivation from the incentive. The high power incentive is likely to more than offset any crowding out of intrinsic motivation. This would imply we would observe a V shape in effort level between incentives in round 2. Once the incentives are removed, we should observe similar levels of crowding out from both the low and high monetary incentives.

However, it is also possible that the low monetary incentive will be low enough that it will not significantly crowd out intrinsic motivation. According to SDT, this would mean that the incentive is not perceived as undermining autonomy or competence, but instead as a token of support. Alternatively, it reinforces the idea that the task is not particularly costly to undertake (as per Bénabou & Tirole, 2003). A high power monetary incentive may have the opposite effect and thus crowd out intrinsic motivation. While the size of the high power incentive may offset any crowding out effect when applied, we would observe reduced effort when it is removed; thus, concluding it has crowded out intrinsic motivation. In the most extreme case we could observe an inverse V shape comparing effort between treatment groups in round 2.

We also consider a non-monetary extrinsic incentive, as the literature shows this type of incentive may be effective, and potentially less prone to crowding out. It is useful to understand how such an incentive interacts with a heterogeneous population. As with monetary incentives, we argue that our charitable donation incentive is unlikely to crowd out motivation in a low type due to the lack of existing motivation to crowd out. Might it crowd in motivation? Again, theory suggests this is unlikely, though it needs empirical testing to confirm. A charitable donation provided as an extrinsic incentive may not directly support feelings of autonomy and competency. However, it may create a positive feeling about the task, for example by associating the task with feelings of relatedness (pro-sociality) to the wider community (as the third innate psychological motivator according to SDT). From the perspective of Bénabou and Tirole (2003), the provision of an extrinsic incentive may reinforce to a low type the idea that the task is costly to undertake. However, we may think that a charity incentive is perceived to provide lower extrinsic motivation than an equivalent monetary incentive. Thus, there is a chance that a charity incentive may be seen as good news about the nature of the task. Therefore, under both these theoretical perspectives we cannot rule out that it may in fact increase intrinsic motivation for low types.

A similar logic applies with regards to high types and the charity incentive. It seems unlikely that a charity incentive would crowd out intrinsic motivation. Given high types are already highly motivated, it also seems unlikely the charity incentive would crowd in motivation as there may be little or no scope for this increase. Or, perhaps intrinsic motivation is increased but it is unobservable as effort is already maximised.

Our three round experimental design (no incentive, incentive, no incentive) allows us to more thoroughly investigate the extent to which extrinsic incentives impact intrinsic motivation. It also provides a baseline measure of intrinsic motivation, which we categorise as low or high. However, there is a possibility that the design may also impact intrinsic motivation; that is, providing the task first with no incentive may influence subjects' relationship with the task, as could removing the incentive. This is not a limitation so much as an element of the nature of the task that should be considered when interpreting the results and comparing to other studies.

The implication for initial intrinsic motivation is that subjects will be more familiar with the task given they have completed a round of the task before receiving an extrinsic incentive. Thus, they will have some pre-conceived ideas about the nature of the task and the cost of undertaking it. However, they are still relatively unfamiliar with it and an extrinsic incentive could still shape their perception of the task or indeed undermine their feelings of competence and autonomy. Impacts here may be less strong than if they were incentivised in round 1, which they are not in our study.

In terms of round 3, again the removal of a previous incentive may impact intrinsic motivation. First, it could reduce negative associations with the task as it has been given back to them without incentivisation, thus increasing intrinsic motivation beyond what it was during round 2 when the incentive was applied. In the opposite direction, intrinsic motivation could be even lower than round 2 due to loss aversion. Thus, we assume round 3 effort is a good proxy for intrinsic motivation levels in round 2, given the direction of further changes in intrinsic motivation from round 2 to 3 are ambiguous. These factors could alter the interpretation of our results, and could be fruitfully explored in future research. For our study, we argue that the interpretation of observed effort outlined in this section is sufficient as the interaction between intrinsic motivation and extrinsic incentives will be the main factor driving our observed effort levels in our experiment. Indeed, this three round within-subject aspect of our design, alongside this specific interpretation, is a core part of the intrinsic motivation and extrinsic incentive literature (eg. Deci, 1971).

In the rest of this section we outline our experimental design. Then, we discuss the specific hypotheses we test and our predictions. Finally, we provide details on our analytical approach to test the hypotheses.

3.2. Experimental procedures and design

The experiment was run over 12 sessions from 6 April to 3 June, 2016, at the Monash Laboratory for Experimental Economics (MonLEE) at Monash University in Melbourne, Australia. Subjects remained in a session for between 1 h and 1 h and 45 min, depending on how fast they completed the final survey. On arrival, subjects were randomly seated, signed consent forms and provided overview instructions in hard copy. Further instructions were handed out as needed and read aloud (see online appendix).

The activities section of the experiment incorporated multiple rounds of a word encoding real effort task (Erkal et al., 2011), programmed using zTree (Fischbacher, 2007). Existing literature suggests that real effort tasks provide subjects with utility and are also designed to give a relatively fine measure of effort on the intensive margin Brüggen and Strobel (2007), Gill and Prowse (2012). Using a real effort task also helps make our results comparable with other recent lab studies on crowding out (eg. DellaVigna & Pope, 2018; Heyman & Ariely, 2004; Kajackaite & Werner, 2015; Pokorny, 2008). We measure relative performance in the real effort task on the intensive margin as it unlikely that we will fully crowd out effort in a lab based real effort task (Araujo et al., 2016; Erkal et al., 2018).

An example screenshot of the task is shown in Fig. 1. The task consists of correctly inputting numbers in the boxes below the 5 randomly selected letters. Once the numbers are correctly inputted and the subject clicks "OK", they are given a new random "word" and a new set of code numbers for each letter of the alphabet. The number pad on the right-hand side of the keyboard, along with the *Tab* keys, were disabled for all subjects in all sessions to remove the advantage a particularly experienced computer user could have in the task. The outcome variable measured from the task is words completed per minute, which we call performance.

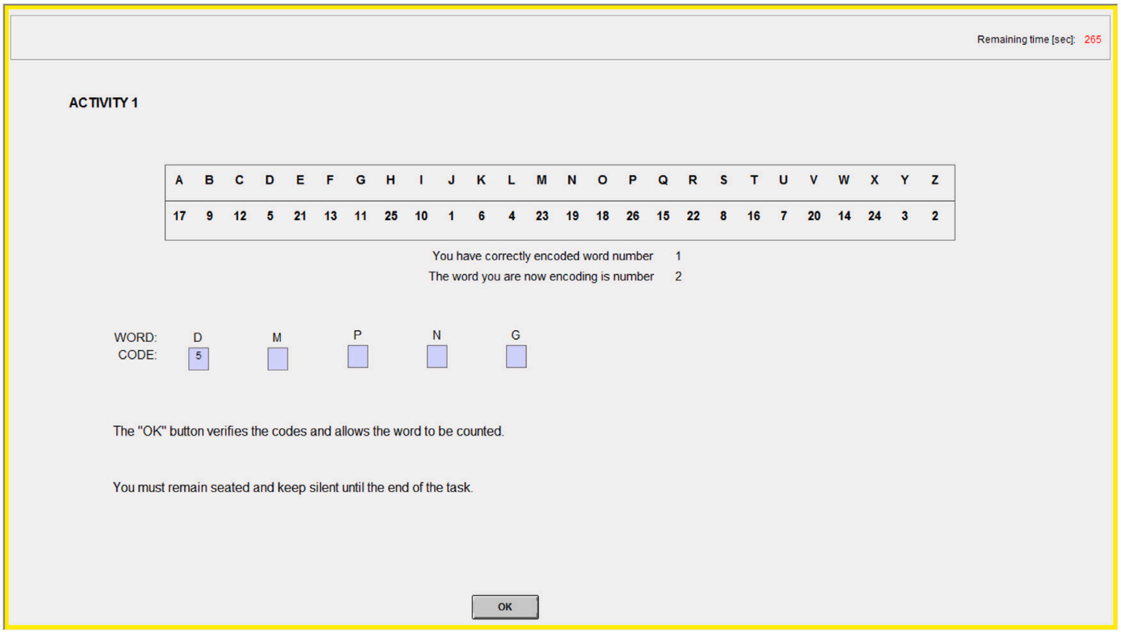
We label our outcome variable *performance* rather than *effort* to be accurate with our terminology. Our task is known in the literature as a real effort task, as it is simple to complete and therefore primarily measures effort over ability and learning. However, as Eriksson et al. (2009) show, mistakes in a real effort task can mean there is a gap between correct submissions (performance/output) and what they label effort (all submissions, correct and incorrect). We expect mistakes to be lower in our real effort task compared to that found by Eriksson et al. (2009), as their task involved adding four two digit numbers together and ours is an easier matching task, though we acknowledge a limitation in our study is that we did not record number of mistakes. The learning effect we find in our control group (see Results section) further demonstrates the importance of acknowledging the distinction between effort and performance, as it may be that effort is remaining similar over time while words completed per minute (performance) is increasing. However, we maintain that the simplicity of the "real effort" task utilised means that our observed performance will be highly correlated with our latent variable of interest, intrinsic motivation. We account for learning in our task over time using a difference-in-differences analytical approach, and discuss whether this approach sufficiently controls for learning in Section 4.3.1.

Details of the activities section are shown in Table 1, and are consistent with the three round design discussed in the theory section. Subjects started with a 2 min practice, to briefly familiarise themselves with the task. The first round was 5 min without incentives or mention of incentives, thus providing a measure of intrinsic motivation. In round 2, subjects were told that they would be given the same task again, for another 5 min. The control group were provided with no incentives nor mention of incentives. All other subjects were given an incentive to complete each word depending on their between subject treatment group — see Section 3.2.1. In round 3, all subjects were given the real effort task for another 5 min, with no incentives, to understand intrinsic motivation after the incentive is applied.

A cash payment at the end of the experiment was made for round 2 (depending on treatment and number of words completed). Thus, the control and charity treatment groups received no payment for their participation in this part of the experiment. The low and high power treatment groups received an average of $1.03 and $21 respectively for their performance in round 2 (see Table 5 in the results section for mean performance by round and treatment). After this portion of the activities were completed, the subjects undertook further tasks which provided the rest of their payments.[3]

A particularly important aspect of the design according to the theory covered in the Background section is subjects' information about the nature of the task, and their information about extrinsic incentives. Thus, the instructions were designed such that the task was framed neutrally, so as to avoid explicitly encouraging or discouraging effort. Furthermore, information about incentives was only provided for the current round of the real effort task. Future rounds were not mentioned to avoid subjects speculating internally about incentives provided in future rounds.

[3] These tasks included a time preference task (earnings of between AUD$10 and AUD$20 in vouchers) and AUD$20 for participating in a survey and discrete choice experiment on health.



Remaining time [sec]: 265

ACTIVITY 1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
17	9	12	5	21	13	11	25	10	1	6	4	23	19	18	26	15	22	8	16	7	20	14	24	3	2

You have correctly encoded word number 1

The word you are now encoding is number 2

WORD:	D	M	P	N	G
CODE:	5				

The "OK" button verifies the codes and allows the word to be counted.

You must remain seated and keep silent until the end of the task.

Fig. 1. Example screen of real effort task given to subjects, with the code for the first letter of the "word" completed.

Table 2
Between subject treatment groups.

Treatment group	Incentive applied (in round 2 only)
Control	None
Low power	$0.05 per word
High power	$1 per word
Charity	Two words plants one indigenous tree within Victoria (equivalent to $1 per word)

To minimise the extent to which subjects were influenced by trying to please the experimenter with their performance (Zizzo, 2010), we made it clear at the commencement of the session that an administration assistant would be paying the subjects in a private, neighbouring room. This arrangement increased the level anonymity of the data collection process, without making it too salient to the subjects to suggest a certain way of acting.

3.2.1. Between subject treatments

The four between subject treatment groups are shown in Table 2. The incentives provided in round 2 include none for the control group, AUD$0.05 per completed "word" completed for low power, and AUD$1 per word for high power. The charity treatment group were told in round 2 that "every 2 words you complete will fund the planting of one indigenous tree in Victoria. A local environmental charity will receive the funds to plant these trees after the experiment". The charity was Tree Project, who state that every AUD$2 donated leads to one tree being planted.[4] Thus, while subjects were not told the monetary amount of their donation until the end of the experiment, it is equivalent to AUD$1 per word completed. To ensure credibility of donations, subjects were also told that a session-level donation receipt would be emailed to them to prove the donation had been made.

3.3. Sample

The sample consists of adults over the age of 18. The sample was restricted by not allowing the typical subject pool, undergraduate students, to participate, to avoid oversampling this population.

Subjects were recruited from Monash University's Centre for Health Economics and the Monash Business Behavioural Laboratory databases, and through other advertisements, including on the Gumtree website (Volunteers Section), the Monash University staff newsletter *The Insider* and the local community newspaper *The Leader*. An example advertisement and email are included in the online appendix. Advertisements were general in nature to not bias the sample towards individuals particularly interested in our research aims; the study was referred to as "an economics experiment aimed at studying behaviours". Subjects signed up by replying to the "BCC" email they were sent, or directly emailing the contact email. They were then given a list of available session times.

Sessions were held on weekdays, at either 12pm or 5:30pm. In order to avoid differences in the composition of the treatment groups, each treatment was assigned to one 12pm session and one 5:30pm session. The aim was to have roughly 50 people in each treatment group. However, the number of no shows in each session had a large variance, meaning it was difficult to reach the required number of subjects in each session. One smaller extra session was run at 12pm for the control treatment to supplement the numbers. Email addresses associated with each session time were recorded in order to send reminders for session times. All data recorded from the actual sessions were anonymised, as per the study's ethics and to reduce potential experimenter demand effects. We collected a number of variables for each subject, allowing us to conduct rigorous balance tests of treatment groups. Overall, we had a strong randomisation procedure, but given the heterogeneity of our subject pool and the variation in no shows each session, we spend the first part of the Results section extensively considering the balance of the treatment groups.

It is standard to compensate subjects for their participation in a study such as this one, given the time and travel costs for participation. Thus, a rough figure of earnings for participation was given to potential subjects at the recruitment stage. Abeler and Nosenzo (2015) find that including potential earnings in a recruitment email for a lab experiment increases sign up rates threefold compared with no mention of monetary reward, but does not impact the measured pro-social or approval motivations of the subjects. The potential earnings for our experiment were not emphasised to subjects after the initial recruitment stage, and at no point were subjects informed that the amount they earned would be linked to their performance, until the relevant point of the

[4] http://www.treeproject.org.au/, accessed 23 August, 2016.

experiment. Additionally, we note that the minimum wage at the time was just under $18 an hour, so the minimum advertised payment is not significantly higher than minimum wage for 1.5 h of time. The campus location would require driving and paying for parking or catching public transport for most people, hence the payment does indeed roughly match time and transport costs for subjects. Thus, it is unlikely we systematically recruited money-motivated subjects.

3.4. Hypotheses

The purpose of this paper is to better understand how extrinsic incentives affect performance and intrinsic motivation across a heterogeneous population. We test three sets of hypotheses across the full sample, plus high and low motivation sub-samples. These three sets are based around the three difference-in-differences: round 2 versus round 1, round 3 versus round 1 and round 3 versus round 2. We compare each incentive treatment to the control, as well as high power versus low, and high power versus charity. As these comparisons suggest a number of hypotheses we summarise our predictions in this section rather than comprehensively list them all.

Many of our hypotheses are tested using difference-in-differences relative to performance in the control group. Thus, it is important to first consider how performance might evolve over the three rounds in the control. We suspect that motivation will remain similar across the three rounds in the control group, but we note the potential for learning effects, as discussed in Section 3.2. Hence, in the results Section 4.3.1, we look for learning effects in the control group, before our formal hypothesis testing. We also consider whether any observed learning effects will impact our hypothesis testing, by looking for evidence of differentiated learning effects between treatment groups. For ease of exposition, in the rest of this section we assume that learning effects are sufficiently controlled for using the difference-in-differences approach.

Our first set of hypotheses consider round 2 performance relative to round 1. Round 1 performance consists of intrinsic motivation only, whereas round 2 performance includes extrinsic motivation from the incentive, plus intrinsic motivation which may be affected by the incentive type and size. Our hypotheses are based on the premise that intrinsic motivation will be crowded out by monetary incentives in the high motivation group, but not in the low motivation group. Hence, we expect that in round 2, the low power monetary incentive will lower performance for the high motivation group, and increase performance for the low motivation group. These counteracting effects lead to an ambiguous effect for the full sample. The high power monetary incentive will increase performance across all groups and the charity incentive will increase performance across all groups. This latter prediction is because we suspect the charity incentive will not crowd out motivation, and indeed may crowd in motivation in the low motivation sub-sample as explained in the theory section above.

Our second set of hypotheses consider round 3 performance relative to round 1. This comparison allows us to measure the change in intrinsic motivation from before the incentive was applied, to after its removal (that is, any crowding out). Following on from above, we predict that relative to the control, the monetary incentives will lower performance to a similar extent in the high motivation group (crowd out). They will have little effect on the low motivation group. The charity incentive will increase performance in the low motivation group (crowd in), and have no impact on the high motivation group.

Our third set of hypotheses consider round 3 performance relative to round 2. This comparison allows us to measure the size of the extrinsic motivation from the incentive, given we expect intrinsic motivation to differ from rounds 1 to 2, but be similar from rounds 2 to 3 (although our design cannot rule this out, as discussed in the theory section). We expect relatively homogeneous effects across the full sample and two sub-samples. However, we note diminishing marginal returns may mean for example that there is little difference in the low versus high power incentives for the high motivation sub-sample. With that point in mind, we expect the high power incentive will see at least as large a drop in effort as the low power and charity incentives. All incentives will have a drop in effort from round 2 to 3 relative to the control.

3.5. Analytical approach

We employ a combination of non-parametric and linear regression models to test our hypotheses, with a focus on the difference-in-differences between rounds, between treatment groups. We model the full sample, plus a high motivation sub-sample (performance is above the median in round 1) and a low motivation sub-sample (performance is at or below the median in round 1).

We first estimate the significance of the difference-in-differences between treatments using a non-parametric Mann–Whitney U test. Then we estimate the following linear regression model using OLS:

$$p_{i,r} = \alpha + \boldsymbol{\beta}_r \boldsymbol{r} + \boldsymbol{\beta}_t \boldsymbol{t}_i + \boldsymbol{\beta}_{rt}(\boldsymbol{r} \otimes \boldsymbol{t}_i) + \epsilon_{i,r}. \tag{1}$$

In this model, $p_{i,r}$ is performance level for individual i in round r, α is the intercept coefficient, and $\boldsymbol{\beta}_r$ is a vector of coefficients on dummy variables for the round, $\boldsymbol{r}$, relative to round 1. These coefficients will show any changes in performance in rounds 2 and 3, relative to round 1, for the base treatment group (control). Thus, they measure how performance evolves over the three rounds for the baseline control treatment.

The vector $\boldsymbol{\beta}_t$ is coefficients on the dummy variables for the treatment group of individual i, $\boldsymbol{t}_i$, relative to the control. This coefficient will test whether there is any difference in round 1 performance (base round) between the treatment groups and the control group (base treatment group). As there is no differences in the treatment groups to this point, these coefficients serve as a randomisation check.

Next, $\boldsymbol{\beta}_{rt}$ is a vector of coefficients on the vector of all interactions between rounds and treatment groups, $(\boldsymbol{r} \otimes \boldsymbol{t}_i)$. These coefficients provide comparisons of the difference-in-differences of the treatment groups and rounds, relative to control. We test the full set of hypotheses using F-test for the relevant combination of coefficients, as explained in the results section.

Finally, $\epsilon_{i,r}$ incorporates the error terms. When estimating this model, standard errors are clustered by individual to account for the panel nature of the data.

4. Results

4.1. Balance tests and data cleaning

Due to the heterogeneous nature of the sample, we start by undertaking a balance test to double check the randomisation procedure. We note that we first drop one subject from all analysis as they did not complete the practice round due to technical issues, leaving an initial sample size of 178. The subject's round 1 performance was low compared to their other rounds, likely due to the lack of a practice round. Table 3 shows a multinomial logit balance test between treatments. The model estimates the likelihood of a subject being in a particular treatment group, using the main observed covariates of the subjects as predictors. Each coefficient shown is relative to the control group (base category). We include impatience, present bias, future bias and waist-to-height ratio, which are all measured from the latter part of the experiment. Overall, the model is not statistically significant; only two coefficients of the 27 are significant at the 5% level. Therefore, the observables do not predict assignment to any specific treatment group, which suggests there are not any systematic differences between the treatment groups.

Next, we test for differences in round 1 performance. There should be no treatment group differences here as all subjects received the same treatment up until this point. Table 4 shows round 1 performance distribution. We perform a non-parametric Kruskal–Wallis test for differences in the distribution of the treatment groups. This yields a p-value of 0.141, again showing no systematic differences between treatment groups.

However, we note the high power treatment does have a higher minimum, 25th percentile and median than the other treatments. This

Table 3
Multinomial logit balance test between treatments.

	Dependent variable: Treatment group (relative to control)		
	Low power	High power	Charity
Constant	−0.804	−1.241	0.858
	(2.911)	(2.988)	(2.928)
Age	−0.006	−0.020	−0.005
	(0.017)	(0.018)	(0.017)
Female	−0.469	−0.011	0.092
	(0.440)	(0.450)	(0.447)
Education years	0.024	0.020	−0.036
	(0.156)	(0.160)	(0.157)
Personal income	−0.004	0.006	0.005
	(0.008)	(0.008)	(0.008)
Impatience	−0.060	−0.019	0.091
	(0.090)	(0.091)	(0.089)
Present bias	0.826	0.754	0.470
	(0.587)	(0.598)	(0.610)
Future bias	1.225**	1.138**	0.505
	(0.541)	(0.542)	(0.539)
Waist-to-height ratio	1.723	2.108	−2.004
	(3.140)	(3.147)	(3.178)
Treatment group N	46	44	44
Total N			178
Log Likelihood			−238.630
LR test p-value			0.881

Notes: Standard errors are in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 4
Distribution of first round performance (words per minute) by treatment on the full data set.

Treatment	*Percentile*				
	Min	25th	Median	75th	Max
All	1	3.2	3.6	4.2	6.4
Control	1.4	2.95	3.6	4	6
Low power	1.4	3	3.4	3.95	5.8
High power	2.4	3.55	3.80	4.2	5.6
Charity	1	3.2	3.6	4.25	6.4

Table 5
Summary statistics.

Statistic	N	Mean	St. Dev.	Min	Max
Demographics					
Age	168	36.3	14.6	19	78
Education years	168	14.7	1.4	11	16
Personal income	168	36.4	32.8	10	138
Female	168	0.58			
Performance (Words per minute)					
Round 1	168	3.7	0.8	2	6
Round 2					
Control	40	3.9	0.8	2	6
Low power	41	4.1	0.8	2.6	6.2
High power	44	4.2	0.7	2.4	5.8
Charity	43	3.9	0.9	2.4	6.6
Round 3					
Control	40	4.0	0.9	2.4	5.8
Low power	41	4.0	0.8	3	6
High power	44	3.9	0.8	2	6
Charity	43	4.1	0.9	2.2	6.6
Total earnings ($)					
Control[a]	40	33.07	3.96	30	40
Low power	41	33.63	3.27	30.80	40.95
High power	44	53.03	4.57	42	65
Charity[a]	43	32.12	3.42	30	40

Notes:

[a]All earnings in these Treatments were from completing subsequent tasks that are not part of this paper.

Personal income is a categorical variable, but is included as a continuous variable using the category mid-points.

difference is picked up in a non-parametric Mann–Whitney U Test, which has a p-value of 0.058 when comparing round 1 performance between the high power and control. This difference shows up in the difference in difference modelling too.

Thus, we take a precautionary approach and drop the lowest 5th percentile of the sample, based on round 1 performance. This removes 10 individuals with a score of 2 words per minute or less in round 1, leaving 40 in the control, 41 in low power, the original 44 in high power and 43 in the charity treatment. Now, the Kruskal–Wallis test for round 1 performance yields a p-value of 0.479 and the Mann–Whitney U Test comparing round 1 performance between the high power and control has a p-value of 0.215. The balance test for this reduced treatment is essentially the same as shown Table 3, with the same level of statistical significance on the same two of 27 coefficients.

Thus, we undertake all further analysis with the marginally reduced sample. We include analysis on the full data set in the appendix and discuss any implication from dropping these individuals in the results in Section 4.4 at the end of the Results.

4.2. Summary statistics

Table 5 summarises the main demographics collected on the study subjects. We do not make any claims of representativeness, rather we aimed to have a heterogeneous subject pool for demographic variables such as age and income. We also aimed to have balanced treatment groups, through random assignment, as covered previously. The subject pool is mostly non-students (72%) and entirely non-undergraduate students. Overall, the sample is younger and better educated compared with recent census data for the study location. The gender split was slightly higher for females, at 58%.

Performance in terms of words completed per minute in each round is summarised in the middle section of Table 5, with rounds 2 and 3 shown at an overall level and by treatment. Mean performance in each round by treatment is shown visually in Fig. 2. Performance of those in the control and charity treatments increases each round, whereas this is not the case for those in the monetary incentive treatments (low and high power). In these two treatments, performance is increasing between each round except for rounds 2 to 3, where there is a decrease, most noticeably in high power.

Summary statistics of total earnings by subject in each treatment is shown at the bottom of Table 5. Note that all earnings in the control and charity treatments came after the real effort task analysed here, as outlined in Section 3.2.

4.3. Hypothesis testing

We present the difference-in-differences modelling to test our hypotheses in Tables 6 and 7, as outlined in Section 3.5. We start by briefly summarising these results, before discussing them in relation to each of our hypotheses.

In the top half of Table 6 we present p-values from the non-parametric Mann–Whitney U test for the difference in performance between rounds, for each treatment compared with control, and each of the treatment groups compared with high power. The first three rows of results are for the full sample, then next three or for the high motivation sub-sample, followed by the low motivation sub-sample. The bottom half of the table repeats the top half, but using the linear model presented in Table 7 to calculate the p-values.

Table 7 shows a difference-in-differences model estimated using OLS, as per Eq. (1) in Section 3.5. Column (1) presents the estimates for the full sample, column (2) the high motivation sub-sample and column (3) the low motivation sub-sample. The dependent variable is performance level in each round, meaning each individual has three observations.

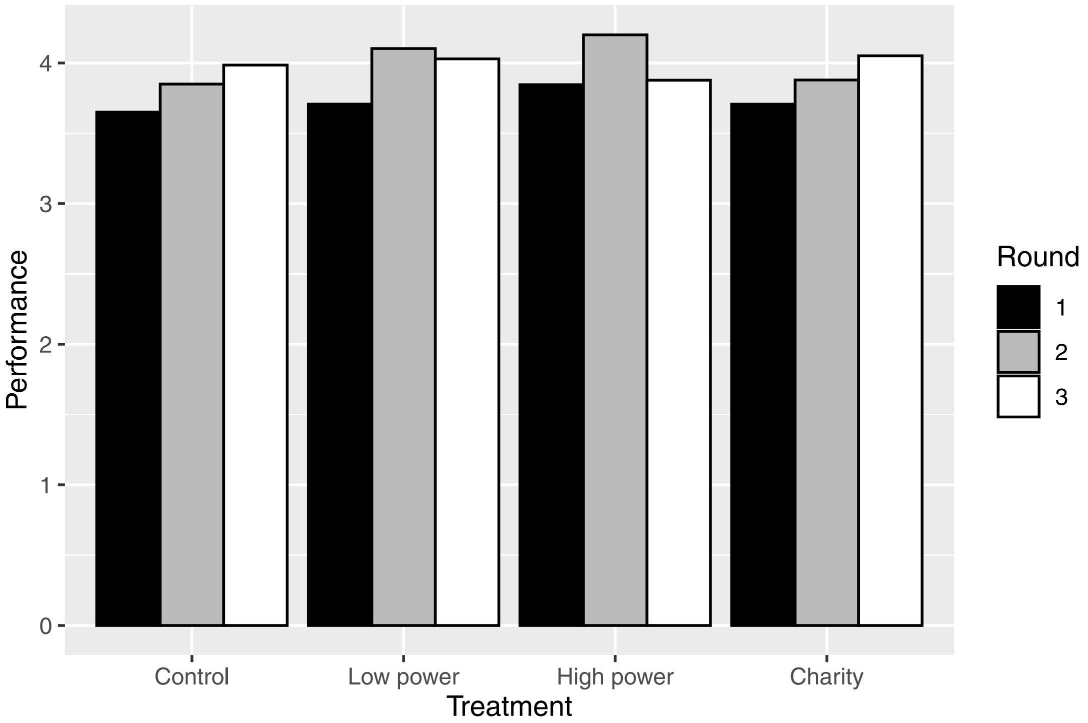


Fig. 2. Average performance (words per minute) in each round, by treatment group.

Table 6
Two-sided p-values for difference-in-differences of performance between each treatment group and control, and each treatment group and high power, between rounds.

	Difference-in-differences: p-value				
	Control vs			High power vs	
	Low power	High power	Charity	Low power	Charity
			Mann–Whitney U test		
Full sample					
R1 to R2	0.007***	0.128	0.646	0.314	0.046**
R1 to R3	0.969	0.005***	0.879	0.002***	0.006***
R2 to R3	0.020**	0.000***	0.469	0.025**	0.000***
R1 > median					
R1 to R2	0.087*	0.833	0.332	0.045**	0.334
R1 to R3	0.976	0.001***	0.299	0.000***	0.021**
R2 to R3	0.300	0.000***	0.794	0.008***	0.000***
R1 ≤ median					
R1 to R2	0.041**	0.006***	0.734	0.297	0.015**
R1 to R3	0.923	0.954	0.231	0.893	0.363
R2 to R3	0.040**	0.022**	0.481	0.566	0.002***
			Linear model (see Table 7)		
Full sample					
R1 to R2	0.004***	0.049**	0.681	0.607	0.022**
R1 to R3	0.873	0.010**	0.919	0.010**	0.009***
R2 to R3	0.007***	0.000***	0.625	0.017**	0.000***
R1 > median					
R1 to R2	0.147	0.779	0.159	0.070*	0.226
R1 to R3	0.993	0.001***	0.291	0.000***	0.015**
R2 to R3	0.162	0.000***	0.976	0.015**	0.001***
R1 ≤ median					
R1 to R2	0.010**	0.002***	0.497	0.221	0.009***
R1 to R3	0.873	0.924	0.283	0.998	0.431
R2 to R3	0.024**	0.015**	0.524	0.339	0.003***

Notes: R1 is shorthand for round 1, etc. Tests for the linear model are t-tests for single coefficients, and F-tests for multiple, using clustered standard errors as per Table 7. *p < 0.1; **p < 0.05; ***p < 0.01.

After the constant term, the first two variables in Table 7 are round dummies, relative to round 1 (base round). These identify the changes in performance between rounds 2 and 3 and round 1 for the control group (base treatment).

The next three variables in descending order are dummies for each treatment. These coefficients identify whether there is any difference in performance between the incentive treatment groups and the control group in round 1 (base treatment and base round). There are no

Table 7
Difference-in-differences models of rounds 1 to 3, including all treatments.

	Dependent variable: Words/minute in each round		
	All (1)	R1 > median (2)	R1 ≤ median (3)
Constant	3.650***	4.242***	3.114***
	(0.121)	(0.127)	(0.104)
Round 2	0.200***	0.242***	0.162**
	(0.047)	(0.071)	(0.063)
Round 3	0.335***	0.379***	0.295***
	(0.062)	(0.096)	(0.083)
Low power	0.057	0.113	0.086
	(0.169)	(0.198)	(0.125)
High power	0.195	0.012	0.141
	(0.158)	(0.161)	(0.136)
Charity	0.057	0.247	0.030
	(0.177)	(0.198)	(0.134)
Low power*R2	0.195***	0.147	0.238**
	(0.067)	(0.101)	(0.092)
High power*R2	0.155**	−0.027	0.394***
	(0.078)	(0.095)	(0.125)
Charity*R2	−0.028	−0.142	0.062
	(0.068)	(0.101)	(0.091)
Low power*R3	−0.013	−0.001	−0.017
	(0.081)	(0.127)	(0.106)
High power*R3	−0.303***	−0.517***	−0.017
	(0.117)	(0.150)	(0.184)
Charity*R3	0.009	−0.146	0.129
	(0.091)	(0.137)	(0.120)
N	168	81	87
Observations	504	241	261
Adjusted R^2	0.018	0.050	0.100
F Statistic	16.382***	6.854***	9.725***

Notes: Standard errors clustered at the individual level are in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

statistically significant coefficients here across any of the columns, as expected with randomised treatment groups.

The three variables following are treatment dummies interacted with the round 2 dummy. These coefficients thus show the difference-in-differences between the treatments and the control (base treatment) group, between round 2 and round 1 (base round). For example, in column (1) we can see that for the full sample the increase in performance from round 1 to 2 is statistically larger for the low power treatment, compared with the control, at the 1% level. Individuals in the low power treatment increased their performance from round 1 to 2 on average by 0.195 more words per minute than the control.

The final three coefficients are the treatment dummies interacted with the round 3 dummy. As above, these show the difference-in-differences between the treatments and the control (base treatment) group, between round 3 and round 1 (base round). Hence, in column (1) we can see that the low power incentive has no statistically significant difference-in-differences between round 3 and round 1 compared with the control group. In order to understand the difference-in-differences between a treatment group and the control for round 3 versus round 2, we take the relevant round 3 interaction term and subtract the round 2 interaction term. For example, for the low power incentive, performance drops by 0.208 relative to the difference observed in the control group. To compare the low power treatment group with the high power, we just compare the differences directly (for example −0.208 versus −0.458 for the full sample). The other component in the change in effort from round 2 to 3, given by the difference between round dummies, occurs for both treatments, so cancels out. Hence, these differences in treatment-round interactions are the parameters we compare statistically to provide the relevant results in Table 6.

4.3.1. *Learning effects*

We first consider how performance evolves over the three rounds in the control. For the control group, we assume motivation is similar across the three rounds and look for any learning, and comment on how it might impact the testing of other hypotheses. We can see the impact of learning by looking at the round 2 and round 3 coefficients in Table 7. These coefficients show changes in performance in these rounds for the control group, relative to round 1. We find that performance for the control group increases in each round, in the full sample and the high and low motivation sub-samples. The learning effects are similar across the high a low motivation sub-samples.

This finding shows the importance of using the difference-in-differences approach, as it helps control for learning effects. We acknowledge the potential for differentiated learning effects between treatment groups; however, we find no evidence of this and we find no evidence of situations where differentiated learning could greatly impact our results. Indeed, we see an increase in performance in the low and high power treatments when they are applied in round 2, which could arguably increase the speed of learning. However, no treatment groups show a higher increase in performance compared with the control from round 1 to 3. While learning could mask crowding out for the low power treatment in round 3, we believe this unlikely as there is no crowding out in round 3 for low power shown across the three columns of Table 7, regardless of whether it increased performance in round 2 (columns (1) and (3)) or not (column (2)). Hence, we believe learning has been sufficiently controlled for and our subsequent analysis reflects this assertion.

4.3.2. *Effect of extrinsic incentives during application*

The non-parametric and parametric models presented in Tables 6 and 7 both support the following finding in relation to two our first set of hypotheses:

Result 1a: The low power incentive has a statistically significant positive impact on performance, relative to control, when it is applied for the low motivation sub-sample, as well as the full sample.

This result is given by the difference-in-differences between the low power treatment group and the control, for round 2 versus round 1 performance, and has a significance at least at the 5% level. The precise non-parametric and linear model p-values are given in Table 6 (R1 to R2). The performance increase for the low motivation sub-sample is 7.3%. There is mixed evidence for the high motivation sub-sample. The Mann–Whitney U test shows it increases performance for the high motivation sub-sample too, though only at the 10% level of significance. The linear regression has a *p*-value of 0.147. This finding is in line with our prediction for the low motivation sub-sample. But, we expected a decrease in relative performance in the high motivation group, due to crowding out of the low power incentive being stronger than its additional extrinsic motivation, which we do not find. Our testing of the second set of hypotheses below helps us unpick whether a low power monetary incentive crowds out motivation.

Next, we find:

Result 1b: The high power incentive is effective at raising performance for individuals in the low motivation sub-sample relative to control, when applied (in round 2).

This result has a *p*-value of 0.002 in the linear regression, and raises performance by 12% over the control. There are mixed results for the full sample, with the linear regression finding performance increases for the full sample with a *p*-value of 0.047, but the *p*-value is 0.128 in the non-parametric test. There is no difference for the high motivation sub-sample. This finding is again only partially in line with our prediction, which was that the high power incentive would be effective at increasing performance in all groups. This finding could be the result of crowding out in the high motivation sub-sample from the high power incentive; again, we test this when we test second set of hypotheses below.

Next, we find:

Result 1c: For the high motivation sub-sample, the low power incentive is more effective at raising performance than the high, when applied (in round 2). For the low motivation sub-sample and full sample, there is no difference between the two monetary incentives.

For the high motivation group, this result follows from the high power vs low power comparison shown in Table 6, and the higher coefficient of Low power*R2 versus High power*R2 in Table 7 (which is the relevant comparison). This finding is at the 5% level for the non-parametric testing and at the 10% level for the linear model, and is consistent with our original prediction noted above. We did however expect the high incentive would have a larger effect for the low motivation group.

Finally, we find:

Result 1d: The charity incentive does not change relative performance when applied in round 2 compared to the control. The charity incentive is less effective at increasing effort when applied than the high power incentive, except for the high motivation group, for which there is no difference.

The first part of this finding is given by the lack of significance across all our testing for the charity incentive (difference-in-differences for round 2 versus round 1 for the charity versus the control, eg. Charity*R2 coefficients). It is contrary to our hypothesis, as we expected the charity incentive to increase performance across all groups. We did however expect the high power incentive to be more effective than the charity incentive at raising effort when applied. The finding for the full sample and low motivation sample is significant at the 5% or 1% level.

Overall, our findings show monetary incentives increase performance in our low motivation sub-sample, but do not necessarily increase them in our high motivation group. Thus, we next test for crowding out in round 3 vs round 1 performance between treatment groups and the control.

4.3.3. Effect of extrinsic incentives after removal

We now look at the second set of hypotheses, to understand no-incentive performance after removal of incentives, versus before their application (round 3 versus round 1):

Result 2a: The low power incentive does not lower performance in round 3 versus round 1 (does not crowd out motivation) for any group, relative to the control.

Result 2b: The high power incentive lowers performance in round 3 versus round 1 on average, compared to both control, low power and charity, driven by the high motivation sub-sample. It does not lower performance in round 3 versus round 1 for the low motivation sub-sample.

These results are consistent across the non-parametric and parametric testing. The p-values showing crowding out from the high power incentive in the full sample and the high motivation sub-sample are 0.00997 and 0.0007 respectively (Table 7, round 3 interaction coefficients). The drop in performance in the high motivation group is around 11% below where we would expect them to be if they were assigned to the control or low power treatments.

Interpreting these results alongside Results 1a and 1b, we can see two effects are at play (or not at play). First, the low power incentive does not crowd out intrinsic motivation. Thus, any increase in effort when it is applied in round 2 reflect only the added extrinsic motivation it provides. Hence, the second effect we see is the low power incentive is most motivating to the low motivation group when it is applied.

The high power incentive also increases performance in the low motivation sub-sample without crowding out motivation. However, the high power incentive shows the two effects simultaneously in the high motivation sub-sample. It both increases performance and crowds out motivation, cancelling out the effect of the incentive when applied in round 2. When it is removed, we see the reduced performance (caused by crowding out) for the highly motivated from the high power monetary incentive.

We also find:

Result 2c: The charity incentive does not affect intrinsic motivation.

This result holds for all groups using both non-parametric and parametric testing. While the charity incentive is not effective in increasing effort (Result 1d) we find it does not crowd out intrinsic motivation either. Hence, Result 2c is consistent with our prediction that a non-monetary incentive would not cause crowding out.

We also look at the third set of hypotheses, to better understand the removal of incentives relative to when they are applied (round 3 versus round 2). This comparison shows us the size of the extrinsic motivation of each incentive in round 2, given we assume intrinsic motivation is similar from round 2 to 3. First, we find:

Result 3a: Performance for both monetary incentives drops relative to the no-incentive control when the incentives are removed. This holds for all groups, except for low power incentive and the high motivation group.

This result hold at the 5 or 1% levels, as shown in Table 6. The change in performance from round 2 to 3 is given in Table 7. For example, we can calculate the change for the low power treatment group, relative to control, by subtracting the Low power*R2 coefficient from Low power*R3. This gives us a drop of 0.255 for the low motivation sub-sample, which is around 7% of the low power treatment group effort in round 2.

This result shows us that the monetary incentives do provide extrinsic motivation to increase performance, after taking into account any crowding out effect they have on intrinsic motivation. Comparing the two monetary incentives with each other, we find:

Result 3b: The drop in performance for high power incentives is larger than for low power incentives for the high motivation group, but there is no difference between the monetary incentives for the low motivation group.

This result holds at the 5 or 1% levels, again shown in Table 6. The result also holds at the 5% level for the full sample, clearly driven by the high motivation sub-sample. For the highly motivated, the change in effort from round 2 to 3 in the low power treatment is −0.148, whereas it is −0.49 for the high power.[5]

For the high motivation group, the result is consistent with our theory that the high power incentive provides more extrinsic motivation than the low power incentive. There is no difference between the incentives in round 2 because the extrinsic motivation from the high power incentive is offset by lower intrinsic motivation. This is not the case for the low motivation group, contrary to our prediction. For this group, while the high power incentive is no worse for crowding out than the low, it also does not provide significantly more extrinsic motivation.

Finally, we find:

Result 3c: There is no difference between the charity treatment and the control group when removing the charity incentive. The drop in performance from removing the high power incentive is statistically different from the charity treatment. These results hold across the full sample and the low and high motivation sub-samples.

This result is consistent with our previous findings for the charity incentive. There is no statistical difference between the control and charity treatment groups, hence the differences between the charity and high power treatments are also statistically the same as the control versus high power.

[5] These calculations are given by the round 3 interaction term (eg. Low power*R3) minus the round 2 interaction term, as described in the first part of Section 4.3. Technically, there is also a slight increase in effort going from round 2 to round 3, given by the difference in the round dummies, of 0.137. However, this increase applies to all treatment groups, so it cancels out for the purpose of comparison between any given treatment groups. This is the learning effect across all treatment groups, already discussed in Section 4.3.1.

4.4. Supplementary results

We briefly comment on the supplementary results in the appendix. First, Table A.1 shows p-values for non-parametric testing of the differences between rounds and the control (rather than difference-in-differences). While much of the literature takes this approach, we argue these results are less informative and they do not account for initial motivation (first round performance). These results reinforce our strong finding that the high power incentive crowds out motivation for the highly motivated (R3 column for R1 > median), and that the monetary incentives are effective for the low motivation sub-sample without crowding out intrinsic motivation (higher effort in R2 column for low motivation sub-sample, no differences in R3).

Table A.2 repeats the regressions of Table 7, but with the original sample without dropping the lowest 5th percentile (as outlined in Section 4.1). We see the rationale for this data cleaning, as the high power treatment has significantly higher performance for the full sample and the low motivation sub-sample ($p = 0.032$ and $p = 0.027$ respectively), due to the lack of representation of individuals with performance below 2.4 words per minute in round 1. Column (2) is identical in this Table A.2 compared with Table 7 as the median round 1 performance is still 3.6 in our supplementary results, thus none of our results for the high motivation sub-sample are affected. Hence, the result that high incentives crowd out motivation for the highly motivated is unaffected by our decision to drop the lowest 5th percentile. We also do not see any crowding out for the low motivation sub-sample. However, including the least motivated reduces the strength of the results on the round 2 performance for the monetary incentives. The 10 individuals dropped across the treatment groups do not appear to respond to incentives, but they are too few observations from which to draw any conclusions. Overall, all our results are robust to the original sample, except Result 1a, for which the evidence is weaker when including the lowest 5% of the full sample, in terms of round 1 performance (initial motivation).

5. Discussion and conclusion

In this paper we present a lab-in-the-field experiment on intrinsic motivation, and how a range of extrinsic incentives interact with baseline intrinsic motivation. We employ a rich within- and between-subject design that allows us to use a difference-in-differences approach to test the main hypotheses and provide new insights. This approach adds to the existing economic literature on crowding out, which generally does not account for baseline intrinsic motivation. Our experimental design also allows us to better understand how both intrinsic and extrinsic motivation contributes to performance.

For the full sample, we find support for “pay — but do not pay too much” (Pokorny, 2008) as both low and high power monetary incentives are effective at raising performance in the real effort task, but the high power incentive crowds out motivation on average when it removed. While we find a similar effect for low motivation individuals when the monetary incentives are applied, we find no crowding out when they are removed. For the highly motivated group, we see little impact from the incentives when applied, but a strong crowding out effect from the high power incentive, which is our most clear finding. Hence, we find “don’t pay the highly motivated too much”. Thus, it appears that more underlying motivation creates more potential for crowding out.

Hence, we do not find a V-shaped response to incentives in either the high or low motivation groups, as found by Gneezy and Rustichini (2000) and Heyman and Ariely (2004). It appears in our case that the low power monetary incentive was so small, and the high motivation sub-sample was motivated enough, that the low power monetary incentive did not crowd out intrinsic motivation. We did not expect to see any meaningful crowding out in the low motivation sub-sample, given their low starting point. This finding thus shows that when small enough, low monetary incentives will not necessarily cause crowding out across a heterogeneous population, which is more consistent with Pokorny (2008) and DellaVigna and Pope (2018). Indeed, further research could look at an even wider range of monetary incentives, as per DellaVigna and Pope (2018), but accounting for initial and subsequent intrinsic motivation for the task, as we have done with our design. This could help establish a finer understanding of the relationship between size of monetary incentive and level of crowding out, by level of initial intrinsic motivation.

We find the charity incentive is not powerful enough to increase performance, despite being of the same monetary value as the high powered monetary incentive. This is in contrast to the findings of Imas (2014) and DellaVigna and Pope (2018). Building on these studies though, with our unique experimental design we also find that the charity incentive does not have a significant crowding out effect, rather it is overall neutral.

Our findings generally suggest future research into targeting incentives at low motivation individuals. This research should investigate what effect common knowledge of this targeting has on both the low and high type individuals, given our subjects had common knowledge that all other subjects in their session faced the same incentives. Research on common knowledge is important when considering policy interventions that are openly targeted at specific individuals. It is an open question as to how low motivation individuals might respond to being singled out on the basis of their previous performance. Finally, work such as Segal (2012) that we highlight in the background section provide important insights into the drivers of intrinsic motivation. While it is beyond the scope of this paper to look at these, it seems an important area for further research.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2022.101972. Included are supplementary results, experimental instructions and recruitment materials.

References

Abeler, J., & Nosenzo, D. (2015). Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, *18*(2), 195–214.

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

Balart, P., Oosterveen, M., & Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review*, *63*, 134–153.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, *70*(3), 489–520.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *The American Economic Review*, *96*(5), 1652–1678.

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

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008.

Cubel, M., Nuevo-Chiquero, A., Sanchez-Pages, S., & Vidal-Fernandez, M. (2016). Do personality traits affect productivity? evidence from the laboratory. *The Economic Journal*, *126*(592), 654–681.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105–115.

Deci, E. L., & Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. In L. Berkowitz (Ed.), *Advances in experimental social psychology. Vol. 13* (pp. 39–80). Academic Press.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268.

DellaVigna, S., & Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies*, *85*(2), 1029–1069.

Dessí, R., & Rustichini, A. (2015). *Strong intrinsic motivation*: *Working Paper TSE - 567*, (p. 24). Toulouse School of Economics.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679–688.

Erkal, N., Gangadharan, L., & Koh, B. H. (2018). Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review*, *101*, 528–545.

Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *The American Economic Review*, *101*(7), 3330–3348.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Gill, D., & Prowse, V. L. (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review*, *102*(1), 469–503.

Giusti, G., & Dopeso-Fernández, R. (2020). Incentive magnitude and reference point shifting: a laboratory experiment. *International Journal of Manpower*, *41*(8), 1157–1177.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics*, 791–810.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Heyman, J., & Ariely, D. (2004). Effort for payment a tale of two markets. *Psychological Science*, *15*(11), 787–793.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119.

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

Kajackaite, A., & Werner, P. (2015). The incentive effects of performance requirements – A real effort experiment. *Journal of Economic Psychology*, *49*, 84–94.

Kautz, T., Heckman, J., Diris, R., ter Weel, B., & Borghans, L. (2014). *Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success*: *Technical Report w20749*, Cambridge, MA: National Bureau of Economic Research.

Lee, W., & Reeve, J. (2017). Identifying the neural substrates of intrinsic motivation during task performance. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(5), 939–953.

Ma, Q., Jin, J., Meng, L., & Shen, Q. (2014). The dark side of monetary incentive: How does extrinsic reward crowd out intrinsic motivation. *NeuroReport*, *25*(3), 194–198.

Pascual-Ezama, D., Prelec, D., & Dunfield, D. (2013). Motivation, money, prestige and cheats. *Journal of Economic Behaviour and Organization*, *93*, 367–373.

Pokorny, K. (2008). Pay—but do not pay too much. *Journal of Economic Behaviour and Organization*, *66*(2), 251–264.

Promberger, M., & Marteau, T. M. (2013). When do financial incentives reduce intrinsic motivation? comparing behaviors studied in psychological and economic literatures. *Health Psychology*, *32*(9), 950–957.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, Article 101860.

Ryan, R. M., Deci, E. L., Vansteenkiste, M., & Soenens, B. (2021). Building a science of motivated persons: Self-determination theory's empirical approach to human experience and the regulation of behavior. *Motivation Science*, *7*(2), 97–110.

Rydval, O., & Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: an illustration. *Economics Letters*, *85*(3), 315–320.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), 1438–1457.

Staněk, R., & Krčál, O. (2019). Time preferences, cognitive abilities and intrinsic motivation to exert effort. *Applied Economics Letters*, *26*(12), 1033–1037.

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.

Taylor, M. P. (2020). Heterogeneous motivation and cognitive ability in the lab. *Journal of Behavioral and Experimental Economics*, *85*, Article 101523.

Titmuss, R. M. (1970). *The gift relationship: from human blood to social policy*. London: Allen & Unwin.

Viceisza, A. C. G. (2016). Creating a lab in the field: Economics experiments for policymaking. *Journal of Economic Surveys*, *30*(5), 835–854.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.