The Dynamic Effect of Incentives on Post-Reward Task Engagement

Indranil Goswami

University of Chicago

Booth School of Business

Oleg Urminsky

University of Chicago

Booth School of Business

*** Draft version: Please contact the authors for an updated version ***

Emails: igoswami@chicagobooth.edu, oleg.urminsky@chicagobooth.edu

Abstract

Although incentives can be a powerful motivator of behavior when they are available, an influential body of research has suggested that rewards can persistently reduce engagement after they end. This research has resulted in widespread skepticism among practitioners and academics alike about using incentives to motivate behavior change. However, recent field studies looking at the longer-term effects of temporary incentives have *not* found such detrimental behavior. We design an experimental framework to study dynamic behavior under temporary rewards, and find that although there is a robust decrease in engagement immediately after the incentive ends, engagement returns to a post-reward baseline that is equal to or exceeds the initial baseline. As a result, the net effect of temporary incentives on behavior is strongly positive. The decrease in post-reward engagement is not on account of a reduction in intrinsic motivation, but is instead driven by a desire to take a 'break', consistent with maintaining a balance between goals with primarily immediate and primarily delayed benefits. Further supporting this interpretation, the initial decrease in post-reward engagement is reduced by contextual factors (such as less task difficulty and higher magnitude incentives) that reduce the imbalance between effort and leisure. These findings are contrary to the predictions of major established accounts and have important implications for designing effective incentive policies to motivate behavior change.

Keywords: post-reward engagement reduction, intrinsic motivation, incentives and rewards, dynamic behavior, decision making, effort-balancing

Incentives can be a powerful tool to influence behavior, in settings as diverse as health, education, employment, and marketing promotions (Gneezy, Meier, & Rey-Biel, 2011; Prendergast, 1999; DelVecchio, Henard, & Freling, 2006) where people's decisions often involve a tradeoff between immediate and delayed benefits (Loewenstein, Brennan, & Volpp, 2007). Providing immediate temporary incentives can help motivate people to take beneficial action, countering the effects of hyperbolic discounting and present bias (Ainslie, 1975; Urminsky & Zauberman, 2015) . However, policy makers are often skeptical about using incentives, based on psychological theories of motivation. Critics contend that economic incentives will "smother people's enthusiasm for activities they might otherwise enjoy" (Kohn, 1999). As a result, people are predicted to engage in less of the incentivized behavior after an incentive ends than if the incentive had not been introduced in the first place (Pink, 2011).

This negative view of incentives stems from a large and influential academic literature on Cognitive Evaluation Theory (Deci & Ryan, 1985) and the Overjustification Hypothesis (Lepper & Greene, 1978) which found that adults did less of a task in lab studies (compared to an non-incentivized control group) immediately after an incentive ended. However, these studies only measured post-reward behavior immediately after the incentives were withdrawn, and did not examine the dynamics of post-reward behavior. Nevertheless, a highly influential meta-analysis of this literature generalized beyond the findings to warn that "if people use tangible [i.e., monetary] rewards, it is necessary that they be extremely careful… about the intrinsic motivation and *task persistence* of the people they are rewarding" (Deci et al., 1999, p. 656, emphasis added). In contrast with the prediction that incentives permanently reduce intrinsic motivation (Dickinson, 1989; Tang & Hall, 1995), recent program-evaluation studies that measured long-term effects after incentives ended have not found evidence of post-reward reduction in

incentivized behaviors, with some even reporting positive long-term effects (e.g., workplace smoking cessation, Halpern et al., 2015; academic performance, Jackson, 2010).

In this research, we investigate the question raised by these conflicting findings: what are the dynamic effects of incentives on people's task engagement once the incentive ends? To answer this question, we have developed an experimental framework that facilitates dynamic measurement of post-incentive task engagement over a series of choices. We find initial reduction in engagement with the incentivized task after the rewards end, consistent with the findings of prior lab experiments. However, this reduction is momentary in duration and the incentive yields no persistent negative effect, or even a positive longer-term effect over multiple rounds, consistent with field studies of post-incentive behavior. Our results suggest that prior theories of intrinsic motivation were incomplete for predicting post-incentive behavior.

Extending the literatures on goal-balancing (Fishbach & Dhar, 2005) and justification in decision-making (Lerner & Tetlock, 1999; Shafir, Simonson, & Tversky, 1993), we suggest that people balance effort (which often yields delayed benefits) with the more immediate enjoyment of leisure. Incentives can play a dual role in this balancing process. First, incentives motivate greater effort, and investing more effort can disturb one's perceived sense of balance, justifying more leisure immediately after the incentives end. Second, incentives are rewarding (Knutson, Adams, Fong & Hommer 2001). Sufficiently generous incentives may increase liking for the incentivized task and reduce the need for leisure, via positive reinforcement (De Houwer, Thomas, & Baeyens, 2001; Razran, 1954). Our account makes testable predictions about how the effect of temporary incentives on post-reward behavior would vary with contextual factors and provides different recommendations for how to design effective temporary incentive policies.

Next, we define post-reward dynamic task engagement and briefly review prior accounts that make relevant predictions. We then introduce an experimental framework which enables us to measure the existence and persistence of a post-incentive reduction in task engagement. We test the effect of ending an incentive in Study 1, as well as in an internal meta-analysis of all data we have collected. In the remaining four studies, we investigate boundary conditions of our findings as a means to understand how the incentive affects task engagement and to test between competing accounts. We vary the presence and type of post-incentive break from the task (Study 2), the perceived magnitude of the incentive (Study 3), the type of task (effortful vs. leisure, Study 4) and how decisions about engagement are made (up-front vs. sequentially, Study 5). We conclude with the implications for future research and for making temporary incentive polices more effective.

Post-Reward Engagement Reduction

Consider a situation in which people repeatedly choose which of two tasks to do for a set period of time (e.g., 30 seconds) – a 'target' task and an alternative task, as illustrated in Figure 1. People's choices of how much of the target task to do without any incentive (in Round 1, the baseline period) are based on their relative intrinsic motivation for the focal task, compared to the alternative task. When some people receive a previously unannounced temporary incentive for choosing to do the target task (in Round 2), they are likely to do more of this incentivized task.

The key question in our research is how much of the target task people will do in Round 3, after the incentive has ended. The established theories predict a persistent *reduction in engagement* in the incentivized task after contingent incentives have ended, with fewer choices

of the incentivized task than if the incentive had never been introduced (e.g., vs. control, as shown in Figure 1). This reduction in engagement is the behavior associated with a reduction in intrinsic motivation, which is predicted when people had been interested in the target task before the incentive and when the incentive does not increase perceived competence (Deci, Koestner, & Ryan, 1999, 2001; Lepper, Henderlong, & Gingras, 1999). The magnitude and duration of this post-incentive reduction, relative to any increase in choosing the task during the incentive period, determines the net effect of a temporary incentive policy on behavior change, but this had not been studied in the prior literature.

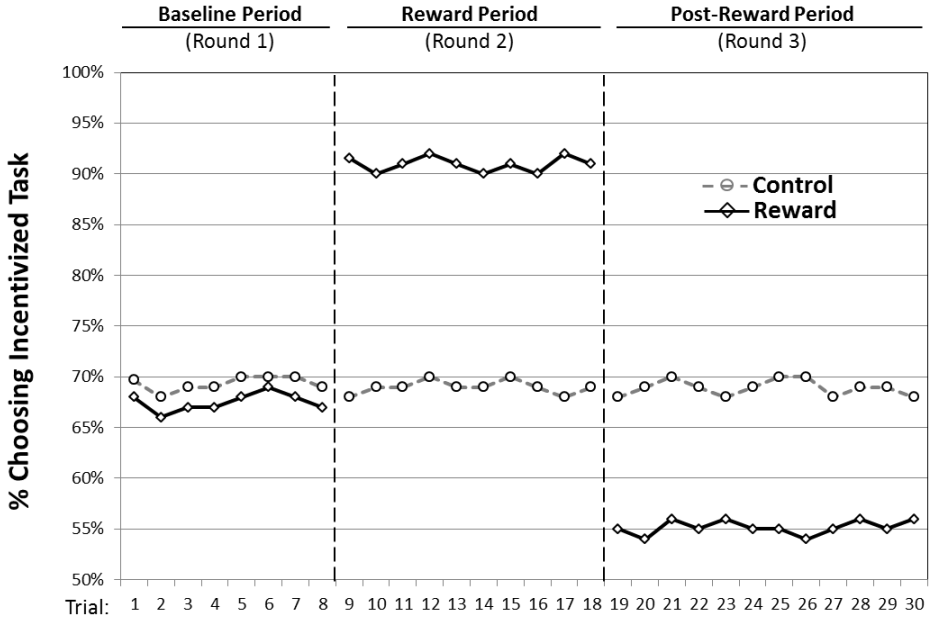


Figure 1: Illustration of persistent post-reward engagement reduction due to temporary incentives, as predicted by prior theories.

Existing Theories of Post-Reward Engagement Reduction

Prior research includes multiple accounts of how incentives will affect post-incentive behavior. While these accounts similarly predict post-incentive decline in engagement, the

underlying processes differ, and the accounts therefore make differing predictions about the circumstances in which the decline will be stronger or weaker.

Undermining Autonomy. Cognitive Evaluation Theory argues that people engage in tasks to fulfill innate needs of competence, exploration, and autonomy (White, 1959; Deci, 1971). An external incentive that is conditional on doing the target task will exert control over people's behavior, and thereby undermine satisfaction from autonomy and exploration (Deci & Ryan, 1985; Ryan & Deci, 2000) . As a result, engagement in the task will decline (e.g. compared to initial engagement or a control group) once an incentive ends, unless the incentive communicates competence, which can countervail the negative effects. This account predicts that because the incentive changes people's interpretation of the target task and the benefit they experience from doing the task, the post-incentive engagement reduction will be persistent (Deci et al., 1999). In this account, the negative effect on intrinsic motivation is stronger either when the task was initially more intrinsically motivating or when the conditional incentive is larger.

Overjustification. An alternative account, based on self-perception (Bem, 1967), posits that incentives impact inferences about one's own preferences. After being paid to work on a target task, people may infer that they were doing the task only because of the external rewards (Kruglanski et al., 1972; Lepper et al., 1973; Lepper & Greene, 1978), discount their own intrinsic motives (Nisbett & Valins, 1971; Kelley, 1973) and infer that they do not like doing the task. Based on this belief about their own preferences, people would then do less of the task after the incentive ends than they would have if the incentive was not introduced. This account predicts that the reduction will persist until another influence causes those preferences to change, and the reduction will be stronger for larger incentives.

Task and Social Inferences. People can form inferences about unknown aspects of the target task from an incentive, based on what they think of the incentive provider's motives. Since incentives are often introduced when tasks are unattractive or difficult, people may conclude that an incentivized task is relatively unappealing (Benabou & Tirole, 2003), reducing subsequent choices to do the task. High incentives should send a stronger signal, resulting in stronger reductions, while prior first-hand experience with the target task would be expected to mitigate the effects.

Expectations. Ending an incentive, without advance notice, can be seen as a violation of norms or expectations, resulting in disappointment (Fehr & Falk, 2002). Relatedly, the incentive may create a reference point, such that then doing the task without an incentive could be viewed as a relative loss (Kalyanaram & Winer, 1995). In both cases, when people learn that their expectation of continuing incentives will not be fulfilled, engagement in the target task can fall (Esteves-Sorensen, Macera, & Broce, 2013). If people cope with disappointment and adjust their reference points over time (Gilbert, Pinel, Wilson, Blumberg, & Wheatley, 1998; Schkade & Kahneman, 1998), these accounts predict that the post-incentive reduction will be temporary, and can be prevented when people have advance knowledge of the temporary nature of the incentive.

Temporal Shift in Effort. When decision makers know that the incentive will be temporary, they can strategically shift effort they would have invested during the post-incentive period, instead doing more of the target task during the incentive period. As a result, there might be little or no change in overall task engagement due to the incentives, but an increase during the incentive period and a decrease afterwards. For example, employees might move up task completion in time to help them earn a bonus (Oyer, 1998) or consumers might stockpile non-perishable goods while a purchase incentive is available (Mela, Jedidi, & Bowman, 1998).

Reduction in engagement due to effort shifting should occur primarily when people know that the rewards are temporary and they can predict the timing of rewards ending. Temporal shifting should be larger and persist for longer when the incentive is larger.

These accounts all predict an increase during the incentive period, followed by a persistent decline when the incentive ends, and stronger effects for larger incentives. Other accounts, based on information and signaling (Dube, Xueming, & Fang, 2015; Frey & Oberholzer-Gee, 1997; Gneezy & Rustichini, 2000), predict situations (e.g., pro-social tasks) for which an incentive may reduce engagement *while* the incentive is available. We constrain our investigation to the dynamic effects of *successful* incentives (i.e. those which increased engagement while available) on post-reward engagement. Our experimental framework, which we describe next, is designed to minimize the potential for task and social inferences, disappointment about the incentive ending, or temporal shifts in effort, in order to effectively test the two prominent motivation-based accounts discussed first.

Experimental Paradigm to Examine the Dynamics of Post-Reward Behavior

In order to examine the longer-run dynamics of post-reward behavior, we developed an experimental paradigm in which we can track repeated decisions about whether to engage in the target task after the incentives end. The "quitting paradigm" often used to study persistence (Deci, 1971; Heyman & Ariely, 2004) uses time spent working on an unsolvable task as the key measure. However, in this approach people may only make one quitting decision, and therefore the measure cannot distinguish between a temporary reduction in motivation to do the target task and permanent disengagement. In our experimental setup, participants make repeated choices between doing a moderately effortful but interesting math problem (the target task) or watching an enjoyable video clip (the alternative task), as illustrated in Figure 2.

The 30 trials were divided into three rounds: pre-incentive baseline (8 trials), randomized incentive (10 trials), and post-incentive (12 trials). The initial baseline round gave participants first-hand experience with the target task, minimizing the potential for participants assigned to the reward group to draw inferences from the incentive about unknown task characteristics or preferences. Participants were randomly assigned to reward (incentive during Round 2) or control (no incentive) groups. At the end of each round, participants in both the control and the reward group were informed that the round was over, and instructed to click "Next" to proceed to the next round. Participants did not know in advance how many trials were in each round, or how many rounds would occur, limiting their ability to strategically shift effort from the post-reward to the reward period.

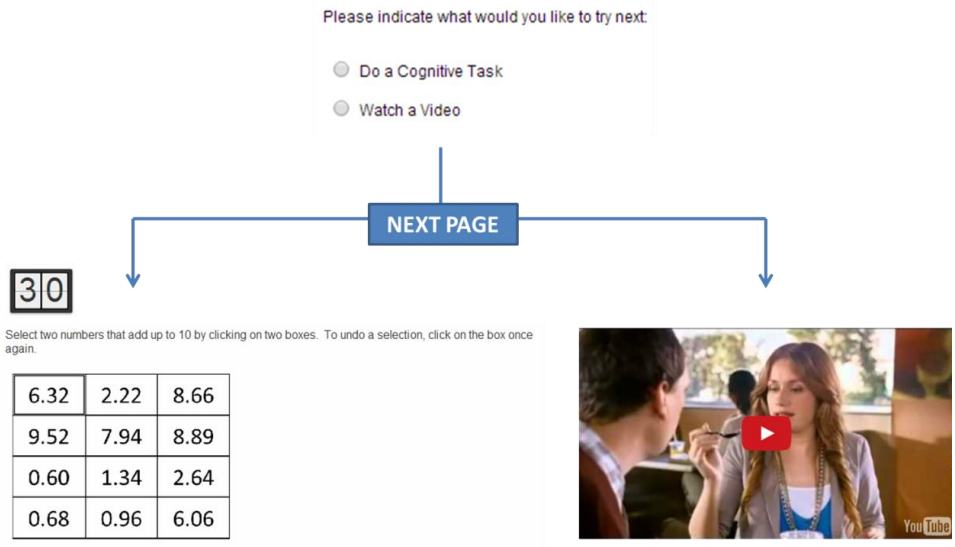


Figure 2: Participants chose between doing a math task or a video task in each trial. The specific task (i.e., math problem or video) was revealed on the subsequent page, after their choice.

The math task used in the studies (except for Study 4), was adapted from Experiment 1 of Mazar, Amir, and Ariely (2008) and involved finding the two numbers in a grid of 12 numbers

that added up to 10. Participants had 30 seconds to solve the problem, the same duration as the video clips in the alternative task.[1] All participants were given the following instructions:

"In this survey you will be a given a series of choices between doing cognitive tasks and watching videos of interesting television advertisements collected from across the world. The cognitive task will train your mental reasoning skills, and we will use your results to calibrate and standardize a training test. You can do as many of them as you want, or can just enjoy the videos."

To investigate the effect of external incentives on intrinsic motivation, it is important to use tasks that are intrinsically motivating. Deci, Koestner, & Ryan, (1999) set average ratings of at least the mid-point on a scale measuring how "interesting and enjoyable" the task is (1=*Low*, 9=*High*) as a criterion. A pre-test (N=47) confirmed that both the math task ($M_{math} = 6.02, SD = 2.19$) and the video task ($M_{video} = 6.64, SD = 2.08$) fulfilled this requirement. In addition, our pre-tests confirmed that people would choose to do some math tasks without any incentive, even with the videos as an alternative.[2]

Participants in all the studies were recruited from Amazon Mechanical Turk (AMT), an online employment marketplace for brief tasks that is used by companies and academic researchers. Potential participants were told that the study would take around 30 minutes to complete, and in addition to the base participation fee of $1.75, they had a possibility of earning additional bonuses. At the end of the baseline Round 1, those participants randomly assigned to the reward condition were told that they could earn bonuses by correctly solving the math problems and the amount. Reward-condition participants were told in advance that the reward opportunity was only available during the next round, so that the end of the incentive would not come as a surprise and cause disappointment (although ultimately that does not seem to matter

[1]Detailed stimuli are available in the Appendix A of the Supplemental Materials. Study 4 used a different math task.
[2]Details for additional pre-test measures of the tasks are in Appendix B.

for our findings, see Study A in Appendix E). At the end of Round 2, participants in the reward condition were told how much bonus they had earned, to be paid at the end of the experiment. For the control (no-incentive) group, a surprise bonus, similar in magnitude to the average bonus earned by the reward group, was announced at the end of the experiment. Data was included for all participants who completed the first round of the task, to minimize selection effects from the experimental manipulation (Zhou & Fishbach, 2016), with those dropping out after the first round coded as not doing the focal task in subsequent trials.

Study 1: The Dynamics of Post-Reward Task Engagement

Method

Adult participants were recruited from Amazon MTurk. The mean meta-analysis effect-size for post-reward reduction in engagement when people are given pre-announced tangible rewards is Cohen's $d = -.36$ (Deci et al., 1999, based on 128 controlled experiments). A two-tailed t test to detect this effect-size with 80% power would require a sample size of 122 in each cell. In our experimental setup, we have a baseline period with repeated measurements for every individual, and we therefore expected substantially higher statistical power and used a smaller sample in this initial study.

A target of 100 participants was requested, yielding 91 surveys. In all studies, the sample size was determined in advance, and we used the same data-cleaning protocol (see Appendix C), implemented prior to data analysis. Records with duplicate IP addresses (3 respondents), or who reported having technical problems with viewing the videos or working on the math task (4 respondents), or who failed a basic attention check at the end of the study (7 respondents) were removed prior to analysis, yielding 77 valid completes. In all studies, we include any participants who completed Round 1 and therefore could have been exposed to an incentive, even if they did

not complete the rest of the study. Our final sample included two participants (2.6%) who completed Round 1, but then dropped-out later in the experiment.[3] Immediately after the end of the first round, participants randomly assigned to the incentive condition were informed that they could earn 5 cents for every math task they chose to do and answered correctly.

Results

In the control condition, the math task was chosen 67% of the time during the baseline period, confirming that participants found the task intrinsically motivating. This is a prerequisite for our intended tests, because tasks which have little or no baseline engagement cannot show substantial reduction in engagement, due to the floor effect. The average engagement (i.e., choices of the math task) remained at around the same level during the subsequent experimental periods in the control condition, as shown in Figure 3.

In the incentive condition, the proportion of math task choices increased to 88% when the incentive was available – significantly higher than in the control condition (61%) during the same period, controlling for the baseline proportion of math task choices in Round 1[4] (β= 0.28, t=5.42, p<.001). Therefore, the incentive was successful at increasing participants' task engagement while the incentive was available, also a precondition of the intended test.

The key question we designed the experiment to test was whether people's engagement with the focal task after the incentive ended was affected by having previously offered the incentive. For participants in the reward condition, the average percentage of math task choices dropped to 53% in the very first trial in Round 3, immediately after the incentives ended, compared to 72% in the control condition. This initial test yielded a borderline significant

[3]All results are unaffected by excluding participants who did not complete the experiment.
[4]In all studies we report both linear regression t-tests that control for the individual baseline levels in Round 1 and results from a hierarchical regression model.

difference ($\beta = -0.19, t = -2.02, p = .047$). The drop in people's choices of the math task after the incentivized Round 2 had ended is consistent with the prior findings in lab studies of reductions in the incentivized behavior immediately after rewards ended.

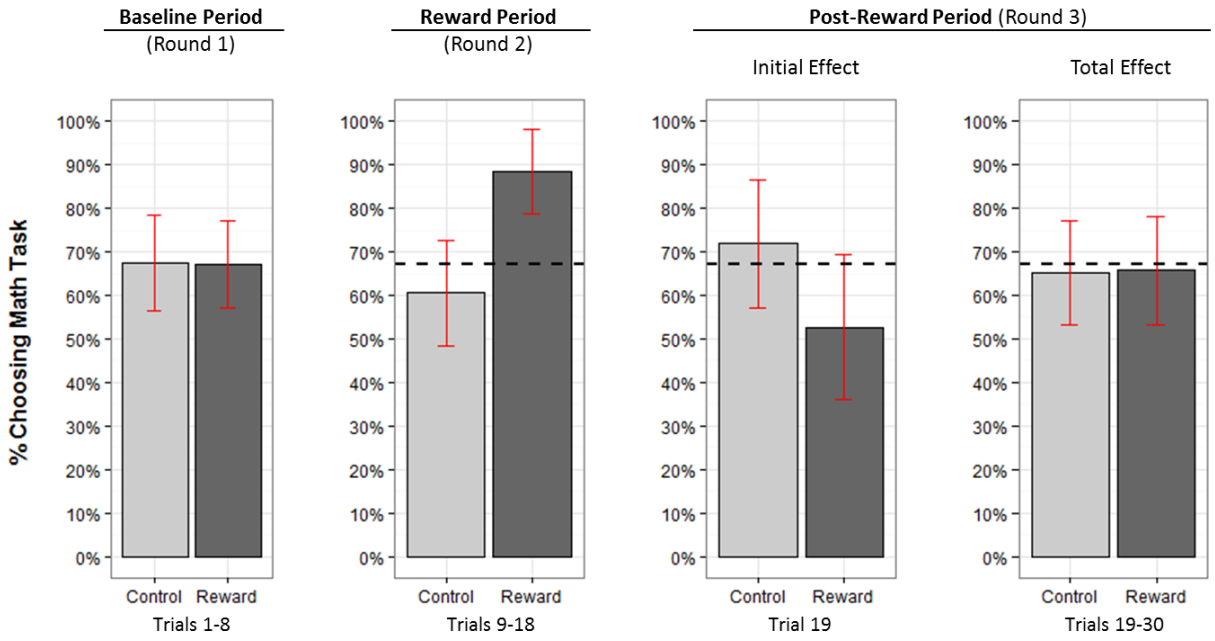


Figure 3: Proportion of math task choices in the control and reward group in each phase of the experiment. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs.

Since we track the dynamics of post-incentive task engagement over repeated trials, we can also distinguish between temporary and permanent disengagement with the target task. Averaging across the entire post-reward period (Round 3), people did a similar number of math tasks in the reward condition as in the control condition during the same period (66% vs. 65%; $\beta = .008, t < 1$). Although we do find an initial decrease in engagement immediately after the incentives ended, this decrease was brief, and the incentive did not result in any overall reduction in post-reward engagement. This result is not consistent with the predictions made by the prior theories of a persistent decrease in engagement due to the incentive.

Because the reduction in engagement was only momentary, the immediate negative post-reward effect of incentives was small in scope relative to the prolonged positive effect during the incentive period. Overall, offering a temporary incentive in the reward condition yielded a significant positive net effect in the study, with people choosing to do more of the math task in the incentive condition than in the control condition during the post-baseline periods, controlling for Round 1 (76% vs. 63%; β =0.13, t = 2.89, p=.005).

Hierarchical Regression Specification

In order to more precisely quantify the magnitude and duration of the post-reward engagement, we used a hierarchical non-linear mixed model (Raudenbush & Bryk, 2002), with the per-trial observations nested under individuals to account for the potentially correlated errors. We model both the long-run post-reward baseline level of engagement and the possibility of lower immediate post-reward engagement. We assume that after any initial decline, the return to the long-run level of engagement is inversely proportional to the number of periods t since the incentive ended. Based on this assumption, we define the probability of individual i choosing to do the math task in the post-reward period (Round 3) during trial t as[5]:

$$P(Y_{ti} = 1) = logit\ (\beta_{0i} + \beta_{Mi}\ \frac{1}{t}) \quad (1)$$

In this model, β_{0i} is a person-specific intercept and β_{Mi} is a person-specific coefficient capturing a non-linear difference in task engagement during Round 3, in effect contrasting the first few trials with the remainder of the trials in the round. These parameters are in turn a function of time-invariant factors: the experimental condition C_i and an individual-level

[5]See Appendix D for full details. We relax this inverse proportionality assumption in several ways and test robustness to different specifications, including a nonparametric test.

covariate X_i, the person's baseline interest in the math task (e.g., the number of math tasks done by person *i* during Round 1). The expected proportion of math tasks chosen in each trial *t* of Round 3 in a given experimental condition can therefore be written as:

$$P(Y_{ti} = 1) = logit\left(\beta_{00} + \beta_{01}\frac{1}{t} + \beta_{02}X_i + \beta_{LT}C_i + \beta_{MOMENTARY}C_i * \frac{1}{t}\right) \quad (2)$$

This model tests for differences in engagement during Round 3 between the experimental condition ($C_i = 1$), compared to the same time periods in the control condition ($C_i = 0$). The coefficients β_{00}, β_{01} and β_{02} capture the levels of engagement during Round 3 in the control condition. The term $\beta_{00} + \beta_{02}X_i$ represents the long-run engagement (i.e., as 1/t goes to zero) for a person i assigned to the control condition with level of Round 1 engagement X_i. In general, $\beta_{00} + \beta_{02}X_i$ will be similar to the mean of the last few trials of Round 3 for control condition participant *i*. The second coefficient, β_{01}, can be thought of as an adjustment, capturing non-linear deviation from the long-run engagement level during the initial trials in Round 3, for control condition participants.

The remaining parameters provide the key hypothesis tests, of the difference in engagement during Round 3 for participants in the experimental condition, relative to control. First, β_{LT} represents the difference in long-run engagement (i.e., as 1/t goes to zero), and will approximate the difference in the means of the last few trials between control and experimental condition participants. When β_{LT} is positive and significant, it means that participants in the experimental condition were completing more math tasks by the end of Round 3 than participants in the control condition.

The second coefficient, $\beta_{\text{MOMENTARY}}$, captures momentary differences between control and experimental conditions in how the first few trials deviate from the long-run level. When the long-run level is similar in incentive and control conditions (e.g., β_{LT} is not significantly different from zero), $\beta_{\text{MOMENTARY}}$ has an additional interpretation. In these cases, a significant and negative $\beta_{MOMENTARY}$ indicates a momentary engagement reduction after the incentive ended, compared to the corresponding trials in the control condition, controlling for individual baseline effort X_i.[6]

This model focused on understanding dynamic post-incentive behavior in Round 3. We also use a similar hierarchical difference-in-difference model to test whether the difference in the overall probability of choosing the math task between two experimental rounds ($R_t = 0$ vs. $R_t = 1$) differs by condition ($C_i = 0$ vs. $C_i = 1$):

$$P(Y_{ti} = 1) = logit\ (\beta_0 + \beta_1 * R_t + \beta_2 * C_i + \beta_3 * R_t * C_i\ ) \quad (3)$$

We can test multiple hypotheses using the key coefficient β_3, depending on how the rounds (R_t) and conditions (C_i) are coded. Using different coding in this model (see Appendix D), we estimate either β_{REWARD} , the effect of incentives on during-reward performance; β_{POST}, the average post-incentive engagement level; or the net effect of incentives (combining the incentive and post-incentive periods), β_{NET}. Next, we use this approach to analyze the results of Study 1, followed by an internal-meta analysis of the same two conditions in all the data we have collected.

Regression Results for Study1

[6] In Appendix N we present an alternative parameterization, which instead estimates immediate post-incentive differences and the non-linear long-run deviation from those immediate differences, and finds very similar results.

There was no significant difference in the long-run engagement during Round 3 between incentive and control ($\beta_{LT} = 0.62, z = 0.87, p = .38$).We do find a significant momentary engagement reduction after the incentive ended ($\beta_{MOMENTARY} = -2.59, z = -2.59, p < .01$). Consistent with the non-significant β_{LT}, there was no average reduction in engagement with the math task after incentives ended from the difference-in-difference model ($\beta_{POST} = -0.003, z = -0.01, p > .250$). Therefore, choices of the math task returned to the pre-reward baseline level after an initial post-reward momentary decrease. Since the incentive did boost math task choices significantly over the baseline level during Round 2 ($\beta_{REWARD} = +4.60, z = +4.22, p < .001$), the net effect of incentives was a significant increase in the number of math tasks chosen over the course of the study ($\beta_{NET} = +1.21, z = +2.75, p = .006$).

Discussion

In Study 1, consistent with the findings of the existing psychological theories, we find a reduction in engagement with the target task immediately after the incentive ends. However, contrary to the predictions of those theories, this reduction in engagement is not persistent. Instead, tracking trial by trial engagement, we find a momentary reduction in engagement, followed by a return to baseline, and no longer-run reduction in engagement. This absence of a longer-term reduction in engagement is consistent with the empirical findings of several recent field studies. These studies did not track engagement over time, but measured people's total behavior in the days, weeks or months after the incentive ended and found either no long-term effect (John et al., 2011; Lacetera, Macis, & Slonim, 2011; Volpp et al., 2006), or a modest increase in engagement (Charness & Gneezy, 2009; Jackson, 2010).

Our findings therefore suggest a potential reconciliation between the immediate post-incentive declines reported in the lab experiments and the lack of long-term post-incentive

decline in the field studies. However, it is important to first establish the robustness of the findings beyond this one study, and to explore heterogeneity in the response to incentives across individuals.

Internal Meta-Analysis: The Dynamics of Post-Reward Task Engagement

In order to test the robustness of the findings, we conducted an internal meta-analysis, combining all data collected in the same reward and control conditions as Study 1 across our studies. The resulting data comes from ten of our studies (including conditions from Studies 1, 2 and 3; see Appendix E for full details), totaling 1,098 participants (530 in control, and 568 in the reward group). The large sample size allows us to examine post-incentive behavior more closely, do multiple robustness checks, and look at heterogeneity in responses.

The raw proportion of math tasks in each trial, plotted separately by experimental condition, is shown in Figure 4. The results reveal a significant reduction in engagement for the first trial (t = 3.96, p < .0001) and a marginally significant reduction for the second trial (t = 1.82, p = .069), in the incentive conditions compared to control (see Appendix G, specification 2). The model reveals both a relatively small but significant positive long-run effect of incentives (β_{LT} = 1.26, z=5.92, p<.001) and a strong significant momentary reduction in engagement relative to that long-run baseline ($\beta_{MOMENTARY} = -3.03, z = -9.15, p < .001$). Overall, comparing the pre- and post-reward rounds, we find a significant average *increase* in choices of the math task post-reward ($\beta_{POST} = +0.36, z = +2.00, p = .045$).

Furthermore, given the momentary nature of post-reward reduction in engagement and the otherwise positive effects of incentives, the combined data also show a strong positive net effect of incentives on the number of math tasks chosen ($\beta_{NET} = +1.21, z = +9.55, p < .001$).

These results are robust to different parameterizations of the post-reward recovery (e.g., t^n with different values of n, $n \leq 0$) and also to excluding participants who dropped out before completing the full study.[7] It is important to note that accuracy in correctly solving the math problems did not vary between reward and control across the rounds. As a consequence, the same results hold if we look at the number of successfully solved math problems, rather than the number of attempts (see Appendix H). Furthermore, the results are not explained by participants' speed in doing the math tasks. The average time taken per math task in Round 2 did not moderate post-reward behavior in Round 3, and there was no difference in the average time to complete a math task during Round 3 between control and reward conditions ($M_{Control}$=19.13 vs. M_{Reward}=20.01, t=0.71, p=.48). Likewise, the post-reward behavior was not moderated by proportion of tasks correctly answered during the reward period (see Appendix J).

Next, we investigate heterogeneity in people's behavior, quantifying how many participants' individual behavior is represented by the average patterns shown in Figure 4. In the reward group, 41% of the participants showed a decrease in engagement in the first trial of Round 3, relative to their own individual baseline level in Round 1, significantly more than in the no-reward control group (41% vs. 26%; $\chi^2\ (1) = 28, p < .0001$). The results are similar if we instead compare the average of the first two or even three trials in Round 3. This suggests that, consistent with the average results, many individual participants in the reward group did show a reduction in engagement during the first three post-incentive trials, significantly more than in the control group. The tendency to reduce engagement was not moderated by initial motivation (i.e., choices of the math task in Round 1, see Appendix I).

[7]Details are in Appendix F of the Supplemental Materials. Appendix G reports estimates of two flexible non-parametric models that also find the same result.

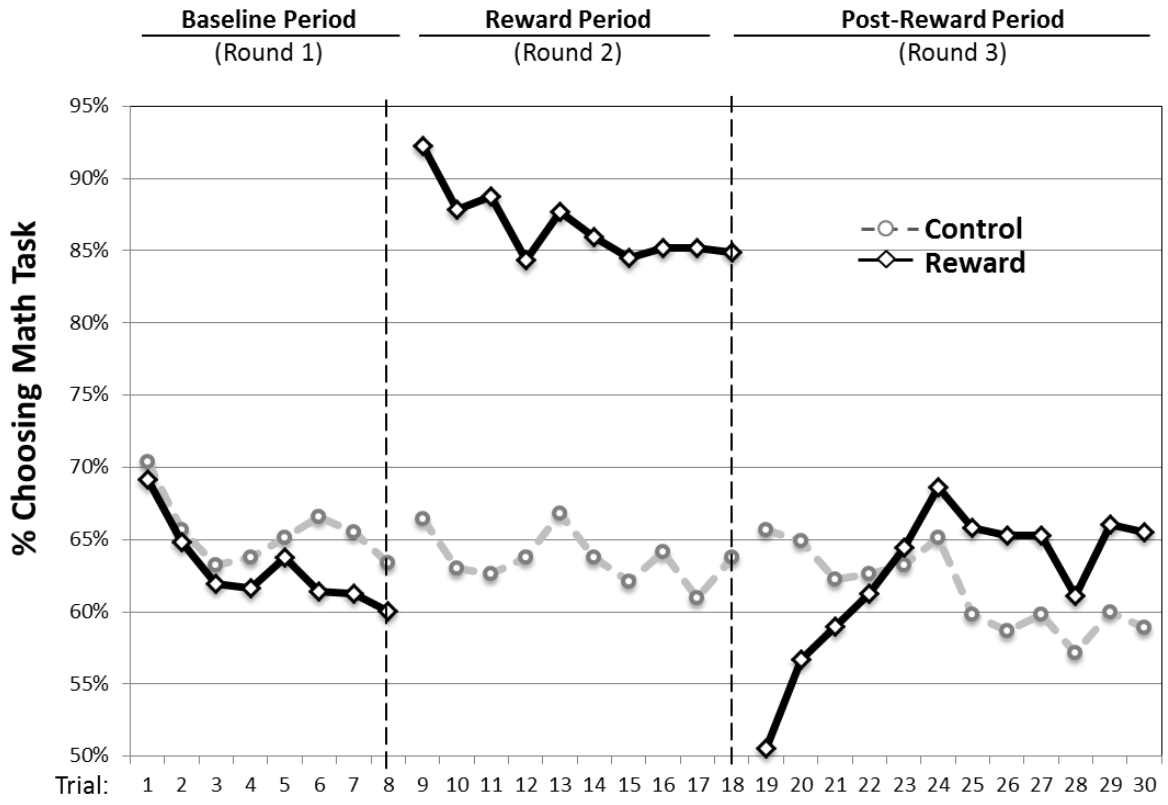


Figure 4: Raw percentage of participants choosing the math task in each trial in the internal meta-analysis (approximate 95% CI = +/- 4%). The difference in engagement between Reward and Control is significant for all trials in Round 2, for Trial 19 (p<.001), marginal for Trial 20 (p=.07), and for Trials 24-30 (all ps <.05)

We also investigated the possibility that the initial post-incentive decrease and subsequent increase over baseline that we observe in Round 3 could be explained by heterogeneity, resulting from averaging the behavior of one set of participants who displayed an initial reduction and another set of participants with a long-term increase. We quantified the degree to which participants had an initial post-reward reduction, by computing the difference in average math tasks in the first three trials of Round 3 versus that person's Round 1 baseline. We also quantified the participants' final post-reward behavior, by computing the difference in average math tasks in the last seven trials of Round 3 versus that person's Round 1 baseline. The relationship between initial and final post-reward (Round 3) behavior did not differ between the reward and control groups ($\beta = 0.01, t$=0.41, p>.250, see Appendix J). Among participants in

the reward group who had an initial reduction, 34% were finally above (settled at a higher post-reward baseline), compared to 13% in the corresponding segment of the control group. Among participants in the reward group who were not initially below (i.e., *did not* show a momentary post-incentive decrease in engagement), 45% were finally above, compared to 31% in the corresponding segment of the control group. Thus, both participants who had an initial reduction and those who did not have an initial reduction showed a similar longer-term increase, relative to control.

Prior theories of intrinsic motivation (Lepper & Greene, 1978; Deci & Ryan, 1985) predict a persistent decrease in engagement after the incentive ends specifically because introducing the incentive would result in a decrease in intrinsic motivation for the target task. Consistent with our finding that the reduction in engagement is only momentary and engagement eventually increases to be higher than the original baseline, self-reported interest in the task at the end of the study, a measure of intrinsic motivation, was actually higher in the reward group compared to the control in our meta-analysis ($M_{reward} = 5.97, SD = 2.35\ vs. M_{control} = 5.59, SD = 2.51;\ t(890) = 2.3, p = .02$). In fact, Deci et al.'s (1999) conclusions about intrinsic motivation were based only on a meta-analysis of task engagement measures, and their analysis likewise did not find the predicted decrease in self-reported task-liking after performance-contingent incentives ended. These results suggest that the momentary reduction in task engagement after rewards end may not be attributable to the changes in intrinsic motivation that they have proposed, but may need to be explained by a different psychological process.

An Effort-Balancing Account of Incentives and Task Engagement

We hypothesize that our findings may be explained by people attempting to strike a balance between investing effort and enjoying leisure. Some intrinsically motivating tasks, such as watching videos, represent leisure and provide immediate benefits with minimal effort. Other intrinsically motivating tasks require immediate effort or cost to gain some combination of immediate and delayed benefits. For example, going to the gym, going to the doctor for a regular health checkup, learning a new skill or practicing an existing skill are intrinsically motivating, in that people are willing to engage in these tasks without any external rewards, but require upfront effort. As a result, people often feel they *should* be engaging in such activities for their future welfare, but may *want* to engage in more enjoyable and less effortful activities (e.g., watching TV). We can think of decisions about whether to engage in these kinds of effortful intrinsically motivating tasks as similar to self-control conflicts between goals representing more immediate and more delayed benefits (Ainslie, 1975; Bazerman, Tenbrunsel, & Wade-Benzoni, 1998; Milkman, Rogers, & Bazerman, 2008).

People often strive to find a balance between competing goals (Dhar & Simonson, 1999; Fishbach & Dhar, 2005), including between effort and leisure (Kool & Botvinick, 2014), and between effort and expected reward (Kivetz, 2003). In the absence of incentives, we can think of people's choices between doing tasks associated with more delayed benefits (e.g., solving math problems) or doing tasks with immediate benefits (e.g., watching videos) as representing their preferred level of balance. In our setting, this would be measured by the pre-reward baseline level of engagement in the target task (vs. the leisure alternative).

Introducing a temporary incentive adds an immediate benefit to any pre-existing intrinsic motivation associated with the task. When goals compete for resources, people are more motivated by the goals that are more salient (Shah & Kruglanski, 2002) and more proximal

(Brown, 1948; Kivetz, Urminsky, & Zheng, 2006; Urminsky & Goswami, 2015), which would favor the incentivized task. When people therefore increase work on the incentivized task, their competing leisure goals are likely to be deferred to avoid interruption (Jhang & Lynch, 2015) until a time deemed more appropriate (Shu & Gneezy, 2010).

As people forego leisure in favor of investing effort into the compensated task, the leisure goal is likely to persist in the background. When the incentive ends, motivation will shift from the completed incentive-earning goal, and thereby from the effortful task, to the neglected competing goal (Fishbach & Dhar, 2005; Förster, Liberman, & Friedman, 2007; Khan & Dhar, 2006; Kruglanski et al., 2002) of pursuing leisure. Recent research on self-control has suggested that depletion-like regulatory failures can represent a motivated switching between labor and leisure to strike a balance between two goals, without necessarily reflecting exhaustion of limited resources (Inzlicht et al., 2014; Inzlicht & Schmeichel, 2012). Choices of immediately gratifying options are also increased by the availability of justifications (Kivetz & Zheng, 2006; Lerner & Tetlock, 1999), such as having just worked on an effortful task. Therefore, immediately after the incentive ends, people will be more motivated to do a leisure alternative and will find it easier to justify doing so.

Prior research has shown that after some level of goal-attainment, the reduction in motivation to pursue that goal is only temporary (Förster, Liberman, & Higgins, 2005). Even minimal satisfaction of an immediate gratification goal can restore people's motivation to pursue longer-term goals (Urminsky & Kivetz, 2011). Hence, after engaging in extra post-reward leisure, any perceived balance between competing goals is eventually restored.

However, introducing temporary incentives more generous than the minimal amount needed to induce additional effort can have further effects. Incentives that are experienced as

rewarding and enjoyable (Knutson, Adams, Fong & Hommer 2001) can likewise help restore the balance, reducing the post-incentive need for leisure. Furthermore, associating a target task with sufficiently high-magnitude rewards can foster positive associations and increase task liking (De Houwer, Thomas, & Baeyens, 2001; Razran, 1954). Findings in the literature on reinforcement behavior also suggests that rewards can sometimes result in eventual positive long-term effects on behavior (Brooks & Bouton, 1993). Generous incentives may therefore both counter the momentary reduction in engagement and bolster the long-run level of engagement.

We propose that people prioritize effortful incentivized tasks when a reward is available, but then are looking to take a break from the effortful task when the incentive ends, making the alternative leisure task temporarily more appealing and easier to justify. After taking a break, by doing more of the leisure task, balance is restored and task engagement settles at a baseline level, which may have been increased relative to the initial baseline by the experience of the incentive. This "effort-balancing" account makes similar predictions as the prior theories about engagement reduction immediately after the incentive, but predicts that the reduction will not persist.

We view this account as identifying one source of common and important influences on task engagement, which may co-occur with other factors affecting task engagement. For example, greater engagement with the target task when the incentive is available may also promote habit-formation (Neal, Wood, & Quinn, 2006), resulting in an even higher post-reward baseline, over and above the effort-balancing account predictions. Conversely, when people encounter highly novel situations involving unfamiliar tasks and incentives, they may subsequently draw negative inferences about their task liking from the previous incentive, which would conflict with the recovery from momentary reduction predicted by our account. In particular, children may be especially likely to draw such inferences, consistent with the prior

evidence from studies of child behavior of persistent over-justification (Kruglanski et al., 1972; Lepper et al., 1973; Greene & Lepper, 1974).

The effort-balancing account makes specific testable predictions about how the post-incentive reduction in engagement will be affected by contextual factors. Giving people a break after the incentive ends should help restore balance and ameliorate the reduction in engagement. Incentivizing a less effortful task or paying a higher incentive (but holding effort constant) should reduce the imbalance arising in the incentive period, likewise lessening the post-incentive reduction in engagement (and potentially fostering a higher post-incentive baseline). These predicted boundary conditions are not consistent with prior theories of intrinsic motivation, enabling us to use these competing predictions to test between accounts in the remaining studies. First, we test the effect of giving people a break after the incentive ends.

Study 2: Providing a Break Eliminates Engagement Reduction

Method

In Study 1, we found an effect size of $d = -.47$ for the momentary reduction in engagement. This means that we need a sample size of 72 per cell to detect a difference with 80% power. In Study 2, our intended focal comparison was two (of 6) combined conditions against another two. We requested 350 participants from Amazon MTurk, yielding 322 surveys. Unusable cases (due to duplicate IPs, technical problems or a failed attention check) were removed prior to analysis, yielding 257 valid completes, or approximately 85 per pair of conditions to be compared. The few participants who completed Round 1 but then dropped-out part way through (5.4%) were coded as not doing the focal task and included in the analysis.

Participants were randomly assigned to one of six conditions. Two of the conditions (no-reward control, 5-cent per correct answer incentive) were the same as in Study 1. In the other four conditions, participants were given the 5-cent conditional incentive in Round 2, and then were asked to do an unrelated activity, potentially providing a break, for 90 seconds immediately after the incentives ended. Ninety seconds was equal to the duration of three trials, the approximate duration of post-reward reduction in engagement observed in the internal meta-analysis.

In the four unrelated-activity conditions, we varied task type (writing vs. logo-matching) and whether or not participants needed to make choices about the activity, in a 2x2 design. In the no-choice conditions, participants were assigned to an activity (either writing about their opinions on an assigned topic or matching assigned brand logos to brand names). Participants were assigned and completed three such tasks, each of 30 seconds duration. In the choice conditions, participants had to first choose which unrelated activity they would do (picking their topics in the writing task or selecting product categories in the logo-matching tasks).

The choice manipulation provides a direct test between the effort-balancing account and prior autonomy-based accounts. Consistent with findings from prior research (Vohs et al., 2014), a pre-test confirmed that requiring people to make choices (in the choice conditions) was more difficult, on the one hand, but that it also provided directionally more of a sense of autonomy (details in Appendix B), compared to the no-choice break. If reduction in engagement occurs because of a loss of autonomy, giving people choices about the unrelated activity (i.e. the choice conditions) may counter this loss of autonomy, ameliorating the reduction in engagement. However, if the reduction occurs because people need a break after exerting effort, then the non-

effortful break provided in the no-choice conditions will be more effective in eliminating the reduction.

Results

We tested two types of unrelated activities (writing and logos) for generalizability, and found no differences. We therefore collapsed these conditions in our analysis and report results comparing the two no-choice breaks with the two choice activity conditions.

On average, participants chose to do the math task 66% of the time during the baseline period in the control condition, and there were no systematic changes in this level over the trials. In the no-break incentive (i.e., the replication) condition, there was a directional decrease in the choice of math tasks immediately after the rewards ended, relative to control (58% vs. 63%). We find an increase in long-run engagement from incentives($\beta_{LT} = 1.91, z = 3.10, p = .002$), and replicate the momentary reduction in engagement relative to the long-run baseline ($\beta_{MOMENTARY} = -2.49, z = -2.55, p = .01$). Overall, we observed a significant average post-reward increase in engagement ($\beta_{POST} = +1.11, z = +2.33, p = .02$), which contributed to the finding reported earlier in our internal meta-analysis.

These results were moderated by providing participants a break after the incentive round. When participants were given a non-effortful break (i.e. an unrelated activity that did not require making choices), participants chose the math task 66% of the time in the immediate post-reward trial – similar to the same trial in the control condition (63%; $\beta = -0.008, t < 1$; Figure 5). There was a marginal long-run increase in engagement($\beta_{LT} = 1.29, z = 1.80, p = .072$, but no momentary reduction relative to the long-run engagement ($\beta_{MOMENTARY} = -1.14, z = -1.04, p > .250$). The non-effortful break successfully arrested the post-reward engagement

reduction found in the replication (no break) incentive condition ($\beta_{MOMENTARY:\ REPLICATION\ VS.\ NON-EFFORT\ BREAK} = +1.60, z = +2.13, p = .03$).

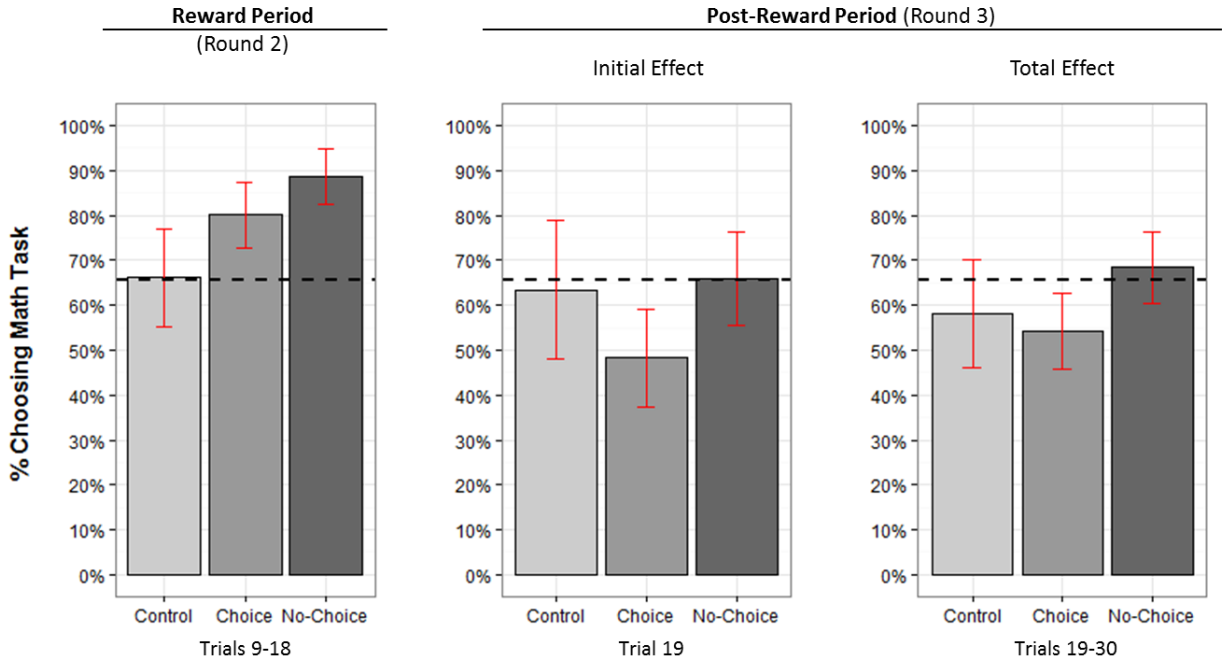


Figure 5: Comparison of effortful unrelated activities (choice) and low-effort break (no-choice) conditions to control in Study 2. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs.

These findings are specific to participants taking a break, rather than to doing any unrelated activity. When participants instead engaged in the same activities by making relatively effortful choices, the momentary reduction in engagement found in the incentive condition was not eliminated. Choices of the math task in the first trial of Round 3 were directionally lower, compared to the same trial in the control condition (48% vs. 63%, $\beta = -0.11, t = -1.35, p = .18$), reflecting a significant overall momentary reduction in engagement ($\beta_{MOMENTARY} = -2.15, z = -2.10, p = .04$) relative to the non-significant long-run baseline ($\beta_{LT} = 0.77, z = 1.24, p = .22$), compared to control. Engaging in the effortful version of the alternative activities yielded a similar post-reward reduction in engagement as in the replication incentive condition ($\beta_{MOMENTARY:\ REPLICATION\ VS.\ EFFORT\ BREAK} = +0.62, z = +0.82, p > .250$).

Compared to the combined replication and the effortful choices conditions (the two conditions in which our account predicts momentary reduction), giving people a non-effortful break significantly arrested the immediate post-reward decrease in engagement ($\beta_{MOMENTARY} = +1.24, z = +1.97, p = .049$), compared to control. However, we did not have sufficient power to detect a significant interaction when directly comparing the effortful break to the non-effortful break. ($\beta_{MOMENTARY: NON-EFFORT\ BREAK\ VS.\ EFFFORT\ BREAK} = +1.04, z = +1.39, p = .165$). These findings are consistent with the proposal that people reduce task engagement after the incentive in order to take a break, and are inconsistent with the autonomy-based prior theories, which would have predicted that giving people choices should be as effective or more effective at eliminating the reduction.

Aside from the initial post-incentive differences, the longer-term engagement in Round 3 was similar overall in both the no-choice break and choice activity conditions. In the choice activity conditions, engagement returned to the pre-reward baseline level (54% vs. 58%, $\beta = 0.004, t < 1; \beta_{POST} = +0.12, z = +0.26, p > .250$). Participants' average engagement with the math task in the no-choice-break conditions was directionally higher after the incentive (68% vs. 58%;$\beta = 0.07, t = 1.46, p = .15$), reflecting a modest increase in engagement($\beta_{POST} = +0.48, z = +1.99, p = .046$).

Discussion

In Study 2, we tested participants' need to restore balance as an explanation for the findings by providing them with a break after the incentive ended. A non-effortful unrelated activity gave participants a break, potentially restoring their perceived balance between effort and leisure, and successfully preventing an immediate post-reward decrease in task engagement.

In contrast, when the same unrelated activities involved making relatively more effortful choices, the activity presumably did not offset the effort exerted during the reward period and did not serve as a break resulting in a post-incentive momentary reduction in engagement.

These findings are difficult to reconcile with prior accounts, particularly Cognitive Evaluation Theory, which explains negative effects of incentives in terms of a reduction in perceived autonomy provided by the task. It is possible that giving participants an unrelated activity could allow them to re-assert their sense of autonomy in general, reducing the momentary need for autonomy, and countering the theorized effect of incentives. If that were the case, however, an unrelated activity involving more of an opportunity to express autonomy after the incentive ended (i.e., by making choices) should be more effective at countering the momentary engagement reduction. Instead, we find the opposite, suggesting that it is the lack of required effort that makes the break effective.

In the next study, we test a prediction of the effort-balancing account about how post-incentive engagement is affected by the magnitude of the incentive. In our account, when incentives are higher but during-incentive efforts are similar, people will feel less of a need to balance out the efforts with leisure or a break, and will feel less justified in doing so. In fact, when people find the task incentive particularly rewarding, relative to effort, long-term engagement in the task may actually be bolstered by the incentive. In contrast, the prior accounts involving intrinsic motivation or inferences predict the opposite, a stronger and more persistent post-reward reduction in task engagement from higher incentives, since larger incentives would be experienced as more controlling and as a stronger basis for self-perception inferences.

Study 3: Large Rewards Do Not Reduce Engagement

Method

Adult participants were recruited from Amazon MTurk. A target of 320 participants was requested, yielding 305 surveys. Unusable cases (duplicate IPs, technical problems, failing the attention check) were removed prior to analysis, yielding 235 valid completes. Participants in this sample who completed Round 1, and then dropped-out part way through (5.5%), were coded as not doing the focal task and included. In this study, we have over 70 per cell, again consistent with the estimated required power from Study 1.

Participants were randomly assigned to one of four conditions: two conditions identical to Study 1 (a no-reward control condition and a 5-cent performance-contingent condition), and two other performance-contingent conditions with different relative reward amounts. In the *low-reward* condition, participants were told that they would earn 1 cent per correct answer in the second round, and that participants in a previous study were paid 5 cents per correct answer, as a basis for comparison. In the *high-reward* condition, participants were told they would earn 50 cents per correct answer, and that previous participants had earned 5 cents per correct answer. At the end of the study, participants were asked a few follow-up questions about their experience, including how much they liked the math task.

Results

On average, participants in the control condition chose the math task 60% of the time during Round 1 and there were no systematic trends across trials. The incentives in all three incentive conditions successfully increased average choices of the math task during Round 2 (1 cent: 70%, t=1.89, p=.06; 5 cent: 89%, t=4.78, p<.001; 50 cent: 89%, t=5.51, p<.001) compared to control (60%). Given that choices of the math task during Round 2 were nearly identical in the

5 cent and 50 cent conditions, participants in the 50 cent condition effectively earned more for the same effort.

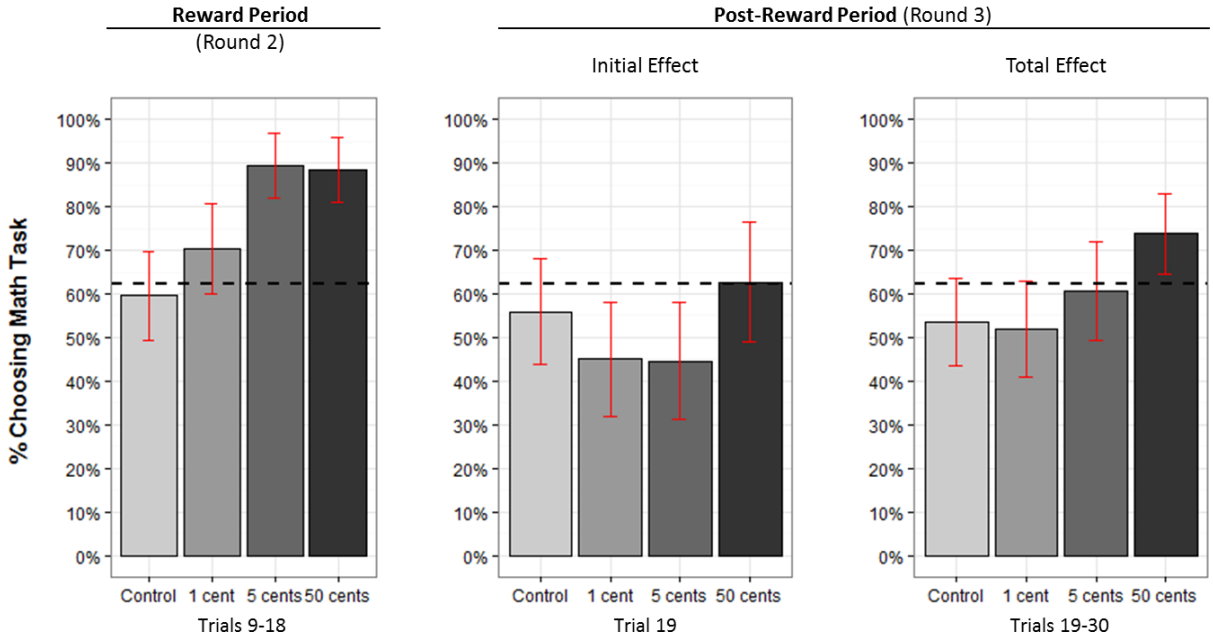


Figure 6: Average math task choices in the control, replication (5 cents), low-reward (1 cent) and high-reward (50) conditions in Study 3. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs.

In the low-reward (1 cent) condition, math task choices were directionally lower in the first post-reward trial than in the same trial for the control condition (45% vs. 56%; $\beta = -0.12, t = -1.47, p = .14$). We found no long-run effect of the one cent incentive ($\beta_{LT} = 0.23\ , z = 0.30, p = .77$), and there was a significant momentary reduction in engagement ($\beta_{MOMENTARY} = -1.94\ , z = -2.04, p = .04$) after the incentive. Overall, there was no average reduction in engagement (52% vs. 53%; $t < 1$; $\beta_{POST} = -0.22, z = -0.34, p > .250$).

Similarly, in the replication condition (5 cents), the math task was only chosen 45% of the time in the first trial after the incentive ended, marginally less than in the control condition

(56%; β= -0.16, *t*= -1.93, *p*=.055). Participants' engagement reflected no long-run effect of the incentive ($\beta_{LT} = 1.19\ , z = 1.42, p = .16$), and significant momentary reduction in engagement ($\beta_{MOMENTARY} = -3.52\ , z = -3.07, p = .002$) after the incentive. As in the previous studies, overall there was no difference in the average effort level in Round 3 between the replication and control conditions (60% vs. 53%;$\beta = 0.01, t < 1$; $\beta_{POST} = +0.26, z = +0.44, p > .250$).

Unlike the momentary reduction observed in the lower reward conditions, directionally more participants chose the math task immediately after the high (50 cent) incentive ended, compared to the same trial in the control condition (63% vs. 56%; β = 0.07, *t*<1). The high incentive increased the level of long-run engagement ($\beta_{LT} = 0.24\ , z = 3.83, p < .001$), and initial choices of the math task were directionally lower than later choices ($\beta_{MOMENTARY} = -0.15, z = -1.68, p = .093$)[8]. Overall, the fifty cent incentive resulted in higher post-incentive average engagement than in the control condition ($\beta_{POST} = +1.97, z = +3.42,\ p < .001$),

The key comparison in this study is between lower and higher rewards. Across the incentive conditions, participants were significantly more likely to choose the math task in the trial immediately after the incentive ended when incentives were higher ($r = .16, t = 2.14, p = .034$). Overall, more math tasks were done after the incentive when the reward per task had been a larger amount in the prior round($\beta_{POST\ vs.REWARD\ SIZE} = +0.05, z = +3.46, p < .001$). Contrary to the predictions of the prior literature, higher perceived rewards did not result in any momentary or persistent post-reward reduction in task engagement.

Discussion

Post-reward Baseline Engagement Level

[8] These results are based on a linear model because of non-convergence in the logit model, so the coefficients are not directly comparable to the other studies.

As noted above, the proportion of math tasks chosen in all of Round 3, after the incentive ended, was higher for larger incentives (Control: 53%, 1 cent: 52%, 5 cents: 60%, 50 cents: 74%, $F(233)=4.36$, $p=.038$). One explanation might be that higher incentives result in more experience with the task during the incentive period, facilitating more habit formation or belief updating about the task, or practice effects. However, the size of the incentive did not influence the time taken to do the math tasks in Round 2 ($\beta = .021$, $p = .38$) or Round 3 ($\beta = -.004$, $p = .89$), and did not influence accuracy in solving the math tasks in Round 2 ($\beta = -.001$, $p = .33$) or Round 3 ($\beta = -.0003$, $p = .80$).

Since the effort expended during the incentive period was identical in the 50-cent and 5-cent conditions (both 89%), this comparison is particularly instructive. Contrary to the habit-formation or learning accounts, which would have predicted the same post-reward behavior in both conditions, there was a significantly higher post-reward baseline level in the high-reward (50 cent) condition than in the lower (5 cent) condition ($\beta_{POST:\,50\ vs.\ 5} = +1.88, z = +2.84, p = .004$). This finding is instead consistent with theories of evaluative conditioning (Razran, 1954; De Houwer et al., 2001), which suggest that highly rewarded activities become more attractive, as people's attention is spontaneously drawn to previously rewarded activities, potentially mediated by the release of dopamine (Anderson et al., 2016).

Research in social psychology has long suggested that paying people for an interesting target task turns play into work (Lepper & Greene, 1975), because of a reduction in intrinsic motivation in the target task. Consistent with the results of our internal meta-analysis reported earlier, we do not find any evidence for this concern. Post-study self-report ratings suggested that the math task actually felt less like work when participants were paid a higher amount (e.g., 50 cents), compared to being paid a lower amount that elicited the same effort level ($M_{50\ cents} =$

$4.32, SD = 2.57\ vs.\ M_{5\ cents} = 5.68,\ SD = 2.48;\ t(100) = 2.7, p = .007$). Participants also rated working on math tasks under high rewards to be a significantly better opportunity for them ($M_{50\ cents} = 8.12, SD = 1.00\ vs. M_{5\ cents} = 6.55, SD = 2.03; t(100) = 4.93, p < .001$).

Distinguishing Between Effort-Balancing and Alternative Accounts

We have proposed that the momentary post-reward reduction in engagement occurs because people want to take a break and engage in leisure to balance out prior efforts, rather than because of a decline in intrinsic motivation. If incentives are seen as controlling and thereby reduce people's intrinsic motivation, we should have found the strongest and most persistent reductions for the largest incentive. Instead, we find the exact opposite. When participants earned more (50 cents vs. 5 cents) for the same effort, we found no momentary reduction, relative to control. Instead, consistent with our account, participants earning high rewards reduced their engagement after the incentive ended back to the same level as the control condition, and then increased their engagement to a new, higher long-run level.

This study also addresses potential concerns that doing more of a focal task (e.g, because of an incentive) could result in fatigue or satiation with the task and a desire for variety. An increase in effort during the incentive period could then directly result in reduction in engagement, rather than because of perceived balance. If this were the case, then we would expect participants in the 50-cent high-reward condition (who did significantly more of the task than in the 1-cent low-reward condition; $89\%\ vs\ 70\%; \beta = 0.18, t = 3.49, p < .001$), to also show a stronger and more persistent reduction in engagement. However, contrary to the fatigue and satiation accounts, we find the opposite. Participants in the 50 cent condition did more math tasks in Round 2 than participants in the 1 cent condition, but then showed less, rather than more, post-incentive reduction in engagement. In fact, the high incentive resulted in a significantly

higher post-reward baseline relative to the low-reward condition ($\beta_{POST:50\ vs.1\ cent} = +2.44, z = +3.33, p < .001$), which is also inconsistent with reference point or disappointment accounts of the findings.

In the studies so far, we have incentivized participants for solving math problems, a moderately effortful task, similar to tasks used in prior research. Incentives are generally deemed unnecessary to motivate people to do leisure tasks, and are rarely applied to such tasks in practice (although in-game rewards and gambling may be considered exceptions). While prior research has not tested the effects of incentives on purely leisure tasks, doing so provides a particularly strong test between our account and alternative accounts. Since leisure tasks are typically more intrinsically motivating, prior accounts would predict a stronger post-reward decrease in intrinsic motivation and therefore stronger engagement reduction when people are paid to do the leisure task (Calder & Staw, 1975; Deci et al., 1999). However, since doing the leisure task involves less effort for the same reward, people would feel less of a need to take a break after being paid for leisure, and we would predict little or no momentary reduction in engagement.

Study 4: Paying for a Leisure Task Does Not Reduce Engagement

Method

Adult participants were recruited from Amazon MTurk. A target of 320 participants was requested, yielding 305 surveys. Unusable cases (duplicate IP addresses, technical problems, failed attention check) were removed prior to analysis, yielding 246 valid completes. Due to the rate of unusable cases, the study is slightly underpowered (67%) relative to the power analysis in

Study 1. Participants who completed Round 1 but then dropped-out (2.4%) were coded as not doing the focal task and included in the analysis.

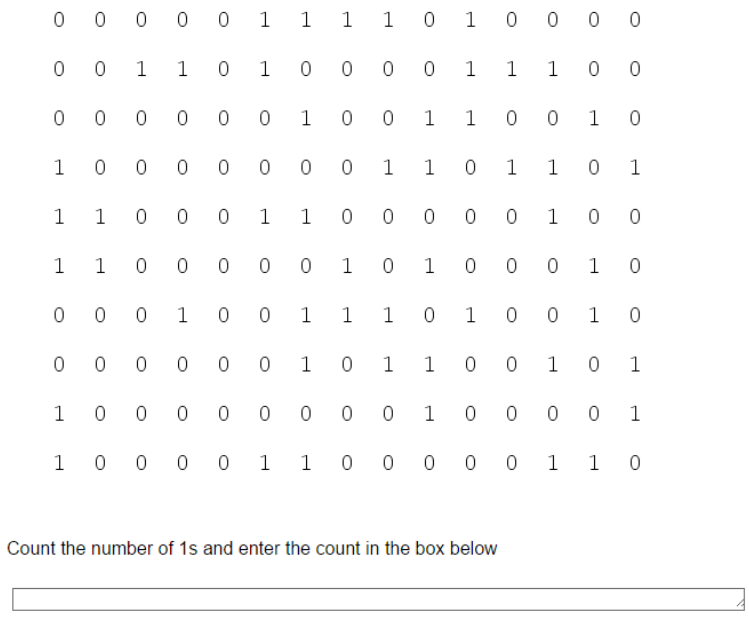


Figure 7: Example of math task used in Study 4, which required participants to count the number of 1s in the grid (of 150 numbers) in 30 seconds.

Participants were randomly assigned to one of four conditions in a 2 (Target task: Math vs. Video) x 2 (Control vs. Incentive) between-subjects design. For generalizability, we use a different cognitive math task in this study (adapted from Abeler et al., 2011) that required participants to count the number of 1s in a grid of 150 1s and 0s (see Figure 7).[9] We pre-tested this math task and found that participants were willing to engage in the task without rewards, even when they had the option of a video task as an alternative choice. In the two incentive conditions, they were either paid 5 cents for correctly completing math tasks, as in the prior studies, or paid 5 cents for each video they watched and rated (1- 5 stars). The two control

[9]We also replicate the findings with the original math task used in earlier studies (Appendix L).

conditions matched the two incentive conditions, highlighting the target task without providing any incentive.

Results

The math task used in this study was intended to be less intrinsically motivating than the video task, and this is reflected in the relatively low level of baseline attempts, both overall (40%) and, in particular, when video watching was the focal task (30% math in Round 1; see the dotted lines across charts in Figure 8). We have two different control conditions in this study, so we compare each incentive condition to the corresponding control condition.

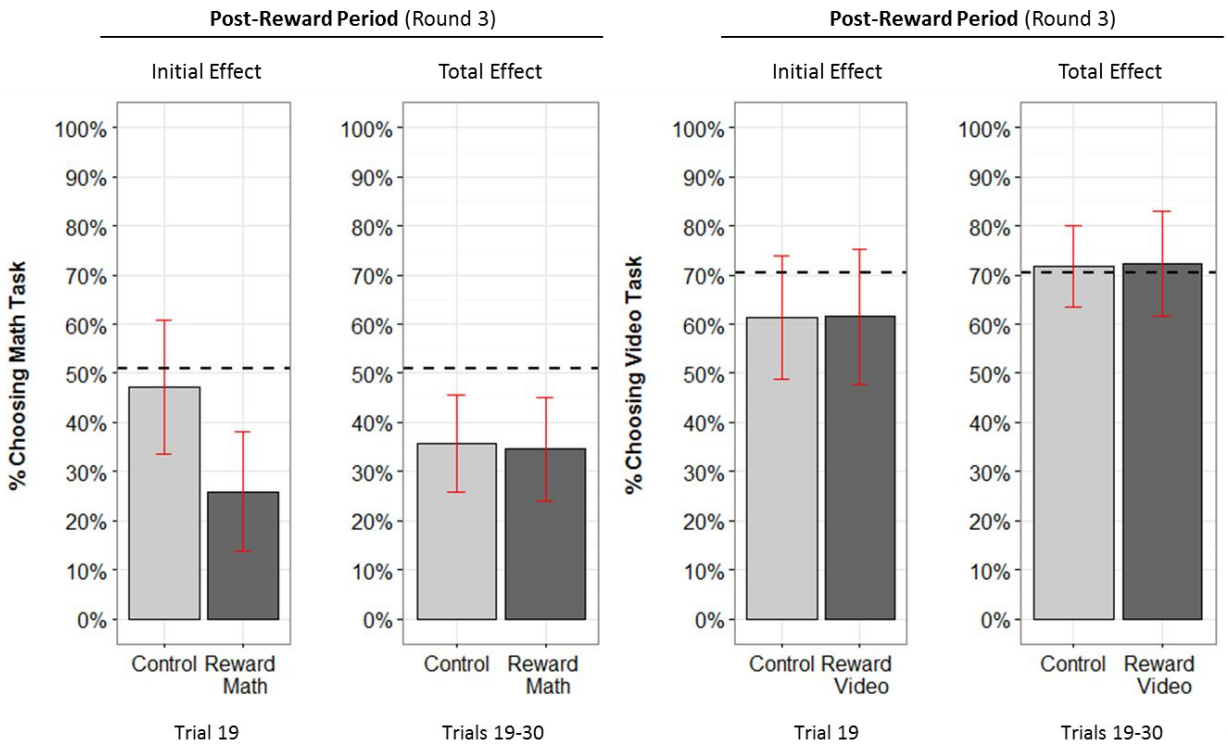


Figure 8: Post-reward engagement when the math task was incentivized (left-panel) or the video task was incentivized (right panel). Dotted lines represent the baseline (average effort level in Round 1), and the vertical lines are 95% CIs.

We replicated the momentary reduction in engagement when people were paid for doing the math task. Fewer people chose the math task in the first trial of Round 3 in the incentive condition after the rewards had ended, compared to in the matching control condition (47% vs. 26%; β = -0.19, *t*= -2.3, *p*=.02; Figure 8). There was no average reduction of effort due to incentives in Round 3 overall, relative to control (35% in both). These results were further confirmed in the hierarchical regressions ($\beta_{LT} = 0.94, z = 1.54, p = .12; \beta_{MOMENTARY} = -4.17, z = -3.75, p < .001;\ \beta_{POST} = -0.15, z = -0.28,\ p > .250$).

When participants were paid for the leisure task instead, we did not find any reduction in engagement. The proportion of video choices in the first trial after the incentive ended was to the same as in the corresponding control condition (61% in both, $\beta = 0.02, t < 1$) and we found no long-term reduction of effort (72% in both). These results were further confirmed by the hierarchical regression models ($\beta_{LT} = 0.05, z = 0.07, p = .95; \beta_{MOMENTARY} = -0.52, z = -0.57, p = .57; \beta_{POST} = -0.06, z = -0.10,\ p > .250$).

In fact, the momentary reduction in engagement observed when incentivizing the math task was completely eliminated when the videos were incentivized instead ($\beta_{MOMENTARY:MATH\ VS.\ VIDEO} = +3.50, z = +2.57, p = .010$). There was no difference between the two conditions in terms of the longer-term post-reward baseline level ($\beta_{LT:MATH\ VS.\ VIDEO} = -0.04, z = -0.06, p > .250$) or the average level of engagement($\beta_{POST:MATH\ vs.\ VIDEO} = +0.36, z = +0.52, p > .250$).

Discussion

These findings provide direct support for our effort-balancing interpretation. Prior motivation-based theories would predict stronger and more persistent post-reward reduction in

engagement when paying people for a more intrinsically motivating task (e.g., watching videos). Furthermore, the incentive in the video task was not performance-based, as participants only had to rate the video and there was no correct answer. Deci et. al. (1999) observed that completion-contingent rewards (as in the video task) tend to have a stronger negative effect on subsequent motivation than performance-contingent rewards (as in the math task). In their account, this could occur because of the potential competence feedback provided by learning about the performance-contingent incentive.

In contrast, we found that completion-contingent incentives for the leisure task yielded no reduction in engagement, and had a significantly different effect from incentivizing the math task. This is inconsistent with the motivation-based theories, as well as with satiation and variety seeking, which would predict similar reductions for both incentives. However, since the leisure task involves little effort, our account suggests that people feel little need to balance out the extra choices of that task after the incentive ended, resulting in no post-reward reduction.

In the studies presented thus far, people made repeated choices about what to engage in during the post-incentive period, allowing us to measure when preferences for the focal task return to (or even exceed) the original baseline level. If people instead make a single binding decision about future engagement immediately after the incentives have ended, then the temporary decrease in motivation for the target task could result in a long-term disengagement. In the next study, we test this possibility, and test a solution predicted by our account, that providing people with a break (as demonstrated in Study 2) would prevent long-term disengagement in locked-in decisions.

Study 5: Taking A Break Prevents Locking-in Engagement Reduction

Method

Adult participants were recruited from Amazon MTurk. A target of 250 participants was requested, yielding 235 surveys. Unusable cases (duplicate IPs, technical problems, failed attention check) were removed prior to analysis, yielding 189 valid completes. Power (73%) was slightly less than suggested by Study 1. Participants who completed Round 1 but then dropped-out afterwards (1.0%) were coded as not doing the focal task and were included in the analyses.

Participants were randomly assigned to one of three conditions in a between-subjects design – a control group, and two incentive condition (5 cents per correct task in Round 2), one with a 90 second non-effortful break between Rounds 2 and 3 (as in Study 2), and one without. The study used the same math task as was used in Studies 1, 2 and 3. At the start of Round 3, participants in all conditions were asked to choose one of the two tasks as the only task they would do during the entire ensuing period (i.e., doing only math tasks or only watching videos). In the break condition, this decision was made after the break. All participants were told that they would not be able to change their decision during the third round.

Results

Choices of which task to do in Round 3 varied significantly across the conditions ($\chi^2(2) = 6.5, p = .04$). In the control condition, when participants were asked to make a binding future decision immediately after the end of Round 2, 60% of the participants choose to only do math tasks in Round 3. This is similar to the average baseline level of engagement in Round 3 in the control condition of previous studies. In contrast, in the no-break incentive condition, only 43% of the participants choose math tasks for Round 3, marginally less than that

in the control condition ($\chi^2(1) = 3.6, p = .057$; Figure 8). This is similar to the differences observed in the first post-incentive trial in prior studies, but has more impact on the net effect because the decision is locked-in for the remainder of Round 3. This locked-in reduction in math tasks in Round 3 counters the incentive-based increase in Round 2, and as a result the incentives in the no-break reward condition yielded no net increase in math tasks across the study ($M_{control} = 4.70\ vs.\ M_{reward,no\ break} = 4.15,\ t(120) < 1$).

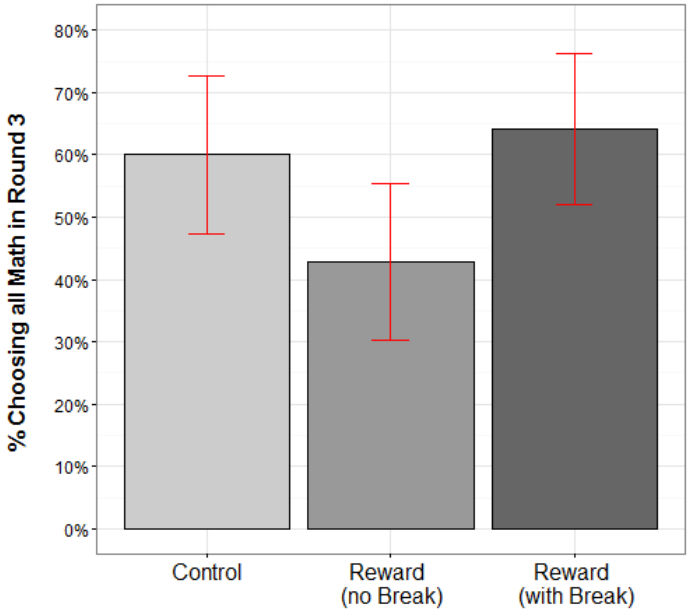


Figure 9: Percentage of participants locking-in math tasks for the post-reward period. The vertical lines are 95% CIs.

However, participants in the with-break incentive condition, who instead faced the same choice after a 90 second post-incentive break, did not lock-in a reduction and, in fact, chose very similarly to the control condition participants, with 64% opting for the math task in Round 3 ($\chi^2(1) = 0.22, p = .60$). The effort-balancing account predicts that adding a brief break immediately after the rewards ended will help people restore their sense of balance between work and leisure after exerting more effort for the incentive. Consistent with this prediction,

people who had a break after the incentive round were willing to commit to doing the math task in Round 3 as much as people in the control condition and significantly more than those in the no-break rewards condition ($\chi^2(1) = 5.7, p = .02$). In fact, simply adding a break resulted in a marginally higher net effect of incentives, compared to rewards without a break ($M_{reward,with\ break} = 5.52\ vs. M_{reward,no\ break} = 4.15,\ t(130) = 1.9, p = .054$). Therefore, a brief break immediately after the end of the reward period successfully precluded a lock-in of the momentary reluctance to do more of the math task after the incentive ended.

Discussion

In everyday life, people often make decisions that are effectively "locked in" (e.g., enrolling in an automatic savings plan, buying a subscription or renewing a trial gym membership), because of the cost or inconvenience of revisiting and changing the decision. Promotions and incentives are often used to increase awareness and generate initial experience, before having people make the "sticky" lock-in decision. Our findings suggest that this strategy may not work well, because of the temporary reduction in interest after the incentive ends. If the lock-in decision is instead structured to occur after a break, the likelihood of choosing long-term engagement is likely to be higher. This practical implication for how to structure the decision environment to maximize the effectiveness of temporary incentives arises directly from our effort-balancing account, and is not predicted by any of the prior accounts.

General Discussion

Effect of Temporary Incentives on Post-Reward Behavior

Temporary incentives can be very effective in motivating beneficial behavior while people are being rewarded. However, a large and influential literature has warned against the use

of such policies because of the possibility that external incentives will undermine intrinsic motivation, resulting in a persistent reduction in engagement after the rewards end (Deci, Koestner, & Ryan, 1999). However, these conclusions with adults were based on observing a single initial behavior, immediately after the incentive ended. Noting this shared limitation of the extant studies with adults, the authors pointed out the need for "studies of interesting [e.g., intrinsically motivating] behaviors that examine repeated administration of rewards over time, have appropriate no-reward control groups, and use reasonable sample sizes" (Deci, Koestner, & Ryan, 1999, p. 650). In this paper, we have attempted to fill this crucial and long-standing gap in the literature.

Using a novel experimental framework to track choice-by-choice engagement in intrinsically motivating tasks, we do find a momentary reduction in task engagement immediately after the task incentive ends, consistent with prior findings. However, consistent with recent field studies of incentives, we do not find the predicted long-term negative effects after the incentive ends. Instead, we find that the immediate post-incentive reduction is brief, people then return to or even exceed baseline levels of effort, and as a result, there is a strong positive net effect of incentives. Thus, our results reconcile the findings of prior lab studies, which measured initial post-incentive engagement, and field studies, which measured long-term engagement. Across the studies, our findings cannot be explained by incentives reducing people's intrinsic motivation (Studies 1, internal meta-analysis, 3 & 4), rewards undermining people's autonomy (Study 2), fatigue and reference points (Study 3), or variety seeking (Studies 3 & 4).

In particular, our results are incompatible with existing accounts of post-incentive intrinsic motivation. A temporary reduction in engagement contradicts the theoretical premise

that intrinsic motivation for the target task is reduced by the incentive. A reduction in intrinsic motivation would have instead resulted in a persistent decrease in engagement with the target task after the incentive ended, since there was no change in information in the decision environment. As noted in our internal meta-analysis, self-reported measures of task interest at the end of our studies were not lower in the incentive conditions, as would be expected if intrinsic motivation had been eliminated. Instead, self-reported task interest was actually higher in the incentive conditions, reflecting a reinforcing effect of incentives. Notably, this lack of evidence for a decline in intrinsic motivation is not unique to our data. As reported in the prior meta-analyses (Deci, Koestner, & Ryan, 1999), self-reported measures also did not demonstrate a significant decrease in task-interest when using performance-contingent incentives.

An Effort-Balancing Account of Dynamic Post-Reward Behavior

In this paper, we outlined an effort-balancing account as a new way of thinking about the effects of incentives on post-reward effort over time. We propose that increased task effort, induced by incentives, can result in a sense of imbalance between effort and leisure, resulting in a desire to take a break from the incentivized task when the incentive period ends. As a result, post-reward task engagement can decrease below the baseline level *momentarily*, but will return to or even exceed the baseline level when balance has been restored. When people are given a non-effortful break after the incentive ends (Studies 2 and 5), their need for a break will be satisfied and the post-reward reduction in engagement can be arrested.

Our account makes testable predictions about how characteristics of the available incentive will affect post-incentive engagement. The larger the incentive, holding effort constant, the less people will need to balance out the effort after the incentive ends, and the less people will feel justified in taking a break. Consistent with this prediction, we found that a relatively

high incentive eliminated the post-reward momentary reduction in engagement, generating an overall post-incentive increase in engagement relative to control (Study 3). Likewise, when a leisure task was incentivized, we predicted that the relatively low required effort during the incentive period would result in little post-reward reduction in engagement, which was confirmed in Study 4. Notably, prior accounts based on intrinsic motivation would have made the opposite prediction in both cases: stronger and more persistent post-reward reductions in task engagement after larger incentives or after incentivizing a more intrinsically motivating leisure task.

In our account, any post-incentive reduction in engagement is brief. Building on prior research on positive reinforcement (Anderson et al., 2016; De Houwer, Thomas, & Baeyens, 2001; Brooks & Bouton, 1993; Razran, 1954), our account further suggests that people's longer term engagement can increase over the baseline when the incentive is sufficiently large. We find general evidence of increase over baseline in our internal meta-analysis, and document stronger increases over baseline for a larger incentive in Study 3. This prediction of our account is also consistent with a comparison across recent field studies. Studies that provided larger incentives tended to find a positive longer-term increase in engagement (Kane et al., 2004; Volpp et al., 2008; Cawley & Price, 2009; Charness & Gneezy, 2009; Halpern et al., 2015), whereas studies in similar domains with smaller incentives (John et al., 2011; Volpp et al., 2006) tended to find no such positive longer-term behavior (see Appendix O for a summary of past field studies that that looked at post-reward behavior).

Implications for Future Research

Our initial findings and the proposed effort-balancing interpretation of these findings relies on data from a single paradigm. Future research should test the generality of our findings,

including in the field and in settings not involving a potentially relevant participation fee, as well as over longer periods of time.

It could be useful to also test whether these findings extend to other types of behavior for which people may be incentivized. Incentives are often used to encourage people to make purchases, for example, and it is not clear if the momentary reduction in engagement after a financial incentive would extend to situations where the effort itself takes the form of spending money. The effect of an incentive may also depend on whether it is seen as an attempt to control people's behavior. Recent research has found that non-incentivized monitoring (e.g., of medical practitioners' hand-washing; Staats et. al., 2016) can result in a long-term reduction in post-intervention behavior. It might be useful to study the effects on behavior when incentives are accompanied by heavy-handed monitoring.

It would also be useful to study the effect of temporary incentives on prosocial behavior. Past research has suggested that people might be less responsive to incentives for such activities (Dickinson, 1989; Benabou & Tirole, 2003; Ariely, Bracha, & Meier, 2009), and that introducing monetary rewards can undermine the effectiveness of incentivizing pro-social behavior (Dube, Xueming, & Fang, 2015; Heyman & Ariely, 2004; Yang, Urminsky, & Hsee, 2015). However, this research has focused on pro-social behavior while the incentive was available, and it would be useful to investigate the effects of incentives on dynamic longer-term post-reward pro-social behavior.

Our findings also suggest that we need a better understanding of the contextual factors that can bolster or undermine the post-reward effectiveness of incentives. We view the proposed effort-balancing account as a key mechanism in how people respond to incentives, but by no means the only mechanism. Although our studies document positive net effects of incentives,

incentives may not always provide a net gain in motivation. It would be useful for future research to investigate whether the findings differ as a function of other factors that have been proposed to affect post-incentive engagement, including task feedback and competence signals, task variety, participants' baseline fatigue, and prior experience with the tasks and incentives. In particular, it would be useful to extend our paradigm to study the effects of incentives on children, whose lack of experience with incentives may result in different responses than adults.

Future research may identify low-cost psychological interventions to bolster the effects of incentives. In our studies, we found that higher incentives and less effortful tasks yielded more post-incentive engagement. Other interventions may be able to generate the same subjective perception, that the incentivized effort represents a rewarding experience. For example, highlighting the fun aspects of an effortful but beneficial task or highlighting the enjoyable aspects of temporarily working in groups may make an experience feel more rewarding, resulting in less initial post-incentive reduction and longer-term positive effects. Conversely, building on our finding that even momentary reductions can have substantial negative effects if the initial decisions are locked-in, future research may identify interventions, beyond providing a break, to counter the initial lock-in effect.

Re-opening the Door to Incentives

Incentives are a cornerstone of economics, and featured prominently in foundational theories developed in the early days of psychology as well. Rewards were considered to be a powerful reinforcer of desirable behavior (Skinner, 1953), an important determinant of motivational force in expectancy-valence theory (Vroom, 1964; Fishbein, 1967), and a means of creating and strengthening expectations of personal efficacy (Bandura, 1977). Despite the reinforcing nature of rewards, incentives have come to be viewed as counter-productive and

either non-psychological or subject to psychological backlash. Our results suggest that these concerns may have been over-stated, and that the ways in which people respond to incentives over time involves different psychological mechanisms than previously thought. Our key insight in this paper is that post-reward engagement reduction may often have more to do with people wanting a temporary break after investing effort in their work, than with having their enthusiasm for work "smothered" by incentives.

Taking a dynamic perspective on the effects of incentives in this research allowed us to uncover the momentary nature of post-reward reductions in engagement, and raises new questions about how incentives and motivation interact over time. What are the psychological factors that extend or inhibit the observed momentary reduction in engagement? What kinds of psychological interventions could be leveraged to make incentives more effective and motivating in the long-term? We believe that our findings should not only re-open the door to testing the use of temporary incentives in consequential domains like health and education, but also to a new re-investigation of this most fundamental psychological driver of human motivation.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, 470–492.

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*(4), 463.

Anderson, B. A., Kuwabara, H., Wong, D. F., Gean, E. G., Rahmim, A., Brašić, J. R., … Yantis, S. (2016). The Role of Dopamine in Value-Based Attentional Orienting. *Current Biology*.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *The American Economic Review*, 544–555.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191.

Bazerman, M. H., Tenbrunsel, A. E., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review*, *23*(2), 225–241.

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

Brooks, D. C., & Bouton, M. E. (1993). A retrieval cue for extinction attenuates spontaneous recovery. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*(1), 77.

Brown, J. S. (1948). Gradients of approach and avoidance responses and their relation to level of motivation. *Journal of Comparative and Physiological Psychology*, *41*(6), 450.

Calder, B. J., & Staw, B. M. (1975). Self-perception of intrinsic and extrinsic motivation. *Journal of Personality and Social Psychology*, *31*(4), 599–605.

Cawley, J., & Price, J. A. (2009). Outcomes in a program that offers financial rewards for weight loss (No. w14987). National Bureau of Economic Research.

Charness, G., & Gneezy, U. (2009). Incentives to exercise. *Econometrica*, *77*(3), 909–931.

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., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1–27.

Deci, E. L., & Ryan, R. M. (1985). Cognitive evaluation theory. In *Intrinsic Motivation and Self-Determination in Human Behavior* (pp. 43–85). Springer.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853–869.

DelVecchio, D., Henard, D. H., & Freling, T. H. (2006). The effect of sales promotion on post-promotion brand preference: A meta-analysis. *Journal of Retailing*, *82*(3), 203–213.

Dhar, R., & Simonson, I. (1999). Making complementary choices in consumption episodes: Highlighting versus balancing. *Journal of Marketing Research*, *36*(1), 29–44.

Dickinson, A. M. (1989). The detrimental effects of extrinsic reinforcement on "intrinsic motivation." *The Behavior Analyst*, *12*(1), 1–15.

Dube, J. P., Xueming, L., & Fang, Z. (2015). Self-Signaling and Pro-Social Behavior: a cause marketing mobile field experiment. *Working Paper, University of Chicago*.

Esteves-Sorensen, C., Macera, R., & Broce, R. (2013). *Do Monetary Incentives Crowd Out Intrinsic Motivation? A Field Test in the Workplace*. mimeo. Retrieved from

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. *European Economic Review*, *46*(4), 687–724.

Fishbach, A., & Dhar, R. (2005). Goals as excuses or guides: The liberating effect of perceived goal progress on choice. *Journal of Consumer Research*, *32*(3), 370–377.

Fishbein, M. E. (1967). Readings in attitude theory and measurement.

Förster, J., Liberman, N., & Friedman, R. S. (2007). Seven principles of goal activation: A systematic approach to distinguishing goal priming from priming of non-goal constructs. *Personality and Social Psychology Review*, *11*(3), 211–233.

Förster, J., Liberman, N., & Higgins, E. T. (2005). Accessibility from active and fulfilled goals. *Journal of Experimental Social Psychology*, *41*(3), 220–239.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review*, 746–755.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*(3), 617.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, 191–209.

Gneezy, U., & Rustichini, A. (2000). Fine is a price, a. *J. Legal Stud.*, *29*, 1.

Greene, D., & Lepper, M. R. (1974). Effects of extrinsic rewards on children's subsequent intrinsic interest. Child development, 1141-1145.

Halpern, S. D., French, B., Small, D. S., Saulsgiver, K., Harhay, M. O., Audrain-McGovern, J., … Volpp, K. G. (2015). Randomized trial of four financial-incentive programs for smoking cessation. *New England Journal of Medicine*.

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

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. Perspectives on Psychological Science, 7(5), 450-463.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. Trends in cognitive sciences, 18(3), 127-133.

Jackson, C. K. (2010). A Little Now for a Lot Later A Look at a Texas Advanced Placement Incentive Program. *Journal of Human Resources*, *45*(3), 591–639.

Jensen, M. C. (2003). Paying people to lie: The truth about the budgeting process. *European Financial Management*, *9*(3), 379–406.

Jhang, J. H., & Lynch, J. G. (2015). Pardon the Interruption: Goal Proximity, Perceived Spare Time, and Impatience. *Journal of Consumer Research*, *41*(5), 1267–1283.

John, L. K., Loewenstein, G., Troxel, A. B., Norton, L., Fassbender, J. E., & Volpp, K. G. (2011). Financial incentives for extended weight loss: a randomized, controlled trial. *Journal of General Internal Medicine*, *26*(6), 621–626.

Kalyanaram, G., & Winer, R. S. (1995). Empirical generalizations from reference price research. *Marketing Science*, *14*(3_supplement), G161–G169.

Kane, R. L., Johnson, P. E., Town, R. J., & Butler, M. (2004). A structured review of the effect of economic incentives on consumers' preventive behavior. *American Journal of Preventive Medicine*, *27*(4), 327–352.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107.

Khan, U., & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research*, *43*(2), 259–266.

Kivetz, R. (2003). The Effects of Effort and Intrinsic Motivation on Risky Choice. *Marketing Science*, *22*(4), 477–502.

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research*, 39–58.

Kivetz, R., & Zheng, Y. (2006). Determinants of justification and self-control. *Journal of Experimental Psychology: General*, *135*(4), 572.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J Neurosci, 21(16), RC159.

Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes*. Houghton Mifflin Harcourt.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 131.

Kruglanski, A. W., Alon, S., & Lewis, T. (1972). Retrospective misattribution and task enjoyment. *Journal of Experimental Social Psychology*, *8*(6), 493–501.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 34, pp. 331–378). Academic Press.

Lacetera, N., Macis, M., & Slonim, R. (2011). *Rewarding Altruism? A Natural Field Experiment* (Working Paper No. 17636). National Bureau of Economic Research.

Lepper, M. R., & Greene, D. (1975). Turning play into work: Effects of adult surveillance and extrinsic rewards on children's intrinsic motivation. *Journal of Personality and Social Psychology*, *31*(3), 479.

Lepper, M. R., & Greene, D. (1978). Overjustification research and beyond. *The Hidden Costs of Reward. Hillsdale, NJ: Erlbaurri*.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129.

Lepper, M. R., Henderlong, J., & Gingras, I. (1999). Understanding the effects of extrinsic rewards on intrinsic motivation—Uses and abuses of meta-anlysis: Comment on Deci, Koestner, and Ryan (1999).

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255.

Loewenstein, G., Brennan, T., & Volpp, K. G. (2007). Asymmetric paternalism to improve health behaviors. *Jama*, *298*(20), 2415–2417.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Mela, C. F., Jedidi, K., & Bowman, D. (1998). The long-term impact of promotions on consumer stockpiling behavior. *Journal of Marketing Research*, 250–262.

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2008). Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science*, *3*(4), 324–338.

Neal, D. T., Wood, W., & Quinn, J. M. (2006). Habits—A repeat performance. *Current Directions in Psychological Science*, *15*(4), 198–202.

Nisbett, R. E., & Valins, S. (1971). *Perceiving the causes of one's own behavior*. General Learning Press.

Oyer, P. (1998). Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *Quarterly Journal of Economics*, 149–185.

Pink, D. H. (2011). *Drive: The Surprising Truth About What Motivates Us*. Penguin.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, *37*(1), 7–63.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Razran, G. (1954). The conditioned evocation of attitudes (cognitive conditioning?). *Journal of Experimental Psychology*, *48*(4), 278.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68.

Sandel, M. J. (2012). *What money can't buy: the moral limits of markets*. Macmillan.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science*, *9*(5), 340.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, *49*(1), 11–36.

Shah, J. Y., & Kruglanski, A. W. (2002). Priming against your will: How accessible alternatives affect goal pursuit. *Journal of Experimental Social Psychology*, *38*(4), 368–383.

Shu, S. B., & Gneezy, A. (2010). Procrastination of Enjoyable Experiences. *Journal of Marketing Research*, *47*(5), 933–944.

Skinner, B. F. (1953). *Science and human behavior*. Simon and Schuster.

Staats, B. R., Dai H., Hofmann D., Milkman K. L. (2016). Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare. *Management Science*

Tang, S.-H., & Hall, V. C. (1995). The overjustification effect: A meta-analysis. *Applied Cognitive Psychology*, *9*(5), 365–404.

Urminsky, O., & Goswami, I. (2015). Impatient to Achieve or Impatient to Receive: How the Goal Gradient Effect Underlies Time Discounting. *Workng Paper, University of Chicago*.

Urminsky, O., & Kivetz, R. (2011). Scope insensitivity and the "mere token" effect. Journal of Marketing Research, 48(2), 282-295.

Urminsky, O., & Zauberman, G. (2015). The Psychology of Intertemporal Preferences. Blackwell Handbook of Judgment and Decision Making, George Wu and Gideon Keren (eds), Wiley-Blackwell.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2014). Making choices impairs subsequent self-control: a limited-resource account of decision making, self-regulation, and active initiative.

Volpp, K. G., Levy, A. G., Asch, D. A., Berlin, J. A., Murphy, J. J., Gomez, A., … Lerman, C. (2006). A randomized controlled trial of financial incentives for smoking cessation. *Cancer Epidemiology Biomarkers & Prevention*, *15*(1), 12–18.

Volpp, K. G., John, L. K., Troxel, A. B., Norton, L., Fassbender, J., & Loewenstein, G. (2008). Financial incentive–based approaches for weight loss: a randomized trial. Jama, 300(22), 2631-2637.

Vroom, V. H. (1964). Work and motivation. 1964. *NY: John Wiley &sons*, 47–51.

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychological Review*, *66*(5), 297.

Yang, A., Urminsky, O., & Hsee, C. K. (2015). Eager to Help yet Reluctant to Give: How Pro-social Effort and Pro-social Choices Diverge. *Working Paper, University of Chicago*.

Zhou, H. & Fishbach, A. (2016) The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (yet False) Research Conclusions, Journal of Personality and Social Psychology, forthcoming.

Supplemental Materials

APPENDIX A: STUDY STIMULI

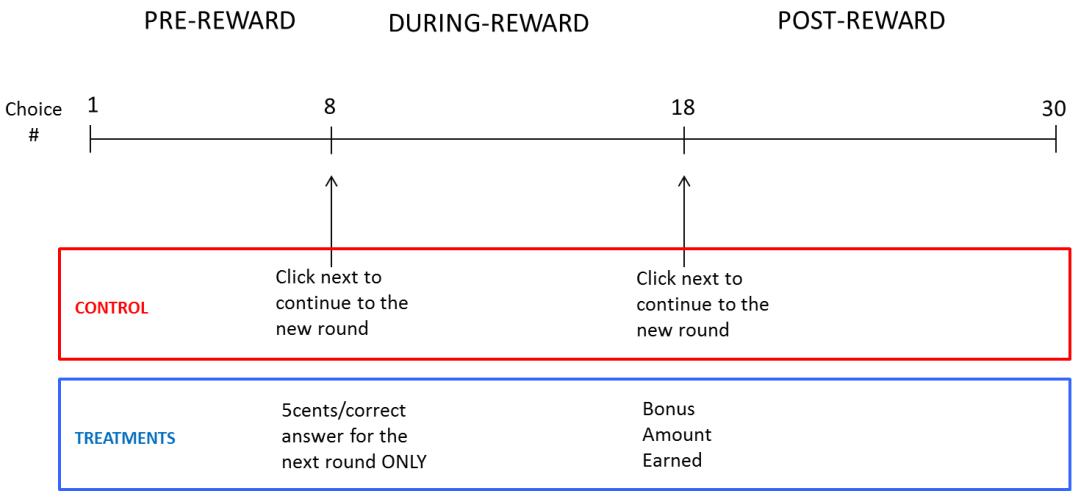


Exhibit A.1: The exhibit shows how the series of 30 repeated choices (or trials) were organized in all the studies. The trials were divided into three rounds for all the experimental groups, with 8, 10, and 12 trials in the pre-, during-, and post-reward rounds respectively. The reward group(s) learnt about the total rewards earned before the start of the third round. However, all payments were made at the very end of the study, and *not* after the immediate end of the second round.

PLEASE READ THE INSTRUCTIONS CAREFULLY.

In this survey you will be a given a series of **choices between doing cognitive tasks and watching videos of interesting television advertisements** collected from across the world.

The cognitive task will train your mental reasoning skills, and we will use your results to calibrate and standardize a training test. **You can do as many of them as you want**, or can just enjoy the videos.

A typical cognitive task in this survey looks like this.

The task will require **searching and selecting two numbers in a grid such that they add up to 10**. You can select a number by clicking on the box containing the number. An example is shown below.

8.63	4.38	2.68
5.72	1.67	7.38
7.32	3.62	1.29
7.02	5.17	1.62

Note:

1. **All** cognitive tasks in this study have an unique pair of answer.

2. You should **select ONLY two numbers in a grid**, and no more, as shown above.

You will get **30 seconds to complete each such tasks** after which the survey will automatically advance to the next screen. If you are done with the task before the time limit, you can click NEXT to proceed.

Alternatively, you can choose to watch a video clip for the same duration, and click NEXT when you are done viewing.

Click NEXT to see a typical video clip we have in this survey. This will also help test if the video loads properly in your browser.

NOTE: These clips are from different years and various countries, and therefore some of them might not be of very high video quality. But you should be able to view them without any issues.

Exhibit A.2: Four-stage instructions provided to all participants before the start of all the studies. Participants were informed that they could do as many of the target tasks (i.e., math tasks) as they wanted, or enjoy the videos during the whole study. They were also informed that all the cognitive tasks had one solution. Finally, they were informed that both tasks were of the same total duration. All participants were shown a sample video and asked if they were able to view and hear it properly - the experiment was aborted for those who reported having a problem.

Task Type = Writing

Please indicate your choice of a topic below.

- ○ Your favorite genre of Music
- ○ Your idea of favorite vacation

Please indicate your choice of a topic below.

- ○ Why are social media sites so important in today's popular culture?
- ○ Why are tattoos so popular in today's society?

Please indicate your choice below about which you would like to give your opinion.

- ○ Should kids be given smartphones?
- ○ Should the minimum age for teenagers to get a driver's license be increased?

Task Type = Brand-name Matching

Please choose the product category for which you would like to do the matching between logos and brand names.

- ○ Automobile
- ○ Sports Good

Please choose the product category for which you would like to do the matching between logos and brand names.

- ○ Oil and Gas Companies
- ○ Luxury and Accessories

Please choose the product category for which you would like to do the matching between logos and brand names.

- ○ Technology Companies
- ○ Banks and Financial Institutions

Exhibit A.3: The exhibit shows the types of tasks used for unrelated activities break in Study 2. Participants did all the three tasks of a particular type (Writing or Brand-name Matching), and each one of these tasks had a time limit of 30 seconds. The exhibit shows the choice condition. In the no-choice condition, one of the options was randomly pre-selected for the participant.

BONUS INFORMATION: PLEASE READ CAREFULLY.

In a previous version of this survey were able to pay as much as 5 cents for every correct answer, but in this version **we are unable to pay more**.

You will get 1 cent for every cognitive task that you answer correctly.

BONUS INFORMATION: PLEASE READ CAREFULLY.

In a previous version of this survey were able to pay only 5 cents for every correct answer, but in this version **we are able to pay a LOT more**.

You will get 50 cents for every cognitive task that you answer correctly.

Exhibit A.4: The exhibit shows the manipulations used to make the perceived reward magnitude salient in Study 3.

Target Task = Math

PLEASE READ THE INSTRUCTIONS CAREFULLY.

In this survey you will be asked to do a task. The task is to **solving cognitive math tasks**. WE WILL USE YOUR RESPONSES TO DESIGN EXPERIMENTAL STIMULI FOR A SPATIAL REASONING STUDY.

Since doing the task can be tiring, you will also have an option of a different task, **evaluating videos of television advertisements,** so that you can take a break.

It is completely up to you to choose which task you want to do in each round.

Target Task = Video

PLEASE READ THE INSTRUCTIONS CAREFULLY.

In this survey you will be asked to do a task. The task is to **evaluate videos of television advertisements**. WE WILL USE YOUR RESPONSES TO DESIGN EXPERIMENTAL STIMULI FOR AN ATTENTION AND PERCEPTION STUDY.

Since doing the task can be tiring, you will also have an option of a different task, **solving cognitive math problems,** so that you can take a break.

It is completely up to you to choose which task you want to do in each round.

Exhibit A.5: The exhibit shows the two types of framing used for the math (work) and the video (leisure) task in Study 4.

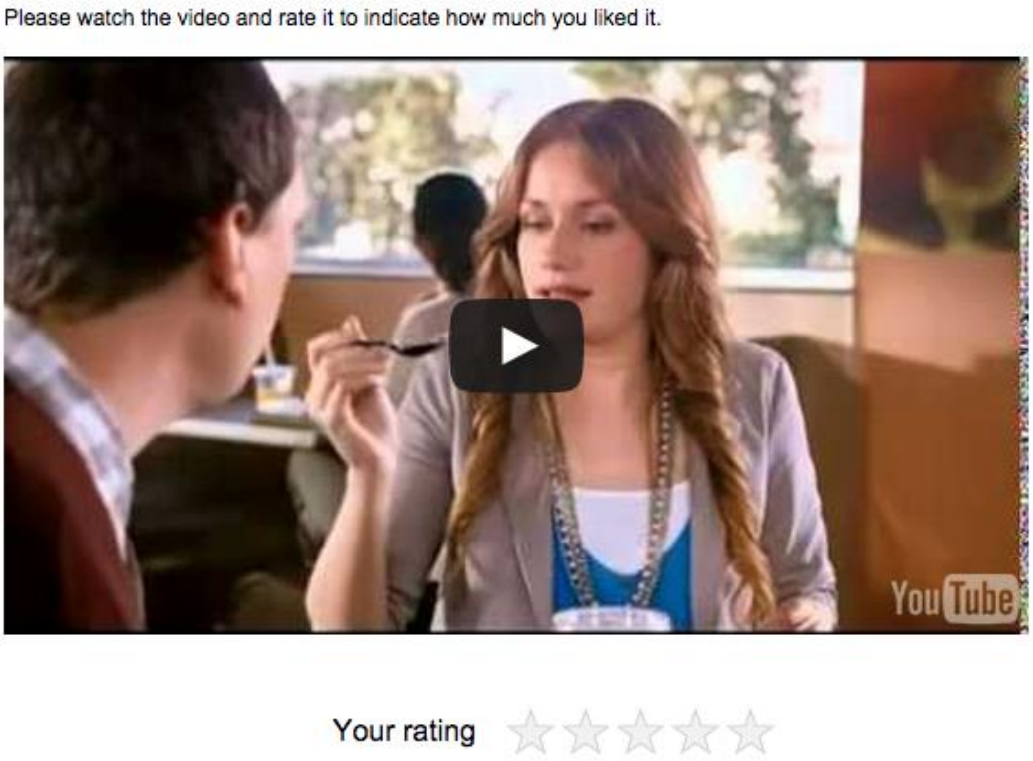


Exhibit A.6: The exhibit shows a typical video *task* used in Study 4. Unlike other studies (Study 1- 3), in this study the participants in the target-task = video condition were asked to rate the video in order to get their rewards.

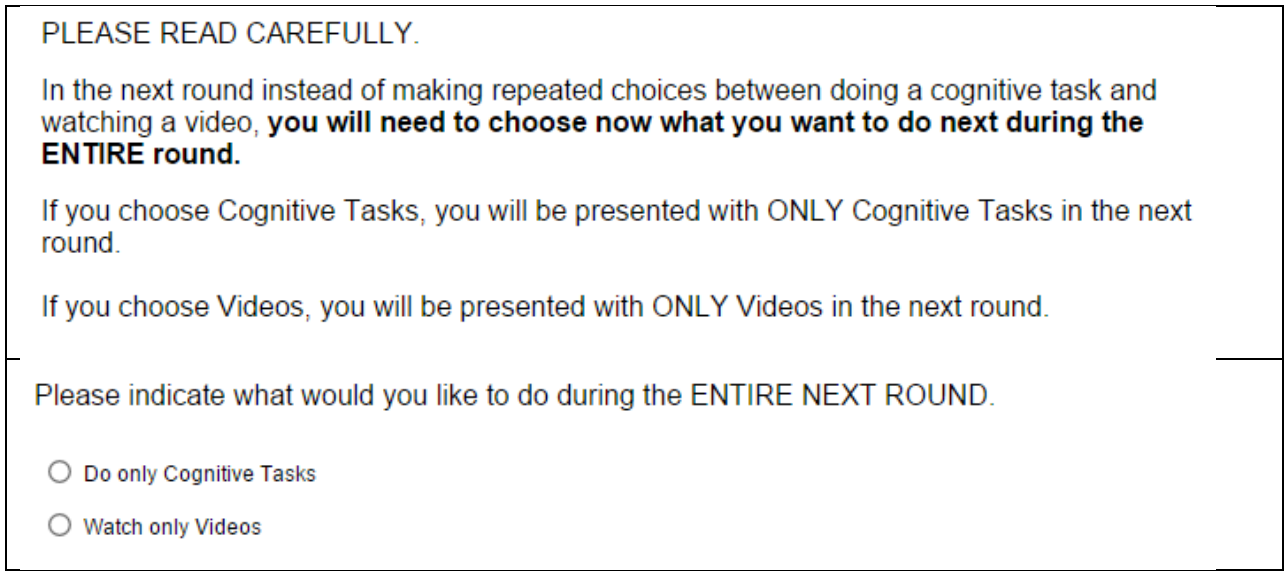


Exhibit A.7: The exhibit shows the instruction given to participants at the end of Round2 in Study 5. Participants were asked to choose which task they would like to do during the entire ensuing period of Round 3.

APPENDIX B: PRE-TEST RESULTS

Task Pretest (Math task, Video task)

A pretest (N=47) was done to examine how people felt about the math task (i.e., the target task) and the video task (i.e., the alternative task). A random sample of participants was chosen from the same population and they judged the two tasks on several attributes. Participants judged the math task as relatively more work-like compared to the video ($M_{math} = 6.87, SD = 2.19\ vs.\ M_{video} = 2.49, SD = 1.96; t(46) = 9.77, p < .001$), on 9-point scales, but considered the video more leisure-like compared to the math ($M_{math} = 3.89, SD = 2.63\ vs.\ M_{video} = 6.78, SD = 2.37; t(46) = 6.24, p < .001$). The math task was also judged as relatively more effortful ($M_{math} = 5.59, SD = 2.44\ vs.\ M_{video} = 1.95, SD = 1.52; t(46) = 9.05, p < .001$) and less entertaining compared to the video ($M_{math} = 4.64, SD = 2.32\ vs.\ M_{video} = 6.43, SD = 2.10; t(46) = 4.26, p < .001$).

Participants felt that the math task had more long-term benefits whereas the videos had higher immediate benefits($M_{math} = 5.38, SD = 1.97\ vs.\ M_{video} = 3.97, SD = 1.65; t(46) = 3.65, p < .001$). Participants also felt that more justification (on a 1 = *Less* to 9 = *More* scale) was needed for choosing to watch the video task over doing the math tasks, than for the opposite choice($M_{choose\ math} = 2.70, SD = 2.28\ vs.\ M_{choose\ video} = 4.72, SD = 2.97; t(46) = 3.53, p < .001$).

Most importantly, both the tasks satisfied the pre-condition required of tasks that can be deemed appropriate for testing theories of intrinsic motivation (Deci, Koestner, & Ryan, 1999). Both task had a rating of higher than the mid-point on a scale measuring how "interesting and enjoyable" the task is (1=*Low*, 9=*High*): math task ($M_{math} = 6.02, SD = 2.19$) and the video task ($M_{video} = 6.64, SD = 2.08$).

The question texts and the scales used are shown below:

Questions Texts	Scale
To what extent did you find the cognitive task interesting and enjoyable?	Extremely uninteresting and unenjoyable (1), (2), (3), (4), Neutral (5), (6), (7), (8), Extremely interesting and enjoyable (9)
To what extent did you find the video interesting and enjoyable?	
To what extent would you describe doing the cognitive task as a work activity?	NOT at all (1), (2), (3), (4), Neutral (5), (6), (7), (8), VERY MUCH like a work activity (9)
To what extent would you describe watching the video as a work activity?	
To what extent would you describe doing the cognitive task as a leisure activity?	
To what extent would you describe watching the video as a leisure activity?	
To what extent did you find the cognitive task effortful and tiresome?	NOT effortful and tiresome at all (1), (2), (3), (4), Neutral (5), (6), (7), (8), EXTREMELY

To what extent did you find the video effortful and tiresome?	effortful and tiresome (9)
To what extent did you find the cognitive task entertaining and relaxing?	Extremely unentertaining and unrelaxing (1), (2), (3), (4), Neutral (5), (6), (7), (8), Extremely entertaining and relaxing (9)
To what extent did you find the video entertaining and relaxing?	
To what extent do you believe that doing the cognitive task has short-term versus long-term benefits?	Has immediate benefits but little long-term benefits (1), (2), (3), (4), Neutral (5), (6), (7), (8), Has long-term benefits but little immediate benefits (9)
To what extent do you believe that watching the video has short-term versus long-term benefits?	
Imagine you had to make a choice between doing the cognitive tasks that will train your mental reasoning skills or watching the videos of interesting television advertisements. To what extent would you feel the need to justify to yourself choosing to do the cognitive tasks?	I will NOT need any justification at all (1), (2), (3), (4), Neutral (5), (6), (7), (8), I will need a LOT of justification (9)
Imagine you had to make a choice between doing the cognitive tasks that will train your mental reasoning skills or watching the videos of interesting television advertisements. To what extent would you feel the need to justify to yourself choosing to watch the videos?	

Breaks Pretest (Study 2)

Fifty-two participants from the same population as Study 2 participated in the online pretest. Participants were either shown the two versions of the writing task (with and without choice on the topic of writing), or the two versions of the logo-matching task (with and without choice on the product category for which the brand names and the logos were required to be matched).

After reading about these tasks, participants indicated which of the versions they thought required more thinking, required more work, was more difficult, or provided more autonomy. For each of the questions, participants were also given the option to indicate if they could not say whether the two versions were different on the attribute in question.

Across these various attributes, the proportion of participants who chose the "Can't Say" option varied from 7% ("required more thinking") to 19% ("granted more autonomy"). We coded this data as not-available. Among participants who gave an answer, 73% indicated that the version with choices required more thinking ($\chi^2(1) = 10.08$, $p = .001$), 68% indicated that the version with choices required more work ($\chi^2(1) = 6.72$, $p = .009$) and 66% indicated that the version with choices was more difficult ($\chi^2(1) = 5.00$, $p = .025$). However, 64% of the participants indicated that the task version with choices gave them more autonomy than the version without choices ($\chi^2(1) = 3.43$, $p = .064$). Therefore, the participant population found the task versions with choices more effortful, even though they felt it granted them marginally more autonomy.

The question text and the measurement instrument used are shown below:

How would you compare the two tasks on the following dimensions?

Please remember these are relative judgments. That is you are indicating, of these two tasks, which description fits one of them better.

However, if you are completely unsure please indicate CAN'T SAY.

	Task without Choice	**Task with Choice**	**Can't Say**
Requires more thinking			
Requires more work			
Is more difficult			
Grants more autonomy			

APPENDIX C: DATA CLEANING PROTOCOL

Every study started with an initial sampling of both types of tasks (math, video) after which participants were asked if they faced any technical problems. If a problem was reported, the study was aborted and data from these participants were discarded from further analysis.

Our experimental paradigm was specifically designed to capture dynamic changes in behavior over time, and could distinguish between temporary and permanent disengagements. A temporary decrease in motivation to do the target task would be reflected in the participant choosing to watch the video for a few trials before choosing to do the math task again. A more persistent decrease in motivation to do the target activity could be reflected in two ways. Participants could "quit" within the study, by repeatedly deciding to only watch the videos for the remaining duration. Alternatively, participants could quit by ending the study part way through and not completing the remaining trials. We tracked all dropouts, and included participants who dropped-out of the study after completing the pre-reward baseline period coding their participation as zero for the target task. The reward for the math task to the treatment group was announced at the end of the pre-reward baseline period, and therefore this analysis strategy ensured that we included anyone whose behavior could have been impacted by the incentives, whether they finished the study or not.

Participant's data containing duplicate IP addresses were removed prior to analysis. Finally, an attention check question was administered at the end of the survey, and data from participants who reached till the end of the survey but failed the attention check were discarded prior to analysis. Participants who quit part way through and therefore did not answer the attention check question were given the benefit of doubt and were included in the analysis.

APPENDIX D: FULL DETAILS OF HIERARCHICAL REGRESSIONS

Model for Momentary post-reward engagement reduction

We capture total momentary post-reward engagement reduction using a functional form assumption about how effort returns to baseline over time in the post-reward period. Assuming a non-linear return of effort (i.e., likelihood of choosing the math task) to baseline over time (the number of periods *t* since the incentive ended), we parameterize momentary post-reward engagement reduction (MOMENTARY) as:

$$MOMENTARY_t = \frac{1}{t} \quad (1)$$

Using this parameterization[10], the probability of individual i choosing to do the math task in post-reward (Round 3) during trial *t* can be written as:

$$P(Y_{ti} = 1) = \phi\ (\beta_{0i} + \beta_{Mi}\ MOMENTARY_t) \quad (2)$$

In our model we set ϕ to the logit link function and β_{0i} is a person-specific intercept and β_{Mi} is a person-specific momentary post-reward engagement reduction behavior. In the hierarchical regression the parameters in Equation (2) are a function of time-invariant individual-level covariates, to account for the repeated observations per person.

$$\beta_{0i} = \beta_{00} + \beta_{01}\ C_i + \beta_{02}\ X_i + u_{0i} \quad (3)$$

The person-specific baseline parameter β_{0i} is a function of the condition that individual *i* has been randomly assigned to experimental condition C_i, the total number of math task choices by individual i in the pre-incentive Round 1 X_i, as well as the population baseline β_{00} and time-invariant person-specific error term u_{0i}.

$$\beta_{Mi} = \beta_{10} + \beta_{11}\ C_i + u_{1i} \quad (4)$$

The person-specific momentary post-reward engagement reduction behavior β_{Mi} is estimated as a function of experimental condition C_i, as well as the baseline β_{10} and the individual-specific error term u_{1i}. The random effects for the intercept and the slope for every individual i, u_{0i}, u_{i1}, are assumed to be bi-variate normal with zero-mean, variances τ_{00}, τ_{11} and common co-variance τ_{01}. This error structure accounts for the potentially correlated repeated-measures for each individual. Combining equations, (2), (3), and (4) yields an "intercepts and slopes-as-outcomes" model (Raudenbush and Bryk, 2002).

The expected proportion of math tasks chosen in each trial *t* of Round 3 is:

[10] We report robustness analysis with other parameterization in Appendix D of the Supplemental Materials. We also report results using a more flexible non-parametric approach in Appendix E.

$$P(Y_{ti} = 1) = \phi\ (\beta_{00} + \beta_{10} MOMENTARY_t + \beta_{01} C_i + \beta_{02} X_i + \beta_{MOMENTARY} C_i * MOMENTARY_t) \quad (5)$$

The coefficient β_{01} is renamed to β_{LT} in equation (5). The coefficient $\beta_{MOMENTARY}$ in equation (5), which is β_{11} from equation (4) renamed for ease of exposition, tests for a difference in the extent of momentary post-reward engagement reduction in the experimental condition ($C_i = 1$), compared to the corresponding time periods in the control condition ($C_i = 0$). A significant and negative $\beta_{MOMENTARY}$ generally indicates momentary post-reward engagement reduction after the incentive ended, compared to the corresponding trials in the control condition, controlling for individual differences in baseline effort X_i. An important exception to such an interpretation arises when there is an overall *increase* in effort of the reward group relative to the control group in the post-reward period as indicated by a significant β_{LT}. In this case a significant negative $\beta_{MOMENTARY}$ might indicate an immediate decrease in the effort of the reward group relative to its longer-run steady-state level, but not a momentary post-reward engagement reduction relative to the contemporaneous control group.

In a similar vein, an estimate of $\beta_{MOMENTARY}$ that is not statistically distinguishable from zero represents a consistent level of effort throughout Round 3, with two very different potential interpretations. A non-significant $\beta_{MOMENTARY}$ could indicate that no post-reward reduction in engagement has occurred or it could represent a consistent overall increase or decrease in engagement in the reward group in Round 3. Hence, it will be important to also estimate the overall effects of incentives on choices in Round 3, in addition to momentary reduction in engagement behavior and interpret these parameters jointly. Next, we describe the tests we use to estimate the overall effects of incentives.

Difference-in-Difference Model for Overall Effects

We use a hierarchical non-linear difference-in-difference model to estimate differences in the overall probability of choosing the math task between two experimental conditions $C_i = 0$ or $C_i = 1$ and between two experimental rounds $R_t = 0$ or $R_t = 1$. The general specification can be written as follows:

$$P(Y_{ti} = 1) = \phi\ (\beta_0 + \beta_1 * R_t + \beta_2 * C_i + \beta_3 * R_t * C_i) \quad (6)$$

The interpretation of the key coefficient β_3 depends on how the rounds (R_t) and conditions (C_i) are coded. To estimate the effect of incentives on during-reward performance β_{REWARD} we compare during-incentive Round 2 ($R_t = 1$) to baseline Round 1 ($R_t = 0$) and exclude Round 3 data. To estimate the overall post-reward engagement level β_{POST} we compare post-reward Round 3 ($R_t = 1$) to baseline Round 1 ($R_t = 0$), and exclude Round 2. The net effect of incentives (e.g., during-reward and post-reward behavior) β_{NET} is estimated by

comparing the combined during-reward Round 2 and post-reward Round 3 trials ($R_t = 1$) to baseline Round 1 ($R_t = 0$).

APPENDIX E: STUDIES USED IN META-ANALYSIS

Studies	Remarks	Sample Size	$\beta_{MOMENTARY}$	SE	z	p
Study 1	Study 1 in paper	C = 39; R = 38	-2.6	1.0	-2.6	.009**
Study 2	Study 2 in paper	C = 41; R = 46	-2.5	0.9	-2.5	.01*
Study 3	Study 3 in paper	C = 68; R = 56	-3.5	1.1	-3.1	.002**
Other Studies with Control and Replication Treatment conditions included in meta-analysis						
Study A	Replication of momentary reduction with advance notice about temporary nature of rewards at the end of Round 2	C = 31; R = 33	-2.2	0.9	-2.3	.02*
Study B	Replication of momentary reduction in engagement	C = 72; R = 74	-1.6	0.9	-1.7	.09
Study C	Replication of Study 2	C = 42; R = 41	-2.3	1.6	-1.4	.15
Study D	Replication of Study 3	C = 35; R = 36	-1.5	1.2	-1.3	.20
Study E	Study with and without a pre-reward round	C = 56; R = 96	-2.6	1.3	-2.0	.048*
Study F	Study with framing math vs both math and video as important (refer Appendix M)	C = 59; R = 52	-4.7	1.2	-3.8	<. 001***
Study G	Replication of Study with incentives for Math vs Leisure (refer Appendix L)	C = 87; R = 96	-3.2	0.7	-4.6	<. 001***

*p<.05; **p<.01; ***p<.001; C = control group, R = Reward group

Notes:

1. Study 4 in the main paper was not included in the meta-analysis because it used a different math task. The results will not change (in fact, will become stronger) if that study is also included. Study 5 was not included because the participants made a single choice for all the trials in Round 3 (as opposed to a series of choices, on per trial).
2. Studies B and E are part of a different project which also had the regular control and replication cells.

APPENDIX F: ROBUSTNESS CHECKS USING META-ANALYSIS DATA

Estimates for Momentary Reduction in Engagement ($\beta_{MOMENTARY}$) in meta-analysis with various specifications

	(1)	(2)	(3)	(4)
	Model with $MOMENTARY_t = \frac{1}{t}$	Model with $MOMENTARY_t = \frac{1}{t^2}$	Model with $MOMENTARY_t = \frac{1}{\sqrt{t}}$	Model with $MOMENTARY_t = \frac{1}{t}$ and excluding dropouts
Constant	-4.8 (0.3)***	-4.8 (0.3)***	-5.1 (0.3)***	-4.4 (0.3)***
$MOMENTARY_t$	1.3 (0.2)***	1.5 (0.3)***	1.5 (0.3)***	1.2 (0.2)***
Condition = Reward	1.3 (0.2)***	0.9 (0.2)***	2.2 (0.3)***	1.3 (0.2)***
Total Attempts in Round 1	1.0 (0.05)***	1.1 (0.05)***	1.0 (0.05)***	1.0 (0.04)***
$MOMENTARY_t$ * Condition	-3.0 (0.3)***	-2.8 (0.4)***	-3.7 (0.4)***	-2.9 (0.3)***
Dropouts Included	Yes	Yes	Yes	No
N	1098	1098	1098	1046
BIC	9881	9899	9881	9526

*p<.05; **p<.01; ***p<.001

Notes:

1. For the sake of parsimony and simplicity, we model $\boldsymbol{MOMENTARY_t}$ as per specification 1 in this paper which also has the lowest BIC.

2. Columns 1-3 compare different functional form specifications for the $\boldsymbol{MOMENTARY_t}$ parameter, and column 4 excludes the dropouts. Dropouts are participants who dropped-out of the study after completing Round 1.The parameter $\boldsymbol{t}$ indicates the post-reward trial or choice number.

APPENDIX G: NON-PARAMETRIC MODEL FOR ESTIMATIMNG MOMENTARY POST-REWARD ENGAGEMENT REDUCTION

Specification 1

We used the following flexible specification for predicting probability of choosing the math task in the post-reward round by individual i at trial t:

$$P(Y_{ti} = 1) = \phi\ (T * C_i + X_i + S)$$

$\phi =$ Logit link function

$T =$ Cubic splines $(df = 3)$ for post-reward trial$t$, $t \in$ [1..12] i.e., 12 post-reward choices

C_i = Experimental condition to which individual i is randomly assigned

$X_i =$ Total number of math attempts by individual i in the pre-incentive round. This is a measure of individual level differences in ability or interest in the target-task of the experiment

$S =$ Study Fixed Effect

The model specification did not use any assumption about how effort (e.g. attempting the math task) returns to baseline after incentives are stopped. Using flexible cubic-splines a piece-wise third order polynomial is used to fit individual post-reward behavior after the incentive ends.

Also, instead of using distributional assumptions to draw statistical inferences, we predicted the probability of attempting the math task for each of the post-reward trials in both the control and the treatment group. Using a 1000-sample bootstrap test, we examined if this difference did not contain zero to infer statistical difference in behavior between these two experimental groups.

Post-reward Choice #	Bootstrapped 95% CI	
Choice 1	-19.6%	-10.6%
Choice 2	-12.6%	-4.8%
Choice 3	-7.8%	0.0%
Choice 4	-4.3%	3.6%
Choice 5	-1.9%	6.1%
Choice 6	-0.2%	7.7%
Choice 7	0.9%	8.7%
Choice 8	1.4%	9.3%
Choice 9	1.7%	9.6%
Choice 10	1.8%	9.7%
Choice 11	2.0%	9.8%
Choice 12	1.7%	10.7%

Figure in the left-panel shows the predicted probability of choosing the math task in the reward group (blue triangular dot) and the control group (red circular dot) in Round 3 based on a typical run of the non-parametric model. The table in the right-panel shows the bootstrapped 95% CI from the non-parametric model for the *difference* in the predicted probability of choosing the math task between the reward and the control condition.

Specification 2

We use an alternative specification that too did not make parametric assumptions about how post-reward behavior varied during Round 3. The specification was as follows:

$$P(Y_{ti} = 1) = \phi\ (TD * C_i + Round2 * C_i + S)$$

ϕ = Logit link function

TD = Trial dummies for each of the post-reward trials $t \in [1..12]$

Round2 = Reward Round dummy

C_i = Experimental condition to which individual i is randomly assigned

S = Study Fixed Effect

Therefore, the interaction terms of post-reward trial dummies with the experimental condition measured the differences in the relative level of post-reward engagement in the reward group versus the control group, relative to the pre-reward baseline level (i.e., Round 1).

	β	SE	*z*	*p*	
Constant	0.45	0.06	7.87	<.001	***
Reward Round	-0.08	0.04	-1.74	0.082	
Trial 1	0.01	0.10	0.10	0.923	
Trial 2	-0.02	0.10	-0.25	0.804	
Trial 3	-0.14	0.10	-1.45	0.146	
Trial 4	-0.12	0.10	-1.28	0.200	
Trial 5	-0.10	0.10	-1.02	0.306	
Trial 6	-0.02	0.10	-0.16	0.871	
Trial 7	-0.24	0.09	-2.56	0.010	*
Trial 8	-0.29	0.09	-3.07	0.002	**
Trial 9	-0.24	0.09	-2.56	0.010	*
Trial 10	-0.35	0.09	-3.75	<.001	***
Trial 11	-0.23	0.09	-2.48	0.013	*
Trial 12	-0.28	0.09	-2.99	0.003	**
Condition = Reward	-0.10	0.04	-2.23	0.026	*
Round 2 x Reward	1.42	0.07	21.52	<.001	***
Trial 1 x Reward	-0.52	0.13	-3.96	<.001	***
Trial 2 x Reward	-0.24	0.13	-1.82	0.069	
Trial 3 x Reward	-0.03	0.13	-0.24	0.811	
Trial 4 x Reward	0.05	0.13	0.36	0.716	
Trial 5 x Reward	0.16	0.13	1.20	0.230	
Trial 6 x Reward	0.27	0.14	1.97	0.049	*
Trial 7 x Reward	0.37	0.13	2.75	0.006	**
Trial 8 x Reward	0.39	0.13	2.94	0.003	**
Trial 9 x Reward	0.34	0.13	2.58	0.010	**
Trial 10 x Reward	0.27	0.13	2.06	0.039	*
Trial 11 x Reward	0.37	0.13	2.75	0.006	**
Trial 12 x Reward	0.39	0.13	2.94	0.003	**

*p<.05; **p<.01; ***p<.001

Note: The regression includes study fixed effects whose estimates are not shown.

APPENDIX H: EFFECT OF TEMPORARY INCENTIVES ON ACCURACY AND NET OUTCOME

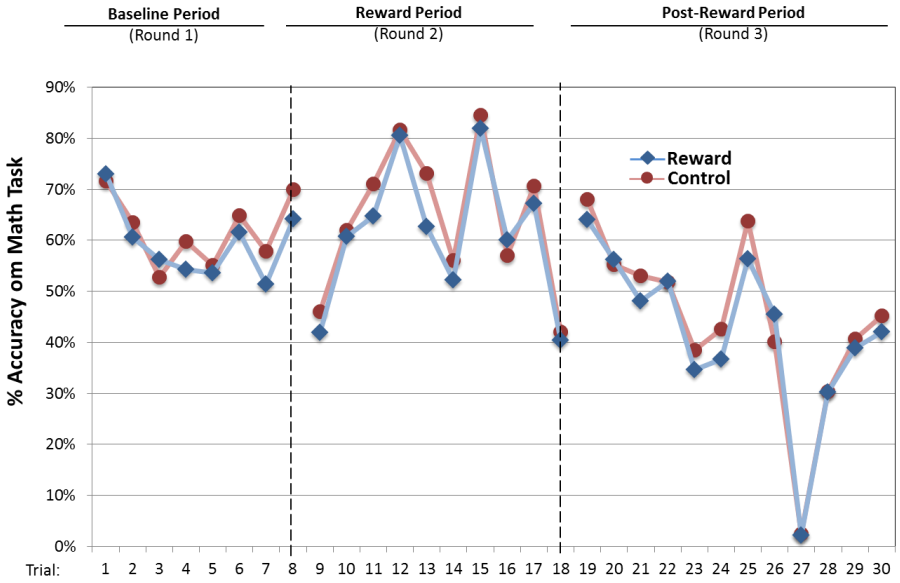


Figure H.1: Raw data showing the percentage of participants accurately solving the math task, conditional on choosing to attempt it. Incentives did not affect accuracy in our experiments.

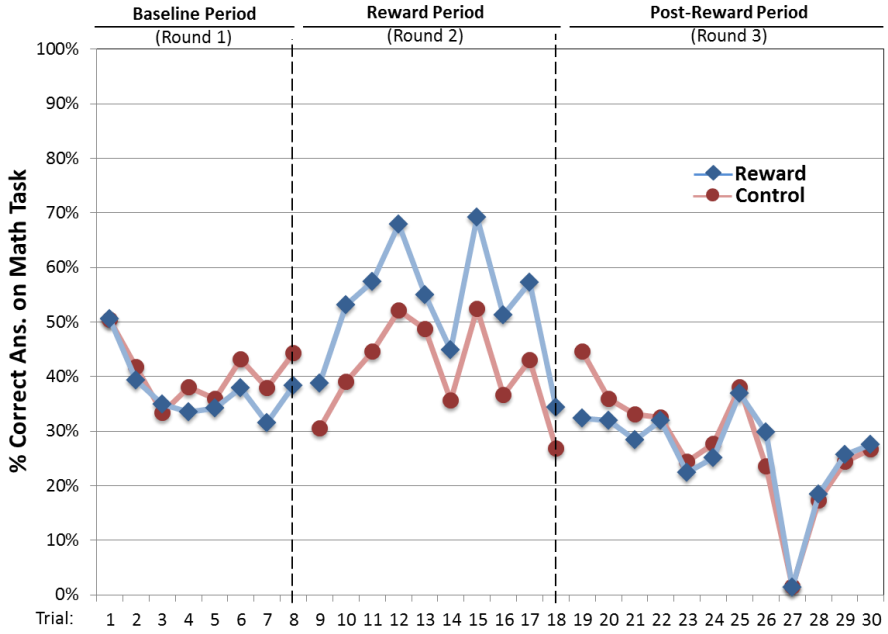


Figure H.2: Raw data showing proportion of correct answers for every trial in each round. Incentives had a net positive effect on the total number of correct answers, driven by the reward period, despite a significant post-reward decrease after incentives ended.

We primarily focused on effort (e.g., choosing to do the target task) as the key variable of interest, consistent with the approach used in the intrinsic motivation literature because effort is a behavioral outcome variable that represents a person's motivation level. Our flexible experimental paradigm also allowed us to also examine the effect of temporary incentives on accuracy (probability of correctly answering the math task after deciding to attempt it) and net outcome (total number of correct answers). The incentive could have resulted in people choosing the math task without being able to answer correctly, resulting in a decrease in accuracy compared to the control condition. However, we did not observe any such effects and temporary incentives did not affect accuracy at all. Therefore, as shown in figures H.1 and H.2, the same conclusions hold for effort and net outcome – a significant positive effect of incentives on the total number of math problems solved correctly ($\beta_{NET} = +1.25, z = +5.21, p < .001$).

APPENDIX I: MODERATION BY INITIAL MOTIVATION

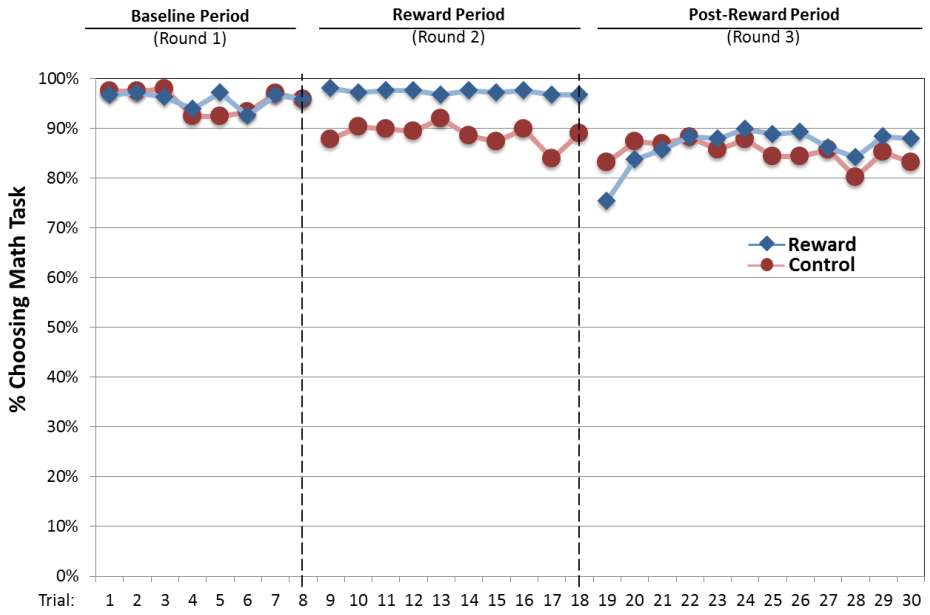


Figure I.1: Post-reward behavior of participants with high initial task interest using the internal meta-analysis data. The figure shows a significant post-reward decrease in engagement after incentives ended.

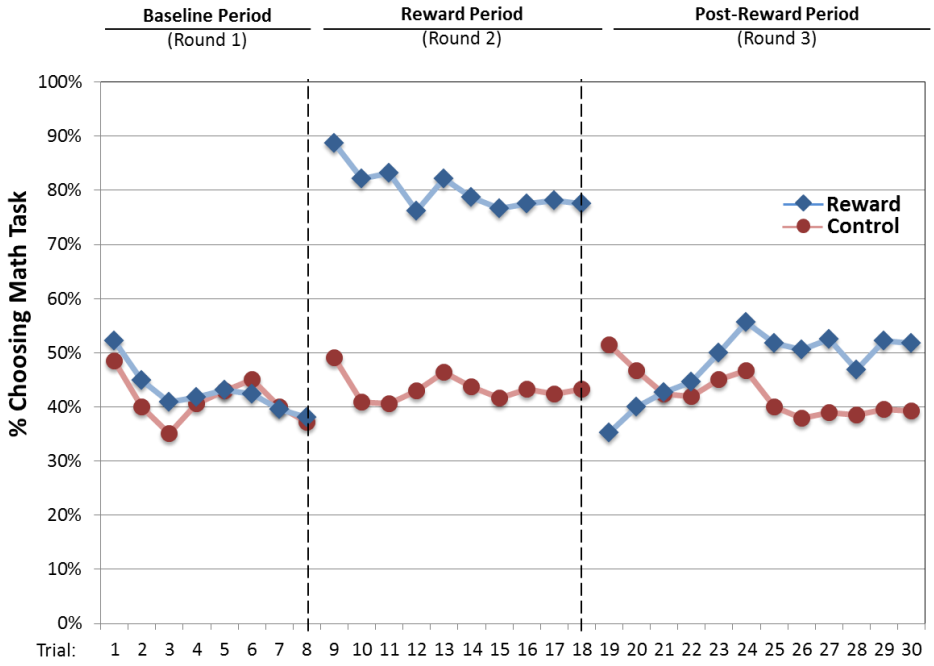


Figure I.2: Post-reward behavior of participants who low initial task interest using the internal meta-analysis data. The figure shows a significant post-reward decrease in engagement after incentives ended.

As shown in figures I.1 and I.2, the result of temporary incentives on post-reward behavior is very similar for the group of participants who exerted more versus less effort in the pre-reward round. This difference in initial effort represents difference in initial intrinsic motivation because in the pre-reward period participants did not know about any impending rewards. Both groups show an increase in effort when the rewards are available, followed by a decrease in the choice of the math task in the immediate post-reward period (low intrinsic motivation: $\beta_{MOMENTARY} = -2.11, z = -2.71, p = .007$; high intrinsic motivation ($\beta_{MOMENTARY} = -3.22, z = -8.84, p < .001$), and initial interest level does not moderate the momentary reduction in engagement behavior ($\beta_{MOMENTARY\ interaction} = .15, z <$ 1).

APPENDIX J: HETEROGENEITY IN POST REWARD BEHAVIOR

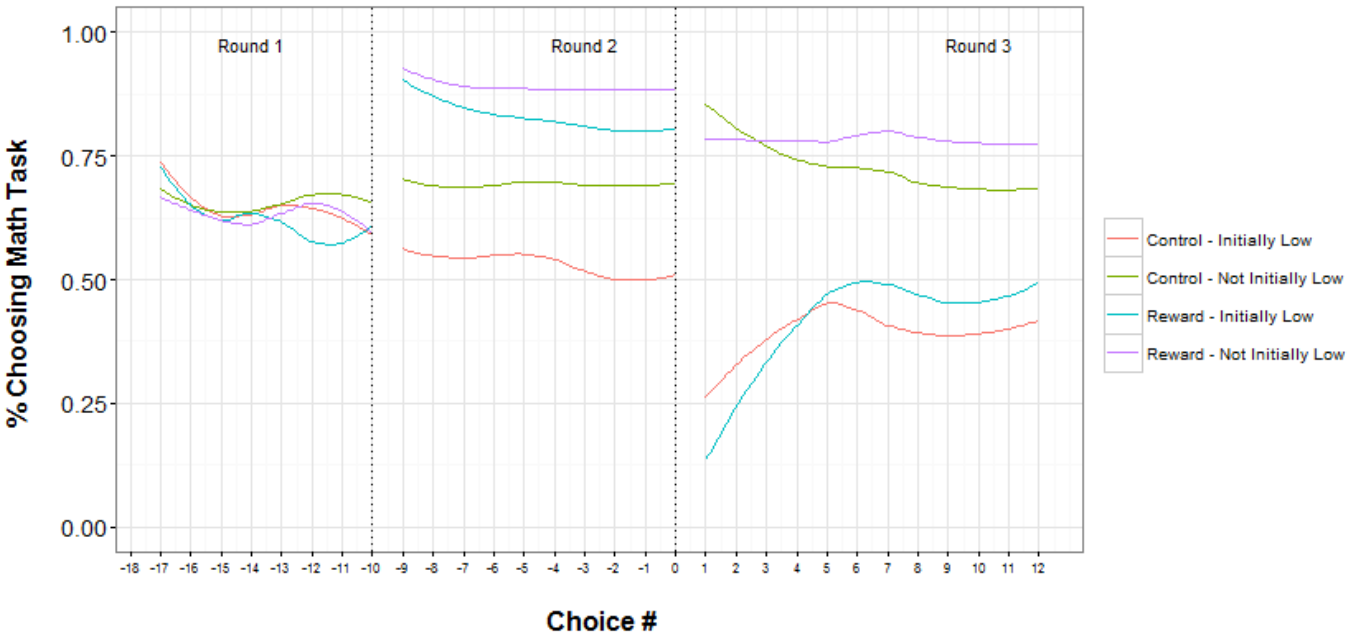


Figure J.1: Post-reward behavior (lowess lines) of participants with different initial behavior after the rewards ended. Both the reward and the control groups are further divided into two sub-groups each based on their initial post-reward behavior (initial post reward behavior lower or higher than their individual pre-reward baseline).

As shown in Figure J.1, the reward group that shows initial post-reward reduction in engagement eventually settles at a level higher than the corresponding control group. This suggests that the aggregate behavior was *not* driven by two types of reward group participants – one that showed a persistent post-reward reduction in engagement as predicted by prior theories (relative to corresponding control) and another that showed a persistent post-reward increase in engagement (relative to corresponding control). As shown in the table below, the final post-reward behavior does not differ between the control and the reward as a function on their initial post-reward behavior.

	DV = Final Post-reward Behavior (Normalized)	
Condition = Reward	0.121***	(-0.014)
Initial Reduction in Engagement (Normalized)	0.627***	(-0.03)
Condition = Reward x Initial Reduction in Engagement	0.016	(-0.039)
Constant	-0.048***	(-0.01)
Observations	1098	
R-Squared	0.503	

Note: *p<0.1; **p<0.05; ***p<0.01; Normalized = Average Initial Reduction - Pre-reward Baseline

Effect of High vs. Low Accuracy on Post-Reward Behavior

	β	SE	z	p	
Constant	-4.21	0.40	-10.48	<.001	***
$MOMENTARY_t$	1.54	0.55	2.82	0.005	**
Condition = Reward	0.90	0.46	1.94	0.053	
Proportion Correct in Round 2	1.70	0.52	3.27	0.001	**
Total Attempts in Round 1	0.81	0.05	17.40	<.001	***
$MOMENTARY_t$ * Condition	-2.50	0.76	-3.29	0.001	***
$MOMENTARY_t$ * Proportion Correct in Round 2	-0.63	0.86	-0.73	0.469	
Reward * Proportion Correct in Round 2	0.19	0.72	0.26	0.794	
$MOMENTARY_t$* Reward * Proportion Correct in Round 2	-0.64	1.16	-0.55	0.582	

*p<.05; **p<.01; ***p<.001

Effect of Average Time Taken in Round 2 on Post-Reward Behavior

	β	SE	z	p	
Constant	-1.27	0.67	-1.90	0.057	
$MOMENTARY_t$	0.41	1.04	0.39	0.696	
Condition = Reward	1.40	0.82	1.70	0.090	
Average Time to do Math in Round 2	-0.10	0.03	-3.52	<.001	***
Total Attempts in Round 1	0.82	0.05	18.13	<.001	***
$MOMENTARY_t$ * Condition	-4.25	1.35	-3.15	0.002	**
$MOMENTARY_t$ * Average Time to do Math in Round 2	0.04	0.05	0.76	0.450	
Reward * Average Time to do Math in Round 2	-0.02	0.04	-0.44	0.660	
$MOMENTARY_t$* Reward * Average Time to do Math in Round 2	0.07	0.06	1.07	0.286	

*p<.05; **p<.01; ***p<.001

Effect of Average Time Taken in Round 3 on Post-Reward Behavior

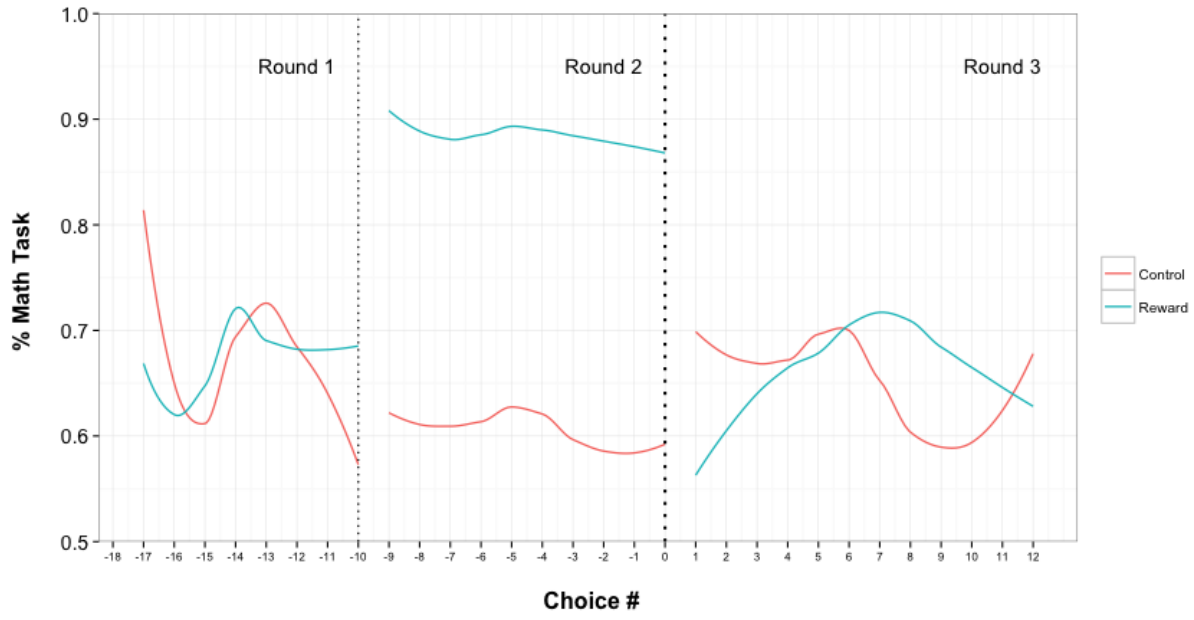
	β	SE	z	p	
Constant	-1.96	0.23	-8.39	<.001	***
$MOMENTARY_t$	1.69	0.38	4.44	<.001	***
Condition = Reward	1.42	0.43	3.29	<.001	**
Average Time to do Math in Round 3	0.00	0.00	-0.43	0.67	
Total Attempts in Round 1	0.59	0.03	17.25	<.001	***
$MOMENTARY_t$ * Condition	-4.33	0.88	-4.91	<.001	***
$MOMENTARY_t$ * Average Time to do Math in Round 3	-0.01	0.01	-0.90	0.37	
Reward * Average Time to do Math in Round 3	-0.01	0.02	-0.33	0.74	
$MOMENTARY_t$* Reward * Average Time to do Math in Round 3	0.07	0.04	1.67	0.09	

*p<.05; **p<.01; ***p<.001

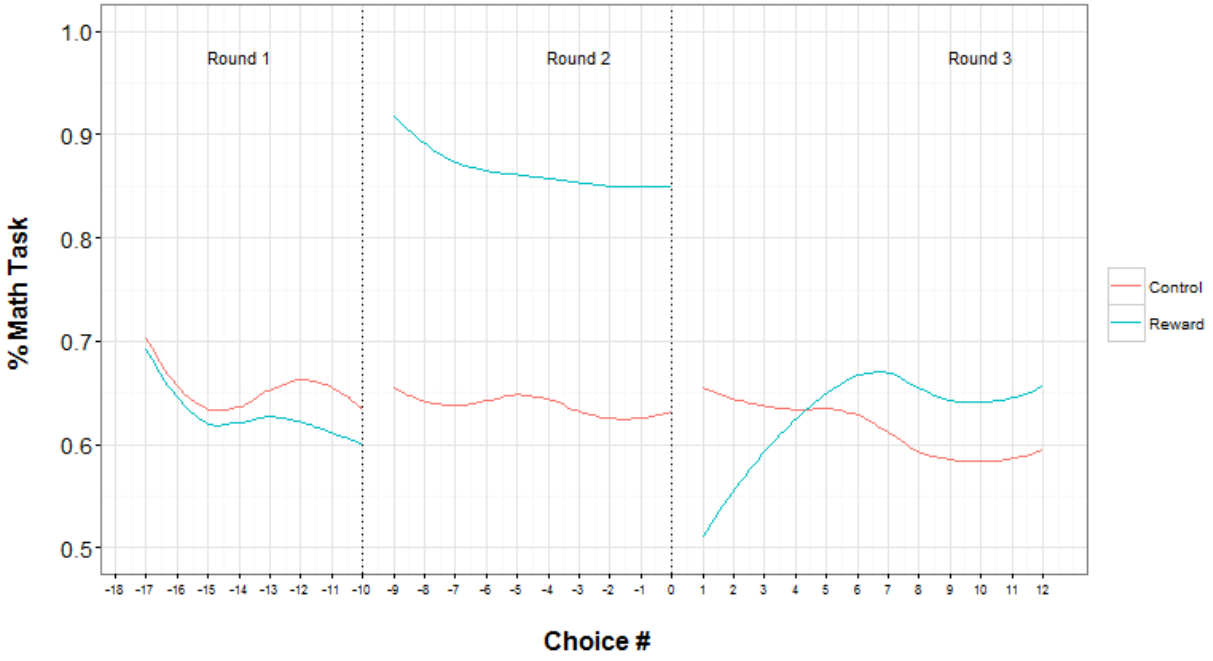
APPENDIX K: ADDITIONAL CHARTS FOR ALL STUDIES

Raw Choice Data of All Studies with Lowess Smoother

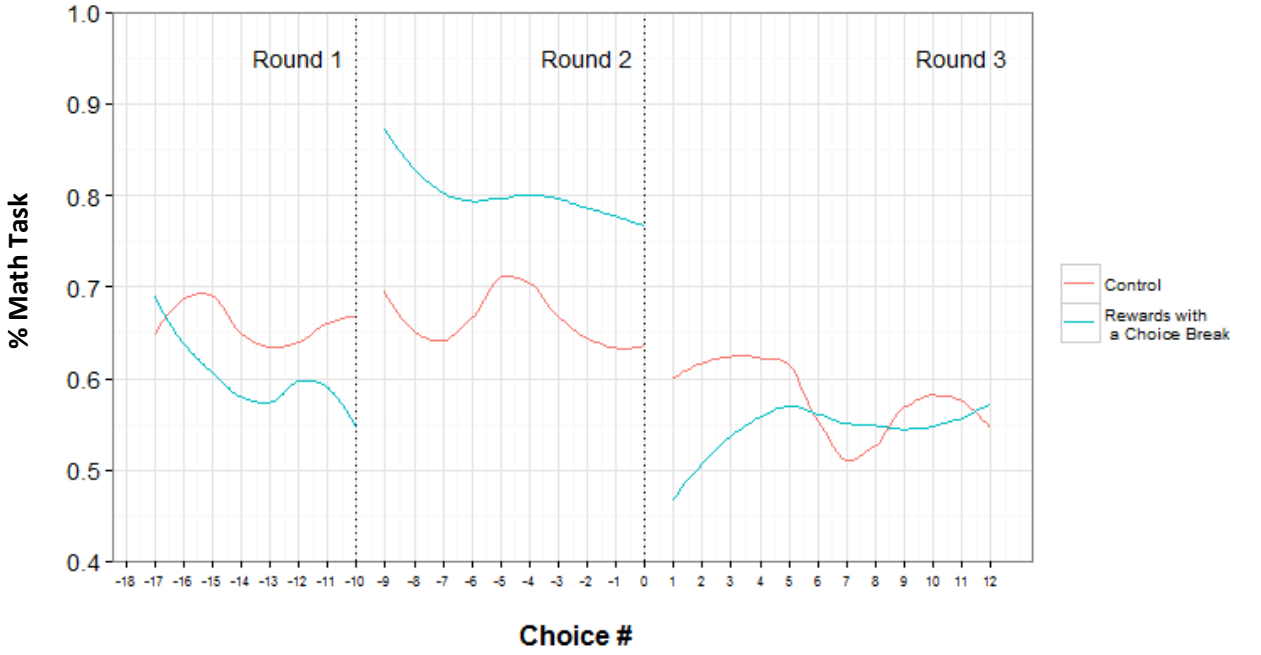
Study 1: The Dynamics of Post-Reward Task Engagement

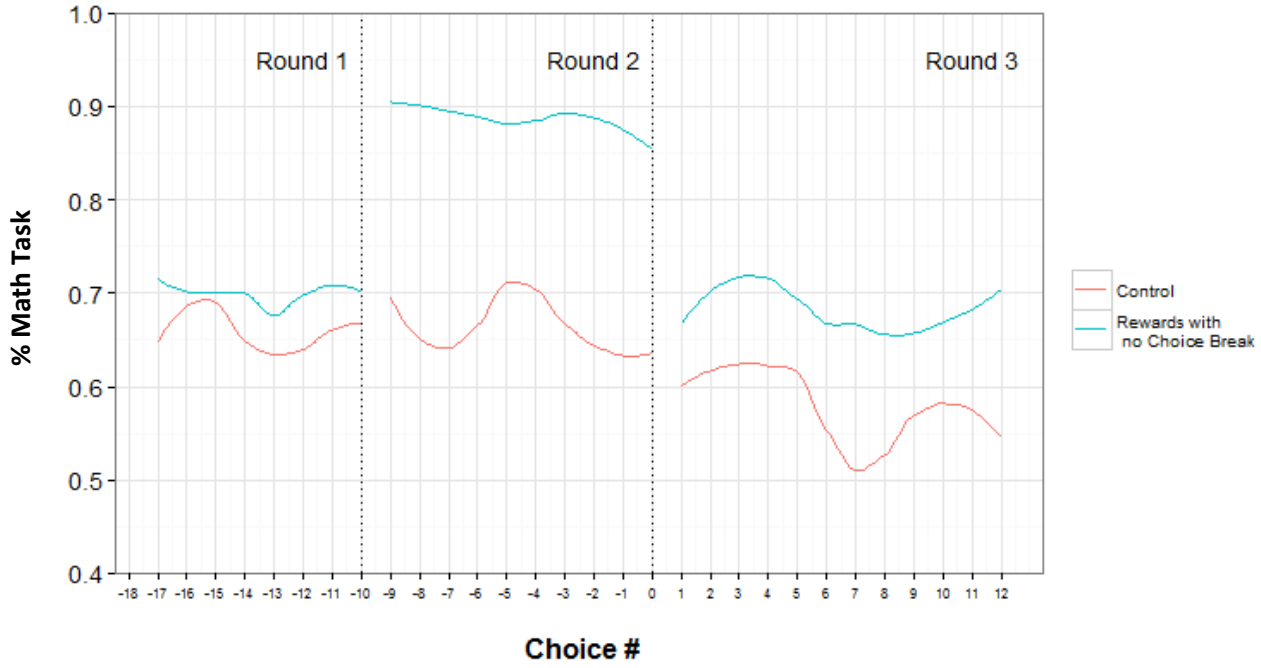


Meta-Analysis: The Dynamics of Post-Reward Task Engagement

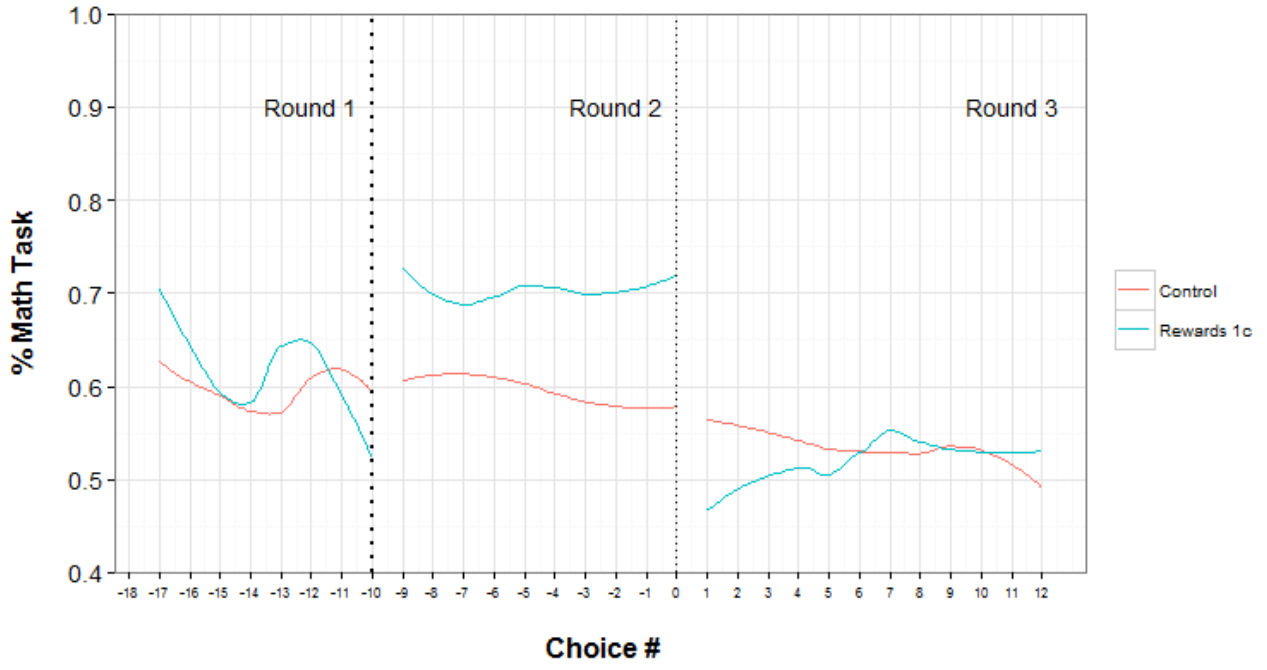


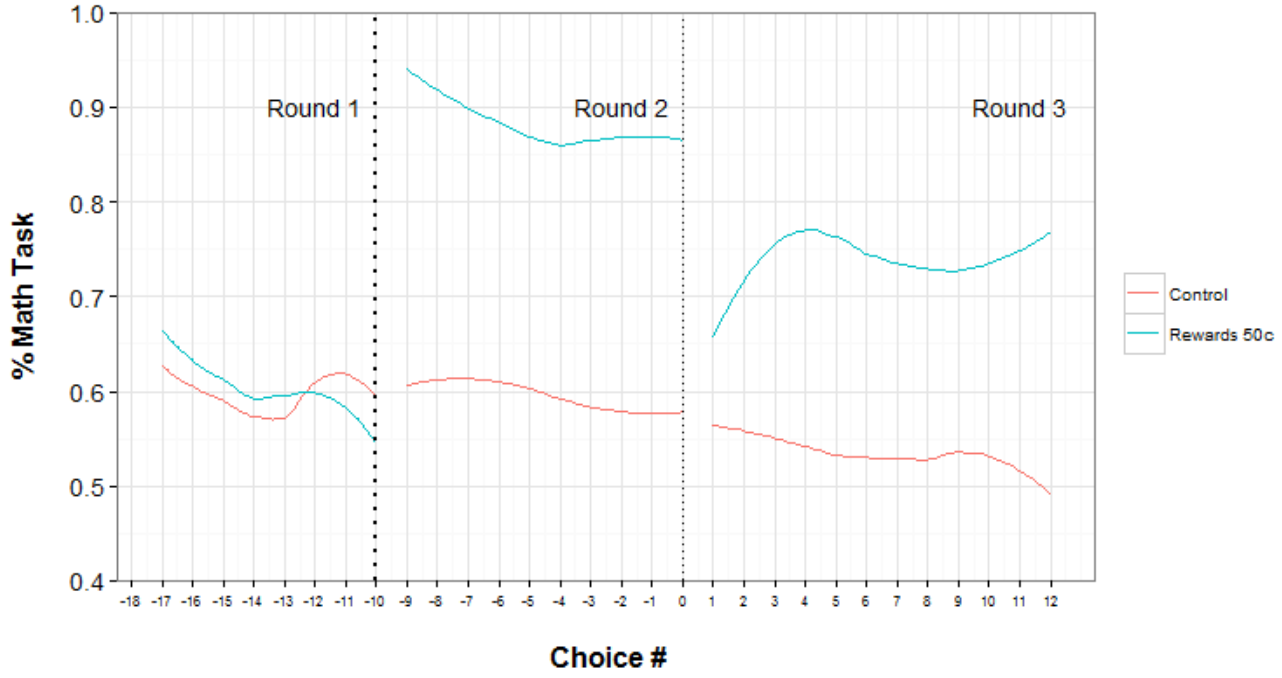
Study 2: Providing a Break Eliminates Engagement Reduction



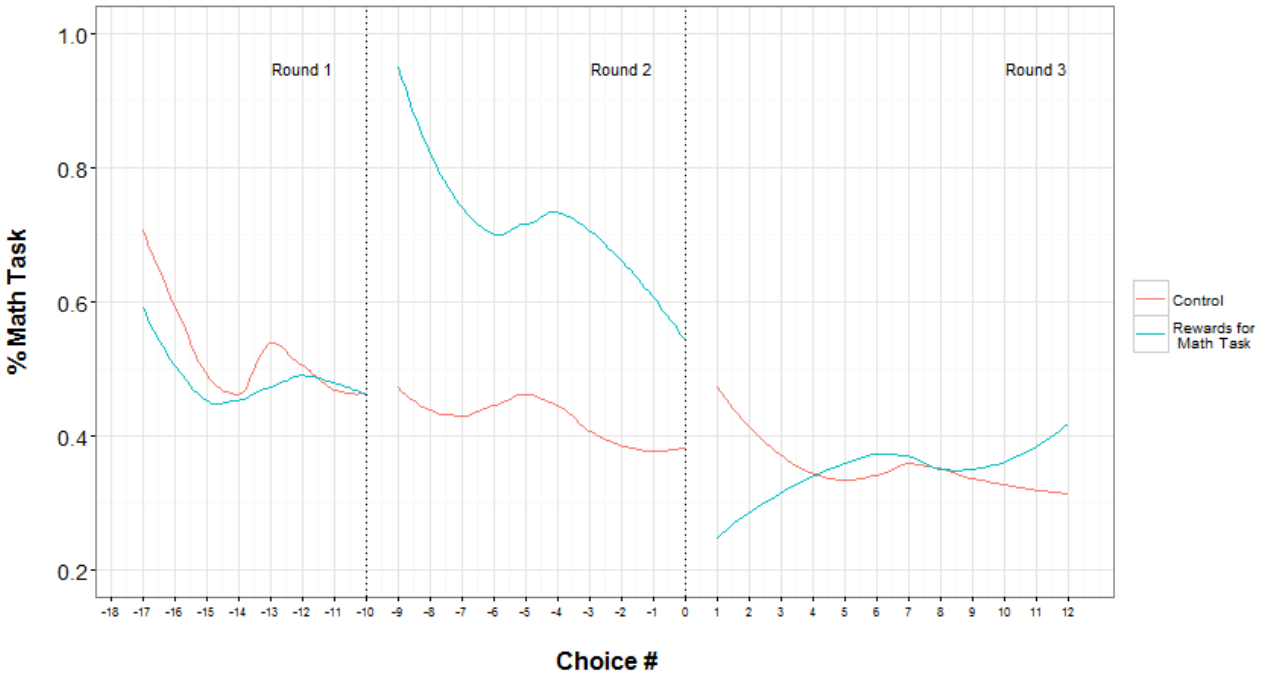


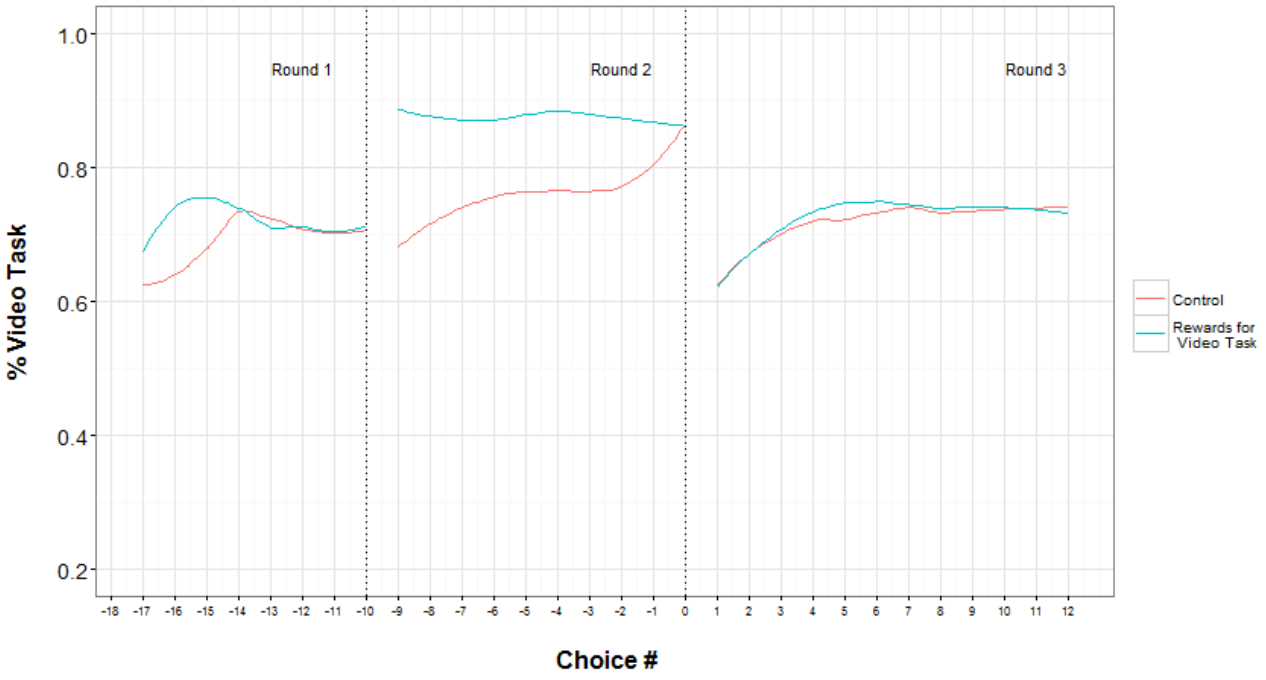
Study 3: Large Rewards Do Not Reduce Engagement





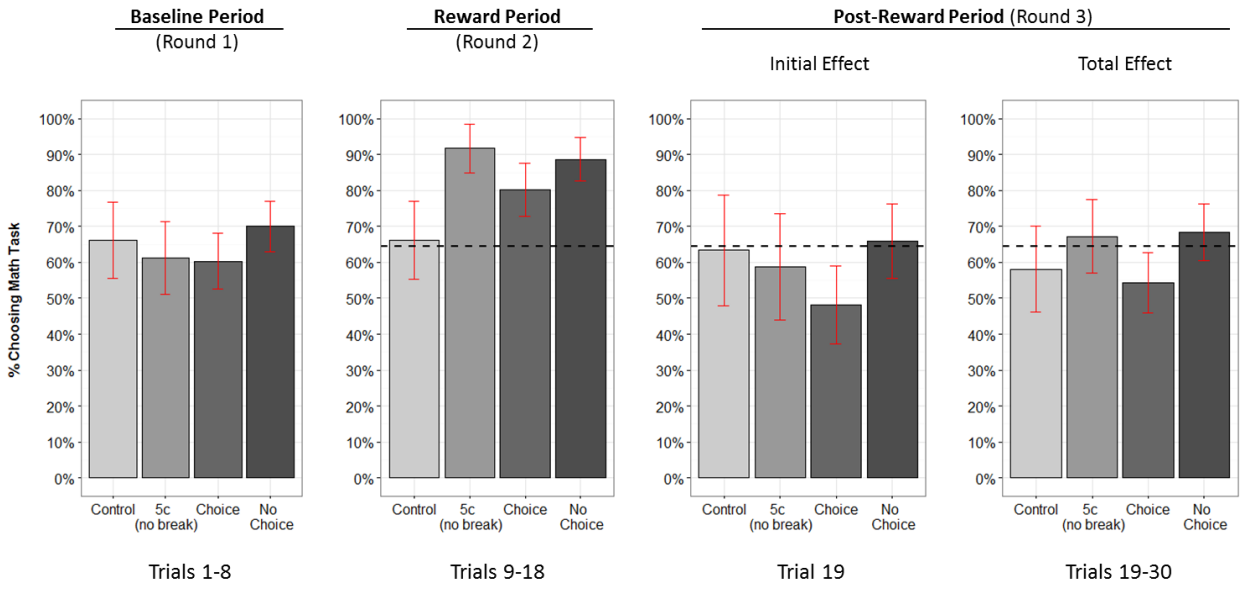
Study 4: Paying for a Leisure Task Does Not Reduce Engagement



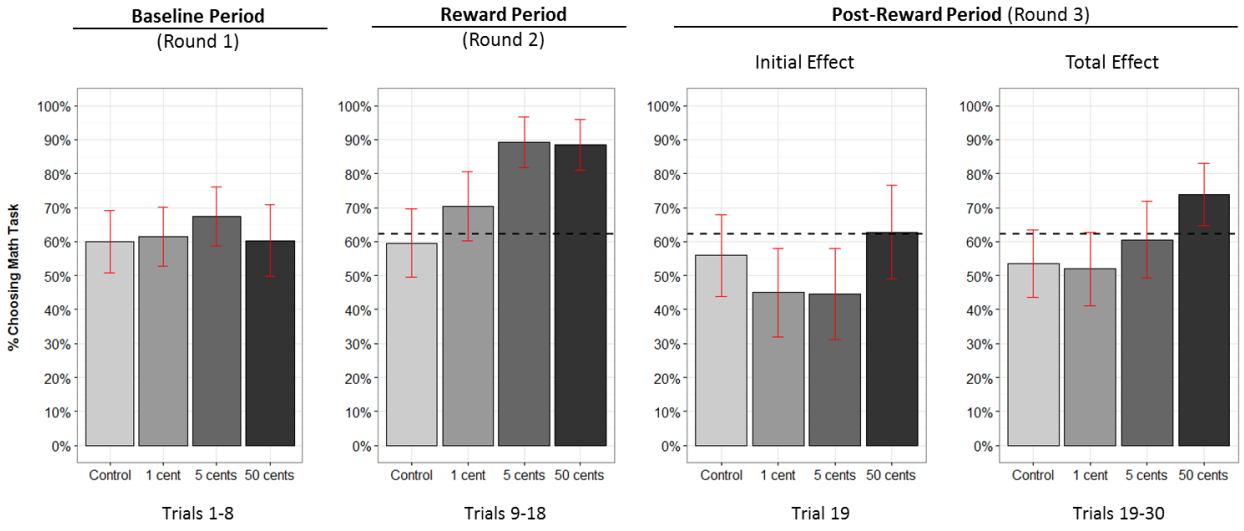


Average Effort by Rounds for All Rounds with 95% CI

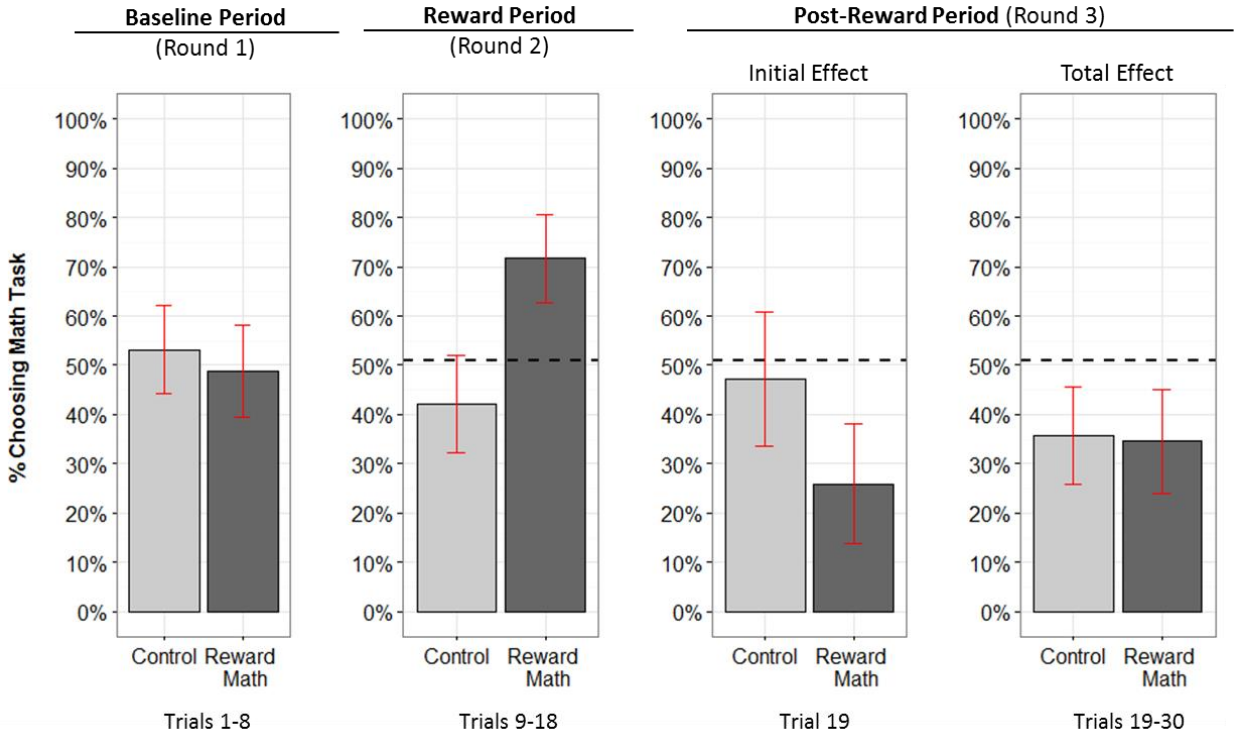
Study 2: Providing a Break Eliminates Engagement Reduction

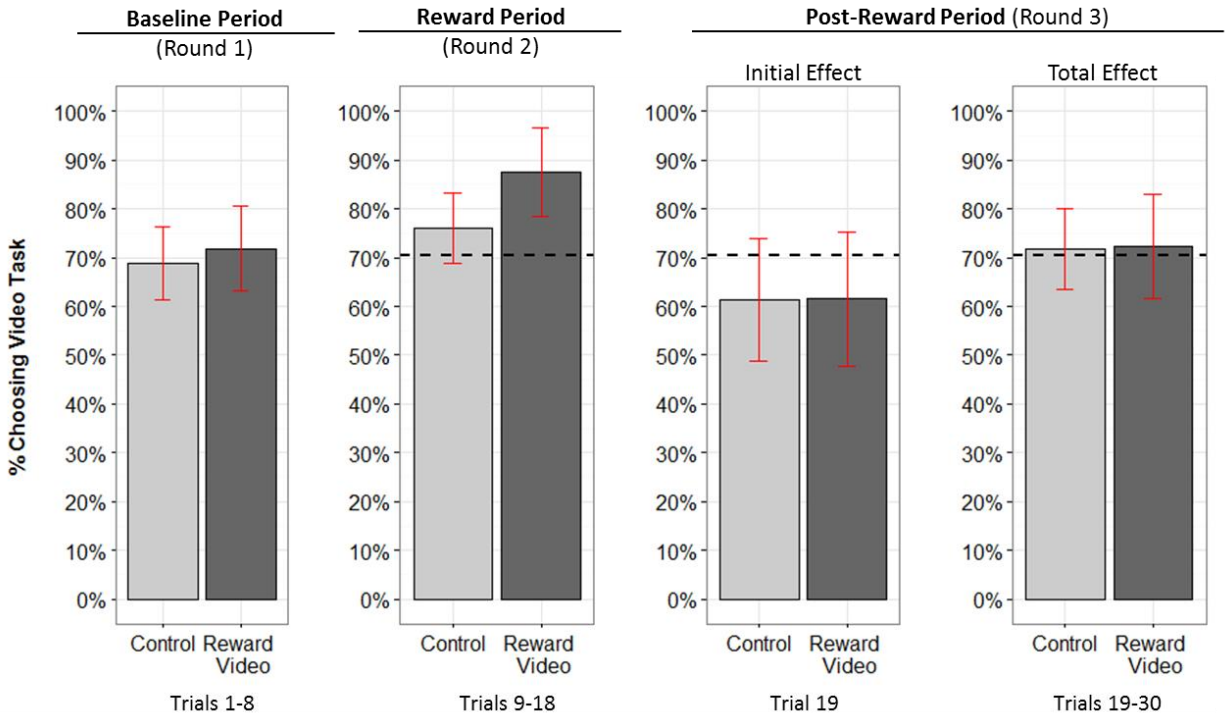


Study 3: Large Rewards Do Not Reduce Engagement



Study 4: Paying for a Leisure Task Does Not Reduce Engagement





APPENDIX L: STUDY WITH PAYING FOR A LEISURE TASK

In the main paper, we reported a study where we varied whether the target task was a cognitive math task or a leisure task like watching and rating videos. In that study we used a different math task that entailed counting that number of 1s in a grid of 150 1s and zeros. Here we report a replication of the same experiment with the original math task that was used in Studies 1-3 and Study 5 of the main paper.

Method

Adult participants were recruited from Amazon MTurk. A target of 480 participants were requested, yielding 477 surveys. Unusable cases (duplicate IP addresses, technical problems, failed attention check) were removed prior to analysis, yielding 340 valid completes.[11] Participants who completed Round 1 but then dropped-out (4.9%) were coded as not doing the focal task and included in the analysis.

Participants were randomly assigned to one of four conditions in a 2 (Target task: Math vs. Video) x 2 (Control vs. Incentive) between-subjects design. In the two incentive conditions, they were either paid 5 cents for correctly completing math tasks, as in the prior studies, or paid 5 cents for each video they watched and rated (1- 5 stars). The two control conditions matched the two incentive conditions, highlighting the target task without any incentive.[12]

Results

We have two different control conditions in this study, so we compare each incentive condition to the corresponding control condition. We replicated the momentary reduction in engagement when people were paid for doing the math task. Fewer people chose the math task in the first trial of Round 3 in the incentive condition after the rewards had ended, compared to in the matching control condition(47% vs. 68%; β = -0.18, t= -2.77, p=.006). There was no long-term reduction of effort due to incentives, relative to control (62% in both). These results were further confirmed in the hierarchical regressions ($\beta_{MOMENTARY} = -3.20, z = -4.55, p < .001$; $\beta_{POST} = +0.06, z = +0.14, p > .250$).

When participants were paid for the leisure task instead, we did not find a reduction in engagement. The proportion of video choices in the first trial after the incentive ended was similar to the corresponding control condition (50% vs. 47% , $\beta = 0.01, t < 1$) and we found no long-term reduction of effort (57% vs. 56%; β=0.0003, t <1). These results were further confirmed by the hierarchical regression models ($\beta_{MOMENTARY} = +0.11, z = +0.11, p > .250$; $\beta_{POST} = +0.10, z = +0.24, p > .250$).

[11] The high number of unusable cases was due to duplicate IPs that happened due because the software did not filter on previous respondents since a few months had passed between this study and the previous studies.

[12] Two different framings of the math and video tasks were used in each condition, but since there were no differences, the results were merged. These framing details are provided at the end of this study.

In fact, the momentary reduction in engagement observed when incentivizing the math task was completely eliminated when the videos were incentivized instead ($\beta_{MOMENTARY\ interaction} = +3.25, z = +2.93, p = .003$). There was no difference between the two conditions in terms of the longer-term post-reward baseline level($\beta_{POST\ interaction} = +0.03, z = +0.06, p > .250$).

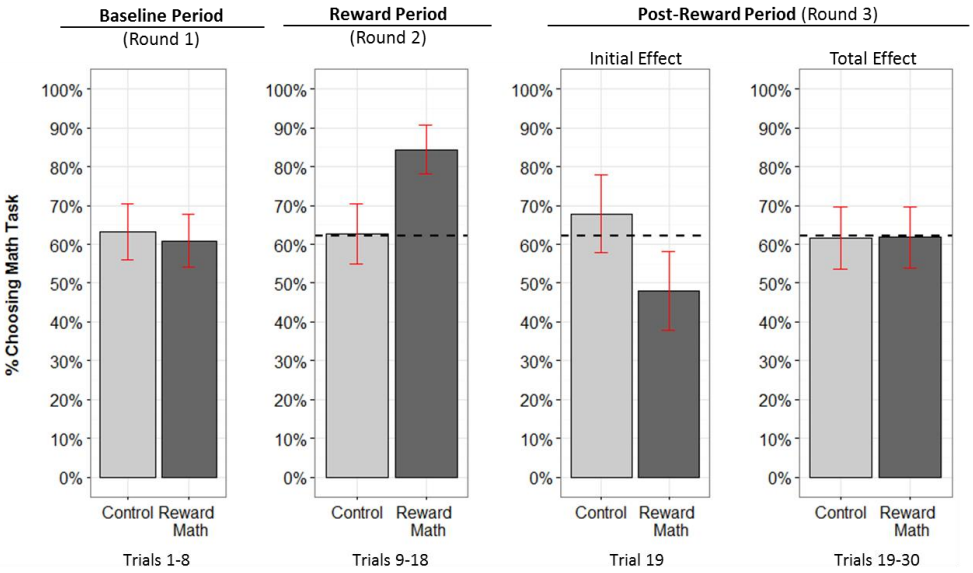


Figure L.1: Results of all rounds when the math task was incentivized. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs

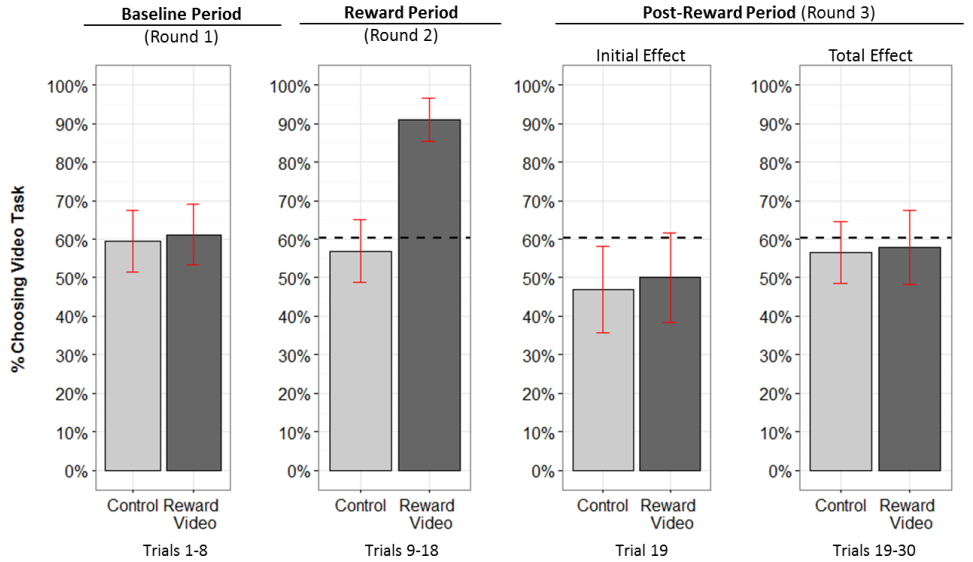


Figure L.2 Results of all rounds when the video task was incentivized. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs

Framing details used in this Study

Target Task = Math
PLEASE READ THE INSTRUCTIONS CAREFULLY. In this survey you will be asked to do a task. The task is to **solving cognitive math tasks**. We will use your results to calibrate and standardize a training task for a reasoning study. Since doing the task can be tiring, you will also have an option of a different task, **evaluating videos of television advertisements,** so that you can take a break. It is completely up to you to choose which task you want to do in each round.
PLEASE READ THE INSTRUCTIONS CAREFULLY. In this survey you will be asked to do a task. The task is to **solving cognitive math tasks**. We will use your results to calibrate and standardize a training task for a reasoning study. Since doing the task can be tiring, you will also have an option of a different task, **evaluating videos of television advertisements,** so that you can take a break. Although the task can feel tedious, you may **learn something useful** from solving the math problems, and **train your mental skills.** It is completely up to you to choose which task you want to do in each round.
Target Task = Video
PLEASE READ THE INSTRUCTIONS CAREFULLY. In this survey you will be asked to do a task. The task is to **evaluate videos of television advertisements**. We will use your responses to design experimental stimuli for an attention and perception task. Since doing the task can be tiring, you will also have an option of a different task, **solving cognitive math problems,** so that you can take a break. It is completely up to you to choose which task you want to do in each round.
PLEASE READ THE INSTRUCTIONS CAREFULLY. In this survey you will be asked to do a task. The task is to **evaluate videos of television advertisements**. We will use your responses to design experimental stimuli for an attention and perception task. Since doing the task can be tiring, you will also have an option of a different task, **solving cognitive math problems,** so that you can take a break. Although the task can feel tedious, you may **learn something useful** from evaluating the videos, and **train your mental skills.** It is completely up to you to choose which task you want to do in each round.

APPENDIX M: STUDY WITH FRAMING BOTH CHOICE OPTIONS AS IMPORTANT

In our studies reported in the paper, the math task was framed as important and potentially beneficial, in part to highlight the self-control tradeoff between goals with more immediate and delayed benefits. This raises the possibility, however, that the framing made participants feel obligated to work on the math task, rather than watch the videos, even after the reward ended. Could this have resulted in a short-lived reduction in engagement? In this study we investigate this potential concern.

Method

Adult participants were recruited from Amazon MTurk to complete an online survey. A target of 300 participants were requested, yielding 291 surveys. Records with duplicate IP addresses, or who reported having technical problems with viewing the videos or working on the math task, or who failed the basic attention check were removed prior to analysis, yielding 219 valid completes. The proportion of participants in this sample who reached until the end of Round 1, but then dropped-out part way through was 2.3%. Participants were randomly assigned to one of four conditions, in a 2 (Control, Reward) x 2 (Math Important, Both Important) between-subjects design. The two replication conditions (Math-Important control and reward) were similar to Study 1. The other two conditions (Both-Important control and reward) were the same, except that participants were told that their data, both from doing math and from the video task was important in the study. Participants were also told that, since the survey was being administered to many people, it was completely up to them to choose what they wanted to do. This framing was designed to remove any signal to the participants that were expected to do the math tasks, and to encourage participants to choose what they truly wanted in do in each round. As a result, if participants' sense of obligation to do the math task had arrested the post-reward reduction in engagement in our studies, we would observe a stronger decrease in engagement in the Both-Important condition.

Results

A manipulation check, collected at the end of the study, confirmed that participants in the Both-Important control condition expressed more agreement that the videos and math task were equally important (on a 9 point scale) than in the Math-Important control condition ($M_{control,both} = 4.43,\ SD = \ 1.61\ vs. M_{control,math} = 3.79, SD = \ 1.75;\ t(110) = 1.98, p = .05$).

Since this study included two differently-framed control conditions, we compared each reward condition to the corresponding control condition. We replicated the momentary reduction in engagement behavior when using the same instructions, in the Math-Important incentive and control conditions. Fewer people chose the math task in the first trial of Round 3 in the reward condition after the incentives had ended, compared with the same trial in the control condition

(38% vs. 64%; β= -0.18, *t*= -2.13, *p*=.04; Figure J.1). There was no longer-term reduction in engagement due to incentives, relative to control (57% vs. 63%; β= - 0.03, *t* <1). These results were further confirmed in the hierarchical regression models ($\beta_{MOMENTARY} = -4.69, z = -4.00,\ p < .001;\ \beta_{POST} = -0.04, z = -0.07, p > .250$).

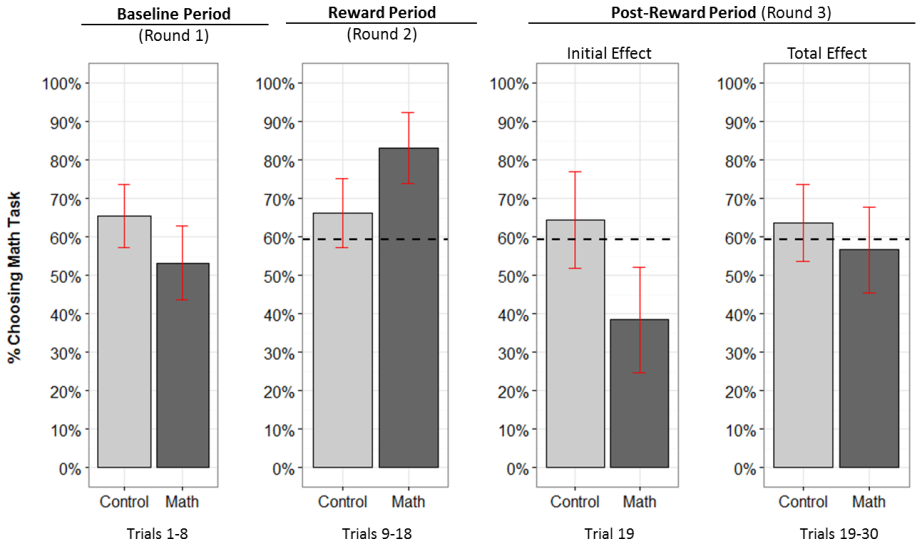


Figure M.8: Results of all rounds in the replication condition when math was the focal task. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs

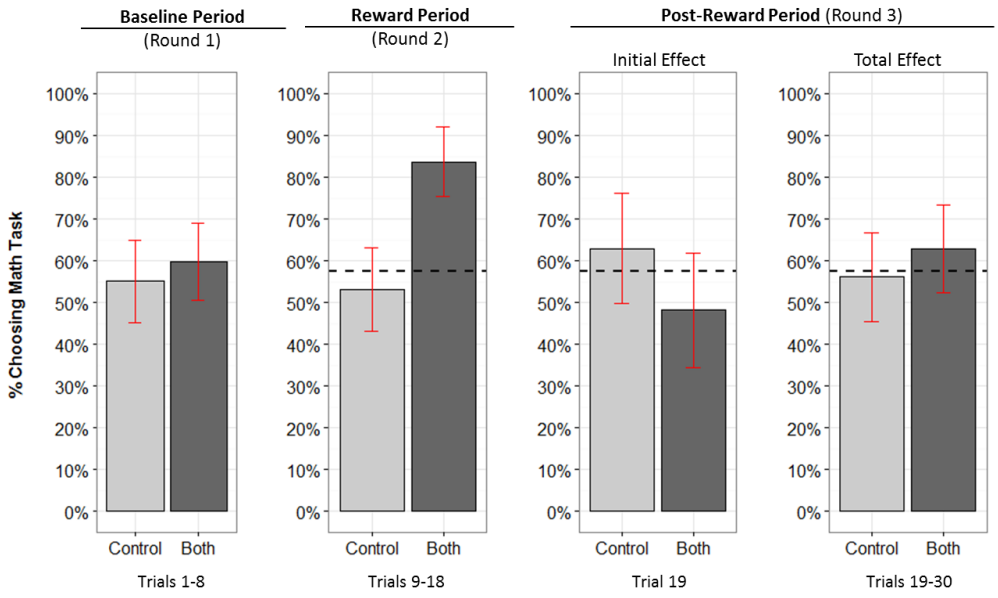


Figure M.9: Results of all rounds when both tasks were framed as equally important to the experimenter. Dotted lines represent the baseline (average effort level Round 1), and the vertical lines are 95% CIs

Likewise, when we instead told participants that both the task options (math and video) are equally important, we again replicate the findings in the Both-Importance incentive and control conditions. Fewer people chose the math task in the first trial of Round 3 in the incentive condition after the incentives had ended, compared to in the control condition (48% vs. 63%; β = -0.18, *t*=-2.42, *p*=.02). There was no longer-term reduction in engagement due to incentives, relative to control (63% vs. 57%; β=0.03, *t* <1). These results were further confirmed in the hierarchical regression ($\beta_{MOMENTARY} = -2.52, z = -3.33, p < .001$; $\beta_{POST} = +0.25, z = +0.51, p > .250$).

A hierarchical regression model confirmed that there was no difference in the extent of momentary reduction in engagement between the two task-framing conditions ($\beta_{MOMENTARY:\, MATH\ VS.\ BOTH} = +1.47, z = +1.08, p > .250$). Likewise, there was no difference in the extent of overall post-reward reduction between the two conditions ($\beta_{POST:\, MATH\ VS.\ BOTH} = +0.27, z = +0.38, p > .250$).

The results of this study suggest that the momentary nature of post-reward reduction in engagement cannot be explained by the experimental instructions inducing a feeling of obligation to do math tasks among the participants. Furthermore, the findings are also inconsistent with a self-signaling account, in which the participants continued with the more challenging math tasks (after a short break) to feel good about themselves.

APPENDIX N: ALTERNATIVE PARAMETERIZATION TO ESTIMATE INITIAL REDUCTION IN ENGAGEMENT IN THE POST-REWARD PERIOD

The following estimates are of regression models that use the same specification as Equation 5 in Appendix D, with $MOMENTARY_t$ replaced by $(MOMENTARY_t - 1)$. $MOMENTARY_t$ is modeled in all equations as $\frac{1}{t}$ where t is the post-reward trial number. In this specification, the estimate of $Condition = Reward$ indicates the initial difference between the reward versus the control group at the start of the post-reward period relative to the long-run post-reward baseline level.

Study 1: Reward vs Control

	β	SE	z	p	
Constant	-3.77	1.01	-3.73	<.001	***
$(MOMENTARY_t - 1)$	1.06	0.75	1.40	.161	
Condition = Reward	-1.97	0.91	-2.18	.029	*
Total Attempts in Round 1	1.09	0.17	6.58	<.001	***
$(MOMENTARY_t - 1)$ * Reward	-2.59	1.00	-2.59	.010	**

*p<.05; **p<.01; ***p<.001

Internal Meta-analysis: Reward vs Control

	β	SE	z	p	
Constant	-3.54	0.33	-10.62	<.001	***
$MOMENTARY_t - 1$	1.30	0.25	5.23	<.001	***
Condition = Reward	-1.76	0.33	-5.34	<.001	***
Total Attempts in Round 1	1.05	0.05	21.87	<.001	***
$MOMENTARY_t - 1$ * Reward	-3.03	0.33	-9.14	<.001	***

*p<.05; **p<.01; ***p<.001

Study 2: Replication (5c) and High-effort Break vs. Low-effort Break

	β	SE	z	p	
Constant	-3.52	0.66	-5.34	<.001	***
$MOMENTARY_t - 1$	0.00	0.52	0.00	.990	
Condition = Reward	-1.20	0.59	-2.02	.043	*
Total Attempts in Round 1	0.95	0.09	10.05	<.001	***
$MOMENTARY_t - 1$ * Reward	-1.24	0.63	-1.97	.049	*

*p<.05; **p<.01; ***p<.001

Study 2: Replication (5c) vs. Control

	β	SE	z	p	
Constant	-3.18	0.94	-3.38	.001	***
$MOMENTARY_t - 1$	0.99	0.77	1.29	.198	
Condition = Reward	-0.59	0.92	-0.64	.523	
Total Attempts in Round 1	0.85	0.13	6.70	<.001	***
$MOMENTARY_t - 1$ * Reward	-2.49	0.98	-2.55	.011	*

*p<.05; **p<.01; ***p<.001

Study 2: High-effort Break vs. Control

	β	SE	z	p	
Constant	-3.34	1.01	-3.30	.001	***
$MOMENTARY_t - 1$	1.23	0.87	1.42	.155	
Condition = Reward	-1.38	0.99	-1.39	.163	
Total Attempts in Round 1	0.93	0.12	7.63	<.001	***
$MOMENTARY_t - 1$ * Reward	-2.15	1.03	-2.10	.036	*

*p<.05; **p<.01; ***p<.001

Study 2: Low-effort Break vs. Control

	β	SE	z	p	
Constant	-4.77	0.98	-4.87	<.001	***
$MOMENTARY_t - 1$	1.06	0.93	1.14	.254	
Condition = Reward	0.15	0.87	0.18	.859	
Total Attempts in Round 1	1.17	0.15	7.77	<.001	***
$MOMENTARY_t - 1$ * Reward	-1.14	1.09	-1.04	.297	

*p<.05; **p<.01; ***p<.001

Study 2: High-effort Break vs. Low-effort Break

	β	SE	z	p	
Constant	-3.78	0.74	-5.14	<.001	***
$MOMENTARY_t - 1$	0.00	0.56	-0.01	.996	
Condition = Reward	-1.45	0.68	-2.13	.033	*
Total Attempts in Round 1	1.01	0.11	8.87	<.001	***
$MOMENTARY_t - 1$ * Reward	-1.04	0.75	-1.39	.165	

*p<.05; **p<.01; ***p<.001

Study 3: Replication (5c) vs. Control

	β	SE	z	p	
Constant	-5.61	1.13	-4.98	<.001	***
$MOMENTARY_{t\ -1}$	0.72	0.74	0.97	.332	
Condition = Reward	-2.33	1.14	-2.04	.042	*
Total Attempts in Round 1	1.35	0.19	6.97	<.001	***
$MOMENTARY_{t\ -1}$ * Reward	-3.52	1.15	-3.07	.002	**

*p<.05; **p<.01; ***p<.001

Study 3: Low Reward (1c) vs. Control

	β	SE	z	p	
Constant	-5.10	1.09	-4.67	<.001	***
$MOMENTARY_{t\ -1}$	0.76	0.64	1.19	.232	
Condition = Reward	-1.71	1.11	-1.54	.123	
Total Attempts in Round 1	1.25	0.18	6.84	<.001	***
$MOMENTARY_{t\ -1}$ * Reward	-1.94	0.95	-2.04	.042	*

*p<.05; **p<.01; ***p<.001

Study 3: High Reward (50c) vs. Control

	β	SE	z	p	
Constant	-4.01	0.96	-4.17	<.001	***
$MOMENTARY_{t\ -1}$	0.67	0.76	0.87	.382	
Condition = Reward	1.31	1.07	1.22	.222	
Total Attempts in Round 1	1.00	0.15	6.76	<.001	***
$MOMENTARY_{t\ -1}$ * Reward	-1.48	1.19	-1.24	.215	

*p<.05; **p<.01; ***p<.001

Study 4: Replication (5c) vs. Control

	β	SE	z	p	
Constant	-4.51	0.92	-4.89	<.001	***
$MOMENTARY_t - 1$	1.52	0.69	2.20	0.028	*
Condition = Reward	-3.21	1.11	-2.89	0.004	**
Total Attempts in Round 1	0.97	0.13	7.38	<.001	***
$MOMENTARY_t - 1$ * Reward	-4.15	1.10	-3.77	<.001	***

*p<.05; **p<.01; ***p<.001

Study 4: Incentives for Video vs. Control

	β	SE	z	p	
Constant	-4.96	1.02	-4.88	<.001	***
$MOMENTARY_t - 1$	-1.42	0.60	-2.36	0.018	*
Condition = Reward	-0.47	0.87	-0.54	0.589	
Total Attempts in Round 1	1.14	0.17	6.59	<.001	***
$MOMENTARY_t - 1$ * Reward	-0.52	0.90	-0.57	0.568	

*p<.05; **p<.01; ***p<.001

APPENDIX O: RESULTS OF FIELD STUDIES THAT HAVE MEASURED BEHAVIOR AFTER CONTINGENT INCENTIVES ENDED

	Domain	Target Group	Incentive size / type	Time-point(s) when post-reward behavior was measured	Finding(s)
Garbarino & Slonim, 2005	Education	University students at a private school	$10 for passing a test	Immediate post-reward behavior (e.g, the subsequent test)	On Average, 2 fewer questions attempted in incentive group vs control ($p<.05$)
Volpp et al., 2006	Smoking Cessation	Smokers at Philadelphia Veteran Affairs Medical Center	$100 for quitting to smoke	6 months after incentive to quit	Quit rates in incentive group (6.5%) not different from control (4.6%, $p>.2$)
Jackson, 2007	Evaluation of Advanced Placement Incentive Program (APIP) in Education	11^{th} and 12^{th} grade students (and teachers) in Texas schools serving underprivileged populations	Between $100 and $500 for getting a score of 3 or more in each eligible test subject	Future test scores and college graduation	13% increase in number of students scoring about 1100/24 on SAT/ACT ($p<.05$) and 5% increase in students matriculating in college ($p<.10$)
Volpp et al., 2008	Warfarin Adherence	Warfarin patients at the UPenn Anticoagulation Management Center	Lottery with daily expected value of $5(Study 1) or $3(Study 2)	Not-reported	Regulation of anticoagulation levels changed from 35% (pre) to 42% (post, w/S, Study 1; ns) and from 65% (pre) to 60% (post, w/S, Study 2; ns)
Volpp et al., 2008	Weight Loss	Healthy adults age 30-70 with a BMI of 30-40 from the Philadelphia VA Medical Center	Lottery incentive (expected value $3/day), or deposit contract with matching incentives (max $8.4/day)	7 months after end of intervention	Both in Lottery ($\Delta \approx$ -9 lbs., $p=.01$) and in deposit contract ($\Delta \approx$ -6 lbs., $p=.03$) participants weighed less than the beginning of the study
Angrist et al., 2009	Education	Entering first-year undergraduates at a primarily commuter school	Merit scholarship or merit scholarship with support service	1 year after end of intervention	0.28 percentage points ($p<.01$) increase in grade points for women; longer-term effect on men non-significant
Cawley & Price, 2009	Weight Loss	Employees from employer that has contract with 'Company X'	Various quarterly monetary rewards or lottery plus refundable bonds	1 year after end of intervention	No significant difference with quarterly rewards w.r.t baseline. For lottery + bonds 3.6 lbs. ($p<.05$) loss w.r.t baseline

Charness & Gneezy, 2009	Gym Attendance	University of Chicago undergrad students	Low Reward: $25 to attend gym once in week. High Reward: $100 to attend gym 8 times in 4 weeks	One attendance measure per week, 7 weeks after intervention	Higher post-intervention gym attendance in high-reward vs control (0.67 visits/week) and vs. low-reward group (0.58 visits/week)
Acland & Levy, 2010	Gym Attendance	Self-reported non-regular gym attenders	$25 to attend gym once in week and then $100 to attend gym 2 times every week for 4 weeks	One attendance measure per week, 5 weeks post-treatment and following weeks into next semester	Higher post-intervention gym attendance in reward vs control (0.256 visits/week)
John et. al, 2010	Weight Loss	Patients at the Philadelphia Veterans Affairs Medical Center with BMIs of 30–40, age 30-70	Deposit contracts in which participants put $0-3 daily of their own money at risk (matched 1:1)	Weigh in 36 weeks after end of intervention	No longer-term difference in weight loss between treatment (1.2 Lbs.) and control (0.27 Lbs; p=.76)
Kimmel et. al, 2012	Warfarin Adherence	Warfarin patients at the UPenn Anticoagulation Management Center	Lottery with daily expected value of $3	6-months after end of intervention	No difference on anticoagulation levels between reward (23%) and control (25.9%; ns)
Royer et. al, 2012	Gym Attendance	Employees from Midwest Fortune 500 company	$10 for visiting company gym (up to 3x per wk.) over 4 wks., free membership, and $20 for new members; w/ or w/o self-funded commitment contracts	Gym use via login records 5-13 weeks and 14-52 weeks after end of incentives	Significantly higher post-intervention gym attendance in incentives vs control (0.11 visits/week; p <.05) in weeks 5-13 after incentives end. The results are directionally positive but *ns* in weeks 14-52 after incentives end
Bareket-Bojmel et. al 2014	Work productivity	Technicians at a global high-tech semiconductor company working at a fabrication plant in Israel.	$25, family pizza meal voucher, verbal reward, or own choice if performance level exceeded base productivity.	Productivity on the first day, second day, and third day after rewards were stopped.	Removal of the cash bonus significantly reduced performance by 13.2% relative to base productivity on day1. However, with verbal praise productivity was 4.2% higher than baseline on day 1.
Sen et. al, 2014	Adherence to medical regimen among diabetes patients	Patients of a Primary Care Medical Home at UPenn	Lottery incentive with expected daily value of $2.80 or $1.40 for daily monitoring	Every month for three months after end of incentives	After three months, adherence rate was 62 % in low, 35 % in high (p=.015) and 27 % in control (p=.002vs. low incentives).
Halpern et. al, 2015	Smoking Cessation	CVS Caremark employees	$800 or refundable deposit of $150	6 months and 12 months after end	Compared to the abstinence rate in control (6.0%), abstinence with individual

			with an opportunity to with $650 in rewards	of incentives	rewards was significantly higher after 6 months (≈15%, p<.01) and after 12 months (≈7.5%, p<.05).
Huffman & Bognanno, 2015	Work Productivity	Workers hired to register people for a company database during street festival	Hourly wage ($18) with a per sign-up monetary bonus ($5/sign-up) for 1 hour	Every hour for three hours after incentive ended	Immediately after incentives ended Reward group recruited more than Control (7% higher)
Mochan et. al, 2015	Purchase of healthy grocery items	Households participating in Points-based Healthy Food program	Forfeiting Heathy Food discount for failing to increase healthy food items purchased by 5% for the month.	Supermarket shopping data in following 6 months after intervention ended	0.49 percentage-points (p<.1) average increase in healthy items purchased in the first three months, and 0.79 percentage-points (p<.01) average increase in healthy items purchased in the next three months
Wang et. al, 2016	Number of hotel nights	Loyalty program members at a major international hotel chain	Bonus points during an 8-month period	Hotel stays in 8 months after the intervention	On Average, compared to the control group, the treatment group stayed one-night more in the post-reward period (p<.01)

Note: Only studies that (a) have measured post-reward behavior (and not self-report), (b) have studied adults (including a Study with 11th and 12th graders), and (c) where the target task is not a pro-social activity (e.g., contribution to charity, blood donation etc.) are included.