Do Monetary Incentives Undermine Performance on Intrinsically Enjoyable Tasks? A Field Test*

Constança Esteves-Sorenson† Robert Broce‡

June 2020

Abstract

Economists have long been intrigued by an influential literature in psychology positing that monetary pay lowers performance on enjoyable tasks by crowding out agents' intrinsic interest in them. But typical experiments in this literature do not report a full set of performance metrics, which might reveal conflicting evidence on crowding out. Further, they may suffer from confounds. To evaluate these issues, we review over 100 prior tests and run a field experiment building on the canonical two-session test for crowding out wherein agents receive pay for an interesting activity in session one that is withdrawn unexpectedly in session two. We test whether pay harms performance using a comprehensive set of performance measures, and if so, whether unmet pay expectations might also contribute to this decline. Our results on output, productivity and quits are most consistent with a standard economics model than with a crowding out one. Additional, though more speculative, evidence suggests that unmet pay expectations may harm output quality.

JEL Codes: C93, D91, J33, M52

Keywords: incentives, intrinsic enjoyment, crowding out, field experiments, expectations

*We thank Rosario Macera for her work in the initial stages of this project. We thank Kaitlyn Croteau, Cathy Le, Tiffany Lin, Karen O'Brien and Ignacio Osorio for excellent research assistance. For their comments, we thank the editor and four anonymous referees, Jason Abaluck, Daylian Cain, Jason Dana, Florian Ederer, Florian Englmaier, Kelsey Jack, Lisa Kahn, Botond Kőszegi, Ian Larkin, Ulrike Malmendier, Stephan Meier, Pedro Rey-Biel, Frédéric Schneider, Olav Sorenson, Amy Wrzesniewski and seminar participants at UC Berkeley, Pontificia Universidad Católica de Chile, Universidad de Chile, University of Warwick, Yale University and the 36th Annual NBER Summer Institute in Personnel Economics. This research was partly funded by the Russell Sage Foundation and the Whitebox Behavioral Grants at Yale University.

†constanca.esteves-sorenson@yale.edu, Yale University.

‡brocer1@southernct.edu, Southern Connecticut State University.

1 Introduction

Boosting worker productivity is a central concern in labor and personnel economics. Although incentive pay is viewed as a key tool for achieving this goal (e.g., Gibbons, 1998; Prendergast, 1999; Lazear, 2000), an influential strand of research in psychology, pioneered by Deci (1971), argues that pay harms performance in an important case: when an agent enjoys a task, that is, engages in it due to "the enjoyment he experiences in the activity" (Deci, 1971, p.105). Incentive pay may feel "controlling" to the agent (e.g., Deci, Koestner, and Ryan, 1999) or create the misperception that the task is not enjoyable (e.g., Lepper, Greene, and Nisbett, 1973), lowering the agent's interest in it and thus harming performance.

The idea that pay harms performance on interesting tasks has received thousands of citations in psychology and, more recently, in economics. And it has become embedded in the standard educational curriculum in human resource and organizational behavior classes at business schools (e.g., Baron and Kreps, 1999; Lazear and Gibbs, 2014).[1]

Evidence for this idea has come primarily from the canonical two-period test in psychology for the crowding out of enjoyment. Subjects engage in an enjoyable task, such as completing interesting puzzles. The treatment group gets an unexpected payment in the first period (e.g., a piece rate per puzzle completed), which is withdrawn unexpectedly in the second period. The control group is not paid in either period.[2] Whereas a standard economics model predicts that the treatment group's effort in the second period should be similar to effort in the control group, the crowding out literature has documented the puzzling finding that the treatment group's outcome *drops below* that of the control group. It argues that the first-period payment undermines agents' interest in the task resulting in a *worse* second-

[1] A book with this idea (Deci and Ryan, 1985) has over 20,000 citations; the psychology meta-analyses and the articles we reviewed had over 20,000; economics' citations exceed 6,000 (source: Google Scholar).

[2] See Deci and Ryan (1985) for an extensive review of all studies using this two-period design.

period outcome than the outcome of agents who were never paid.

One issue with this literature, however, is that different tests report different outcomes and this may determine whether the evidence supports or does not support crowding out. An example is the seminal Deci (1971) paper. In his first test he found that the removal of pay decreased the amount of time subjects spent solving puzzles. But he did not report whether the treatment group solved fewer puzzles (decreased output) or solved them more slowly (decreased productivity). In his second test he found that pay slowed the rate of writing news headlines (reduced productivity) but does not report time spent on the task, the first test's measure, or the number of headlines (output). Since productivity is the ratio of output over time, could the reduced productivity in the second test result from an increase in time spent on the task that left output unchanged? If so, though the productivity evidence was consistent with crowding out, the output and time evidence would not be. Further, the evidence on time spent on the task would conflict with that of the first study.

The lack of joint reporting of performance metrics, in particular of output, productivity and time spent on the task, which might reveal contradictory evidence on crowding out, is not limited to Deci (1971). It occurs in all of the over 100 tests described in 82 papers we reviewed.[3] This has contributed to conflicting meta-analytic findings on crowding out (e.g., Cameron, Banko, and Pierce, 2001) and to demands for more tests (e.g., Gibbons, 1998; Prendergast, 1999; Rebitzer and Taylor, 2011).[4]

Further, crowding out, if detected, could be confounded with other phenomena. The

[3]We discuss these studies in Section 2.2 and in Appendix E.

[4]For example, the psychology meta-analysis of Deci, Koestner, and Ryan (1999) concluded that the bulk of the evidence supported crowding out whereas the one by Cameron, Banko, and Pierce (2001) determined that it did not. Economists have demanded more evidence. Prendergast (1999, p. 18) claimed that "while this idea holds some intuitive appeal, it should be noted that there is little conclusive empirical evidence (particularly in workplace settings) of these influences." Gibbons (1998, p. 130) argued that "field experiments [...] would be especially useful"; Rebitzer and Taylor (2011, p. 765): "Although [...] the evidence is not yet conclusive, we are intrigued by the notion that extrinsic rewards can undermine intrinsic motives." For a review see, for example, Kamenica (2012).

surprise pay in the first period for the treatment group, for example, may create the expectation of another payment in the second period. If this expectation is unfulfilled, effort may decline as agents retaliate or lose morale (e.g., Bewley, 1999).

We add to this research in two ways. We first review over 100 tests in the psychology literature and document features including the tasks performed, the sample sizes per condition and the outcomes reported. We find that no test jointly reported output, productivity and time spent on the task, interconnected measures important for assessing crowding out. The same issue arose in the few related studies in economics. Further, the median sample size by condition in psychology tests was 15 subjects.

Second, we implement a field experiment with two goals: (1) to test whether monetary incentives harm performance on an enjoyable task, by studying and reporting several measures of economic interest, such as output, productivity, and quits; and (2) to investigate the extent to which unmet pay expectations might contribute to any observed decline in performance. We find that results on output, productivity, and quits are most consistent with a standard model: pay boosts performance and its withdrawal does not lead to the pathological underperformance relative to the control group. Additional, though more speculative, results on output quality suggest that unmet pay expectations may worsen performance. This evidence could also be consistent with crowding out, but under stronger assumptions.

The field experiment replicates the canonical two-period test using a simple market research task. Students on two campuses volunteered to blind-taste and rate cookies alone in a room for an anonymous principal, for two sessions, each one week apart, for no monetary pay. They were only offered more cookies as a thank-you for participating.

We chose blind tasting as an activity for three reasons. First, it is a common task. Food manufactures and market research firms often solicit such blind tastings as a means to

improve their products.

Second, it is a task that mainly benefits a principal and an agent, and not a third party, as is typical in employment. It addition, it offers key performance metrics, such as output (number of cookies tasted and evaluated); productivity (minutes per cookie tasted and evaluated); and quits (percentage of tasters who quit after being paid, leveraging our two-session design). It also yields time spent on the task, the most used metric in this literature. We report it, though we view it as a more fragile indicator of performance if not paired with output, for example: if pay reduces agents' time on the task but does not reduce output (thus boosting productivity), then it is unclear that pay harms performance.

Third, and most importantly, tasting is an enjoyable activity that attracts recruits in the absence of monetary compensation. This is because we relied on self-selection into the activity, in the absence of money pay, to reveal liking for it. Ensuring that subjects like at least one component of the task—the one targeted for payment—is important since proponents of crowding out have argued that this is a necessary condition for pay to erode performance (e.g., Deci, Koestner, and Ryan, 1999). For example, subjects may enjoy completing a puzzle (first component) but they may like or dislike searching for pieces (second component). But it is key that they at least enjoy the first component—completing puzzles—so that paying a piece rate per puzzle completed reduces performance. Blind-tasting has two main components: tasting a cookie (e.g., inspecting it and taking a bite) and evaluating it (rating it on several dimensions). Selection revealed whether tasters at least enjoyed tasting: if they disliked tasting cookies (besides disliking evaluating), it is unlikely they would volunteer to taste and evaluate them for no pay.

We randomly assigned 91 subjects who self-selected into the blind tasting to three groups. Those in Control blind-tasted in both sessions and got only thank-you cookies at the end,

as promised. Those in UNANTICIPATED, our main treatment, were surprised with a $0.75 piece rate per cookie rated in the first session (but got no information about a second-session payment). At the start of the second session they were informed that the piece rate would not be granted. These first two conditions therefore mirror those in the canonical design.

We added another treatment, ANTICIPATED, to explore the extent to which an environment with no surprises would yield deficits in performance similar to crowding out. *After recruiting*, but one week before the first session, tasters were informed that they would receive a $0.75 piece rate in the first session but not in the second.

Further, as the increased effort in the first period could lead to fatigue or satiation in the second period, a potential confound with crowding out in most prior tests, we improved on the canonical design by separating the two sessions by one week. We further staved off satiation by emphasizing that a subject did not have to eat a whole cookie but merely sample it (e.g., take a bite), a recommendation subjects generally followed.

Our results are most consistent with a standard model. The piece rate in the first session boosted output by at least 62% and productivity by at least 27% in both treatments, consistent with a standard model. After the withdrawal of pay, output in both treatments fell to a level similar to (though slightly higher) than CONTROL's. These results hold even after accounting for quits. Productivity for ANTICIPATED was also similar to CONTROL's. Also consistent with a standard model, quits in UNANTICIPATED were low at 4%, congruent with subjects' expecting pay in session two, but higher in ANTICIPATED, at 22%, in line with subjects' *not* expecting pay in session two.

Only one result is inconsistent with a standard model: productivity in UNANTICIPATED *exceeded* that in CONTROL in the second session. But this result is also inconsistent with crowding out: for example, Deci (1971) argues that crowding out led to *lower* productivity

than that of the control group after the unexpected removal of the piece rate. But since he did not report time spent on the task or output, it is unclear what caused the lower productivity: could subjects have spent more time on the task while producing the same output (both incongruent with crowding out)?

In our case, the puzzling excess productivity was due to UNANTICIPATED subjects' producing slightly higher output than CONTROL's (inconsistent with crowding out), but spending slightly less time on the task (qualitatively consistent with crowding out, though statistically insignificant). We conjectured that this result could be due to subjects, displeased by the withdrawal of pay, "tasting their cookies and departing as quickly as possible", leaving in their wake sloppier ratings. Our more speculative evidence on evaluation quality, measured by ratings' randomness, suggests this might have occurred: these subjects rated cookies more at random than those in CONTROL. Thus, although the surprise withdrawal of pay did not lead to a deficit in output or productivity it appears to have led to a deficit in quality.

This paper presents, to our knowledge, the first extensive review of studies in this literature documenting that no prior test has jointly reported output, productivity, and time spent on the task.

This paper also offers the first test in economics or in psychology using several performance metrics to explore, as transparently as possible, whether paying agents to perform an enjoyable task that primarily benefits them and a principal (as is typical in employment) undermines performance. Importantly, our test differs from others in economics, which have mainly studied whether pay dampens outcomes on tasks that mainly benefit a third party, such as donating blood or to a charity (named "prosocial" tasks). Pay may harm outcomes in these tasks (e.g., lower donations) by spoiling agents' signal of being prosocial (Bénabou and Tirole, 2006).[5]

[5]The evidence that pay undermines outcomes on prosocial tasks is also mixed. Ariely, Bracha, and

2 Evidence for Crowding Out of Intrinsic Enjoyment

To motivate our extensive literature review, the field experiment design, and performance measures, we start by describing the leading evidence in psychology for the crowding out of enjoyment.

2.1 Leading Evidence

In the most cited paper in this literature, Deci (1971), investigated whether pay dampened performance on enjoyable tasks with two tests. In the first, students solved puzzles in three consecutive sessions to fulfill a class requirement. Deci believed that students would enjoy this activity, and indeed they rated it as highly enjoyable on a 9-point scale. He randomly assigned subjects to 12-person control and treatment groups. In each session they completed puzzles in front of a monitor for about one hour. Those in the treatment group got a surprise $1 piece rate per puzzle completed in the second session (the reward session). But at the start of the third session they were told the piece rate would no longer be paid due to lack of funds (the non-reward session). Those in the control were never paid.

Deci measured enjoyment as the amount of time subjects spent solving puzzles during eight minutes in which the monitor left the room (the "free-choice window"). He found that subjects given the (unexpected) piece rate spent more time than control subjects trying to complete puzzles during this window. But after the payment withdrawal, these subjects spent less time than those in the control solving puzzles during the free-choice window (significant

Meier (2009) found that incentive pay induced contributions to a charity only when these incentives could not be publicly observed. And though Mellström and Johannesson (2008) found that pay reduced the supply of blood among women (but not among men), Lacetera, Macis, and Slonim (2012) found no blood supply reductions. Ashraf, Bandiera, and Jack (2014) also found that financial incentives did not undermine performance among volunteers for a task with a prosocial component: the sale of female condoms for HIV prevention. Chetty, Saez, and Sandór (2014) also found that incentives did not undermine prosocial behavior in the provision of referee reports. For a review of the effect of incentives in non-employment settings, such as education and contributions to public goods, see Gneezy, Meier, and Rey-Biel (2011).

at the 10% level in a one-tailed test). Deci did not report output (number of puzzles solved) or productivity (minutes spent solving each puzzle). Subjects did not receive the option of quitting after being paid. Interestingly, he also noted that pay *did not* dampen self-reported enjoyment in the activity, even after the withdrawal of pay.

In Deci's second test, a field experiment, he split eight students staffing a college newspaper into four-person treatment and control groups. The treatment group received $0.50 per headline written over three weeks. At the end of this period, they were informed that funds had been exhausted and they would no longer be paid. Meanwhile, the four subjects in the control group were never paid.

Deci assessed intrinsic enjoyment via the minutes spent per headline (productivity). The faster subjects wrote a headline, that is, the more productive they were, the more they must enjoy the activity. Deci (1971) interpreted the lower productivity in the treatment group—increase in the minutes to write a headline—following the removal of pay as evidence for crowding out (noting that the control only had two subjects due to attrition). He also interprets the increase in the quit rate in the treatment after the withdrawal of pay as additional evidence for crowding out, although it is also consistent with a standard income effect. Time spent on the task (writing headlines) and output (number of headlines) were not reported. This experiment and Boal and Cummings (1981) are the only two field tests with adults among all the tests we reviewed in the psychology literature.

Treatment group outcomes below those of the control during the non-reward period have been viewed as evidence of the perverse effect of pay on enjoyment. The two leading explanations for this phenomenon have been cognitive evaluation theory and the overjustification hypothesis. The first proposes that individuals construe rewards as unpleasant controllers of behavior undermining "intrinsic motivation, which refers to doing something because it

is inherently interesting or enjoyable" (Ryan and Deci, 2000, p. 55). The second postulates that a person paid to perform an interesting activity may "infer that his actions were basically motivated by the external contingencies [...], rather than by an intrinsic interest in the activity itself" (Lepper, Greene, and Nisbett, 1973, p. 130).[6]

Tests for crowding out based on these two theories require a reward period followed by a non-reward one. Tests for overjustification need an initial, paid, period wherein subjects can (mis)attribute their interest in the task to the reward followed by a non-reward period in which the results of the misattribution become apparent. Tests of cognitive evaluation theory also need two periods (Deci and Ryan, 1985, p.184). In the first, paid, period there is a trade-off between the displeasing nature of the reward and its incentive effect, and it is difficult to disentangle which dominates. Thus any harmful effect of the reward is only visible in the second period, when pay is removed. Hence, crowding-out tests have used a two-period, between-subjects design, in which the treatment receives an unexpected reward in period one that is withdrawn in period two and in which the control is unpaid in both.[7]

This two-period design, however, introduces some potential confounds. On the one hand, subjects may become fatigued (non-separable cost of effort) or satiated (declining marginal utility) with the activity. As the non-reward session immediately follows the reward session in most psychology studies for crowding out (Deci, Koestner, and Ryan, 1999, p. 650), fatigue or satiation could account for the findings in these studies (Cameron and Pierce, 1994).

On the other hand, unmet pay expectations may also contribute to a decline in performance. Pay in the first period may lead subjects to expect pay in the second period. When

[6]Lepper, Greene, and Nisbett (1973), the second most cited paper in this literature, uses this hypothesis to explain why nursery school children who were surprised with a prize for drawing spent less time drawing during the subsequent non-reward period than those children who were never rewarded.

[7]The only exception to the two-period design is the three-period setup in Deci (1971).

they do not receive it, they may become disappointed or angry, reducing their effort.[8,9]

Importantly, tests of crowding out require that subjects like the activity. If not, then pay will not harm performance because "there is little or no intrinsic motivation to crowd out" (Deci, Koestner, and Ryan, 1999, p. 633). To this end, researchers have used reasonable, but arbitrary, cut-offs on enjoyment scales or on the time spent on the task prior to the start of the experiment to assess enjoyability. But there is debate on whether those who rate their enjoyment as a 5 find the task enjoyable while those who rate it as a 4 do not, or if those who spend 4 minutes on the task find it interesting whereas those who spend 3 minutes do not. As a result, failures to replicate crowding out have led to discussions on whether subjects liked the activity in the first place (Deci, Koestner, and Ryan, 1999; Cameron, Banko, and Pierce, 2001).

2.2 Review of Papers on the Crowding Out of Enjoyment

As noted above, Deci (1971) reported different outcomes for each experiment. This may not clearly allow the researcher to assess whether crowding out occurred. Output, productivity and time spent are interrelated and may yield conflicting evidence. For example, pay may reduce productivity (consistent with crowding out) due to increasing time spent on the task while leaving output unchanged (both inconsistent with crowding out). Or pay may reduce time on the task (congruent with crowding out), but leave output unchanged, thus increasing

[8]Calder and Staw (1975) raised this possibility soon after Deci (1971). But the literature has generally dismissed it (e.g., Deci and Ryan, 1985) by arguing that subjects should not be upset with the removal of pay because they started the experiment not expecting to be paid.

[9]A anonymous referee wondered whether gift exchange is similar to crowding out because both gift exchange tests, such as Gneezy and List (2006), and crowding out research show an initial increase in effort that decays over time. These two phenomenon are, however, dissimilar. A crucial difference is that these gift exchange tests and the theory that underpins them do not show or predict the pathological underperformance versus the control group postulated by crowding out. Rather, they reveal that pay raises yield *weakly higher* performance than the control group. Further, in these gift exchange experiments pay is not withdrawn from subjects, in contrast to crowding-out tests. There are other differences as well. For a review of gift exchange tests and their mixed results see, for example, Esteves-Sorenson (2018).

productivity (both incongruent with crowding out).

To assess the extent to which the incomplete reporting of outcomes might be an issue with the prior literature, we conducted an extensive review of prior tests. We first analyzed over 100 tests described in 79 papers in psychology on the crowding out of enjoyment. Table 1 in Appendix E summarizes these experiments, including tasks, experimental conditions, numbers of subjects per condition, and outcomes reported.

The most notable result documented in Table 1 is that no study jointly reports the three interrelated measures. Tests for crowding out use many measures, from time spent on the task (the most common) to output, productivity and even willingness to supply more work. But the typical experiment only reports one or two performance-related outcomes, such as productivity, output, or quits, and none jointly reports output, productivity and time.[10]

The incomplete reporting of results may partially account for the conflicting evidence on crowding out. Meta-analyses of the literature in psychology have failed to arrive at similar conclusions. Deci, Koestner, and Ryan (1999) determined that the balance of the evidence supported crowding out. But Cameron, Banko, and Pierce (2001) concluded it did not. Each meta-analysis used its own set of studies and thus analyzed a different set of outcomes. They also disagreed on whether effect sizes were properly computed, among other issues.

The partial study of outcomes also raises the concern that false positives may populate the crowding-out literature. Outcome choice has been flagged as a reason for the proliferation of false positives in psychology (Simmons, Nelson, and Simonsohn, 2011; Simonsohn, Nelson, and Simmons, 2014a,b).

In the economics literature, the few tests on the potential role of crowding out in tasks that primarily benefit agents and a principal have mostly used single performance measures

[10]Studies in psychology often report other metrics that do not measure performance per se. Rather they attempt to gauge subjects' feelings for an activity, such as their self-perceived liking for or competence in the activity, to uncover the cognitive mechanisms that underpin subjects' behavior.

and have found conflicting results. For example, Gneezy and Rustichini (2000) found that students paid a fixed monetary fee for a one-time 45-minute laboratory session answering IQ questions had fewer correct answers—the single outcome measure—if paid an additional, but low, piece rate. In contrast, Hossain and Li (2014) found that paying subjects for a task framed as regular data-entry work in one session did not reduce subjects' willingness to work in a second session or output or quality in the second session. Huffman and Bognanno (2014) found that subjects given a piece rate for signing up people for a database reduced sign-ups (output, their performance measure) after the removal of pay. They viewed the decline in output as not fully consistent with crowding out.[11]

Another issue in this literature appears to be small sample sizes. Table 1 in Appendix E shows that the median number of subjects per experimental condition across more than 100 experiments was 15 (row 80) raising the issue of replicability due to low power. Recent large-scale replication tests in psychology show (i) that only 36% of effects were statistically significant and in the same direction as those in original studies and (ii) that effect sizes were generally inflated (OpenScienceCollaboration, 2015). This is partially due to publication bias: tests often lack enough power to detect true effects and thus only those studies that, by chance, have large effects are able to reach statistical significance and thus be published. As a result, subsequent replications yield no or smaller effects (e.g., Ioannidis, 2008). Crowding-out tests appear underpowered given the small samples per condition.[12]

Thus we designed a field experiment that would allow us to report several performance

[11] Gneezy and Rustichini (2000) also did not investigate other performance metrics, or explore how pay affects subsequent performance, or task enjoyability. Hossain and Li (2014) did not report time spent on the task or productivity or explore whether the task was enjoyable. For theory, in economics, of the effects of crowding out on performance on tasks that benefit mainly principals and agents, see for example, Bénabou and Tirole (2003).

[12] A replication study of experimental tests in economics found larger replication rates — statistically significant effects in the same direction of the original studies in 61% of cases—but that the effect sizes also tended to be inflated (Camerer et al. 2016).

measures of economic interest for a principal, such as output, productivity and quits, and that would also yield time spent on the task: though this is weaker indicator of performance, it is the one most used in prior tests and complements the output and productivity analyses. We also targeted a larger sample than is typical in this literature and extended the canonical design to assess the role of confounds, such as unmet pay expectations, in case we observed pay harming performance.

3 The Field Experiment

The field experiment comprised one leg on college campus A in January 2012 and three additional legs on campuses A and B in April, June, and July of 2012. We used two campuses in Connecticut to gather a larger sample and to assess whether the results held across two separate environments.

Our experiment builds on blind tasting as a task. A common activity in market research for drinks and foods, including cookies, blind tasting mainly benefits the agent (who tastes the goods) and the principal (who receives the evaluations). Thus this is not a prosocial task, like donating blood or money, where the main motivator is the benefit to others.

(1) **Recruiting subjects who enjoy the task**. The blind tasting was advertised through flyers and electronic mailing lists as a two-session activity. Interested students contacted a research assistant, who described the task as follows:

> You need to taste and evaluate cookies in two sessions, exactly one week apart. You will taste alone, filling out an evaluation form rating each cookie's flavor, aroma and other characteristics. You will not be paid for the task and can taste as many or as few different cookies as you like for up to three hours due to room availability constraints. At the end of the second session, you will receive a luxury Godiva cookie tin as a thank-you gift.

As is common in this type of market research, the principal who commissioned the tasting remained anonymous so as not to bias tasters' ratings. Tasters tasted cookies alone and were unaware that they were participating in a study on incentives.

Blind tasting involves two main components: (i) tasting (e.g., inspecting cookies and taking a bite) and (ii) evaluating (filling in a form rating each cookie on several dimensions). Tasters may like tasting but dislike evaluating. Or they might like both tasting and evaluating.[13] For the purposes of the field experiment, it was only necessary that tasters liked to taste cookies, as they would later be incentivized to do so.

We relied on self-selection into the blind tasting for no monetary pay to ascertain whether subjects enjoyed the task. We thus used a revealed-preference approach instead of relying on more arbitrary measures, such as a rating in a self-reported enjoyment scale. We explicitly offered no monetary pay during recruiting to avoid having individuals sign up for the money rather than from the enjoyment of the activity. However, because this form of market research usually offers participants a thank-you gift (e.g., a gift certificate), we offered one as well. But we chose one that would not undermine self-selection into the study: more cookies. We assumed that the more individuals liked tasting cookies the more they would enjoy receiving cookies as a thank-you. An implication of this monotonic relationship is that if agents did not like tasting cookies, then they would also not appreciate the thank-you cookies and thus would not enroll in the blind tasting because of them. Further, the cookies were perishable and hard to resell, reducing the chance that tasters disliked tasting but joined in to resell

[13]Research suggests that individuals who like a topic exert more effort in evaluations (e.g., answer more questions, write longer answers to open-ended questions) due to a "halo effect" (Groves, Presser, and Dipko, 2004; Holland and Christian, 2009).

the thank-you cookies.[14,15]

Those who selected into the blind tasting therefore revealed that they enjoyed tasting cookies enough that this intrinsic utility and that of the thank-you cookies outweighed the cost of evaluating (if they disliked evaluating) and the opportunity cost of other uses of their time (the model in Appendix F formalizes this intuition).

(2) **Dealing with fatigue, satiation and other confounds**. Tasters filled out an evaluation form after tasting each cookie, as is typical in blind tasting. They rated cookies on a scale of 1 (Excellent) through 5 (Poor) along seven major dimensions: Appearance (e.g., "Does it look chewy?"), Aroma (e.g., "Does it smell home-baked?"), Snap (e.g., "Does it break easily?"), Texture (e.g., "Is it chalky?"), Start (e.g., "Does the flavor develop quickly?"), Flavor (e.g., "Does it have a minty flavor?"), and Overall Rating ("What is the overall rating of this cookie?"). See Appendix H for the cookie evaluation sheet. At the start of session one, tasters signed a consent form ensuring, for example, that they were aware of allergens (e.g., some cookies had nuts), and they answered a short demographic questionnaire (see Appendix I for the protocol).

To allow for variability in outcomes, tasters were fairly unconstrained in how many cookies they could taste and in how long they spent tasting. They could try up to 70 cookies per session in up to three hours (due to site availability).

To minimize satiation and fatigue, two potential confounds with crowding out, we scheduled the two tasting sessions one week apart.[16] If subjects became satiated from tasting

[14]It is also possible that tasters disliked tasting cookies but still selected into the task because the thank-you cookies conveyed other benefits, such as serving as a gift to a friend. The results we show later document that this scenario seems unlikely: tasters engaged substantially with the task (e.g., those in CONTROL spent, on average, 1 hour and 19 minutes tasting cookies), even though the thank-you cookies were not contingent on time spent tasting or amount tasted.

[15]Source: There were no listings on eBay for the whole of North America, accessed on January 2012, for the resale of the type of thank-you cookies offered in our test.

[16]The separation of the two sessions in time to ward off fatigue has also been used in other recent studies

and/or fatigued with the physical or mental effort of filling out the evaluations in the first session, they had a week to recover. Further addressing potential satiation, subjects were informed both by the on-site research assistant and in writing on each evaluation sheet that they could merely partially taste each cookie (see Appendices I and H).

Further, because declining marginal utility from eating cookies could also be confounded with crowding out, we offered non-overlapping sets of 70 cookies in each session.

We also ensured that differences in outcomes across the three conditions could not result from unobserved differences in cookies or in research assistants. Within each leg, campus and session, all subjects in all three groups were given the same 70 cookies. All subjects also interacted with the same research assistant who was blind to the research hypothesis.

Tasters also worked alone and had no contact with other tasters to avoid peer effects on outcomes (e.g., Mas and Moretti, 2009).[17]

(3) **Treatments.** After being recruited, but before their first session, tasters were randomly assigned, without their knowledge, to one of three groups: CONTROL, UNANTICIPATED and ANTICIPATED. Those in CONTROL performed as agreed upon recruitment. They came to their assigned rooms for the two sessions, tasted and evaluated cookies, and got thank-you cookies at the end of session two. CONTROL thus established baseline outcomes in the absence of incentive pay.

UNANTICIPATED. Those in this condition were surprised, at the start of session one before beginning tasting, with information that they would get $0.75 per cookie tasted and evaluated. There was no mention of pay for session two. One week later, at the start of session two, they were informed that they would not be paid the piece rate.[18] This main

on crowding out (e.g., Huffman and Bognanno, 2014).

[17]To ensure tasters had no contact with each other we had them come at different times to the tasting site, taste in a room with the door closed, and enter and exit through different doors.

[18]We offered no cover story for the withdrawal of the payment as that would have entailed deception.

treatment therefore replicates the canonical two-period design in psychology, in which a first-session surprise payment is withdrawn in the next session.

Crowding out can undermine performance in this condition on one or two margins: (i) tasting or (ii) tasting and evaluating. If subjects enjoy tasting but not evaluating, the per-cookie piece rate undermines output and/or productivity by eroding the marginal intrinsic utility in tasting while leaving the marginal cost of effort from evaluating unchanged. If tasters like both tasting and evaluating, the per-cookie piece rate undermines output or productivity by eroding the marginal intrinsic utility in both tasting and evaluating, leaving the marginal cost of effort associated with other aspects of the task unchanged (e.g., the physical effort of holding the pen). Even enjoyable activities must have effort costs otherwise agents' effort would be unbounded (see the model in Appendix Section F).

Anticipated. We noted above that unmet pay expectations may also contribute to second-period shortfalls in performance after the withdrawal of incentives. To investigate this possibility we added a "no-surprises" treatment to the design. Tasters received a telephone call or email one week before the first session informing them that they would get $0.75 per cookie tasted and evaluated in the first session but not in the second. Importantly, the information about the payment structure was offered *after* recruiting and random assignment were completed, so that the incentive scheme would not influence enrollment (see Appendix I for the protocol for each treatment).

Thus, in this treatment, agents were not surprised with pay in either session. As a result, they should behave according to standard model, except if there is crowding out. This idea builds on expectations-based reference-dependent preferences (Kőszegi and Rabin, 2006, 2007) wherein, absent deviations from expectations, agents behave as consumption utility maximizers. As we show later, agents in this "no-surprises" environment did indeed

behave in keeping with a standard model in both sessions and on all measures.

Though ANTICIPATED is not directly comparable to UNANTICIPATED as a treatment—for example, the session one pay is not a surprise in the former but it is in the latter—ANTICIPATED is still useful. It tests whether pathological underperformance occurs in session two in the absence of surprises for agents.[19]

4 Results

This section describes the results on our three main performance measures—output (number of cookies tasted and evaluated), productivity (minutes per cookie tasted and evaluated) and quits (percentage of tasters who quit after being paid in session one)—and discusses whether our findings are more consistent with a standard or with a crowding-out model.

The predictions of these models are straightforward. Under a standard model, output and productivity for either treatment should exceed CONTROL's in session one. Agents work harder when paid the piece rate. Then outcomes should return to CONTROL's level in session two, when the piece rate incentive is withdrawn.

Crowding out, by contrast, predicts that output and productivity in either treatment may exceed CONTROL's in session one depending on whether the incentive effect of the piece rate dominates over crowding out. But in the second session these outcomes should fall *below* CONTROL's, in line with output and productivity findings in prior crowding-out research (propositions 1 and 2 in Appendix F formalize this intuition).

[19] Although ANTICIPATED would more closely mirror UNANTICIPATED if subjects had been surprised in the first session with pay and then informed, sometime before the second session, that pay would be withdrawn, we felt this approach had a crucial disadvantage. It could *increase* the likelihood that agents expected a reward in the second session. Since agents had already been told once that they would not be paid and then had been paid anyway, they might expect the same surprise again. This expectation if unrealized, could depress effort and be confounded with crowding out. We hoped that giving advance warning of the pay scheme and following through in session one—paying the piece rate as promised and thus not surprising agents—would help establish the principal as a reliable promise keeper, reducing the expectation of a surprise payment in session two.

Quits do not necessarily allow distinguishing between the two models. Quits under a standard model stem from the trade-off between the income effect of the piece rate in session one (leading to higher consumption of leisure, inducing subjects to subsequently quit) and expectations of a piece rate in session two (inducing them to return). Thus the income effect of the piece rate cannot be disentangled from crowding out: both impel subjects to quit in session two.

Since the crowding-out literature most commonly reports time spent on the task (see Appendix E), we also describe it briefly. But, if not paired with other outcomes, such as output, this measure is somewhat uninformative of performance.

We show that evidence from our point estimates, graphical analysis, OLS, fixed effects, and from a multiple hypothesis testing procedure is most consistent with a standard model.

4.1 Sample and summary statistics

Sample and summary statistics. We recruited 91 participants for the four legs of our experiment. Random assignment placed 37 subjects in CONTROL, 27 in UNANTICIPATED and 27 in ANTICIPATED. Most subjects (76) came from campus A, where the facilities could accommodate more people (see Appendix Tables D.1 and D.2 for the breakdown by campus, treatment and legs). Most (81) attended session two: 34, 26, and 21 in CONTROL, UNANTICIPATED, and ANTICIPATED, respectively.[20]

[20]Prior to running the experiment, we ran a small, 9-person pilot to assess the response to the \$0.75 piece rate, using fewer cookies (60 or fewer) and a shorter evaluation sheet. We found that subjects often reached the upper bound of cookies to taste, which could lead to low variability in outcomes in the main experiment. So we increased the number of cookies to 70 and extended the questionnaire length. We also dropped two subjects from UNANTICIPATED because the research assistant gave them the wrong instructions in session two, forgetting to tell them they would not be paid.

Relatedly, our final sample size resulted from a compromise between the costs of running the experiment and an assessment of what seemed like a reasonable number of subjects based on the literature and our pilot. Because no prior studies used a similar task or piece rate, we had no prior data for precise power calculations. However, the median sample size was 15 subjects in the over 100 experiments we reviewed (and Deci, 1971, had even fewer). Further, in prior research documenting crowding out (e.g., Deci, 1971), the boost performance from the piece rate was *smaller* than the ensuing shortfall once the piece rate was

Subjects engaged substantially with the task. Table 1 shows that in the 172 subject-sessions (91 in session one and 81 in session two), tasters tasted an average of 35.3 cookies (with a minimum of 4 and a maximum of 70); they tasted and evaluated each cookie in 2.6 minutes, on average; and they spent on average 80.7 minutes tasting (with a minimum of 12 and a maximum of 182).[21] Of the cookies tasted, 70% were partially eaten, indicating that subjects listened to the instructions that they could merely take a bite.

4.2 Performance in UNANTICIPATED: Evidence and Discussion

We start by analyzing and discussing the results for UNANTICIPATED, our main treatment, which mirrors that in the canonical crowding-out test.

Output and productivity evidence for the first (reward) session. Table 2, columns (1) and (2), reports summary statistics per condition, outcome, and session. The piece rate boosted both average output and productivity for those in UNANTICIPATED. Subjects tasted and rated 18.3 more cookies (62%) than those in CONTROL (48.0 versus 29.7 in Panel B, column (1)) and did so 0.93 minutes (29%) faster (2.23 minutes versus 3.15 in Panel C, column (1)).

Importantly, the substantially higher average output and productivity were not driven by a few outliers, but by shifts in entire cumulative distribution functions (CDFs) of output and productivity relative to CONTROL (left panels of Figures 1A and 1B).

We had sufficient statistical power to detect these gains at the 1% to 3.1% level across

withdrawn, suggesting that responsiveness to the piece rate in session one would yield a high likelihood of detecting the crowding-out shortfall in performance in session two. Given these facts, the variability of the data, and the high responsiveness to the piece rate in session one in our pilot, we thought it reasonable to expect that our larger sample would be able to detect crowding out (our rough power analyses suggest that our starting sample had 80% power to detect effect sizes of 10 cookies on output and 0.58 minutes per cookie on productivity at the 5% level). As we see later, many of our effect sizes turned out be larger than these magnitudes, rendering them significant at the 1% level.

[21] Although productivity is typically the ratio of output to time we use the ratio of time to output (minutes per cookie tasted and evaluated) to be consistent with the literature (e.g., Deci, 1971, measures minutes per headline) and for ease of exposition.

three different estimation methods: unadjusted OLS, fixed effects, and multiple hypotheses testing (MHT) using the methodology in List, Shaikh, and Xu (2019). Our first method estimates, via OLS, unadjusted (simple) differences between the treatments and CONTROL in the two sessions. The outcome for subject i, in conditions $t1$ (CONTROL), $t2$ (UNANTICIPATED), and $t3$ (ANTICIPATED) in campus c, leg l, and session s is thus:

$$outcome_{i,t,c,l,s} = \alpha_{1,1} + \alpha_{1,2} t_1 s_2 + \sum_{\tau=2}^{3} \sum_{j=1}^{2} \beta_{\tau,j} t_\tau s_j + \epsilon_{i,t,c,l,s} \tag{1}$$

The parameters of interest are $\beta_{t,s}$, identifying unadjusted differences in average outcomes between the treatments and CONTROL per session. For example, $\beta_{2,1}$ identifies the difference between condition two (UNANTICIPATED) and CONTROL in session one. The parameter $\alpha_{1,1}$ identifies the outcome for the baseline category: CONTROL in session one.

This specification usefully pools subjects' outcomes for sessions one and two, allowing us to cluster standard errors at the subject level. Clustering accounts for serial correlation in outcomes for each subject across sessions, yielding more conservative standard errors (Bertrand, Duflo, and Mullainathan, 2004).[22]

Table 3 reports estimates of the output and productivity differences versus CONTROL using specification (1). The gains in output and productivity yielded by the piece rate (of 18.3 and 0.93, respectively) are statistically significant at the 1% level (Table 3, columns (1) and (3)). They remain significant at the 1% level using fixed-effects and MHT estimation, despite the power losses associated with these two methods (see Section 4.4).

Are the first (reward) session findings consistent with a standard or with a crowding-out model? The excess output and productivity are consistent with a standard model: agents work more when paid more. They could also be consistent with crowding out:

[22]Running separate regressions for sessions one and two would yield the same point estimates but fail to account for serial correlation (we would obtain smaller standard errors than with clustering).

the piece rate may have crowded out the intrinsic marginal utility for tasting (or tasting and evaluating), but its incentive effect dominated, resulting in a net gain in output and productivity (Proposition 1 in Appendix F).

Output, productivity and quits evidence for the second (non-reward) session. Few subjects quit. Only 3 out of 37 (8%) in CONTROL failed to come to the second session (Table 2, Panel A, columns (1)-(2)). This is the baseline quit rate. UNANTICIPATED also had a low quit rate, 1 in 27 tasters (4%), statistically indistinguishable from that in CONTROL (p-values of 0.455, 0.536 and 0.732 with, respectively, unadjusted OLS, fixed-effects and MHT estimation in Appendix Table A.1).

Average output in UNANTICIPATED was close to (though slightly higher than) CONTROL's by 3.6 cookies but this difference was statistically insignificant (Table 3, column 2, row 2). These averages again reflect the behavior of the whole distribution of outcomes (right panel of Figure 1A). And even though both groups had low and similar quit rates, we document, as a robustness check, that even if we had considered quitters as producing zero output, UNANTICIPATED's output would still slightly exceed CONTROL's by 4.6 cookies (Table 2, Panel B, column (3)). But this difference is statistically insignificant (Appendix Table A.2, Panel A).

Average productivity in UNANTICIPATED also exceeded that in CONTROL. On average, these subjects spent 0.78 fewer minutes (28%) per cookie. This productivity difference is statistically significant in the unadjusted OLS specification (1) (p-value=0.002 in Table 3, column (4), row (2)), with fixed-effects and with MHT estimation (p-values of 0.013 and 0.031, respectively, as shown in Section 4.4). This gap again reflects shifts in entire CDFs rather than the behavior of a few outliers (right panel of Figure 1B).[23]

[23]We do not attempt to input the productivity (individual minutes spent on the task/individual output) of quitters since they supplied zero time and zero output (0/0 is undefined).

Are the second (non-reward) session findings consistent with a standard or with a crowding-out model? All findings, except the extra productivity vis-à-vis Control are consistent with a standard model. The similar (though slightly higher) output found in Unanticipated versus Control is consistent with a standard model: in the absence of the piece rate, output returns to Control's level as the two groups face the same incentives (their output is mainly driven by the marginal utility of tasting or of tasting and evaluating cookies). However, it is not consistent with the crowding-out model, which predicts lower output than Control.

The higher productivity in Unanticipated is not consistent with either a standard or a crowding-out model. A standard model predicts a decline in productivity to the level of Control after the piece rate removal. Crowding out predicts lower productivity than in Control. We discuss this finding in more detail in Section 5.

The quit rate of 4%, which is lower but not statistically different from that in Control, is consistent with a standard model. The income effect of the piece rate in session one induces subjects to quit, but the expectation of the piece rate in session two entices them to return, curbing quits. But it could also be consistent with crowding out: the piece rate eroded interest in the task, but its expectation in session two leads tasters to come back.

Brief summary of time spent on the task in both sessions. Unanticipated subjects spent more time on the task than Control ones during the first session: 102.1 versus 79.4 minutes (Table 2, Panel D, column (1)). This difference is statistically significant at the 5% level using specification (1) and MHT (Appendix Table A.3). During the second session, however, these subjects spent 5.5 fewer minutes on the task than those in Control: 56.2 versus 61.7 minutes (Table 2, panel D, column (2)), but the difference is statistically insignificant (Appendix Table A.3). This shortfall of 9% in time spent on the task combined

with a 15% excess output (29.2 versus 25.5 evaluations) approximates the above-noted 28% excess productivity relative to CONTROL. This 5.5-minute deficit shrinks to 2.6 minutes (or 5% relative to CONTROL) if one assumes quitters supply zero time (Table 2, panel D, column (3)). Neither deficit is statistically significant (Appendix Tables A.2 and A.3).

The undersupply of time on the task for UNANTICIPATED in the second session is congruent with findings of prior crowding-out research. Our effect, however, is somewhat small and not statistically significant. Our smaller effect size dovetails with recent research in psychology and economics showing that experimental effect sizes are smaller in replications and in larger samples because of, for example, publication bias (e.g. OpenScienceCollaboration 2015; Camerer et al. 2016). The median sample of 15 subjects per condition in our review raises concerns that past reported effect sizes could be substantially inflated.

Further evidence for engagement with the task. Since a necessary condition to test crowding out is that subjects enjoy the task targeted for pay, one concern is that subjects might dislike tasting (or both tasting and evaluating) cookies, yet join to get the thank-you cookies. We documented above that CONTROL subjects spent ample time tasting cookies (one hour and 19 minutes in session one and one hour and two minutes in session two) and tasted plentiful amounts (an average of 30 cookies in session one and of 26 in session two), despite the thank-you cookies' not being contingent on the time spent or numbers tasted. These findings suggest that subjects liked the task. It thus appears unlikely they enrolled while disliking it.

4.3 Performance in ANTICIPATED: Evidence and Discussion

The previous section outlined the results of CONTROL and UNANTICIPATED conditions that mirror the canonical test's, and documented that the results are most consistent with a standard model. The patterns of the CDFs of output and productivity, the average effects

on these two measures, and the evidence on quits, are all consistent with this model. The only exception is the excess productivity in session two, which is also inconsistent with crowding out. Hence, on these three measures we found no evidence congruent with crowding out but not with standard model. Rather, we found the opposite.

We next consider the results of our secondary treatment, ANTICIPATED. In the absence of unmet wage expectations and of crowding out (both could yield the anomalous deficit in performance relative to CONTROL after reward withdrawal), agents should behave as in a standard model. This is what occurred, as we now describe.

Output and productivity evidence for the first (reward) session. The piece rate here again boosted both output and productivity. Those in ANTICIPATED tasted and rated 21.4 more cookies (72%) than those in CONTROL: 51.1 versus 29.7 (Table 2, Panel B, column (1)). They did so 0.85 minutes (27%) faster than those in CONTROL: in 2.31 minutes versus 3.15 (Panel C, column (1)). Once again shifts in entire CDFs of output and productivity drive these averages (left panel of Figures 1A and 1B). Again, these differences are statistically significant at the 1% level in specification (1) (Table 3, columns (1) and (3), row (3)), with fixed-effects and with MHT estimation (see Section 4.4).

Are the first (reward) session findings consistent with a standard or with a crowding-out model? Once again, the excess output and productivity under piece-rate pay are consistent with a standard model: agents work harder under incentive pay. But they could also be consistent with crowding out: the incentive effect of the piece rate dominated over its crowding out one yielding a net increase in performance.

Output, productivity, and quits evidence for the second (non-reward) session. ANTICIPATED had a much higher quit rate than the other two conditions: 6 in 27 tasters (22%) (Table 2, Panel A, columns (1)-(2)), exceeding that in CONTROL by 14 percentage

points. This difference, however, ends up being statistically insignificant (p-values of 0.134, 0.148, and 0.530 with unadjusted OLS, fixed-effects and MHT estimation, respectively, in Appendix Table A.1).

The average output for those who came to the second session (78% of tasters) exceeded that of CONTROL by 6.2 cookies, though this difference is statistically insignificant (Table 3, column (2)). Even if we assume that quitters supply zero output, average output still exceeds that in CONTROL by 1.2 cookies (Table 2, Panel B, column 3), a statistically insignificant difference (Appendix Table A.2, Panel A, column 2).

The average productivity for those who came to the second session was similar to that of CONTROL: they spent a statistically insignificant 0.1 fewer minutes (6 seconds) per evaluation than CONTROL's (Table 3, column 4, row (3)).

Are the second (non-reward) session findings consistent with a standard or with a crowding-out model? All findings for ANTICIPATED are consistent with a standard model. First, output slightly higher (but statistically insignificant) than CONTROL's, even assuming quitters supplied zero output, supports a standard model, but not crowding out, which predicts lower output.

Second, the similar productivity in the second session of those in CONTROL and the 78% non-quitters in ANTICIPATED also fits a standard model: conditional on returning, ANTICIPATED tasters display the same productivity as those in CONTROL as there is no longer the piece-rate incentive to boost their productivity. However, this similarity does not rule out crowding out: for example, had the 22% of quitters come to the second session they could have had very low productivity — be extremely slow tasting and rating—dragging average productivity below that of CONTROL's, consistent with the crowding out.

Though not statistically significant, the substantially higher quit rate relative to CON-

TROL would also be consistent with a standard model. The income effect of the piece rate increased the consumption of leisure in the second session and the expectation of the piece rate is no longer present to induce subjects to return, as was the case for UNANTICIPATED. But this quit pattern would also be consistent with crowding out: the piece rate in session one eroded interest in the task, and its expected removal in session two did not entice subjects to return.

Brief summary of time spent on the task in both sessions. Subjects in ANTICIPATED also spent more time on the task than CONTROL subjects during the first session— 111.6 versus 79.4 minutes— and those who returned also spent more time on the task in the second session: 77.1 versus 61.7 minutes (Table 2, Panel D, columns (1)-(2)). Even considering the worst-case scenario of quitters supplying zero time in the second session, time spent on the task in this condition still exceeds that in CONTROL by 3.3 minutes: 60.0 versus 56.7 minutes (column (3)).

4.4 Robustness Checks: Fixed Effects and Multiple Hypothesis Testing

We now document that all of the statistically-significant results in the previous sections, such as the increase in output and productivity associated with the piece-rate pay, also hold in more conservative tests incorporating fixed effects and multiple hypothesis testing.

Fixed-Effects Estimation. We randomized tasters into each condition within each leg and campus. For example, in leg two, we randomized subjects in campus A into the three conditions and did the same for campus B. Thus unobserved campus, leg, and session factors might have affected differences between the treatments and CONTROL.

Hence, for our first robustness check, we add time-invariant unobserved campus, leg, and session factors to specification (1). Our fixed-effects specification for the outcome for subject i, in the $t1$ (CONTROL), $t2$ (UNANTICIPATED) and $t3$ (ANTICIPATED) conditions in campus

c, leg l, and session s is:

$$outcome_{i,t,c,l,s} = \alpha_{1,1} + \alpha_{1,2} t_1 s_2 + \sum_{\tau=2}^{3} \sum_{j=1}^{2} \beta_{\tau,j} t_\tau s_j + \lambda_c \times \lambda_l \times \lambda_s + \epsilon_{i,t,c,l,s} \qquad (2)$$

The interaction of campus, leg, and session, $(\lambda_c \times \lambda_l \times \lambda_s)$, conservatively captures unobservable, time-invariant, campus, leg, and session determinants of outcomes. Campus fixed effects control, for example, for unobserved heterogeneity in health consciousness by campus, which could influence the response to pay within campus. Leg fixed effects control, for example, for the unobserved temperature, which could also affect the response to incentives.[24] The interaction conservatively captures these effects on outcomes within a given campus, leg, and session.[25]

The causal parameters of interest are the $\beta_{t,s}$ identifying pooled differences in outcomes between the treatments and CONTROL for each session, but now within a campus and leg. The parameter $\alpha_{1,1}$, identifying the outcome for the baseline category (CONTROL in session one) cannot be separately estimated from the fixed effects, as usual.

Table 4, Panel A, reports the fixed-effects estimates from specification (2) for output and productivity. It shows similar effect sizes to those from unadjusted OLS. But the estimates become noisier, though still significant at the 1.3% level or less.[26] The larger standard errors result from the fact that the modest increases in model fit from the introduction of the interaction of campus, leg and session dummies (for example, the R^2 increases slightly from 0.26 in Table 3 to 0.35, in Table 4, Panel A) are dominated by the loss of degrees of freedom

[24]Examples of these potential campus and leg time-invariant unobservables could be, respectively, tasters on a more health-conscious campus not increasing consumption and thus not producing more evaluations in response to the piece rate, and cookies being less appealing in an experimental leg in a hot month.

[25]This interaction is conservative in that it subsumes standalone campus, leg or session fixed effects and their two-way interactions.

[26]Appendix Table A.1, column (2) and Appendix Table A.3, columns (3)-(4) show that estimates on, respectively, quits and time spent on the task, remain similar after these adjustments but also noisier.

resulting from the addition of 14 new dummy variables.[27]

MHT estimation. As a second robustness check, we use the multiple hypothesis testing procedure in List, Shaikh, and Xu (2019). It corrects the p-value of a single hypothesis test to account for the joint testing of, for example, multiple treatments and multiple outcomes. It thus reduces the chance of obtaining false-positive estimates. This procedure reduces the probability of rejecting the null while offering greater power than classical approaches, such as Bonferroni (1935).

In our case, we jointly tested three conditions (CONTROL, UNANTICIPATED, and ANTICIPATED) and three outcomes in session one: output, productivity and, time spent on the task. And we jointly tested the three conditions and four outcomes in session two: output, productivity, quits into session two, and again, time spent on the task. Though time spent on the task is a secondary outcome, whose analysis is relegated to the appendix, we added it to the MHT estimation, since it renders a more conservative test (generates larger p-values).

Table 4, Panel B, shows differences in output and productivity relative to CONTROL and their respective p-values (the method does not report standard errors). Some of our estimates substantially lose precision (e.g, the p-value of 0.002 on excess productivity of -0.78 in Table 3, column (4), row (2), jumps considerably to 0.031 in Table 4, Panel B) but all the previously statistically significant results remain so at the 3.1% level or less.[28]

5 UNANTICIPATED Second-Session Excess Productivity

Only one result so far is inconsistent with a standard model. Whereas a standard model predicts that productivity in UNANTICIPATED should decline to the level of CONTROL's

[27]The first leg, with the two sessions, was run on only one campus (2 dummy variables). The remaining three legs were run on two campuses, each with two sessions (12 dummy variables).

[28]Appendix Table A.1, column (3) and Appendix Table A.3, columns (5) and (6) report the MHT estimates for, respectively, quits and time spent on the task, and show that they also become noisier.

in the second session, when subjects are no longer paid, subjects in this treatment had higher productivity than those in CONTROL in session two. Since it seemed odd that the surprising withdrawal of pay would boost productivity we now discuss evidence, though more speculative, for why this occurred.

One possible explanation is learning. The increase in productivity in the first session due to the piece rate had lasting effects into the second session through learning, which withstood the withdrawal of pay. Though CONTROL controls for such learning, we considered this explanation, finding three pieces of evidence against it. For example, Appendix B shows that the 78% non-quitters in ANTICIPATED had the same high productivity during the first session, but their productivity declined to the same level as CONTROL's in the second session.

Another possibility is that the boost in productivity might entail hidden costs for the principal. Displeased upon learning they were not going to be paid, subjects may have tasted their cookies (producing the same output as CONTROL) but decamped as soon as possible, by reducing their care in filling out the evaluations. This explanation would be in keeping with Mas (2006) and Kube, Maréchal, and Puppe (2013), who found that when wage expectations are not met, workers shirk on one or several dimensions of the task (e.g., output or quality of the output) motivated by, for example, retaliation or loss of morale (e.g., Bewley, 1999).[29]

We find suggestive evidence for this possibility. One qualification on the analysis that follows, however, is that assessing the quality of subjective ratings is less straightforward than measuring output, productivity, and quits, and thus includes caveats.

Quality measure. It is hard to assess subjective surveys where there is no "correct"

[29]For example, Mas (2006) documented that when policemen fail to receive expected higher wages disputed under arbitration, their performance declines to levels below those prior to arbitration; Kube, Maréchal, and Puppe (2013) showed that when workers hired for a forecasted wage were told, upon arriving at the work site, that they would be paid less, they underperformed vis-à-vis a control group.

answer. Quality is therefore often analyzed in terms of dispersion of responses. Subjects can economize on effort by choosing more responses at random, thus increasing dispersion. Or they can "straight-line" giving the same answers/ratings over consecutive questions, hence reducing or eliminating dispersion (e.g., Krosnick, 1991, 1999).

We therefore used the dispersion in ratings for a fixed cookie to assess rating quality. Conditional on either strategy (straight-lining or responding more randomly)—leading to the same economy of effort, straight-lining had a higher probability of detection, since the research assistant scanned the evaluations before paying the piece rate or conferring the thank-you cookies. Thus it seemed more likely subjects would shirk by responding more randomly. Indeed, straight-lining (e.g., giving a 2 on all dimensions) occurred rarely in our data (e.g., only 3.6% of the total number of evaluations across all subjects and sessions had the same rating on all dimensions). We also found a positive association between the speed of completing evaluations and rating dispersion, consistent with speed leading to less thoughtful evaluations (see Appendix C).

Our measure of dispersion is the standard deviation in the ratings of a cookie.[30] For each evaluation, we computed the standard deviation in ratings across the seven dimensions: Appearance, Aroma, Snap, Texture, Start, Flavor, and Overall Rating. The scale for each dimension ran from 1 (Excellent) to 5 (Poor). Therefore, a cookie rated 1, 3, 4, 2, 5, 5, 3 had a standard deviation of 1.5.

Sample. We assessed whether the dispersion in ratings for a fixed cookie (e.g., an Oreo Chocolate) could differ across the three conditions. It is important to hold the cookie fixed across the three conditions to ensure that differences in dispersion are not due to differences in cookies themselves.

Thus, our analysis focuses on cookies that were sampled in the three conditions at a given

[30]The use of other measures, such as, variance and range, yielded qualitatively similar results.

campus, leg, and session so that we can compare *that cookie's* dispersion in ratings across the three conditions *within* that campus, leg, and session. Although all subjects within a campus, leg, and session received the same 70 cookies, they did not all taste exactly the same ones. Subjects tasted different numbers of cookies and, within those, some cookies were more likely to be tasted, as they had been randomly placed in trays closer to a subject.[31]

Graphical evidence. The top left panel in Figure 2 depicts the empirical CDFs of the standard deviation in ratings for an evaluation for the first (reward) session without holding a specific cookie fixed. It shows that for the most part the CDFs overlap or lie close to each other. But during the second (non-reward) session the CDF for tasters surprised by the withdrawal of pay (UNANTICIPATED) stochastically dominates that for those in CONTROL and ANTICIPATED.

Estimation specification. The graphical evidence, though suggestive, does not hold a given cookie *fixed*. To estimate the average dispersion in ratings for the *same cookie* tasted within a campus, leg and session, we use the following specification: the dispersion in the ratings of a fixed cookie k, for subject i, in the $t1$ (CONTROL), $t2$ (UNANTICIPATED), $t3$ (ANTICIPATED) conditions at campus c, leg l, and session s is

$$dispersion_{k,i,t,c,l,s} = \alpha_{1,1} + \alpha_{1,2} t_1 s_2 + \sum_{\tau=2}^{3} \sum_{j=1}^{2} \beta_{\tau,j} t_\tau s_j + \lambda_c \times \lambda_l \times \lambda_s \times \lambda_k + \epsilon_{k,i,t,c,l,s}$$

[31] We excluded cases in which the same cookie identification number could refer to more than one specific cookie, such as cookie assortments. Thus the analysis drops 5 and 12 subjects in sessions one and two, respectively. For example, we dropped a subject who tasted 15 cookies in session two, of which 8 were from assortments and the remaining 7 were not tasted by subjects in the two other conditions at his campus, leg, and session. To extrapolate the dispersion analysis from this restricted sample to the full sample, we need to assume that had these tasters not been excluded the dispersion results would have been similar. Although we cannot test whether dropping these subjects biases the dispersion results, we can do so for output and productivity. Appendix Table A.4 shows that output and productivity differences across conditions and sessions on the restricted and full samples are similar, suggesting the exclusion of these subjects did not bias output and productivity results. Thus, this exclusion might not bias the dispersion results. Nonetheless, any extrapolation should be viewed in light of these caveats.

The interaction of campus, leg, session, and cookie fixed effects ($\lambda_c \times \lambda_l \times \lambda_s \times \lambda_k$) conservatively captures unobserved time-invariant, campus, leg, session, and cookie determinants of standard deviation. Campus fixed effects control, for example, for whether subjects on one campus are less conscientious than those on the other, with higher propensity to respond at random. Leg fixed effects address, for example, whether during a hot summer leg, some cookies have a higher tendency to melt, leading to more dispersion in their ratings for this leg than for others (e.g., their Appearance rating would be worse but their Flavor rating would remain unchanged). Cookie fixed effects control for unobserved time-invariant differences in cookie characteristics that could lead to differences in dispersion, such as a cookie's having a good appearance but a bad flavor. The interaction addresses whether these unobservables could differentially affect outcomes within campus, leg, session, and cookie.[32]

The parameters of interest are $\beta_{t,s}$, which identify, for a given campus, leg, and session, how the dispersion for the *same cookie* differs between the treatments and CONTROL. For example, $\beta_{2,1}$ identifies the difference in the standard deviation for a cookie in treatment two in session one versus the standard deviation for that *same cookie* in CONTROL in session one, by pooling all these differences within each campus and leg for this fixed cookie. As usual, $\alpha_{1,1}$ identifies the baseline category: the outcome for CONTROL in session one, which cannot be separately identified from the fixed effects.

We conservatively cluster the standard errors by individual to address the potential correlation in ratings' dispersion for a given subject within a session—because each individual produced several evaluations in each session—and across sessions— because dispersion in filling out the evaluations is likely correlated across sessions for the same individual (Bertrand, Duflo, and Mullainathan (2004).

[32]For example, whether a hot summer leg affects the dispersion of a given cookie more on one campus than on the other.

Dispersion during the first (reward) session. Table 5 reports the results of our estimation. For both UNANTICIPATED and ANTICIPATED, the standard deviation *for the same cookie* rated at a given campus and leg during the reward session is similar to that in CONTROL: only 0.03 and 0.01 higher and not statistically significant (column (3)). Though the piece rate induced tasters to work faster, it did not appear to reduce the quality of their output.

Dispersion during the (second) non-reward session. Subjects surprised by the payment's withdrawal increased the dispersion in ratings for the same cookie relative to CONTROL in the second session. Column (4) shows that the standard deviation *for the same cookie* rated at a given campus and leg was 0.10 higher than CONTROL, a difference that more than tripled relative to that in the reward session. This difference is significant at the 5.1% level. By contrast, subjects who expected the reward to be withdrawn (ANTICIPATED) did not exhibit an increase in the dispersion of their ratings. Their dispersion was slightly smaller than in CONTROL and statistically significant (column (4)).

This more tentative evidence on dispersion shows that the excess randomness in the ratings only occurs for those surprised by the withdrawal of the piece rate in session two (UNANTICIPATED). This pattern suggests that the higher productivity in UNANTICIPATED after the removal of pay may stem from this group's providing lower-quality evaluations. This pattern could also be consistent with crowding out, if one assumes, for example, that rating or tasting thoughtfully was the *only* enjoyable component of the task (subjects did not enjoy tasting cookies per se) and that pay undermined that enjoyment.

6 Conclusion

This paper reviews more than 100 tests in the literature on whether pay harms performance in enjoyable tasks and describes an experiment, based on the canonical two-period test for

crowding out, in which we analyzed output, productivity, quits, time spent on the task, and more speculative evidence on quality.

Our extensive review on whether pay damages performance on enjoyable tasks that benefit primarily agents and principals reveals that no prior experiment has jointly tested output, productivity, and time spent on the task, which may yield competing evidence on crowding out. The absence of complete and consistent performance measures may help account for the inconsistent evidence on this phenomenon. Incomplete reporting of outcomes combined with small samples (the median sample size by condition documented by our review was 15 subjects) raises concerns that this research may include many false positives.

We also ran a field experiment with three key features. First, we report all the metrics above to obtain a more transparent view of the effect of pay on performance. We focus on measures of interest to a principal, such as output, productivity, and quits. But we also describe time spent on task, since it is the most used metric in crowding-out research. Second, we recruited larger numbers of participants to each condition than one typically sees in this literature. Third, we aimed to assess the potential role of unmet pay expectations in case we observed that pay harmed performance.

With one exception, our results across output, productivity, and quits are consistent with a standard economics model. Output and productivity results stem from shifts in CDFs of these outcomes rather from a small number of outliers. They are also robust to power losses from fixed-effects and multiple hypothesis testing estimation.

The exception is that productivity for those who have unexpectedly had their pay withdrawn exceeds that in Control. This result is inconsistent with both a standard and a crowding-out model. More tentative evidence suggests that this excess productivity stemmed from these subjects' supplying lower-quality output. This pattern appears more consistent

with subjects' losing morale or retaliating, as prior research suggests (e.g., Bewley, 1999; Mas, 2006; Kube, Maréchal, and Puppe, 2013). It could reflect crowding out, but under fairly strong assumptions,

Our review of the literature and test point to the importance of reporting multiple performance measures to assess the effect of pay on enjoyable tasks. The interrelated measures discussed in this paper, such as output, productivity, and time spent on the task, can lead to different conclusions. Although reporting them may render the interpretation of whether pay harms performance more nuanced and difficult—for example, how should we interpret, in the context of crowding out, a reduction of time spent on the task that leaves output unchanged, boosting productivity?— it would provide a richer and more transparent picture of this phenomenon.

Our experiment also indicates that evidence for crowding out is not as easily detectable as suggested by this literature, even using the canonical test. These effects may be rarer than previously thought.

It is beneficial to have a richer understanding of whether pay hurts performance on jobs agents find interesting. While most jobs entail a mix of enjoyable and unenjoyable tasks it appears desirable that employees largely enjoy their work. All else equal, they are more productive. For example, managers with higher scores on the Enjoyment-of-Work Scale (McMillan, Brady, O'Driscoll, and Marsh, 2002) performed better (Graves, Ruderman, Ohlott, and Weber, 2012).[33] However, crowding out indicates that pay may harm performance in these situations.

Our literature review and experiment disclosing conflicting outcomes contribute to our understanding of this issue. There is, nonetheless, room for improvement. Though our

[33]This scale assesses agreement with statements, such as "My job is so interesting is does not seem like work" or "I do more work than expected of me strictly for the fun of it".

experiment had sufficient power to detect responses to incentives and their removal, our samples were not very large as we relied on prior literature—which finds effects even in small samples—for power calculations. Future studies should aim for larger samples and study other enjoyable undertakings to further knowledge of the extent to which rewarding agents to perform jobs they enjoy may harm their performance.

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011): "Reference Points and Effort Provision," *American Economic Review*, 101, 470–492.

Ariely, D., A. Bracha, and S. Meier (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1), 544–555.

Ashraf, N., O. Bandiera, and B. K. Jack (2014): "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery," *Journal of Public Economics*, 120, 1–17.

Baron, J., and D. Kreps (1999): *Strategic Human Resources: Frameworks for General Managers*. Wiley, first edition edn.

Bell, D. (1985): "Disappointment in Decision Making Under Uncertainty," *Operations Research*, 33(1), 1–27.

Bénabou, R., and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70(3), 489–520.

Bénabou, R., and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652–1678.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-In-Differences Estimates?," *Quarterly Journal of Economics*, 119(1), 249–275.

Bewley, T. F. (1999): *Why Wages Don't Fall During a Recession*. Harvard University Press.

Boal, K. B., and L. Cummings (1981): "Cognitive EvaluationTheory: An Experimental Test of Processes and Outcomes," *Organizational Behavior and Human Performance*, 28(3), 289–310.

Bonferroni, C. (1935): "Il Calculo della Assicurazioni su Gruppi di Teste," *Tipografia del Senato*.

Calder, B. J., and B. M. Staw (1975): "Interaction of Intrinsic and Extrinsic Motivation - Some Methodological Notes," *Journal of Personality and Social Psychology*, 31(1), 76–80.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Science*.

Cameron, J., K. Banko, and W. Pierce (2001): "Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues," *The Behavior Analyst*, 24(1), 1.

Cameron, J., and W. D. Pierce (1994): "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis," *Review of Educational Research*, 64(3), 363–423.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), 817–869.

Chetty, R., E. Saez, and L. Sandór (2014): "What Policies Increase Prosocial Behavior? An Experiment with Referees at the Journal of Public Economics," *NBER Working Paper 20290*.

Crawford, V., and J. Meng (2011): "New York City Cab Drivers Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 101(5), 1912–1932.

Deci, E. (1971): "Effects of Externally Mediated Rewards on Intrinsic Motivation," *Journal of Personality and Social Psychology*, 18(1), 105–115.

Deci, E., R. Koestner, and R. Ryan (1999): "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation," *Psychological Bulletin*, 125, 627–668.

Deci, E., and R. Ryan (1985): *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

Ericson, K., and A. Fuster (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *Quarterly Journal of Economics*, 126, 1879–1907.

Esteves-Sorenson (2018): "Gift Exchange in the Workplace: Addressing the Conflicting Evidence with a Careful Test," *Management Science*, 64(9), 3971–4470.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54(2), 293–315.

Gächter, S., and A. Falk (2002): "Reputation and Reciprocity: Consequences for the Labour Relation," *Scandinavian Journal of Economics*, 104, 1–26.

Gibbons, R. (1998): "Incentives in Organizations," *The Journal of Economic Perspectives*, 12(4), 115–132.

Gill, D., and V. Prowse (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *American Economic Review*, 102(1), 469–503.

Gneezy, U., and J. List (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, 74(5), 1365–1384.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives*, 25(4), 191–210.

Gneezy, U., and A. Rustichini (2000): "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 115(3), 791–810.

Graves, L., M. Ruderman, P. Ohlott, and T. Weber (2012): "Driven to Work and Enjoyment of Work: Effects on Managers' Outcomes," *Journal of Management*, 38(5), 1655–1680.

Groves, R., S. Presser, and S. Dipko (2004): "The Role of Topic Interest in Survey Participation Decisions," *Public Opinion Quarterly*, 68(1), 2–31.

Gul, F. (1991): "A Theory of Disappointment Aversion," *Econometrica*, 59(3), 667–686.

Holland, J., and L. Christian (2009): "The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys," *Social Science Computer Review*, 27(2), 196–212.

Hossain, T., and K. K. Li (2014): "Crowding Out in the Labor Market: A Prosocial Setting is Necessary," *Management Science*, 60(5), 1148–1160.

Huffman, D., and M. Bognanno (2014): "Does Performance Pay Crowd Out Worker Non-Monetary Motivations? Evidence from a Real Work Setting," Working Paper.

Ioannidis, J. P. A. (2008): "Why Most Discovered True Associations Are Inflated," *Epidimiology*, 19, 640–648.

Kahneman, D., and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47(2), 263–291.

Kamenica, E. (2012): "Behavioral Economics and Psychology of Incentives," *Annual Review of Economics*, 4(1), 427–452.

Kőszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Perspectives*, 52(4), 1075–1118.

Kőszegi, B., and M. Rabin (2006): "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 121(4), 1133–1165.

——— (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97(4), 1047–1073.

Krosnick, J. A. (1991): "Response Strategies for Coping With the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5(3), 213–236.

——— (1999): "Survey Research," *Annual Review of Psychology*, 50(1), 537–567.

Kube, S., M. Maréchal, and C. Puppe (2013): "Do Wage Cuts Damage Work Morale? Evidence From a Natural Field Experiment," *Journal of the European Economic Association*, 11(4), 853–870.

Lacetera, N., M. Macis, and R. Slonim (2012): "Will There Be Blood? Incentives and Displacement Effects in Pro-Social Behavior," *American Economic Journal-Economic Policy*, 4(1), 186–223.

Lazear, E. (2000): "Performance Pay and Productivity," *American Economic Review*, pp. 1346–1361.

Lazear, E., and M. Gibbs (2014): *Personnel Economics in Practice*. Wiley, third edition edn.

Lepper, M., D. Greene, and R. Nisbett (1973): "Undermining Children's Intrinsic Interest with Extrinsic Reward-Test of the Overjustification Hypothesis," *Journal of Personality and Social Psychology*, 28(1), 129–137.

List, J., A. Shaikh, and Y. Xu (2019): "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, Forthcoming.

Loomes, G., and R. Sugden (1986): "Disappointment and Dynamic Consistency in Choice under Uncertainty," *Review of Economic Studies*, 53(2), 271–282.

Macera, R., and V. te Velde (2016): "On the Power of Gifts," .

Mas, A. (2006): "Pay, Reference Points, and Police Performance," *Quarterly Journal of Economics*, 121(3), 783–821.

Mas, A., and E. Moretti (2009): "Peers at Work," *American Economic Review*, 99(1), 112–145.

McMillan, L., E. Brady, M. O'Driscoll, and N. Marsh (2002): "A multifaceted validation study of Spence and Robbins' (1992) Workaholism Battery," *Journal of Occupational and Organizational Psychology*, 75, 357–368.

Mellström, C., and M. Johannesson (2008): "Crowding Out in Blood Donation: Was Titmuss Right?," *Journal of European Economic Association*, 6(4), 845–863.

OpenScienceCollaboration (2015): "Estimating the Reproducibility of Psychological Science," *Science*, 349(6251).

Pope, D. G., and M. E. Schweitzer (2011): "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes," *American Economic Review*, 101, 129–157.

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

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, pp. 1281–1302.

Rebitzer, J. B., and L. J. Taylor (2011): "Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets," *Handbook of Labor Economics*, 4, 701–772.

Ryan, R., and E. Deci (2000): "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemporary Educational Psychology*, 25, 54–67.

Shalev, J. (2000): "Loss Aversion Equilibrium," *International Journal of Game Theory*, 29(2), 269–287.

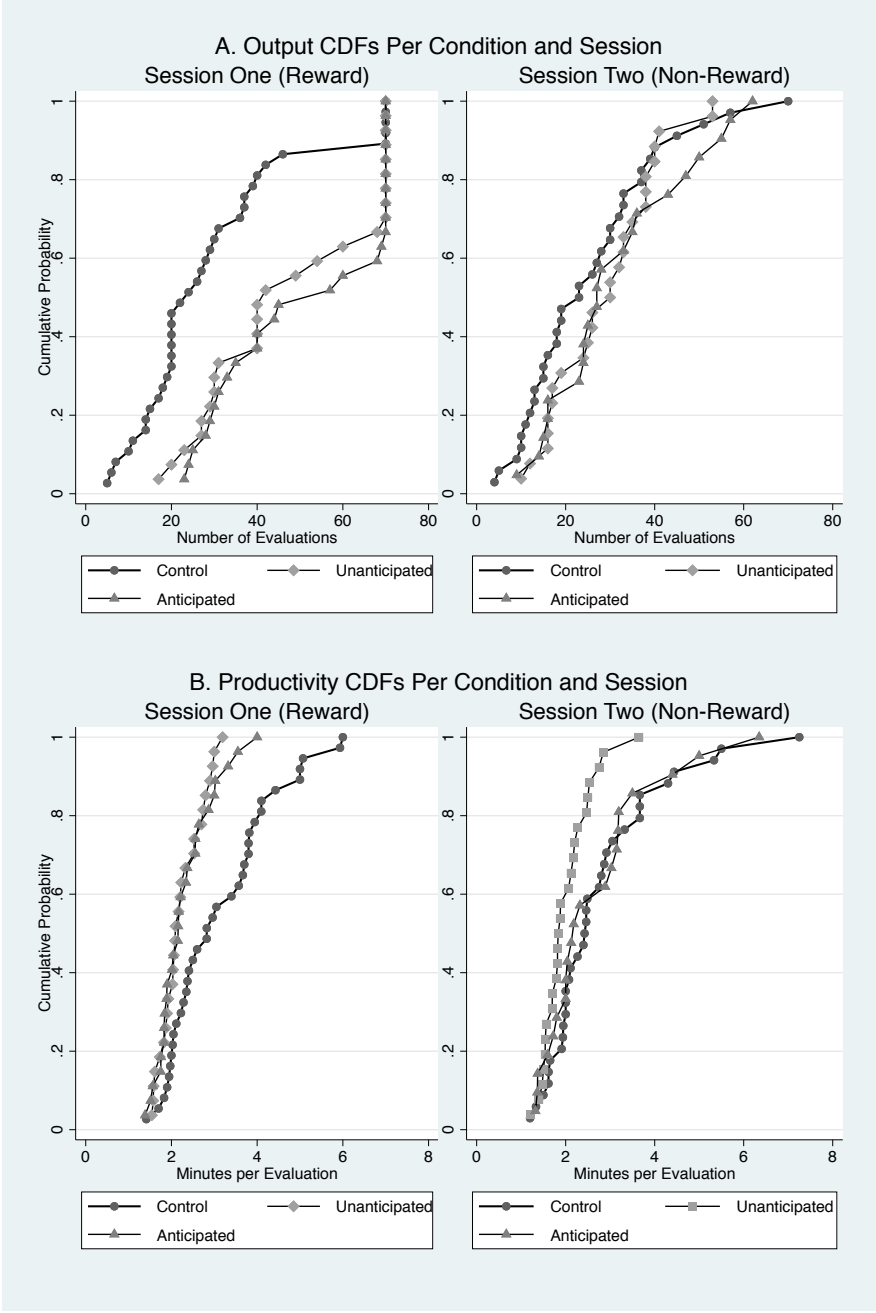
Simmons, J., L. Nelson, and U. Simonsohn (2011): "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22(11), 1359–1366.

Simonsohn, U., L. Nelson, and J. Simmons (2014a): "P-Curve: a Key to the File Drawer," *Journal of Experimental Psychology: General*, 143(2), 534–547.

——— (2014b): "P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results," *Perspectives on Psychological Science*, 9(6), 666–681.

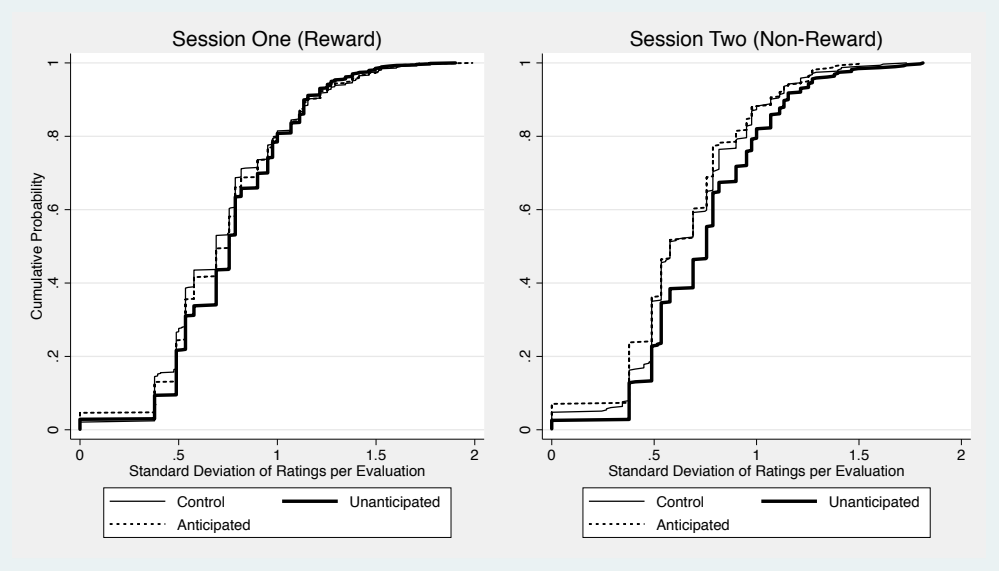
Figures

Figure 1: Cumulative Distributions Functions for Output and Productivity



Notes: Figure 1A depicts the cumulative distribution function (CDF) of output for Control, Unanticipated and Anticipated for the first (reward) session (left panel) and second (non-reward) session (right panel). Figure 1B depicts the same information, but for productivity.

Figure 2: CDFs for the Standard Deviation of Evaluations' Ratings



Notes: Cumulative distribution function (CDF) for the standard deviation of ratings for the cookies evaluated in Control, Unanticipated and Anticipated in the first (reward) session (left panel) and second (non-reward) session (right panel).

Table 1: Summary Statistics Across All Conditions and Sessions

	Number of Subject X Sessions	Mean	Standard Deviation	Min	Max
Output (Number of cookies tasted and evaluated)	172	35.3	19.5	4	70
Productivity (Minutes per cookie tasted and evaluated)	172	2.6	1.1	1.2	7.3
Time spent on the task (Minutes spent tasting and evaluating)	172	80.7	39.3	12	182
Proportion of partially eaten cookies	172	0.7	0.3	0.0	1.0

Table 2: Disaggregated Summary Statistics by Condition and Session

	Unadjusted Data		Data assuming quitters would have supplied zero output and time
	Session		Session
	One (Reward)	Two (Non-Reward)	Two (Non-Reward)
	(1)	(2)	(3)
PANEL A: Sample (number of subjects)			
(1) Control	37	34	37
(2) Unanticipated	27	26	27
(3) Anticipated	27	21	27
N Total (Subjects)	91	81	91
Panel B: Output (Number cookies tasted and evaluated)			
(1) Control Mean	29.7	25.5	23.5
	(19.1)	*(15.3)*	*(16.3)*
(2) Unanticipated Mean	48.0	29.2	28.1
	(19.4)	*(11.8)*	*(12.8)*
(3) Anticipated Mean	51.1	31.7	24.7
	(18.9)	*(15.4)*	*(19.0)*
Panel C: Productivity (Minutes per cookie tasted and evaluated)			
(1) Control Mean	3.15	2.79	-
	(1.21)	*(1.32)*	
(2) Unanticipated Mean	2.23	2.01	-
	(0.48)	*(0.54)*	
(3) Anticipated Mean	2.31	2.69	-
	(0.65)	*(1.30)*	
Panel D: Time on Task (Total minutes spent tasting and evaluating)			
(1) Control Mean	79.4	61.7	56.7
	(37.4)	*(33.5)*	*(36.3)*
(2) Unanticipated Mean	102.1	56.2	54.1
	(37.2)	*(21.2)*	*(23.4)*
(3) Anticipated Mean	111.6	77.1	60.0
	(36.1)	*(38.1)*	*(46.7)*

Columns (1) and (2) summarize the unadjusted (raw) data from the experiment. Column (3) summarizes the output and time on the task for the second (non-reward) session, assuming quitters would have supplied zero output and zero time. It does not summarize productivity (individual minutes spent on the task/individual output) under these assumptions as 0/0 is undefined. Standard deviations are in parentheses.

Table 3: Output and Productivity by Condition and Session-Unadjusted OLS

Specification:	**Unadjusted OLS**			
Dependent Variable:	**Output: Number of Cookies Tasted and Evaluated**		**Productivity: Minutes per Cookie Tasted and Evaluated**	
	Session		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
	(1)	**(2)**	**(3)**	**(4)**
(1) Control	29.7	25.5	3.15	2.79
	(3.2)***	(2.6)***	(.20)***	(.23)***
Difference vs. Control				
(2) Unanticipated	18.3	3.6	-0.93	-0.78
Standard Error	(4.9)***	(3.5)	(.22)***	(0.25)***
p-value	[0.000]	[0.304]	[0.000]	[0.002]
(3) Anticipated	21.4	6.2	-0.85	-0.10
Standard Error	(4.8)***	(4.3)	(.24)***	(0.36)
p-value	[0.000]	[0.150]	[0.001]	[0.782]
N (observations)	172		172	
Number of clusters	91		91	
R-squared	0.26		0.14	

Notes. The results for columns (1) and (2) are from the unadjusted OLS specification (1) where the outcome is output. The results for columns (3) and (4) for the same specification, but where the outcome is productivity. Standard errors are in parentheses and clustered by subject. **Significant at the 5% level; ***Significant at the 1% level. p-values in square brackets shown for easier comparison with the p-values yielded by MHT. All tests are two-tailed.

Table 4: Output and Productivity by Condition and Session-Fixed Effects and MHT

Dependent Variable:	**Output: Number of Cookies Tasted and Evaluated**		**Productivity: Minutes per Cookie Tasted and Evaluated**	
	Session		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
	(1)	**(2)**	**(3)**	**(4)**
PANEL A	**Specification: Fixed Effects**			
Difference vs. Control				
(1) Unanticipated	15.8	1.0	-0.95	-0.80
Standard Error	(5.3)***	(4.1)	(.26)***	(0.32)**
p-value	[0.004]	[0.808]	[0.000]	[0.013]
(2) Anticipated	21.6	5.1	-0.86	-0.12
Standard Error	(4.8)***	(4.0)	(.24)***	(0.39)
p-value	[0.000]	[0.196]	[0.001]	[0.754]
CampusXLegXSession Dummies	Yes		Yes	
N (observations)	172		172	
Number of clusters	91		91	
R-squared	0.35		0.16	
PANEL B	**Multiple Hypothesis Testing Estimation**			
Difference vs. Control				
(1) Unanticipated	18.3	3.6	-0.93	-0.78
p-value	[0.000]***	[0.743]	[0.000]***	[0.031]**
(2) Anticipated	21.4	6.2	-0.85	-0.10
p-value	[0.000]***	[0.509]	[0.001]***	[0.796]
N (observations)	91	91	91	91

Notes. Panel A columns (1) and (2) show the results from the fixed effects specification (2) where the outcome is output. Columns (3) and (4) show the results for the same specification, but where the outcome is productivity. Standard errors are in parentheses and clustered by subject. p-values in square brackets shown for easier comparison with p-values yielded by MHT. Panel B shows the results from MHT estimation. Columns (1) and (3) show the differences in output and productivity between the treatments and CONTROL in session one from jointly testing output, productivity and time spent on the task for all conditions. Columns (2) and (4) show differences in output and productivity between the treatments and CONTROL in session two from jointly testing output, productivity, quits into session two and time spent on the task for all conditions. The number of observations in session two is 91 instead of 81 because quits into session two (represented by a dummy variable) is part of the joint estimation in session two. All tests are two-tailed. Significance levels highlighted next to the p-values in square brackets:**Significant at the 5% level; ***Significant at the 1% level. Note that the MHT method does not report standard errors but rather p-values based on the p-values from the actual data and from simulated samples. It also does not report the sample mean for the control group nor the p-value for the null hypothesis that the true mean for the control group is zero.

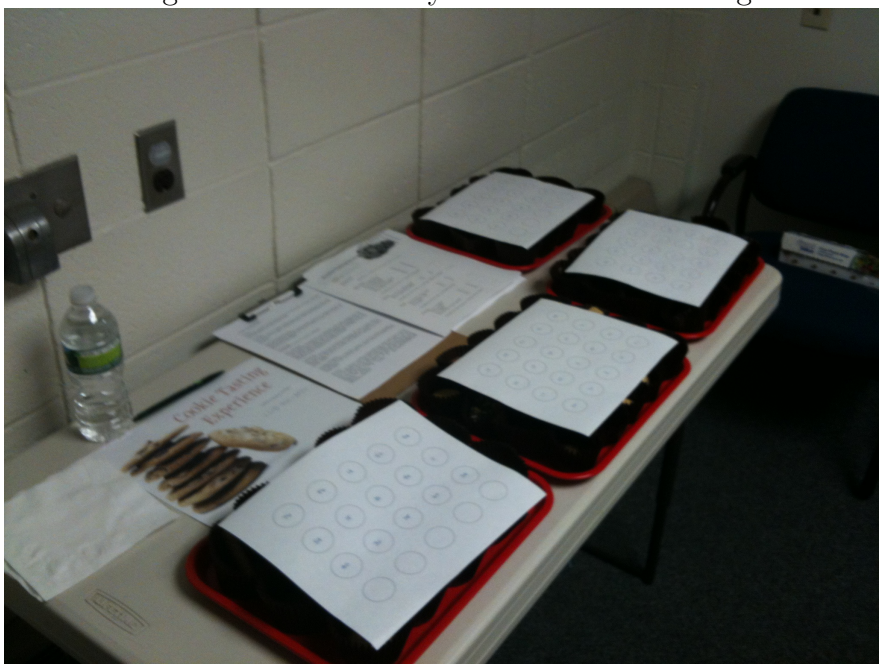
Table 5: Standard Deviation of Evaluation Ratings by Condition and Session

Dependent Variable:	**Standard Deviation of Ratings for a Given Cookie**			
Specification:	**Unadjusted OLS**		**Fixed Effects**	
	Session		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
	(1)	**(2)**	**(3)**	**(4)**
(1) Control	0.74	0.67	-	-
	(0.04)***	(0.03)***		
Difference vs. Control				
(2) Unanticipated	0.03	0.08	0.03	0.10
	(0.04)	(0.05)*	(0.04)	(0.05)*
(3) Anticipated	0.00	-0.03	0.01	-0.02
	(0.05)	(0.06)	(0.05)	(0.05)
CampusXlegXsessionXcookie Dummies	No		Yes	
N (cookieXsubject observations)	3,064		3,064	
Number of clusters	86		86	
R-squared	0.02		0.15	

Notes. The results for columns (1) and (2) are from specification in Section 5 without the fixed effects (unadjusted OLS). The results for columns (3) and (4) are from the full specification displayed in Section 5, that is, with the full set of fixed effects to estimate differences in dispersion for the *same cookie* within a campus, leg, and session. Standard errors are in parentheses and clustered by subject. *Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed.

A Online Appendix Tables and Figures

Figure A.1: Room Layout for Cookie Tasting



Notes: Picture depicting the layout of the tasting room for a tasting session. It shows the trays with the 70 cookies in the individual tasting cups (underneath the paper covers), napkins, water to drink while the subject tasted, the consent forms that subjects signed and the cookie evaluation sheets.

Table A.1: Quit Rates in the Second (Non-Reward) Session by Condition

Dependent Variable:	**Quit Rate into Session Two (Non-Reward Session)**		
Specification/estimation:	**Unadjusted OLS**	**Fixed Effects**	**MHT**
	(1)	**(2)**	**(3)**
(1) Control	0.08		
	(0.05)*		
Difference vs. Control			
(2) Unanticipated	-0.04	-0.05	-0.04
Standard Error	(0.06)	(0.08)	
p-value	[0.455]	[0.536]	[0.732]
(3) Anticipated	0.14	0.14	0.14
Standard Error	(0.09)	(0.10)	
p-value	[0.134]	[0.148]	[0.530]
CampusXLeg Dummies	No	Yes	-
N (observations)	91	91	91
R-squared	0.06	0.09	

Notes. Column (1) shows unadjusted OLS estimates for quit rates into the second session. Whether subject i, on campus c and in leg l quit is specified as $quit_{i,c,l} = \alpha_1 + \alpha_2 t_2 + \alpha_3 t_3 + \epsilon_{i,c,l}$ where *quit*=1 if the subject did not show up for session two and is zero otherwise; t_2 and t_3 identify differences in quit rates between UNANTICIPATED and ANTICIPATED, respectively, and CONTROL. The parameter α_1 identifies the baseline quit rate: the quit rate in CONTROL. Column (2) shows the estimated differences in quit rates between the treatments and CONTROL using fixed-effects estimation: $quit_{i,c,l} = \alpha_1 + \alpha_2 t_2 + \alpha_3 t_3 + \lambda_c \times \lambda_l + \epsilon_{i,c,l}$ where λ_c and λ_l are campus and leg fixed effects, respectively. Column (3) shows estimated differences between the treatments and CONTROL using the MHT procedure (jointly testing output, productivity, quits into session two and time spent on the task for all conditions in session two). Robust standard errors are in parentheses. *Significant at the 10% level. p-values in square brackets. We show p-values in brackets in columns (1) and (2) for ease of comparison with the p-values from MHT in column (3).

Table A.2: Average Number of Evaluations and Time Spent on the Task in the Second (Non-Reward) Session Considering Quitters as Supplying Zero Output and Time

Dependent Variable:	**Output: Number of Cookies Tasted and Evaluated Assuming Quitters in Session Two Supply Zero Output**		**Time: Minutes Spent Tasting and Evaluating Assuming Quitters in Session Two Supply Zero Time**	
	Session		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
PANEL A	**Specification: Unadjusted OLS**			
	(1)	**(2)**	**(3)**	**(4)**
Control	29.7	23.5	79.4	56.7
	(3.2)***	(2.7)***	(6.2)***	(6.0)***
Difference vs. Control				
(1) Unanticipated	18.3	4.6	22.7	-2.6
	(4.9)***	(3.7)	(9.5)**	(7.5)
(2) Anticipated	21.4	1.2	32.2	3.3
	(4.8)***	(4.5)	(9.3)***	(10.8)
CampusXLegXSession Dummies	No		No	
N (observations)	182		182	
Number of clusters	91		91	
R-squared	0.28		0.27	
PANEL B	**Specification: Fixed Effects**			
	(1)	**(2)**	**(3)**	**(4)**
Difference vs. Control				
(1) Unanticipated	15.8	2.9	16.2	-6.2
	(5.3)***	(4.3)	(10.7)	(8.8)
(2) Anticipated	21.6	0.8	32.3	1.9
	(4.8)***	(4.4)	(8.9)***	(10.4)
CampusXLegXSession Dummies	Yes		Yes	
N (observations)	182		182	
Number of clusters	91		91	
R-squared	0.35		0.36	

Notes. Panel A, columns (1) and (2) show the results of the unadjusted OLS specification (1) where the outcome is output and we assume that quitters into the second session would have supplied zero output. Panel A, columns (3) and (4) show the results for the same specification, but where the outcome is time spent on the task and we assume that quitters into the second session would have supplied zero time. Panel B shows the same information as panel A, but using the fixed effects specification (2). Standard errors in parentheses and clustered by subject. **Significant at the 5% level; ***Significant at the 1% level.

Table A.3: Time Spent on the Task (Tasting and Evaluating) by Condition and Session

Dependent Variable:	**Time Spent on the Task**					
Specification:	**Unadjusted OLS**		**Fixed Effects**		**MHT**	
	Session		**Session**		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
(1) Control	79.4	61.7	-	-		
	(6.2)***	(5.8)***				
Difference vs. Control						
(2) Unanticipated	22.7	-5.6	16.2	-11.5	22.7	-5.6
Standard Error	(9.5)**	(7.1)	(10.7)	(8.1)		
p-value	[0.019]	[0.437]	[0.133]	[0.157]	[0.024]**	[0.832]
(3) Anticipated	32.2	15.4	32.3	12.3	32.2	15.4
Standard Error	(9.3)***	(10.1)	(8.9)***	(9.1)		
p-value	[0.001]	[0.131]	[0.000]	[0.178]	[0.002]***	[0.559]
CampusXLegXSession Dummies	No		Yes			
N (observations)	172		172		91	91
Number of clusters	91		91			
R-squared	0.25		0.37			

Notes. Columns (1) and (2) show results from the unadjusted OLS specification (1) where the outcome is time spent on the task. Standard errors are in parentheses and clustered by subject. Columns (3) and (4) show the results from the fixed effects specification (2). Columns (5) and (6) show the results from the MHT estimation. Column (5) shows the difference in time spent on the task between the treatments and CONTROL in session one from jointly testing output, productivity, and time spent on the task for all conditions. Column (6) shows the difference in time spent on the task between the treatments and CONTROL in session two from jointly testing output, productivity, quits into session two and time spent on the task for all conditions. The number of observations in session two for MHT are 91 instead of 81 because quits into session two (represented by a dummy variable) is part of the joint estimation in session two. Standard errors in columns (1)-(4) are in parentheses and clustered by subject; p-values in square brackets shown for easier comparison with p-values yielded by MHT. **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed.

Table A.4: Comparison of Main Outcomes in Full and Restricted Sample for the Analysis of Ratings Dispersion

Specification:	**Fixed Effects**							
Dependent Variables:	**Number of Cookies Tasted and Evaluated per Subject**				**Minutes per Cookie Tasted and Evaluated per Subject**			
Sample:	Full Sample		**Restricted Sample**		Full Sample		**Restricted Sample**	
	Session		**Session**		Session		**Session**	
	One	Two	**One**	**Two**	One	Two	**One**	**Two**
	(1)	(2)	**(3)**	**(4)**	(5)	(6)	**(7)**	**(8)**
Diff. vs. Control								
Unanticipated	15.8	1.0	14.6	0.0	-0.95	-0.80	-0.89	-0.81
	(5.3)***	(4.1)	(5.3)***	(4.5)	(.26)***	(0.32)**	(0.26)***	(0.35)**
Anticipated	21.6	5.1	19.8	4.7	-0.86	-0.12	-0.81	-0.16
	(4.8)***	(4.0)	(4.9)***	(4.4)	(.24)***	(0.39)	(0.25)***	(0.40)
CampusXLegXSession Dummies	Yes		Yes		Yes		Yes	
N (subjectXsession obs.)	172		155		172		155	
Number of clusters	91		86		91		86	
R-squared	0.35		0.31		0.16		0.15	

Notes. The results in all columns are from the fixed-effects specification (2) which is more conservative than unadjusted OLS. Columns (1), (2), (5), and (6) replicate the prior analysis on the full sample. Columns (3), (4), (7) and (8) do the same analysis but only on the sample of tasters used in the analysis of ratings dispersion. This table documents that the results are similar for both samples. Standard errors are in parentheses and clustered by subject. *Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed.

B Why learning is unlikely to explain the excess productivity in Unanticipated during the second (non-reward) session

In this section we explore whether the boost in productivity caused by the piece rate in the first session led to a permanent increase in productivity through learning, which subsequently persists even in the absence of pay. Though Control controls for such learning, we consider whether working at high levels of productivity in the first session could have led tasters in Unanticipated to reach a permanently higher productivity threshold, which persists in the absence of pay.

Three pieces of evidence suggest this was not the case. First, the piece rate yields similar boosts in productivity in the first (reward) session for the subjects in Unanticipated and for the non-quitters in Anticipated (subjects in Anticipated in session one who returned in session two). However, during the second (non-reward) session, only those in Unanticipated continue working at a faster rate than those in Control: Anticipated non-quitters go back to working at a rate similar to Control's. Therefore, it does not seem that higher productivity during session one leads to higher permanent productivity that endures despite the removal of pay.

Specifically, Table B.1, Panel B, columns (1) and (3), document that all subjects in Unanticipated have similar productivity to the 78% Anticipated non-quitters during the first session. The two groups tasted and rated each cookie 0.93 and 0.85 minutes faster, respectively, than Control using the unadjusted OLS specification (1) (the p-value for the difference is 0.59) and 0.94 and 0.85 minutes faster, respectively, than Control using the fixed-effects specification (2) (the p-value for the difference is 0.61). However, in the absence of the piece rate, these two groups behaved very differently. Whereas productivity in Unanticipated still exceeded that in Control by 0.78 and 0.80 minutes per cookie

using, respectively, unadjusted OLS and fixed effects (statistically significant magnitudes at the 5% level), productivity for non-quitters in ANTICIPATED declined to a level similar to that of CONTROL (a statistically insignificant difference of -0.10 and 0.12 minutes per cookie using, respectively, unadjusted OLS and fixed effects), as one would expect in the absence of pay (Panel B, columns (2) and (4)). Further, this 0.68 difference was statistically significant (p-values of 0.026 and 0.038 in, respectively, the unadjusted OLS and fixed-effects specifications).

Our graphical evidence further buttresses the point above. The similarity in average productivity during the first (reward) session between those in UNANTICIPATED and the non-quitters in ANTICIPATED is not due to a few outliers but to the behavior in the whole distribution of tasters' outcomes (bottom left Figure B.1). However, in the absence of the piece rate, these two groups behaved very differently: the top right panel shows that the whole CDF of productivity for UNANTICIPATED continues to the left of CONTROL's whereas the CDF for ANTICIPATED non-quitters basically overlaps with that of CONTROL, as one would expect in the absence of pay.

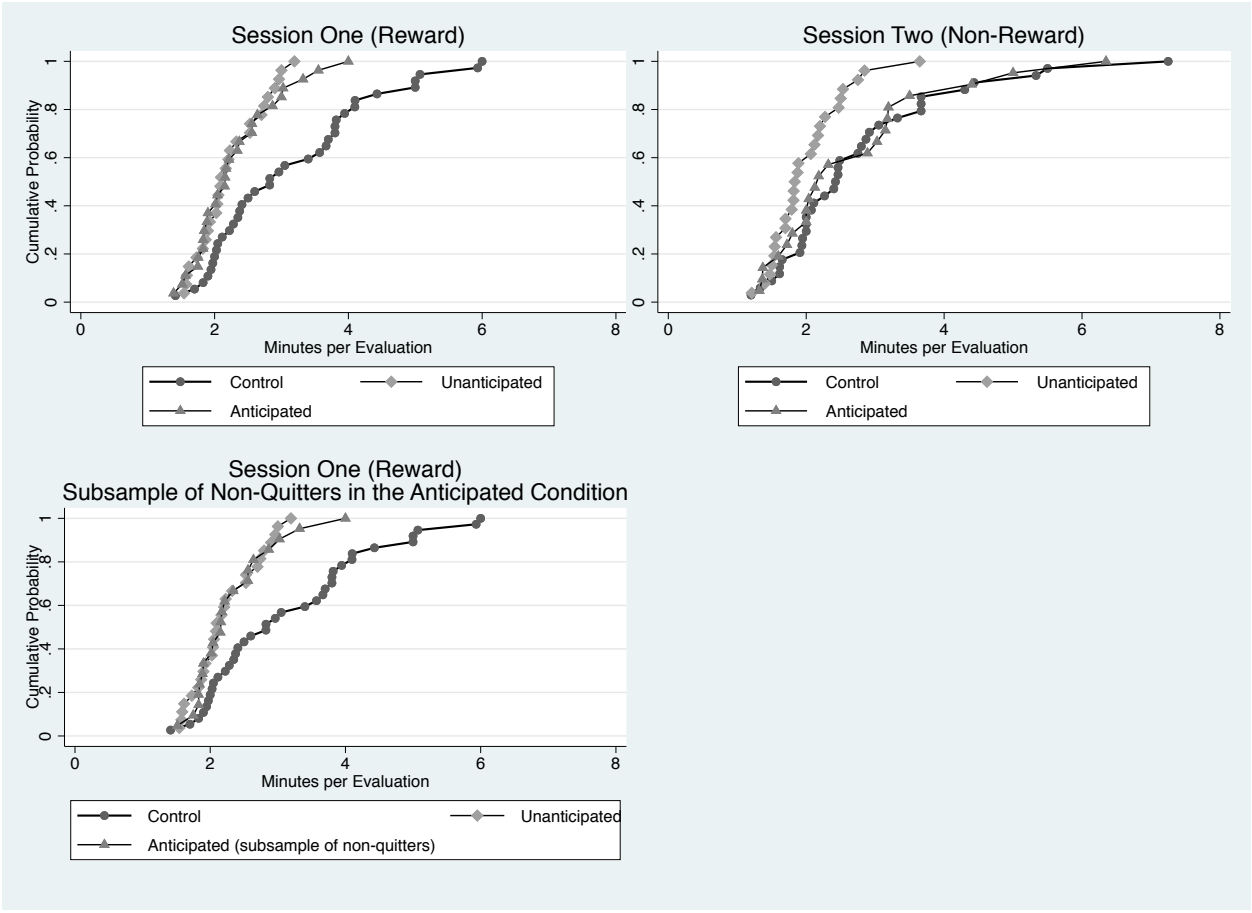
Second non-quitters in ANTICIPATED turned in slightly more evaluations than those in UNANTICIPATED during the reward session (3.1, the difference between 21.4 and 18.3, in Panel A, Column (1)) suggesting they could have had even more experience with tasting and evaluating, though this difference is not statistically significant (p-value of 0.58). The fixed-effects estimation in Column (3) also shows that the non-quitters in ANTICIPATED also turned in 5.7 more evaluations than those in UNANTICIPATED (also not statistically significant, with a p-value of 0.27). But only UNANTICIPATED subjects continued working faster than those in CONTROL in the subsequent non-reward session, as documented above.

Table B.1: Output and Productivity of Non-Quitters in ANTICIPATED

Specification:	**Unadjusted OLS**		**Fixed Effects**	
PANEL A:	**Dependent Variable: Number of Cookies Tasted and Evaluated per Subject**			
	Session		**Session**	
	One (Reward)	**Two (Non-Reward)**	**One (Reward)**	**Two (Non-Reward)**
Sample:	**All subjects in Session One except Session-Two Quitters in the Anticipated Condition**	**Full Sample in Session Two**	**All subjects in Session One except Session-Two Quitters in the Anticipated Condition**	**Full Sample in Session Two**
	(1)	**(2)**	**(3)**	**(4)**
(1) Control	29.7	25.5	-	-
	(3.2)***	(2.6)***		
Diff. vs. Control				
(2) Unanticipated	18.3	3.6	15.6	1.0
	(4.9)***	(3.5)	(5.3)***	(4.1)
(3) Anticipated	21.4	6.2	21.3	5.1
	(5.2)***	(4.3)	(5.0)***	(4.0)
CampusXLegXSession Dummies	No		Yes	
N (observations)	166		166	
Number of clusters	85		85	
R-squared	0.25		0.35	
PANEL B:	**Dependent Variable: Minutes per Cookie Tasted and Evaluated per Subject**			
	(1)	**(2)**	**(3)**	**(4)**
(1) Control	3.15	2.79	-	-
	(0.20)***	(0.23)***		
Diff. vs. Control				
(2) Unanticipated	-0.93	-0.78	-0.94	-0.80
	(0.22)***	(0.25)***	(0.26)***	(0.32)**
(3) Anticipated	-0.84	-0.10	-0.85	-0.12
	(0.24)***	(0.36)	(0.25)***	(0.39)
CampusXLegXSession Dummies	No		Yes	
N (observations)	166		166	
Number of clusters	85		85	
R-squared	0.15		0.16	

Notes. Panel A, columns (1) and (2), show the results for output in both sessions using the unadjusted OLS specification (1). These estimates are for the whole sample except the six tasters in ANTICIPATED who later quit into session two. Therefore, our sample comprises 166 subjectXsession observations (85 subjects in the first session and 81 in the second session) instead of the usual 172 (91 subjects in session one and 81 subjects in session two). Panel (A), columns (3) and (4), show the same analysis but using the fixed-effects specification 2. Panel B shows the same analysis as that in Panel A but for productivity. Standard errors are in parentheses and clustered by subject. **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed.

Figure B.1: Productivity CDFs by Condition and Session—Session One Behavior of Non-Quitters in Anticipated



Notes: the top left panel shows the cumulative distribution functions (CDFs) of productivity across the three conditions for the first (reward) session for the whole sample that came to the first session (91 subjects in total). The bottom left panel shows the same information except that it excludes the 6 subjects in Anticipated who did not come to the second session (thus the productivity CDF for this condition is for the sample of 21 non-quitters). The top right panel shows the CDFs of productivity across the three conditions for the second (non-reward) session for the whole sample that came to the second session (81 subjects).

Third, Anticipated quitters were not faster raters than non-quitters. It could be that the faster raters in the first (reward) session of Anticipated were the ones who quit and that is why the average productivity in the second session for non-quitters is lower, at the level

of CONTROL. However, Appendix Table B.2 shows that during the first session quitters and non-quitters displayed the same productivity (2.3 minutes/evaluation, p-value=0.64) and the same output (51 cookies, p-value=0.90).

Table B.2: Average Output and Productivity During the First (Reward) Session for Quitters and Non-Quitters in ANTICIPATED

	Session One (Reward Session)			
	Full Sample	Non-Quitters	Quitters	Outcomes Non-Quitters=Quitters? p-value
	(1)	(2)	(3)	(4)
Anticipated Condition				
(1) Number of Subjects	27	21	6	
(2) Average Output (Average Number of Cookies Tasted and Evaluated)	51	51	51	0.90
(3) Average Productivity (Average Minutes per Cookie Tasted and Evaluated)	2.3	2.3	2.3	0.64

Given all the evidence laid out in this section, learning to work at higher levels of productivity is unlikely to account for the extra productivity of those in UNANTICIPATED in session two after their pay was unexpectedly withdrawn.

C Correlation Between Speed in Evaluation Completion and Dispersion in Ratings

Studies in psychology and education have shown that changes in the dispersion of subjective answers—reducing dispersion by "straight-lining" or increasing dispersion by answering more randomly—are usually associated with faster response times. We document that, in our setting, there is a negative correlation between the time spent filling out each evaluation and dispersion, suggesting that agents chose the increased-dispersion route to economize on effort.

Empirical method. To document this correlation, we estimated the dispersion (measured by the standard deviation as before) of cookie k, tasted by person i, in campus c, leg l and session s, for the same sample of cookies in Section 5, as follows:

$$sd_{k,i,c,l,s} = \theta_1 + \theta_2 m + \sum_{\tau=1}^{70} \beta_\tau d_\tau + \psi_i \times \psi_s + \lambda_c \times \lambda_l \times \lambda_s \times \lambda_k + \epsilon_{i,k,s,t,c} \tag{3}$$

m is the minutes spent tasting each cookie, the difference between the time at which a subject started eating a cookie and the time at which he or she started eating the next cookie.[34] The average time tasting each cookie is 2.03 minutes in this restricted sample, which is similar, though slightly smaller than the 2.6 minutes per cookie for the overall sample, reported in the descriptive statistics. The coefficient θ_2, which captures the partial correlation between the time to fill out each evaluation and the dispersion in its ratings, will be the parameter of interest throughout.

Other variables control for additional determinants of dispersion and the time spent per evaluation. The sum $\sum_{\tau=1}^{70} \beta_\tau d_\tau$ corresponds to dummies that control for a cookie's ordinal position—when, relative to the order cookies, it was tasted—because subjects might both

[34] We used this measure because it is a closer approximation to the time subjects spent tasting each cookie, given that, for example, the end time was missing on several evaluation sheets or matched the start time.

spend less time and increase their dispersion on later evaluations.

The term $\psi_i \times \psi_s$ is the interaction of individual and session fixed effects, which control for unobserved individual time-invariant differences in ability. When combined with experience with the task, these can affect the time spent tasting each cookie and the dispersion in its ratings. For example, some cookies may be tasted by subjects who have higher ability for the task—e.g., can rate cookies faster without increasing their dispersion—and who get differentially better at it from one session to the next.

As in specification (2), the term $\lambda_c \times \lambda_l \times \lambda_s \times \lambda_k$ controls for unobserved time-invariant campus, leg, session, and cookie unobservables, which may affect speed and ratings' dispersion. For example, a cookie may have a good flavor that takes time to develop (slow Start). Thus, it takes both a longer time to rate and has higher dispersion, due to its high-scoring Flavor but low-scoring Start.

Table C.1 documents that for a given cookie, tasted in a given campus, leg, and session, the longer it takes to taste, the lower its dispersion.

Column (1) shows that a one minute increase in the time it takes to evaluate a cookie increases the standard deviation of the ratings by 0.0027. Column (2) adds controls for the time-invariant unobserved ability of the taster, which both increases the fit of the model to 24% and starts showing the inverse correlation between speed and dispersion: an increase in the time spent per evaluation decreases its standard deviation in ratings by 0.0027. Column (3) adds controls for a cookie's ordinal position, increasing the fit of the model to 26% and yielding an estimate of -0.0029. Column (4) adds the experience of the taster, increasing the magnitude of this estimate to -0.0034, while increasing the fit of the model to 29%.[35] Column (5) adds controls for the interaction of campus, leg, session, and cookie fixed effects, which increase the fit of the model to 40%. Therefore, holding constant the cookie tasted

[35]The interaction of subject and session fixed effects subsumes individual standalone fixed effects.

within a given campus, leg, and session, the ability and experience of the rater of that cookie, and when that cookie was tasted, an increase in one minute in the time it takes to taste that cookie decreases the standard deviation in its ratings by -0.0050. Despite the large number of controls, this estimate is statistically significant at the 10% level.

Table C.1: Correlation Between the Standard Deviation and Time Spent per Evaluation

Dependent Variable: Standard Deviation of Ratings in an Cookie Evaluation					
	(1)	(2)	(3)	(4)	(5)
Minutes spent in evaluation	0.0027	-0.0027	-0.0029	-0.0034	-0.0050
	(0.0039)	(0.0030)	(0.0031)	(0.0033)	(0.0029)*
Controls:					
Subject fixed effects	-	Yes	Yes	-	-
Cookie order fixed effects	-	-	Yes	Yes	Yes
SubjectXsession fixed effects	-	-	-	Yes	Yes
CampusXlegXsessionXcookie fixed effects	-	-	-	-	Yes
R-squared	0.00	0.24	0.26	0.29	0.40
N (cookieXsubjectXsession obs.)	2,961	2,961	2,961	2,961	2,961
Additional information					
Number of subjectXsession observations	155	155	155	155	155

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level. All tests are two-tailed. This sample has 103 fewer observations (2,961 versus 3,064 in Table 5) because the starting times for the tasting of each cookie where unreadable on evaluations for some cookies.

D Additional Tables

Table D.1: Summary Statistics per Campus, Session and Condition on Output and Productivity

	Conditions					
	Control		Unanticipated		Anticipated	
	Session One	Session Two	Session One	Session Two	Session One	Session Two
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Campus A					
Number of subjects	29	26	24	23	23	18
Average number of evaluations	31	26	47	29	53	33
Average minutes per evaluation	3.2	2.8	2.2	2.0	2.2	2.7
	Panel B: Campus B					
Number of subjects	8	8	3	3	4	3
Average number of evaluations	25	25	60	33	41	21
Average minutes per evaluation	3.1	2.7	2.1	1.8	2.9	2.5
Total number of subjects	37	34	27	26	27	21

Table D.2: Summary Statistics per Leg on Output and Productivity

	Conditions					
	Control		Unanticipated		Anticipated	
	Session One	Session Two	Session One	Session Two	Session One	Session Two
Leg 1						
Number of subjects	7	7	2	1	6	4
Average number of evaluations	23	23	19	24	40	22
Average minutes per evaluation	3.3	2.9	2.7	1.8	2.0	2.9
Leg 2						
Number of subjects	13	12	8	8	12	10
Average number of evaluations	31	27	54	32	50	35
Average minutes per evaluation	3.1	2.8	2.0	2.0	2.6	2.7
Leg 3						
Number of subjects	11	10	12	12	6	4
Average number of evaluations	30	23	55	28	57	26
Average minutes per evaluation	3.1	2.8	2.2	1.9	2.3	2.7
Leg 4						
Number of subjects	6	5	5	5	3	3
Average number of evaluations	33	31	33	28	67	41
Average minutes per evaluation	3.2	2.8	2.5	2.4	1.9	2.3
Number of subjects	37	34	27	26	27	21

E Review of the Psychology Literature

E.1 Review of 79 papers on the effect of pay on performance on enjoyable tasks

This section summarizes over 100 experiments described in 79 papers in psychology on the effect of pay on performance on enjoyable tasks. It describes the papers along several dimensions:

- Column (1): Paper id
- Column (2): Paper authors
- Column (3): Number of citations in Google Scholar on August 2019
- Column (4): Studies/Experiments in the paper
- Column (5): Whether the condition is a treatment ("Experimental") or a control condition ("Control"). If there are several control conditions, these are numbered as "Control I", "Control II", etc.
- Column (6): Description of the payment and/or manipulation
- Column (7): Types of tasks in each experiment
- Column (8): Types of subjects
- Column (9): Number of subjects in the treatment conditions (N_t)
- Column (10): Number of subjects in the control conditions (N_c)
- Column (11): Whether the experiment is a laboratory ("lab") or a field experiment

- Column (12): Whether the experiment was held a laboratory (in a laboratory space, classroom or trailer) or in the field (natural versus artificial field experiment). If the experiment was
  - held in a regular laboratory space, it has a dash "-".
  - held in a trailer that is taken to an specific location outside the laboratory (i.e., "moving laboratory") it is coded as "Trailer".
  - held in a classroom, it is coded as "Classroom".
  - in the field, using an current task in the subjects' environment (e.g., Deci (1971) with the newspaper headlines field experiment) it is coded as "Current".
  - in the field, using a task created for the purpose of the experiment, it is coded as "New".
- Column (13): Outcome(s) measured in the experiment or study
- Column (14): Explanation of the outcome(s) measured in the experiment or study

The most notable finding of this review, outlined in Table 1, is that of the over 100 experiments described in these papers, none has jointly reported output, productivity, and time spent on the task. Further, the median sample size per condition was 15 subjects (see row 80).

E.2 Review of 14 additional papers in psychology

We also examined an additional 14 papers in psychology on whether feedback on outcomes harms performance on enjoyable tasks. This research studies, for example, whether verbal feedback can be perceived as controlling thus crowding out intrinsic interest in the task.

Though these papers are outside the scope of our research, which focuses on pay, for completeness, we also describe these experiments. Table 2 documents that, again, no experiment has jointly reported output, productivity and time spent on the task. And the median sample size per condition was limited here as well, at 18 subjects (see row 15).

TABLE I

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	Amabile, T. M., Hennessey, B. A., & Grossman, B. S. (1986)	790	**S1**	Experimental	Reward (taking two pictures with an instant camera)	(1) Test of Artistic Creativity, (2) Test of Verbal Creativity, (3) Problem-solving test of creativity	Students (boys and girls) ranging in age from 5 to 10 years	115		Field	New	Quality Time spent on task	Artistic Creativity, Verbal Creativity, and Puzzle Creativity scored on 5-point scale Mean number of minutes spent on each task
			S2	Experimental		(1) Collage-making, (2) Story Telling		80		Field	New	Quality Self-reported satisfaction	Student output rated by judges on a 40-point creativity scale; Students rated their own satisfaction after completion of task
			S3	Experimental	Monetary Reward	Collage Making	Undergraduate adult women	60		Lab	-	Quality Self-reported enjoyment and creativity	Student output rated on a 40-point creativity scale by judges, Students rated their own enjoyment and creativity after completion of the task
2	Anderson, R., Manoogian, S. T., & Reznick, J. S. (1976)	395	**E1**	Experimental	Monetary Reward	Free-style drawing with multicolored, felt-tipped pens	Lower socioeconomic preschool children	19		Field	New	Time spent on task	Mean number of minutes spent drawing
					Symbolic Reward			18					
					Positive Verbal Reinforcement			17					
				Control I	(Experimenter present but ignored the child)				18				
			E2	Control I	(Experimenter present but ignored the child)				9	Field	New	Time spent on task	Mean number of minutes spent drawing
				Control II	(Experimenter present and attentive but not directly reinforcing the child for drawing)				9				
				Control III	(Time control with no treatment between the pretest and posttest measurements)				9				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
3	Arkes, H. R. (1979).	25	**E**	Experimental	Monetary reward, Easy task	Puzzle-solving (Half of subjects did a difficult task and half an easy task)	Undergraduate males	16		Lab	-	Time spent on task	Mean number of seconds spent touching blocks
				Experimental	Monetary reward, difficult Task			16				Self-reported competence; Self-reported satisfaction	Subjects used Likert scales to assess their own competence and satisfaction after completing a few puzzles
				Control	No reward, easy task				16				
				Control	No reward, difficult task				16				
4	Arnold, H. J. (1976).	135	**E**	Experimental	Monetary reward (On session 1)	Star trek videogame playing	Undergraduate males and females	17		Lab	-	Non-quits/Quits	Number of people/percentage of people who participated in each subsequent session
					Monetary reward (On session 2)			17				Self-reported Satisfaction/Enjoy-ment	Subjects rated their own satisfaction/enjoyment at the conclusion of each session
					Monetary reward (On session 3)			19				Self-reported Competence	Subjects rated their own competence at the conclusion of each session
5	Arnold, H. J. (1985).	101	**E**	Control	No reward	Star trek videogame playing	Undergraduate males and females		16	Lab	-	Output	Number and percentage of enemy starships destroyed
				Experimental	Fixed monetary reward			13				Self-reported enjoyment/compe-tence; Self-reported attribution of performance	Subjects used Likert scales to assess their own competence and satisfaction at the conclusion of the experiment; Subjects rated several variables' effect on their performance at the conclusion of the experiment
				Experimental	Performance contingent monetary reward			13				Willingness to supply further work	Number of sessions subject signed up for beyond the initial three
6	Boal, K. B Cummings, L. L. (1981).	59	**E**	Experimental	Performance contingent monetary reward	Calculating and transcribing data from a property rental record onto a coding form	Caucasian adults	21		Field	New	Time spent on task, time late for work; time on break	Mean number of minutes participant was late to work/took a break on the job/time at which the subject quit working
				Experimental	Fixed monetary reward			21				Self-reported assessment	Participants responded to a questionnaire asking them to rate their performance on several metrics, including competence, self-determination, external locus of causality
				Control	No Reward				21				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
7	Boggiano, A. K., Harackiewicz, J. M., Bessette, J. M., & Main, D. S. (1985).	19	**E**	Experimental	Task contingent reward (5 colorful stickers), More Salient	Puzzle/Maze Solving	Male and female kindergarteners	13		Lab	Trailer	Time spent on task	Mean number of seconds spent solving puzzles during "free" time
				Experimental	Task contingent reward (5 colorful stickers), Less Salient			13					
				Experimental	Performance contingent reward (5 colorful stickers), More Salient			13					
				Experimental	Performance contingent reward (5 colorful stickers), Less Salient			13					
				Control	No reward				13				
8	Boggiano, A. K., & Hertel, P. T. (1983).	25	**E**	Experimental	High Interest, Monetary Reward	Memory task	Undergraduate males and females	15		Lab	-	Recall	Mean number of positive, neutral and negative words recalled in a memory exercise
				Experimental	High Interest, No Reward			15					
				Experimental	Low Interest, Monetary Reward			15					
				Experimental	Low Interest, No Reward			15					
				Control	Monetary Reward				15			Interest	Participants rated their interest in the forthcoming task
				Control	No reward				15				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
9	Boggiano, A. K., & Ruble, D. N. (1979).	251	**E**	Experimental	Task contingent reward (2 Hershey's Kisses)	"Finding the Hidden pictures" Game	Preschool and middle elementary school children	48		Lab	Trailer	Time spent on task	Percent of time (6 minutes) spent on target task
				Experimental	Performance contingent reward (2 Hershey's Kisses)			48				Output	Number of select figures circled
				Control	No reward				48				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
10	Boggiano, A. K., Ruble, D. N., & Pittman, T. S. (1982).	53	**E**	Experimental	Reward (Spalding ball); competence Information; easy task	"Finding the Hidden pictures" Game	Fourth-grade children (predominantly white, middle-class backgrounds)	14		Lab	Trailer	Self-reported expectation of fun; Self-reported difficulty and competency	Students rated how fun they expected the forthcoming task to be; Students rated the just-completed task in terms of difficulty and how good they were at it
				Experimental	Reward (Spalding ball); competence information; medium task			13					
				Experimental	Reward (Spalding ball); competence information; difficult task			14				Time spent on task	Proportion of time (9 minutes) spent on target task
				Experimental	Reward (Spalding ball); no competence Information; easy task			13					
				Experimental	Reward (Spalding ball); no competence Information; medium task			14					
				Experimental	Reward (Spalding ball); no competence Information; difficult task			14					
				Control	No reward; competence Information; easy task				14				
				Control	No reward; competence Information; medium task				13				
				Control	No reward; competence Information; difficult task				14				
				Control	No reward; no competence Information; easy task				14				
				Control	No reward; no competence Information; medium task				14				
				Control	No reward; no competence Information; difficult task				14				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
11	Brennan, T. P., & Glover, J. A. (1980).	20	**E**	Experimental	Reward Group (point bonus on next exam for working on the puzzle)	Puzzle-solving (SOMA)	Undergraduate males and females	19		Lab	-	Time spent on task (three rounds)	Researcher covertly recorded what activity the participant picked and for how long (in seconds) he/she/they did it
				Experimental	Directions Group (No reward for working on the puzzle)			19					
				Control	Control Group (No reward and not asked to work on the puzzle)				20				
12	Brockner, J., & Vasta, R. (1981).	21	**E**	Experimental	Monetary reward	Puzzle-solving (SOMA)	Male Undergraduates	22		Lab	-	Time spent on task; Output	Number of 15-second intervals spent occupied with the puzzle (out of 32); Number of puzzles solved (out of 4)
				Control	No reward				30			Self-reported motivation; Self-reported interest; Self-reported performance; Self-reported hypothesis	On a scale of 1-41: Subjects rated their intrinsic/extrinsic motivation; their interest in the task; their performance in the task. Subjects were also asked what they thought was the true purpose of the study
13	Calder, B. J., & Staw, B. M. (1975).	691	**E**	Experimental	Monetary reward	Solving 15 jigsaw-type blank puzzles	Male Undergraduates	10		Lab	-	Self-reported rating of puzzle task; Self-reported satisfaction	Mean ratings of the puzzle task using 7-point scales and 13 adjective pairs; After task completion, subjects rated how enjoyable they found the task on a 17-point scale
				Control	No reward				10				
				Experimental	Monetary reward	Solving 15 jigsaw-type picture puzzles		10				Self-reported perception of situation	Mean scores of awareness of extrinsic/intrinsic motivation and perceived effort (out of 11)
				Control	No reward				10			Willingness to supply more work	Average number of additional minutes of volunteer labor participants signed up for

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
14	Crino, M. D., & White, M. C. (1982).	10	**E**	Experimental	(Contingent vs. Noncontingent feedback) X (Single reinforcement vs. Multiple reinforcement)	Solving blank puzzle (low motivation)	Male Undergraduates	20		Lab	-	Self-reported interest	Interest assessed by mean score on a series of 7-point semantic differentials scales
				Control	No feedback				5			Self-reported interest; feelings about task ; and enjoyment	Scored using nine-point faces scales at the conclusion of the experiment
				Experimental	(Contingent vs. Noncontingent feedback) X (Single reinforcement vs. Multiple reinforcement)	Solving picture puzzle (high motivation)		20				Willingness to supply further work	Number of minutes (from 0-120) the subject volunteered for a subsequent siimilar study
				Control	No feedback				5				
15	Davidson, P., & Bucher, B. (1978).	40	**E**	Experimental	Rewarded with tokens for playing with Clown or House Machine during reinforcement session; reward for putting the marble in the correct hole	Operating two child-operating teaching machines and a marble dropping apparatus	Children (4-5 years old)	3		Lab	Trailer	Output; Productivity; Self-reported Choice; Effectiveness of motivator	Mean number of responses to the clown or house apparatus in each session; Average number of seconds the subject took to respond on the machine; Respondents stated which machine they preferred at the conclusion of the experiment; Effectiveness of motivator assessed by taking the percentage of responses made to the reinforced hole (S+) after the first reinforced response

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
16	Daniel, T. L., & Esser, J. K. (1980).	103	**E**	Experimental	Monetary reward	Solving blank puzzle (low interest; high task Structure)	Undergraduate males and females	8		Lab	-	Times spent on task	Average number of seconds spent on task
				Experimental	Monetary reward	Solving blank puzzle (low interest; low task structure)		8					
				Control	No reward	Solving blank puzzle (low interest; high task structure)			8			Self-reported behavioral measure	Participants rated the task using eight 9-point graphic rating scales with bipolar adjectival anchors
				Control	No reward	Solving blank puzzle (low interest; low task structure)			8				
				Experimental	Monetary reward	Solving picture puzzle (high interest; high task structure)		8				Self-reported willingness to engage with task	Participants rated their interest in participating in similar study on two 5-point graphic rating scales
				Experimental	Monetary reward	Solving picture puzzle (high interest; low task structure)		8					
				Control	No reward	Solving picture puzzle (high interest; high task structure)			8			Productivity	Average number of seconds it took to complete 10 puzzles
				Control	No reward	Solving picture puzzle (high interest; low task structure)			8				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Experimental	Cognitive group 1 (produced one dichotomy and failed subsequent tasks)	Playing at centers. Center 1 (sorting), Center 2 (class inclusion) and/or Center 3 (combinatorial reasoning)	Children from kindergarten and grades 1, 2, and 4	30				Time spent on each task	Average number of seconds participant chose to spend at each center
			E1	Experimental	Cognitive group 2 (produced 3 dichotomies and failed subsequent tasks)			30		Lab	Trailer	Self-reported interest and difficulty	After completion of study, participants ranked each task on a 5-point scale of interest and difficulty
				Experimental	Cognitive group 3 (only failed combinatorial reasoning)			30					
				Control	No reward/Cognitive Group 1				10				
				Control	No reward/Cognitive Group 2				10				
				Control	No reward/Cognitive Group 3				10			Time spent on each task	Average number of seconds participant chose to spend at each center
17	Danner, F. W., & Lonky, E. (1981).	240		Experimental	Verbal praise (positive verbal feedback)/Cognitive Group 1			10					
				Experimental	Verbal praise (positive verbal feedback)/Cognitive Group 2	Playing at centers. Center 1 (sorting), Center 2 (class inclusion) and/or Center 3 (combinatorial reasoning)	Children from kindergarten and grades 1, 2, and 5. (Each group had aassigned equal number of children from each cognitive conditions)	10					
			E2	Experimental	Verbal praise (positive verbal feedback)/Cognitive Group 3			10		Lab	Trailer	Self-reported interest and difficulty	After completion of study, participants ranked each task on a 5-point scale of interest and difficulty
				Experimental	Extrinsic reward (good work certificate)/Cognitive Group 1			10					
				Experimental	Extrinsic reward (good work certificate)/Cognitive Group 2			10					
				Experimental	Extrinsic reward (good work certificate)/Cognitive Group 3			10					

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
18	Deci, E. L. (1971).	5,437	**E1**	Experimental	Monetary reward	Puzzle-solving (SOMA)	Undergraduate males and females	12		Lab	-	Time spent on task	Measure seconds spent solving puzzles;
				Control	No reward				12			Self-reported enjoyment	Self-rating of enjoyment of task on a 9-point scale
			E2	Experimental	Monetary reward	Writing headlines (newspaper)	College students	4		Field	Current	Productivity	Measure average minutes per headline written
				Control	No reward				4			Quits	Measure percentage of people who did not come to future sessions
			E3	Experimental	Verbal Reward (positive feedback)	Puzzle-solving (SOMA)	Undergraduate males and females	12		Lab	-	Time spent on task Enjoyment	Measure seconds of free choice time spent solving puzzles AND Enjoyment self-assessment (scored-based system, out of 9)
				Control	No reward				12			Arts vs Technical students – time spent on the task	Measure seconds of free choice time spent solving puzzles, compare based on the student's field of study
19	Deci, E. L. (1972).	1,012		Experimental	No verbal Reinforcement and Money after	Puzzle-solving (SOMA)	Undergraduate males and females	16				Time spent on task	Average number of seconds of free time spent working on the puzzle
				Control	No verbal Reinforcement and no Money				16				
			E	Experimental	No verbal reinforcement and money before			16		Lab	-		
				Experimental	Verbal Reinforcement and Money after			16					
				Control	Verbal Reinforcement and no Money				16				
				Experimental	Verbal reinforcement and money before			16					

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
20	Dollinger, S. J., & Thelen, M. H. (1978).	102	**E**	Experimental	Tangible reward (pretzels)	Mazes and geometric design problems solving	Children enrolled in summer programs in day-care centers and nursery or elementary schools.	10		Lab	Trailer	Latency	Average number of seconds before attempting maze play
				Experimental	Verbal Reward (positive feedback)			10				Output attempts	Average number of mazes attempted (out of ten)
				Experimental	Symbolic Reward (star)			10				Time spent on task	Average number of seconds spent on mazes
				Experimental	Self administered symbolic reward			10				Self-reported Intrinsic Motivation	Participants rated their self-perceived competence, interest in the task, and researcher's interest in the task on 5-point scale
				Control	No reward				10			Self-reported Difficulty Assessment; Self-reported motivating factors	Participants explained which mazes they would like to do over again based on a 3-point scale; Participants were rated dichotomously on preference for quality over quantity, performance anxiety, and preference of free time activity
21	Earn, B. M. (1982).	29	**S1**	Experimental	Large monetary reward (task engagement contingent)	Solution of anagrams	Undergraduate students with High or Low level of locus of control	20		Lab	-	Time spent on task	Average number of seconds spent on task
				Experimental	Small monetary reward (task engagement contingent)			20				Output	Average number of puzzles solved in 20 minutes
				Control	No reward				20			Self-reported satisfaction with task; Self-reported satisfaction with pay/competency	Subjects rated their interested in the task on four 7-point semantic differential scales; Subjects rated their satisfaction with their pay and their own competency on a 5-point scale
			S2	Experimental	Big monetary reward (performance contingent)			20		Lab	-	Time spent on task	Average number of seconds spent on task
				Experimental	Small monetary reward (performance contingent)			20				Output	Average number of puzzles solved in 20 minutes
				Control	No reward				20			Self-reported Satisfaction with Task; Self-reported Satisfaction with Pay/competency	Subjects rated their interested in the task on four 7-point semantic differential scales; Subjects rated their satisfaction with their pay and their own competency on a 5-point scale

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
22	Enzle, M. E., Roggeveen, J. P., & Look, S. C. (1991).	23	**E**	Experimental	Self Administered Monetary Reward	Spill & Spell crossword game; ambiguous behavior standard	Male and female undergraduates	10		Lab	-	Time spent on task	Average number of seconds of free time spent on game
						Spill & Spell crossword game; Unambiguous behavior standard		10					
				Experimental	Other Administered Monetary Reward	Spill & Spell crossword game; Ambigous Behavior Standard		10					
						Spill & Spell crossword game; Unambigous Behavior Standard		10					
				Control	No reward	Spill & Spell crossword game; Ambigous Behavior Standard			5				
						Spill & Spell crossword game; Unambigous Behavior Standard			10				

#	Authors	Cites	Experiment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
23	Fabes, R. A. (1987).	21	E1	Experimental	Task contingent reward (toy: small rubber animal)	Play a building with blocks game	Preschoolers from 37 to 66 months	19		Lab	-	Time spent on task	Average number of seconds spent building with blocks while performing in the experiment
				Experimental	Performance contingent reward (toy: small rubber animal)			18				Compliance	Proportion of children who built using only the correct, allowed blocks
				Control	No reward				19			Time spent on task	Average number of seconds spent building with blocks in free-choice period
			E2	Experimental	Reward (toy: small rubber animal)		Preschoolers from 39 to 59 months	14		Lab	-	Time spent on task	Average number of seconds spent building with blocks in free-choice period; Average number of seconds spent building with blocks while performing in the experiment
				Control	No reward				14			Choice for Researcher	Number/percent of participants who preferred to be escorted back to class by the primary or secondary researcher
24	Fabes, R. A., Eisenberg, N., Fultz, J., & Miller, P. (1988).	24	E	Experimental	Nonreward (positive mood)	beanbag game (beanbag tosses into a clown's mouth)	Preschool children	14		Lab	-	Self-recorded Emotional Measure	Mean happiness score: before, after, and thrice during the experiment, participants rated their happiness on a pictographic 5-point scale
				Experimental	Nonreward (neutral mood)			14				Time spent on task	Average number of seconds the child spent playing the beanbag game
				Experimental	Nonreward (negative mood)			14				Output quality	Mean level of difficulty attempted (distance from target) and the total number of successful tosses for each participant.
				Control	Reward (neutral mood)				14			Self-recorded Choices	Preference for beanbag task and preference for experimenter rated in 5-point scale

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
25	Fabes, R. A., Fultz, J., Eisenberg, N., May-Plumlee, T., & Christopher, F. S. (1989).	178	E	Experimental	Nonreward (actor condition)	Put pieces of paper into piles according to color	Children from 80 to 140 months	14		Lab	-	Output Choice	Average number of pages sorted in two minutes; Percent of children who opted to sort paper during free time
				Experimental	Small toy as reward (actor condition)			14				Recognition of Extrinsic/Intrinsic Motivational Factors	Mean difference score from recognition of extrinsic/intrinsic factors in a story
				Experimental	Nonreward (observer condition)			14				Self-recorded Parenting Technique	Average score on a list of 15 items of parenting techniques (possible score from 10 to 50)
				Experimental	Small toy as reward (observer condition)			14				Prosocial Behavior Scale	Average score of child (by parent) on a prosocial behavior scale from 1-7
				Control	Nonreward (control condition)				14				
26	Fabes, R. A., McCullers, J. C., & Horn, H. (1986).	11	E	Experimental	Reward (small inexpensive toy)	Play Maze games	Children predominantly white, middle class	12		Lab	Trailer	Self-recorded Interest	Average score on a pre-test and post-test of interest in the task (in scale from 1-5)
				Control	No reward				12			Self-recorded Interest Task Enjoyment and Difficulty	Average score on tests of Task Enjoyment and Difficulty (in scale from 1-5)
				Experimental	Reward (small inexpensive toy)	Play Block Design games		12				Time spent on task	Average number of seconds engaged with the task
				Control	No reward				12			Output (attempts) Output Quality	Average items attempted, average items completed, mean difficulty for items attempted, average score on task
27	Feingold, B. D. & Mahoney M. J. (1975).	162	E	Experimental	Reward (candies, toys, small books) according to Dot-to-dot performance during reinforcement session	Connecting the dots in Follow-the-dots books	Second grade children (boys and girls)	5		Lab	Classroom	Quality	Measure number of dots correctly connected in Follow-the-Dot books. The total number of connected dots--both correcty or incorrectly--is not reported.

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
28	Freedman, S. M., & Phillips, J. S. (1985).	68	**E**	Experimental	Contingent monetary reward, high interest task, nonconstrained conditions	Proofreading task	Male and female undergraduate students	13		Lab	-	Output	Measure number of errors corrected in a manuscript
				Experimental	Contingent monetary reward, high intrest task, constrained conditions			12					
				Experimental	Contingent monetary reward, low intrest task, nonconstrained conditions			12				Interest	Questionnaire on interest and engagement and enjoyment and satisfaction with task using 7-point scale
				Experimental	Contingent monetary reward, low intrest task, constrained conditions			12					
				Experimental	Noncontingent monetary reward, high interest task, nonconstrained conditions			13				Self-determination and competence; perceived constraints	Questionnaire self-perceived performance and control over performance in 7-point scale; whether felt had little time to do task
				Experimental	Noncontingent monetary reward, high intrest task, constrained conditions			12					
				Experimental	Noncontingent monetary reward, low intrest task, nonconstrained conditions			12					
				Experimental	Noncontingent monetary reward, low intrest task, constrained conditions			12					
				Control	No monetary reward, high interest task, nonconstrained conditions				13				
				Control	No monetary reward, high intrest task, constrained conditions				12				
				Control	No monetary reward, low interest task, nonconstrained conditions				12				
				Control	No monetary reward, low intrest task, constrained conditions				12				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
			S1	Control	No reward			Unclear (77 for both conditions)			-	Choice of activity	Percentage of subjects that picked solving math problems over watching interesting videos
								39					
				Experimental	5c piece rate for solving each math problem	Solving math problems and watching interesting videos			39			Output	Number of math problems solved
												Productiviy	Average time to solve a math problem
												Quality	Accuracy of math problems solved
												Self-reported enjoyment	Self-reported value on an enjoyment scale
			S2	Control	No reward			Unclear (257 for the 6 conditions)				Choice of activity	Percentage of subjects who chose solving math problems over watching interesting videos
								43					
				Experimental	5c piece rate for solving each math problem				43				
				Experimental	5c piece rate for solving each math problem; asked to write opinions	Solving math problems, writing opinions, matching brand logos			43				
				Experimental	5c piece rate for solving each math problem; asked to match brand logos				43				
				Experimental	5c piece rate for solving each math problem; chose to write opinions				43				
29	Goswami, I. & Urminsky, O. (2017)	22		Experimental	5c piece rate for solving each math problem; chose to match logos		Mturk subjects		43				
			S3	Control	No reward			Unclear (235 for the 4 conditions)				Choice of activity	Percentage of subjects who chose to solve math problems
								59					
				Experimental	5c piece rate for solving each math problem	Solving math problems			59			Productivity	Time to complete each math problem
				Experimental	1c piece rate for solving each math problem				59			Quality	Accuracy in solving math problems
				Experimental	50c piece rate for solving each math problem				59				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
			S4	Control	No reward for solving math problems			Unclear (223 for the 4 conditions)				Choice of activity	Percentage of subjects who chose to solve math problems
								56					
				Control	No reward for watching and rating videos	Solving math problems and watching and rating videos			56				
				Experimental	5c piece rate for solving each math problem				56				
				Experimental	5c piece rate for watching and rating each video				56				
			S5	Control	No reward			Unclear (189 for 3 conditions). Assumed sample sizes below.				Choice of activity	Percentage of subjects who chose to solve math problems
								63					
				Experimental	5c piece rate for solving each math problem; no break between rounds	Solving math problems			63				
				Experimental	5c piece rate for solving each math problem; a break between rounds				63				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
30	Greene, D., & Lepper, M. R. (1974).	316	E	Experimental	Expected award (golden star with ribbon) / Low performance demand	Free-style drawing with multicolored, felt-tipped pens	Preschool children of predominantly white, middle-class backgrounds	15		Field	New	Time spent on task	Transformed mean percent of free-choice time subjects spent with the target activity in their classrooms
				Experimental	Expected award (golden star with ribbon)/ High performance demand			15				Output	Mean number of pictures drawn
				Experimental	No expected award (golden star with ribbon) / Low performance demand			15				Quality	Rated drawings for quality (rated by naive judges)
				Experimental	No expected award (golden star with ribbon) / High performance demand			14					
				Control	No award				14				
31	Hamner, W. C., & Foster, L. W. (1975).	104	E	Experimental	Noncontingent pay	Boring task: Code and transfer scores from a recent math survey to a Fortran work sheet	Undergraduate students	16		Lab	-	Output	Measure number of items scored out of maximum 900
				Experimental	Contingent pay			16				Quallity	Measure number of items scored incorrectly
				Control	No pay				16			Interest	Subjects rated their interest from 1 to 5 (extremely interesting)
				Experimental	Noncontingent pay	Interesting task: Code and transfer scores from a recent sexual attitude survey of college females to a Fortran work sheet		16				Satisfaction with pay	Subjects were asked about their satisfaction with the pay they earned
				Experimental	Contingent pay			16					
				Control	No pay				16				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
32	Harackiewicz, J. M., Abrahams, S., & Wageman, R. (1987).	197	**E**	Experimental	Evaluation condition: Anticipated performance evaluation/Social comparison focus	Paper-and-pencil word game: Construct as many words as possible from a letter matrix	High school students	13		Lab			
				Experimental	Evaluation condition: Anticipated performance evaluation/task focus			13				Output	Measure number of words found in puzzles
				Experimental	Reward condition: Anticipated performance evaluation and performance contingent reward (free movie pass)/Social comparison focus			13					
				Experimental	Reward condition: Anticipated performance evaluation and performance contingent reward (free movie pass)/task focus			13				Self-reported Competence	Pre-task engagement, competence valuation created by responses to two 7-point scale questions
				Control	Feedback control condition: Neither expected evaluation nor were promised a reward/Social comparison focus				13				
				Control	Feedback control condition: Neither expected evaluation nor were promised a reward/Task focus				13			Self-reported competence and task thoughts	Responses to scale-based questions about subjects' thoughts about their competence and the task during the experiment

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
33	Harackiewicz, J. M., & Manderlink, G. (1984).	171	**E**	Experimental	Performance contingent reward (Fast food restaurant gift certificate)	Cartoon-style drawings in which the name Nina is hidden several times	High school students	Exact number of subjects per condition not reported		Lab	-	Output	Measure number of "Ninas" found in cartoon drawing or number of words formed from matrix
				Control	Ego-involvement condition (subjects read that good performance reflect important skills and abilities)							Importance of doing well, anticipated and perceived performance	Answered questions on 7-point scales
				Experimental	Performance contingent reward (Fast food restaurant gift certificate)	Hidden words puzzle in which the subject is to form as many words as possible from contiguous letters						Enjoyment	Questionnaire based on a 7-point scale
				Control	Ego-involvement condition (subjects read that good performance reflect important skills and abilities)								

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
34	Harackiewicz, J., Manderlink, G., & Sansone, C. (1984).	328		Experimental	Performance- contingent reward (movie pass)	Playing a pinball game	Male undergraduates	32		Lab	-	Output Time spent on task Task enjoyment Task concern	Measured number of balls played Amount of time playing pinball Questionnaire on whether subjects find the task enjoyable Questionnaire on the personal importance of doing well at pinball
			S1										
				Experimental	Performance Evaluation but no reward			32				Quality	Scores in the game (1 ball can yield a higher score than others)
				Control	No evaluation nor reward				32			Forecasted performance and perceived performance relative to other subjects	Subjects asked to rate on a 10-point scale how well they though they would do and how they thought they did relative to other students.
				Experimental	Expected performance - contingent reward (movie pass)			15		Lab	-	Same as Study 1 except time spent on the task	Same metrics as in study one, except for time spent on the task.
			S2	Experimental	Unexpected performance contingent reward (movie pass)			15					
				Control	Neither expected nor received a reward				15				
				Experimental	Performance- contingent reward (movie pass)			26		Lab	-	Thoughts about pinball and about their own competence	Questionnaire on items such as those concerning distraction, thoughts about the game, and thoughts about subjects' own competence
			S3	Experimental	Performance Evaluation but no reward			26				Output	Number of balls played
				Control	No evaluation nor reward				26				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Experimental	Immediate monetary reward			15				Time on the task	Measured time (minutes) spent on task beyond the required 10 minutes
35	Hitt, D. D., Marriot, R. G., & Esser, J. K. (1992).	53	**E**	Experimental	Delayed monetary reward	Highly interesting computer game	Male and female undergraduates	15		Lab	-	Interest	Post-experimental questionnaire: did you find the computer game to be of interest?; how interesting did you find the computer game?; how would you rate the game you played on entertainment value?
				Control	No reward				15				
				Experimental	Immediate monetary reward	Boring Computer game		15					
				Experimental	Delayed monetary reward			15					
				Control	No reward				15				
36	Hom, H. L. (1987).	19	**E1**	Experimental	Payment	Pursuit-rotor task	College students	26		Lab	-	Time on the task	Measured seconds the subject stayed engaged in the previous task
				Control	No payment				26				
			E2	Experimental	Positive feedback	Solution of anagrams		17		Lab	-	Output (attempted)	Measured number of anagrams attempted (does not measure anagrams solved)
				Experimental	Positive-negative feedback			17					
				Control	No feedback				17				
37	Karniol, R., & Ross, M. (1977).	126	**E**	Experimental	Performance-relevant reward (marshmallows)	Play the "slide game"	Children aged 4-9	20		Lab	Trailer	Time on the task	Measured duration of children's play at the slide game in seconds
				Experimental	Performance-irrelevant reward (marshmallows)			17					
				Control	No-reward				17				

#	Authors	Cites	Experiment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
38	Kast, A., & Connor, K. (1988).	78	**E**	Experimental	Informational Feedback	Four word-search puzzles	Children from lower-middle-and working-class families	20		Lab	-	Output	Measure number of words children could identify in grid
				Experimental	Controlling Feedback			20				Perception	Children asked to rate messages on scale from 1-4 in manner of Harter (1981)
				Experimental	Mixed Feedback			20				Interest	Interest questionnaire - asked degree of liking for each of the four different games on a 4-point scale.
				Control	No Feedback				20				
39	Kruglanski, A. W., Alon, S., & Lewis, T. (1972).	176	**E**	Experimental	Prize (plastic puzzle game)	Five successive games: (1) "Follow the leader", (2) "Word construction", (3) "Song Matching", (4) "Discover the rhyme" and (5) "speed writing"	Elementary school students	68		Field	New	Localization of causality	Open-ended questionnaire: (1) asking why you decided to participate in the games in your class; (2) Same question with 3 choices: because I like to compete, because I wanted to win prizes, or because I find group games interesting
				Control	No prize				64			Enjoyment	4-point scale on questionnaire: to what extent did you enjoy the games
40	Kruglanski, A. W., Friedman, I., & Zeevi, G. (1971).	561	**E**	Control	No-incentive	Five tasks: (1) Suggesting titles, (2) Composition of a story, (3) reading a newspaper story, (4) Test of nonsense syllables and (5) Zeignarik measures.	Male and female aging from fifteen to sixteen		16	Field	New	Recall (2 tasks); Creativity (2 tasks);	<u>Creativity:</u> measured number of titles suggested or number words (out of 50 given) used in composition of a story; <u>Recall:</u> measured number nonsense words recalled or correct answers to informative questions based on newspaper article
				Experimental	Extrinsic-incentive (reward: guided tour)			16				The tendency to differentially recall interrupted and completed activities (Zeignarik) Enjoyment	Ratio (differential tendency to recall interrupted to completed activities) 5-pt scale of enjoyment rating

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
41	Kruglanski, A. W., Riter, A., Amitai, A., Margolin, B., Shabtai, L., & Zaksh, D. (1975).	149	**E1**	Experimental	Payment Present	Coin-toss guessing game (Monetary intrinsic condition)	Boys from 14 to 15 years old	12		Lab	-	Perceived interest	7-point scales: Asked to what extent did you find the game interesting?; to what extent do you think boys your age will find this interesting?
				Control	Payment absent				12			Perceived Choice for the activity over alternatives	7-point scale: Asked to what extent do you think you'd play this in the future during leisure with friends?
				Experimental	Payment Present	Constructions of models according to pictures using		12					
				Control	Payment absent				12				
			E2	Experimental	Payment Present	Novel arithmetic game (Stock market game in Money intrinsic condition)	15- and 16-year old high school students	20		Lab	-	Quits/ Choice for game over alternative	Asked to choose between continuing game and playing a new one with same chance of earning a monetary payment
				Control	Payment absent				20			Interest and perceived attractiveness of game	Asked four questions, each with 7-point scales
				Experimental	Payment Present	Novel arithmetic game (Athletics game in Money extrinsic condition)		20					
				Control	Payment absent				20				
42	Lepper, M. R., Greene, D., & Nisbett, R. E. (1973).	3,474	**E**	Experimental	Expected-award (certificate with a gold seal and ribbon)	Drawing activity	Preschool children predominantly from white, middle-class backgrounds	18		Field	New	Percentage of time spent on task	Measured the percentage of time that the child chose to play with the experimental activity out of the total time he was present while materials were available
				Experimental	Unexpected-award (certificate with a gold seal and ribbon)			18				Quality	Naïve judges blind to condition rated children's drawings
				Control	No award				15				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
43	Lepper, M. R., Sagotsky, G., Dafoe, J. L., & Greene, D. (1982).	167	**E1**	Experimental	Means-end condition	Listen to a short story and answer questions about it	Preschool children (male and female)	16		Lab	-	Judgment and opinion (direct comparative question)	Children asked what food the protagonist of story would prefer and why
				Control	Control condition				12				
			E2	Experimental	Means-end condition	Watch a slide show dipicting a scenario in which another child was asked to engage in two hilghy similar drawing activities under one of two conditions		20		Lab	-	Judgment and opinion Choice and interest	Children asked what activity the protagonist of story would prefer and why Children asked what activity they would prefer and why; children aske to rate each of the two activities on a 4-point visual scale
				Control	Control condition				20				
			E3	Experimental	Means-end condition	Two drawing activities: felt-tipped magic marker pens and artist's oil pastel crayons		32		Lab	Classroom	Quality Time spent on task	Measured amount of time spent with each of the 2 activities during the experimental session; measured proportion of time spent with each of the 2 target activities over other alternatives in the classroom several weeks after the experimental session
				Control	Control condition				32				
44	Loveland, K. K., & Olley, J. G. (1979).	140	**E**	Experimental	Expected reward ("Good player" award)	Drawing activity	Preschool children	12		Field	New	Timem spent on task	Measured (log of) the amount of time spent drawing
				Control	No reward				12			Output Quality	Measured number of sides paper sheets used and quality of drawings (in terms of form diversity)
45	Luyten, H., & Lens, W. (1981).	54	**E**	Experimental	Task contingent (monetary) reward	Constructions of models according to pictures using wooden blocks	Female and Male Undergraduates	10		Lab	-	Time spent on task	Measure the seconds of free time spent building models in presence of attractive alternative (popular magazines)
				Experimental	Performance contingent (monetary) reward			10				Interest Willingness to engage in future similar study	Interest questionnaire: rated interest, challenge, and willingness to return on 7-point scales
				Experimental	No reward/Task-contingent (monetary) reward			10					
				Control	No (monetary) reward				10				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
46	McGraw, K. O., & McCullers, J. C. (1979).	364	**E**	Experimental	Task contingent (monetary) reward	Water-jar problems	Male and female undergraduates	36	36	Lab	-	Productivity	Time to reach a solution to the problem (in seconds)
				Control	No reward							Task attitude and willingness to engage in future similar study even if they not paid	Measured task and research attitude, and asked subjects to rate how likely they were to return for another study, even if unpaid
47	McLoyd, V. C. (1979).	139	**E**	Experimental	High value reward (toy or tool)	Reading from one of six storybooks	Second and third grade children	18	18	Lab	Classroom	Time spent on task	Measure number of seconds of contact with the book during free-choice period
				Experimental	Low value reward (toy or tool)			18				Output	Measure number of words read during free-choice period
				Control	No reward				18			Interest	Observed whether first object contacted during free-choice period was book; observed whether child thought reading the book was the most fun thing in the room
48	Morgan, M. (1981).	68	**S1**	Experimental	Tangible reward (edibles)	Jigsaw puzzle solving	5-, 8- and 11 years-old children	60	60	Field	New	Time spent on task	Measured the amount of time spent on the target activity (puzzle) when other activities were available
				Control	No reward							Liking	Expressed degree to which they liked the activity on a 6-point scale
												Enjoyment	Rating by experimenter on a 9-point scale - rated the extent to which they judged subjects to have enjoyed the task during experimental interval
			S2	Experimental	Tangible reward (edibles)		8-years old children	20	20	Field	New	Time spent on task	Measured the amount of time spent on the target activity (puzzle) when other activities were available
				Control	No reward							Liking	Expressed degree to which they liked the activity compared to another activity of their choice (selected from five other options)
												Enjoyment	Rating by experimenter on a 9-point scale - rated the extent to which they judged subjects to have enjoyed the task during experimental interval

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
												Output	Number of jigsaw puzzles solved during the experimental session (not free-choice session). If a puzzle was incomplete the experimenter recorded the fraction to the nearest one thenth of the puzzle that had been completed.
49	Morgan, M. (1983).	34	**E1**	Experimental	Involved reward (chocolates or marshmallows)	Jigsaw puzzle solving	Children from 5 to 10 years old	40		Field	New	Time spent on task	Measured number of seconds spent on target activity during free-choice period
				Experimental	Observer reward (chocolates or marshmallows)			40				Choice for activity	Observed which activity subject started and ended on during free-choice period
				Control	Involved control				40			Liking	Liking of target activity obtained for 10-year olds 2 to 4 days after experimental sessions (on 6-point scale)
			E2	Experimental	Involved reward (chocolates or marshmallows)		Children aged 8-10 years	20		Field	New	Time spent on task	Average time spent on puzzles
				Experimental	Observer no reward			20				Liking	Liking of puzzle solution rated on 6-point scale
				Control	Involved control				20			Output	Measured number of puzzles solved during experimental session
				Control	Observer control				20				
50	Mynatt, C., Oakley, T., Arkkelin, D., Piccione, A., Margolis, R., & Arkkelin, J. (1978).	22	**E**	Experimental	High-Base-Rate reward (M&M's)	Target games playing	Children aged 6-7 years from a first grade class	5		Field	Current	Choice for activity	Game choice after baseline period of the target games were recorded: frequency with which these games were chosen was measured
				Control	High-Base-Rate control				5				
				Experimental	High-Base-Rate control			5					
				Control	Low-Base-Rate control				5				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
51	Newman, J., & Layton, B. D. (1984).	39	**E**	Control	Zero reward	Two tasks: (1) High interest level toy (multicolored plastic pieces), (2) Low interest level (jigsaw puzzle)	Boys and girls from first or second grade		27	Lab	Trailer	Time spent on task	Measured number of 5-second time segments (out of 60) where the child was seen to be both touching and looking at toy
				Experimental	Small reward (1 M&M)			27					
				Experimental	Large reward (15 M&M's)			27					
52	Pallak, S. R., Costomiris, S., Sroka, S., & Pittman, T. S. (1982).	36	**E**	Experimental	Unexpected verbal reward	Drawing activity	Boys and girls aged 5-7 years	16		Lab	Classroom	Time spent on task	Measured duration of drawing during free play period
				Experimental	Expected verbal reward			14				Quits	Measured proportion of children drawing during free play period
				Experimental	Unexpected symbolic reward (surprise box)			15					
				Experimental	Expected symbolic reward (surprise box)			15					
				Control	No reward				12				
53	Perry, D. G., Bussey, K., & Redman, J. (1977).	18	**E**	Experimental	Initial performance contingent reward and final not-contingent reward	Drawing activity	7 year old (second grade) children	16		Lab	Classroom	Time spent on task	Measured duration of drawing (in seconds) during free period
				Experimental	No initial performance contingent reward and final not-contingent reward			16				Latency	Measured children's time until they started drawing during the free-time period
				Experimental	Initial performance contingent reward and no final not-contingent reward			16					
				Control	No initial performance contingent reward and no final not-contingent reward				16					

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
54	Pittman, T. S., Cooper, E. E., & Smith, T. W. (1977).	52	**E**	Control	No reward-no cue condition	Playing game "Gravitation" (moving a ball as far up as possible in an inclined plane)	Male and female undergraduate students		20	Lab	-	Output	Measure number of trials during the free-choice period
				Experimental	Monetary reward-no cue condition			20					
				Experimental	Monetary reward-intrinsic cue condition			20				Self-perception of arousal	Questionnaire with 91-point scale on how much subjects' levels of arousal were due to their interst in the game and how hard subjects tried to do their best during the first 10 trials.
				Experimental	Monetary reward-extrinsic cue condition			20					
55	Pittman, T. S., Emery, J., & Boggiano, A. K. (1982).	192	**E1**	Experimental	Task-contingent reward (surprise box)	Playing a shape matching game	Male and female second grade children	10		Lab	Classroom	Time spent on task Choice	Measured amount of time (in seconds) spent with target activity during free-choice period; Measured preference for complexity of activity
				Experimental	Task-noncontingent reward condition (surprise box)			10					
				Control	No reward			10					
			E2	Experimental	Reward (rubber ball)	Playing the find the hidden figure game	Fourth grade students	27		Lab	Trailer	Time spent on task	Measured amount of time spent playing with simple/intermediate versions of hidden-figure game during free-choice period
				Control	No reward				27				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Control	No money-control				10			Productivity prior to free-choice period	Measured amount of time required to solve puzzles (seconds) prior to the free-choice period
				Control	Money-control				10			Time spent on task during the free-choice period	Measured amount of time spent on puzzles during free-choice period (instead of reading or doing nothing)
56	Porac, J. F., & Meindl, J. (1982).	53	**E**	Experimental	No money-extrinsic	Puzzle-solving (SOMA)	Male undergraduates	10		Lab	-		
				Experimental	Money-extrinsic			10				Self-perception of whether performance was due to intrinsic or extrinsic motivation	Questionnaire leading the indiviudal to introspect on whether task performance was intrinsically or extrinsically motivated
				Experimental	Money-intrinsic			10					

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Experimental	Promised a reward (lottery ticket)/No performance feedback			10				Time spent on task	Time spent on puzzles
				Experimental	Promised a reward (lottery ticket)/Positive feedback			10					
				Experimental	Promised a reward (lottery ticket)/Negative feedback			10					
				Experimental	Unexpected reward (lottery ticket)/No performance feedback			10					
			E1	Experimental	Unexpected reward (lottery ticket)/Positive feedback	Puzzle-solving (SOMA)		10		Lab	-		
				Experimental	Unexpected reward (lottery ticket)/Negative feedback			10					
				Control	Not promised a reward/No performance feedback				10				
				Control	Not promised a reward/Positive feedback				10				
				Control	Not promised a reward/Negative feedback				10				
57	Pretty, G. H., & Seligman, C. (1984).	110					Male and female undergraduate students						
				Experimental	Promised a reward (lottery ticket)/Neutral self-statements			10				Time spent on task	Time spent on puzzles
				Experimental	Promised a reward (lottery ticket)/Positive self-statements			10					
				Experimental	Promised a reward (lottery ticket)/Negative self-statements			10					
				Experimental	Unexpected reward (lottery ticket)/Neutral self-statements			10					
			E2	Experimental	Unexpected reward (lottery ticket)/Positive self-statements	Puzzle-solving (SOMA)		10		Lab	-		
				Experimental	Unexpected reward (lottery ticket)/Negative self-statements			10					
				Control	Not promised a reward/Neutral self-statements				10				
				Control	Not promised a reward/Positive self-statements				10				
				Control	Not promised a reward/Negative self-statements				10				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
58	Reiss, S., & Sushinsky, L. W. (1975).	210	**E1**	Experimental	Promise of reward (Play with an attractive doll)	Listening to songs	First grade girls aged 6 to 7 years	16		Field	New	Time spent on task	Time spent listening to songs
				Control	No promise of reward				16				
			E2	Experimental	Each child was rewarded with tokens tradable for toys, for listening to a target song (different for every child)		Boys and girls from a kindergarten class	9		Field	New	Time spent on task	Time spent listening to songs (seconds)
59	Rosenfield, D., Folger, R., & Adelman, H. (1980).	171	**E**	Experimental	High pay; contingent/competency feedback	Working on a crossword game "Ad-lib"	Female Undergraduates	15		Lab	-	Time spent on task	Time spent playing game during free time
				Experimental	High pay; contingent/no-feedback			15					
				Experimental	High pay; noncontingent/no-feedback			14					
				Experimental	Low pay; contingent/competency feedback			15					
				Experimental	Low pay; contingent/no-feedback			15					
				Experimental	Low pay; noncontingent/no-feedback			14				Willingness to supply further work	Questionnaired on willingness to return to work on the same task for experimental credit only
				Control	No-pay; high competence feedback				15				
				Control	No-pay; low competence feedback				15			Liking	Sum of responses to survey measuring liking for task

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Experimental	Non-salient reward condition (received prize consisting of assorted candies and chocolates from the experimenter)			20				Choice	Toy chosen by child
			E1	Experimental	Salient reward (received prize consisting of assorted candies and chocolates from under a box)		Male and female children aged 47-59 months	20		Lab	Classroom	Time spent on task	Time spent playing with drum
				Control	Neither promised nor given a reward				20			Liking	What toy the child said was most fun
60	Ross, M. (1975).	343				Playing a drum							
				Experimental	Think-reward (thinking about marshmallows during task)			17				Time spent on task	Time spent playing with drum
				Experimental	Non-ideation (promised marshmallows but not instructed to ideate in any way)			19				Memory	Asked if they could remember initial question
			E2				Male and female children aged 42-60 months			Lab	Classroom		
				Experimental	Distraction (thinking about something different from the prize during the task)			16				Liking; Choice	What toy the child said was most fun; what toy the child picked first
				Control	Control (neither promised nor awarded marshmallows)				14				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
61	Ross, M., Karniol, R., & Rothstein, M. (1976).	72	**E**	Experimental	Task-contingent reward (candies)	Drawing activity	First, second and third grade children	12		Lab	Trailer	Time spent on task	Time spent drawing with Bic Bananas
				Experimental	Wait-contingent reward (candies)	Waiting in room for experimenter		12				Output	Number of separate drawings child drew
				Control	No reward				12			Quality	Two adult raters rated the drawings on a 7-point scale
62	Ryan, R. M., Mims, V., & Koestner, R. (1983).	1,155	**E**	Experimental	Monetary reward/Informational	Hidden figures task	Undergraduate students	16		Lab	-	Time spent on task	Time spent working on puzzles during free-choice period
				Experimental	Monetary reward/Controlling			16				Interest/Enjoyment	Subjects rated their Interest and enjoyment on 7-point scales
				Experimental	Monetary reward/Neutral			16				Sefl-perceived tension and pressure experienced; self-eperienced degree of the effort; and extent to which felt task was worthwhile	Questionnaire on these topics using 7-point scales
				Control	No reward/Informational				16				
				Control	No reward/Controlling				16				
				Control	No reward/Neutral				16				
63	Salancik, G. R. (1975).	50	**E**	Experimental	Monetary reward/Poor Performance	Playing with a road-set	Male undergraduates	19		Lab	-	Time spent on task	Time spent playing with toy cars
				Experimental	Monetary reward/Good Performance			19					
				Control	No reward/Good Perfomance				19			Enjoyment and Interest	Verbal questions on a 21-point bipolar scale
				Control	No reward/Poor Performance				19			Satisfaction with own performance	Verbal questions on a 21-point bipolar scale
64	Sarafino, E. P. (1984).	6	**E**	Experimental	Low delay (contingent reward: inexpensive plastic trinkets)	Riddle game	Boys and girls from middle-class day-care centers	29		Field	New	Choice	Choice of activity during free play
				Experimental	Moderate Delay (contingent reward: inexpensive plastic trinkets)			29					
				Experimental	high delay (contingent reward: inexpensive plastic trinkets)			29					
				Control	No reward				14				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
65	Shapira, Z. (1976).	183	**E**	Experimental	Pay (monetary reward) for successfully completing the task within 15 minutes	Solving seven puzzles (SOMA) of varying difficulties	Male and female undergraduates	30		Lab	-	Choice	The subject's first choice when it came to the puzzles and the order of the preferred configurations
				Control	No pay				30			Perceived Competence	Questionnaire that asked about difficulty and probability of their success
66	Smith, T. W., & Pittman, T. S. (1978).	66	**E**	Experimental	Reward per trial held constant; 10 trials	Skill game: Labyrinth (rolling a steel ball in a small maze platform)	Male and female undergraduates	11		Lab	-	Output	Number of attempts (trials) at the game during the free-choice period
				Experimental	Reward per trial held constant; 25 trials			11					
				Experimental	Reward per trial held constant; 50 trials			11				Interest	Rated interest in activity in 91-point scale
				Experimental	Reward per trial held varied 10 trials			11					
				Experimental	Reward per trial held varied 25 trials			11				Interest	Rated interest in activity in 91-point scale
				Experimental	Reward per trial varied 50 trials			11					
				Experimental	Non-reward distraction per trial; 10 trials			11					
				Experimental	Non-reward distraction per trial; 25 trials			11					
				Experimental	Non-reward distraction per trial; 50 trials			11					
				Control	No reward/ no reward distraction; 10 trials				11				
				Control	No reward/ no reward distraction; 25 trials				11				
				Control	No reward/ no reward distraction; 50 trials				11				

#	Authors	Cites	Experiment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
67	Sorensen, R. L., & Maehr, M. L. (1976).	31	**E**	Experimental	Tangible reinforcement (tokens)	Playing a puzzle like game ("Atoms", Creative Playthings)	Preschool and third grade children	20		Lab	Trailer	Time spent on task	Time spent working on puzzle during free time
				Control	No reinforcement				20				
68	Staw, B. M., Calder, B. J., Hess, R. K., & Sanderlands, L. E. (1980).	107	**E**	Experimental	Norm-of-payment condition	Solving 15 jigsaw type-puzzles	Undergraduate males	50		Lab	-	Satisfaction	Questionnaire used to create satisfaction index
				Control	Norm-of-no-payment condition				43			Willingness to supply further work	The amount of time (in minutes) volunteered to come back to perform task
69	Swann, W. B., & Pittman, T. S. (1977).	261	**E1**	Experimental	Task-contingent reward (good player award) / Child Decision	Drawing activity	Elementary school students	12		Field	New	Choice	Proportion of children choosing drawing first
				Experimental	Task-contingent reward (good player award) / Adult Decision			12				Time spent on task	Seconds spent drawing
				Experimental	Task-noncontingent reward (good player award) / Child Decision			12					
				Experimental	Task-noncontingent reward (good player award) / Adult Decision			12					
				Control	No reward				12				
			E2	Control	Decision irrelevant: No reward				13			Choice	Proportion of children choosing drawing first
				Experimental	Child decision: no reward			13					
				Experimental	Child decision: task contingent reward			13				Time spent on task	Seconds spent drawing
				Experimental	Child decision: task contingent reward plus star			13					
				Experimental	Child decision: task contingent reward plus praise			13					

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
70	Taub, S. I., & Dollinger, S. J. (1975).	24	**E**	Experimental	Reward	Coding task	Fourth and fifth grade children	124		Field	New	Output Quality	Number of figures coded correctly and incorrectly
				Control	No reward				124				
71	Thompson, E. P., Chaiken, S., & Hazlewood, J. D. (1993).	113	**E**	Experimental	Monetary reward; high NFC/DC score	Brainstorming tasks	Undergraduate students	18		Lab	-	Time spent on task	Time spent playing game
				Experimental	Monetary reward; low NFC/DC Score			19				Self-perception of intrinsic motivation	Post-examination survey where subjects assessed their perceived intrinsic motivation, self-determination, and effort expended while engaged in the task
				Control	No reward; high NFC/DC score				18				
				Control	No reward; low NFC/DC score				19				
72	Tripathi, K. N., & Agarwal, A. (1988).	13	**E**	Experimental	Performance contingent reward	Algorithmic task: block building, code substitution, sorting of cards and numerical problem.	Undergraduate students	10		Lab	-	Time spent on task	Time spent on block building, code substiution etc.
				Experimental	Task contingent reward			10				Involvement with task	Incidental learning score (level of retention of information about the task.
				Control	No reward				5			Interest	Questionnaire gauging interest in task
				Experimental	Performance contingent reward	Heuristic task: candle problem, pattern drawing tasks, water jar problems.		10					
				Experimental	Task contingent reward			10				Willingness to engage further in the task	Questionnaire measuring subjects' willingness to persist in the task
				Control	No reward				5				
73	Vasta, R., Andrews, D. E., McLaughlin, A. M., Stirpe, L . A., & Comfort, C. (1978).	41	**E1**	Experimental				6		Lab	-	Choice	Activity chosen by child
			E2	Experimental	Reward (star) for playing in the coloring task	Three activities: (a) geo-blocks, (b) cardboard puzzles, (c) Dittoed copies of coloring book pages	Kindergarten and first grade children	6		Lab	Classroom	Choice	Activity chosen by child

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
74	Vasta, R., & Stirpe, L. A. (1979).	47	**E**	Control	No reinforcement	Three activities: (a) math problems, (b) letter-number code problem, (c) number sequence problems	Third/fourth grade students	5		Lab	Classroom	Time spent on task	Time spent engaging in each activity
				Experimental	Reward (star) for solving math problems			5				Output Quality	Number of pages of each activity Accuracy of each activity (measured by the number of correct and incorrect problems)
75	Weiner, M. J. (1980).	32	**E**	Experimental	Monetary reward	Solution of anagrams	Male and female undergraduates	24		Lab	-	Output	Amount (number of pages) completed of each activity
				Control	No reward				24				Number of anagrams solved during the treatment period and free-time
76	Weiner, M. J., & Mander, A. M. (1978).	64	**E**	Experimental	Contingent monetary reward, high competency manipulation	Decoding words within cartoons	Undergraduate females	10		Lab	-	Quality Enjoyment; Time spent task; Willingness to volunteer for a similar experiment in the future	Number of words correctly decoded Questionnaire answered by subjects
				Experimental	Contingent monetary reward, low competency manipulation			10					
				Experimental	Contingent monetary reward, average competency manipulation			10					
				Experimental	Noncontingent monetary reward			10					
				Experimental	Noncontingent monetary reward, low competency manipulation			10					
				Experimental	Noncontingent monetary reward, average competency manipulation			10					
				Control	No reward, high competency manipulation				10				
				Control	No reward, low competency manipulation				10				
				Control	No reward, average competency manipulation				10				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
77	Wicker, F. W., Brown, G., Wiehe, J. A., & Shim, W.-Y. (1990).	18	**E**	Experimental	Money	Think Tac Toe Puzzle solving	Male and female undergraduates	29		Lab	-	Time spent on task	Free-time spent on puzzles
				Experimental	Pressure			29					
				Control	Control				29				
78	Williams, B. W. (1980).	82	**E**	Experimental	Attractive reward (comic book)	Four games: Gravitation, SOMA, Mercury Maze, Twister	Male and female fifth-grade students	12		Lab	Trailer	Time spent on task	Number of seconds spent on target activity
				Experimental	Unattractive reward (comic book)			12					
				Experimental	Request condition (not rewarded but asked to perform the task)			12					
				Control	No reward nor request				12				
79	Wimperis, B. R., & Farr, J. L. (1979).	48	**E**	Experimental	Contingent pay	Enriched task condition: building whole models	Undergraduate students	8		Lab	-	Output Quantity	Total number of nut-and-bolt "connections" made in a session
				Experimental	Noncontingent pay			8				Quality	Two experimenters rated the quality of the models resulting from the nut-and-bolt connections
				Control	No pay				8				
				Experimental	Contingent pay	Unenriched task condition: building subunits		8				Non-quits/Quits	Number of subjects that volunteered for an extra session for no pay
				Experimental	Noncontingent pay			8				interest	Self-reported measures of interest in the activity
												Perceived effort, and performance	Self-reported measured based on a 2-item Likert-type scale.
				Control	No pay				8				
80	**Total citations (papers 1-79)**	**20,922**					**Median sample size per condition**	**15**					

Table II

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
				Experimental	Positive feedback, supportive perceptions of self-determination cue			10				Time spent on task	Mean number of seconds spent solving puzzles
				Experimental	Positive feedback, supporitve perceptions of being controlled			10					
1	Anderson, S., & Rodin, J. (1989)	71	**E**	Experimental	Mildly negative feedback, supportive perceptions of self-determination	Brain -teasers	Undergraduate students	10		Lab	-	Self-reported satisfaction	Subjects rated their feelings using a 7-point scale and asked to predict future efficacy specific to brain teaser questions
				Experimental	Mildly negative feedback, supportive perceptions of being controlled			10					
				Control	No feedback				10			Intrinsic motivation score	Sum of standardized mood and target activity scores. Two standardized scores were given equal weights and simply added

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
2	Blanck, P. D., Reis, H. T., & Jackson, L. (1984).	89		Experimental	Praise male (female) subject, male (female) experimenter, male (female) sex linkage task	Word Creation in a word-cube game	Undergraduate males and females	9		Lab		Time spent on task	Mean number of seconds spent solving puzzle, with and without administrator present
				Experimental	Praise male (female) subject, male (female) experimenter, female (male) sex linkage task			9					
				Experimental	Praise male (female) subject, female (male) experimenter, male (female) sex linkage task			9					
				Experimental	Praise male (female) subject, female (male) experimenter, female (male) sex linkage task			9				Self-reported enjoyment/perfor-mance assessment	Participants responded to a questionnaire asking them to rate their performance on several metrics
				Control	No Praise male (female) subject, male (female) experimenter, male (female) sex linkage task				9				
				Control	No Praise male (female) subject, male (female) experimenter, female (male) sex linkage task				9				
				Control	NoPraise male (female) subject, female (male) experimenter, male (female) sex linkage task				9				
				Control	No Praise male (female) subject, female (male) experimenter, female (male) sex linkage task				9				
				Experimental	Verbal Feedback	Puzzle-solving (SOMA)	Undergraduate females	12		Lab	-	Time spent on task	Mean number of seconds spent solving puzzle without an administrator present
			E2	Control	No Verbal Feedback				12			Self-reported enjoyment/perfor-mance assessment	Participants responded to a questionnaire asking them to rate their performance on several metrics

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
3	Butler, R. (1987).	1,040	**E**	Experimental	Comments Feedback group	Divergent thinking uses test (Session 1 and 3) and Different circles test (Session 2)	Fifth and sixth grade Jewish Israeli pupils (boys and girls, mean age 11,1 years)	50		Field	New	Quantity and Quality	Researchers scored the responses based on number of responses, categories, elaborated responses, and original responses
				Experimental	Grades Feedback group			50				Self-reported interest and enjoyment; Willingness to supply further work	After Sessions 1 and 3 (of 3) participants rated their interest and enjoyment on a 7-point scale; also asked how many additional tasks (from one to seven) they would like to receive
				Experimental	Praise Feedback Group			50				SSSelf-reported success	After Session 3, students were asked to rate their performance on a scale of 7
				Control	No Feedback group				50			Self-reported attributions; Self-reported impact of evaluation	After Session 3, students rated possible motivating factors and effects on the effort they put in on a 7-point scale; students also reported the impact of evaluation on six different factors on a 7-point scale
4	Deci, E. L., Cascio, W. F., & Krusell, J. (1975).	276	**E**	Experimental	Positive Feedback / Male Experimenter	Puzzle-solving (SOMA)	Male Undergraduates	8		Lab	-	Productivity	Average number of seconds spent completing each puzzle
				Experimental	Positive Feedback / Female Experimenter			8				Free time spent on task	Average number of seconds of free time spent solving puzzles
				Control	No Feedback / Male Experimenter				8				
				Control	No Feedback / Female Experimenter				8				
				Experimental	Positive Feedback / Male Experimenter			8					
				Experimental	Positive Feedback / Female Experimenter		Female Undergraduates	8					
				Control	No Feedback / Male Experimenter				8				
				Control	No Feedback / Female Experimenter				8				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
5	Koestner, R., Zuckerman, M., & Koestner, J. (1987).	246	**E**	Experimental	Ability Praise	Working on hidden-pictures puzzles	College students	17		Lab	-	Time on Task	Measured duration (seconds) subjects spent working on puzzles during free-choice period
				Experimental	Effort Praise			17				Interest	Interest questionnaire - 7-point Likert scale (interest, fun, competence, effort, pressure-tension, freedom)
				Control	No Praise				19				
6	Pittman, T. S., Davey, M. E., Alafat, K. A., Wetherill, K. V., & Kramer, N. A. (1980).	229	**E**	Experimental	Verbal reward, informational cue, low surveilance group	Puzzle-solving (SOMA)	Male and female undergraduate students	12		Lab	-	Time spent on task	Measured proportion of free-choice period spent doing new Soma puzzles over attractive alternative (magazines)
				Experimental	Verbal reward, informational cue, medium surveilance group			12					
				Experimental	Verbal reward, informational cue, high surveilance group			12					
				Experimental	Verbal reward, controlling cue, low surveilance group			12					
				Experimental	Verbal reward, controlling cue, medium surveilance group			12					
				Experimental	Verbal reward, controlling cue, high surveilance group			12					
				Control	No reward and no cue				12				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
7	Ryan, R. M. (1982).	2,697	**E**	Experimental	Self-administered informational feedback	Hidden figures task (either ego or task involving induction)	Undergraduate students	32		Lab	-	Time spent on task	Seconds spent solving puzzle during free time
				Experimental	Self-administered controlling feedback			32				Interest	Self-reported questionnaire with 1-7 scale
				Experimental	Informational feedback administered by Experimenter			32				Self-reported tension, pressure, degree of effort, and feelings of whether the task important or worthwhile	Self-reported questionnaire with 1-7 scale
				Experimental	Controlling feedback administered by Experimenter			32					
8	Sansone, C. (1986).	209	**S1**	Experimental	Normative only (Feedback)	Trivia game	Male undergraduates	11		Lab	-	Enjoyment	Subjects ratings on seven items in questionnaire (e.g, enjoyment of task, whether it was fun, whether it was absorbing etc.) used to form an enjoyment scale
				Experimental	Task only (Feedback)			11					
				Experimental	Normative plus task (Feedback)			11					
				Experimental	Raw score only (Feedback)			11					
				Control	No feedback				11				
			S2	Experimental	Negative-Normative feedback/Ego-involvement statement		Male and female undergraduate	20		Lab	-	Enjoyment	Subjects ratings on seven items in questionnaire (e.g, enjoyment of task, whether it was fun, whether it was absorbing etc.) used to form an enjoyment scale
				Experimental	Negative-Normative feedback/No ego-involvement statement			20					
				Experimental	Positive-Normative feedback/Ego-involvement statement			20					
				Experimental	Positive-Normative feedback/No ego-involvement statement			20					Verbal questions on a 21-point bipolar scale
				Experimental	Task feedback/Ego-involvement statement			20					
				Experimental	Task feedback/No ego-involvement statement			20				Self-perception of performance	Asked subjects to rated how well they thought they did relative to other subjects in 7-point scale

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
												Enjoyment	Subjects ratings on seven items in questionnaire (e.g, enjoyment of task, whether it was fun, whether it was interesting etc.) used to form an enjoyment scale
				Experimental	Positive-Normative feedback (choice)			20					
9	Sansone, C. (1989).	84	**E**	Experimental	Positive-Normative feedback (no choice)	Playing a game that involves identifying the names of specific parts of common objects (puzzle could or could not be chosen)	Male undergraduates	21		Lab	-		
				Experimental	Task feedback (choice)			20				Self-perception of performance	7-point scale rating of for, example, perceived competence, tension, and perceived autonomy
				Experimental	Task feedback (no choice)			21					
				Control	No feedback (choice)				20				
				Control	No feedback (no choice)				21				

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
10	Sansone, C., Sachau, D. A., & Weir, C. (1989).	170	S1	Experimental	Positive feedback; Instruction	Playing computer game Zork (subjects either received or didn't receive instruction on how to play)	Male and female undergraduates	20		Lab	-	Latency	Amount of time it took for the subjects to start playing the game
				Experimental	Positive feedback; No instruction			20				Self-perception of performance	Questionnaire consisting of a 25-items in which subjects rated on 7-point scales the degree in which they agreed with statements reflecting competence, perceived self-determination, and other performance concerns
				Experimental	Negative feedback; Instruction			20				Time spent on task	Amount of time subjects played the game after researcher left the room
				Experimental	Negative feedback; No instruction			20				Enjoyment	Questionnaire based on 5-item enjoyment scale
				Control	No feedback; Instruction				20				
				Control	No feedback; No instruction				20				
			S2	Experimental	Specific instruction; fantasy condition	Playing computer game Zork (subjects either received specific, general, or no instructions at all)		19		Lab	-	Interest	Whether the subjects took a brochure advertising similar games
				Experimental	Specific instruction; skill-emphasis			19				Enjoyment	Questionnaire based on 5-item enjoyment scale
				Experimental	General instruction; fantasy condition			19					
				Experimental	General instruction; skill-emphasis			19				Self-perception of performance	Questionnaire consisting of a 25-items in which subjects rated on 7-point scales the degree in which they agreed with statements reflecting competence, perceived self-determination, and other performance concerns
				Control	No instruction; fantasy condition				19				
				Control	No instruction; skill-emphasis				19			Attention to the task	Reaction time in responding to the buzzer and the individual's estimate of how much time had passed before the buzzer sounded

#	Authors	Cites	Experi-ment / Study	Group	Payment / (Manipulation)	Task(s)	Type of Subject(s)	Nt	Nc	Field/ Lab	In Classroom vs. trailer; existing vs. new task	Outcome measure	Explanation of outcome measure
(1)	(2)	-3	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
11	Shanab, M. E., Peterson, D., Dargahi, S., & Deroian, P. (1981).	44	**E**	Experimental	Positive feedback	Puzzle-solving (SOMA)	College students	20		Lab	-	Interest	Questionnaire gauging interest
				Experimental	Negative feedback			20				Choice	Free-time activity they chose
				Control	No feedback				20				
12	Vallerand, R. J. (1983).	156	**E**	Experimental	6 Positive Verbal Reinforcements	Specifically constructed task to assess hockey players' decision-making abilities	French speaking male elite hockey players with ages ranging from 13 to 16 years	10		Lab	-	Intrinsic motivation	Questionnaire using as scale with 23 questions, each scored on a 7-point scale.
				Experimental	12 Positive Verbal Reinforcements			10				Perceived competence	7-point scale questionnaire gauging self-perceived competence
				Experimental	18 Positive Verbal Reinforcements			10					
				Experimental	24 Positive Verbal Reinforcements			10					
				Control	No verbal reinforcement				10				
13	Vallerand, R. J., & Reid, G. (1984).	738	**E**	Experimental	Positive verbal Feedback	Stabilometer motor task	Male undergraduates	38		Lab	-	Intrinsic motivation	Questionnaire using as scale with 23 questions, each scored on a 7-point scale.
				Experimental	Negative verbal feedback			38				Perceived competence	7-point scale questionnaire gauging self-perceived competence
				Control	No feedback				38				
14	Zinser, O., Young, J. G., & King, P. E. (1982).	27	**E**	Experimental	High Verbal Reward	Play "Hidden Pictures" game	Male and female third-graders children	32		Lab	-	Time spent on task	Time spent playing with cards during free play
				Experimental	Low Verbal reward			32					
				Control	No verbal reward				32				
15	**Total citations (papers 1-14)**	**6,076**					**Median sample size per condition**	**18**					

References

Amabile, T., B. Hennessey, and B. Grossman (1986): "Social Influences On Creativity-The Effects Of Contracted-For Reward," *Journal of Personality and Social Psychology*, 50(1), 14–23.

Anderson, R., S. Manoogian, and J. Reznick (1976): "The Undermining and Enhancing of Intrinsic Motivation in Preschool Children," *Journal of Personality and Social Psychology*, 34(5), 915–922.

Anderson, S., and J. Rodin (1989): "Is Bad News Always Bad - Cue and Feedback Effects on Intrinisc Motivation," *Journal of Applied Social Psychology*, 19(6, 1), 449–467.

Arkes, H. (1979): "Competence and the Overjustification Effect," *Motivation and Emotion*, 3(2).

Arnold, H. (1976): "Effects of Performance Feedback and Extrinsic Reward Upon High Intrinsic Motivation," *Organizational Behavior and Human Performance*, 17(2), 275–288.

——— (1985): "Task-Performance, Perceived Competence, and Attributed Causes of Performance as Determinants of Intrinsic Motivation," *Academy of Management Journal*, 28(4), 876–888.

Blanck, P., H. Reis, and L. Jackson (1984): "The Effects of Verbal-Reinforcement of Intrinsic Motivation for Sex-Linked Tasks," *Sex Roles*, 10(5-6), 369–386.

Boal, K., and L. Cummings, LL (1981): "Cognitive EvaluationTheory: An Experimental Test of Processes and Outcomes," *Organizational Behavior and Human Performance*, 28(3), 289–310.

Boggiano, A., and P. Hertel (1983): "Bonuses and Bribes - Mood Effects In Memory," *Social Cognition*, 2(1), 49–61.

Boggiano, A., and D. Ruble (1979): "Competence and the Overjustification Effect - Developmental Study," *Journal of Personality and Social Psychology*, 37(9), 1462–1468.

Boggiano, A., D. Ruble, and T. Pittman (1982): "The Mastery Hypothesis and the Overjustification Effect," *Social Cognition*.

Boggiano, AK, A., J. Harackwiecz, JM, J. Bessette, and D. Main (1985): "Increasing Childrens Interest Through Performance-Contingent Reward," *Social Cognition*, 3(4), 400–411.

Brennan, T., and J. Glover (1980): "An Examination of the Effect of Extrinsic Reinforcers on Intrinsically Motivated Behavior: Experimental and Theoretical," *Social Behavior and Personality*, 8(1), 27–32.

Brockner, J., and R. Vasta (1981): "Do Causal Attributions Mediate the Effects if Extrinsic Rewards on Intrinsic Interest," *Journal of Research in Personality*, 15(2), 201–209.

Butler, R. (1987): "Task-Involving and Ego-Involving Properties of Evaluation - Effects of Different Feedback Conditions on Motivational Perceptions, Interest, and Performance," *Journal of Educational Psychology*, 79(4), 474–482.

Calder, B., and B. Staw (1975): "Self-Perception of Intrinsic and Extrinsic Motivation," *Journal of Personality and Social Psychology*, 31(4), 599–605.

Crino, M., and M. White (1982): "Feedback Effects in Intrinsic/Extrinsic Reward Paradigms," *Journal of Management*, 8(2), 95–108.

Daniel, T., and J. Esser (1980): "Intrinsic Motivation as Influenced by Rewards, Task Interest, and Task Structure," *Journal of Applied Psychology*, 65(5), 566–573.

Danner, F., and E. Lonky (1981): "A Cognitive-Developmental Approach to the Effects of Rewards on Intrinsic Motivation," *Child Development*, 52(3), 1043–1052.

Davidson, P., and B. Bucher (1978): "Intrinisc Interest and Extrinsic Reward - Effects of a Continuing Token Program on Continuing Nonconstrained Preference," *Behavior Therapy*, 9(2), 222–234.

DECI, E. (1971): "Effects of Externally Mediated Rewards on Intrinsic Motivation," *Journal of Personality and Social Psychology*, 18(1), 105–&.

Deci, E. (1972): "Effects of Contingent and Noncontingent Rewards and COntrols on Intrinsic Motivation," *Organizational Behavior and Human Performance*, 8(2), 217–229.

Deci, E., W. Cascio, and J. Krusell (1975): "Cognitive Evaluation Theory and Some Comments on the Calder and Staw Critique," *Journal of Personality and Social Psychology*, 31(1), 81–85.

Dollinger, S., and M. Thelen (1978): "Overjustification and Childrens Intrinsic Motivation - Comparative Effects of 4 Rewards," *Journal of Personality and Social Psychology*, 36(11), 1259–1269.

Earn, B. (1982): "Intrinsic Motivation as a Function of Extrinsic Financial Rewards and Subjects Locus of Control," *Journal of Personality*, 50(3), 360–373.

Enzle, M., J. Roggeveen, and S. Look (1991): "Self-Versus Other-Reward Administration and Intrinsic Motivation," *Journal of Experimental Social Psychology*, 27(5), 468–479.

Fabes, R. (1987): "Effects of Reward Contexts on Young Children's Task Interest," *Journal of Psychology*, 121(1), 5–19.

Fabes, R., N. Einsenberg, J. Fultz, and P. Miller (1988): "Reward, Affect, and Young Childrens Motivational Orientation," *Motivation and Emotion*, 12(2), 155–169.

Fabes, R., J. Fultz, N. Eisenberg, T. Mayplumlee, and F. Christopher (1989): "Effects of Rewards on Childrens Pro-Social Motivation - A Socialization Study," *Developmental Psychology*, 25(4), 509–515.

Fabes, R., J. Mccullers, and H. Hom (1986): "Childrens Task Interest and Performance - Immediate Versus Subsequent Effects of Rewards," *Personality and Social Psychology Bulletin*, 12(1), 17–30.

Feingold, B., and M. Mahoney (1975): "Reinforcement Effects on Intrinsic Interest - Undermining Overjustification Hypothesis," *Behavior Therapy*, 6(3), 367–377.

Freedman, S., and J. Phillips (1985): "The Effects of Situational Performance Constraints On Intrinsic Motivation And Satisfaction - The Role Of Perceived Competene and Self-Determination," *Organizational Behavior and Human Decision Processes*, 35(3), 397–416.

Goswami, I., and O. Urminsky (2017): "The Dynamic Effect of Incentives on Postreward Task Engagement," *Journal of Experimental Psychology*, 146(1), 1–19.

Greene, D, D., and M. Lepper (1974): "Effects of Extrinsic Rewards on Childrens Subsquent Intrinsic Interest," *Child Development*, 45(4), 1141–1145.

Hamner, W., and L. Foster (1975): "Are Intrinsic and Extrinsic Rewards Additive - Test of Decis Cognitive Evaluation Theory of Task Motivation," *Organizational Behavior and Human Performance*, 14(3), 398–415.

Harackwiecz, J., S. Abrahams, and R. Wageman (1987): "Performane Evaluation and Intrinsic Motivation - The Effects of Evaluative Focus, Rewards, and Achievement Orientation," *Journal of Personality and Social Psychology*, 53(6), 1015–1023.

Harackwiecz, J., and G. Manderlink (1984): "A Process Analysis of the Effects of Performance - Contingent Rewards on Intrinsic Motivation," *Journal of Experimental Social Psychology*, 20(6), 531–551.

Harackwiecz, J., G. Manderlink, and C. Sansone (1984): "Rewarding Pinball Wizardry - Effects of Evaluation and Cue Value On Intrinsic Interest," *Journal of Personality and Social Psychology*, 47(2), 287–300.

Hitt, D., R. Marriott, and J. Esser (1992): "Effects of Delayed Rewards and Task Interest on Intrinsic Motivation," *Basic and Applied Social Psychology*, 13(4), 405–414.

Hom, H. (1987): "A Methodological Note - Time of Participation Effects On Intrinsic Motivation," *Personality and Social Psychology Bulletin*, 13(2), 210–215.

Karniol, R., and M. Ross (1977): "Effect of Performance-Relevant and Performance-Irrelevant Rewards on Childrens Intrinsic Motivation," *Child Development*, 48(2), 482–487.

Kast, A., and K. Connor (1988): "Sex and Age-Diffrences in Response to Informational and Controlling Feedback," *Personality and Social Psychology Bulletin*, 14(3), 514–523.

Koestner, R., M. Zuckerman, and J. Koestner (1987): "Praise, Involvement, and Intrinsic Motivation," *Journal of Personality and Social Psychology*, 53(2), 383–390.

Kruglanski, A., S. Alon, and T. Lewis (1972): "Retrospective Misattribution and Task Enjoyment," *Journal of Experimental Social Psychology*, 8(6), 493+.

Kruglanski, A., A. Riter, A. Amitai, B. Margolin, L. Shabtai, and D. Zaksh (1975): "Can Money Enhance Intrinsic Motivation - Test of Content-Consequence Hypothesis," *Journal of Personality and Social Psychology*, 31(4), 744–750.

Kruglanski.AW, F. I, and G. Zeevi (1971): "Effects of Extrinsic Incentive on Some Qualitative Aspects of Task Performance," *Journal of Personality*, 39(4), 606–&.

Lepper, M., D. Greene, and R. Nisbett (1973): "Undermining Children's Intrinsic Interest with Extrinsic Reward-Test of the Overjustification Hypothesis," *Journal of Personality and Social Psychology*, 28(1), 129–137.

Lepper, M., G. Sagotsky, J. Dafoe, and D. Greene (1982): "Consequences of Superfluous Social Contraints - Effects on Young Childrens Social Inferences and Subsequent Intrinsic Interest," *Journal of Personality and Social Psychology*, 42(1), 51–65.

Loveland, K., and J. Olley (1979): "Effect of External Reward on Interest and Quality of Task-Performance in Children of High and Low Intrinsic Motivation," *Child Development*, 50(4), 1207–1210.

Luyten, H., and W. Lens (1981): "The Effect of Earlier Experience and Reward Contingencies on Intrinsic Motivation," *Motivation and Emotion*, 5(1).

McGraw, K., and J. McCullers (1979): "Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set," *Journal of Experimental Social Psychology*, 15(3), 285–294.

McLoyd, V. (1979): "Effects of Extrinsic Rewards of Differential Value on High and Low Intrinsic Interest," *Child Development*, 50(4), 1010–1019.

Morgan, M. (1981): "The Over-Justification Effect - A Developmental Test of Self-Perception Interpretations," *Journal of Personality and Social Psychology*, 40(5), 809–821.

——— (1983): "Decrements in Intrinsic Motivation Among Rewarded and Observer Subjects," *Child Development*, 54(3), 636–644.

Mynatt, C., T. Oakley, D. Arkkelin, A. Piccione, R. Margolis, and J. Arkkelin (1978): "An Examination of Overjustification Under Conditions of Extended Observation and Multiple Reinforcement: Overjustification or Boredom," *Cognitive Therapy and Research*, 2(2), 171–177.

Newman, J., and B. Layton (1984): "Over-Justification - A Self-Perception Perspective," *Personality and Social Psychology Bulletin*, 10(3), 419–425.

Pallak, S., S. Costomiris, S. Sroka, and T. Pittman (1982): "School Experience, Reward Characteristics, and Intrinsic Motivation," *Child Development*, 53(5), 1382–1391.

Perry, D., K. Bussey, and J. Redman (1977): "Reward-Induced Decreased Play Effects - Re-Attribution of Motivation, Competing Responses, or Avoiding Frustration," *Child Development*, 48(4), 1369–1374.

Pittman, T., E. Cooper, and T. Smith (1977): "Attribution of Causality and Overjustification Effect," *Personality and Social Psychology Bulletin*, 3(2), 280–283.

Pittman, T., M. Davey, K. Alafat, K. Wetherhill, and N. Kramer (1980): "Informational Versus Controlling Verbal Rewards," *Personality and Social Psychology Bulletin*, 6(2), 228–233.

Pittman, T., J. Emery, and A. Boggiano (1982): "Intrinsic and Extrinsic Motivational Orientations - Reward-Induced Changes in Preference for Complexity," *Journal of Personality and Social Psychology*, 42(5), 789–797.

Porac, J., and J. Meindl (1982): "Undermining Over-Justification - Inducing Intrinsic and Extrinsic Task Representations," *Organizational Behavior and Human Performance*, 29(2), 208–226.

Pretty, G., and C. Seligman (1984): "Affect and the Over-Justification Effect," *Journal of Personality and Social Psychology*, 46(6), 1241–1253.

Reiss, S., and L. Sushinsky (1975): "Overjustification, Competing Responses, and Acquistion of Intrinsic Interest," *Journal of Personality and Social Psychology*, 31(6), 1116–1125.

Rosenfield, D., R. Folger, and H. Adelman (1980): "When Rewards Reflect Competence - A Qualification of the Over-Justification Effect," *Journal of Personality and Social Psychology*, 39(3), 368–376.

Ross, M. (1975): "Salience of Reward and Intrinsic Motivation," *Journal of Personality and Social Psychology*, 32(2), 245–254.

Ross, M., R. Karniol, and M. Rothstein (1976): "Reward Contingency and Intrinsic Motivation in Children - Test of Delay of Gratification Hypothesis," *Journal of Personality and Social Psychology*, 33(4), 442–447.

Ryan, R. (1982): "Control and Information in the Intrapersonal Sphere: An Extensive of Cognitive Evaluation Theory," *Journal of Personality and Social Psychology*, 43(3).

Ryan, R., V. Mims, and R. Koestner (1983): "Relation of Reward Contingency and Interpersonal Context to Intrinsic Motivation - A Review and Test Usiing Cognitive Evaluation Theory," *Journal of Personality and Social Psychology*, 45(4), 736–750.

Salancik, G. (1975): "Interaction Effects of Performance and Money on Self-Perception of Intrinsic Motivation," *Organizational Behavior and Human Performance*, 13(3), 339–351.

Sansone, C. (1986): "A Question of Competence - The Effects of Competence and Task Feedback on Intrinsic Interest," *Journal of Personality and Social Psychology*, 51(5), 918–931.

——— (1989): "Competence Feedback, Task Feedback, and Intrinsic Interest - an Examination of Process and Context," *Journal of Experimental Social Psychology*, 25(4), 343–361.

Sansone, C., D. Sachau, and C. Weir (1989): "Effects of Instruction on Intrinsic Interest - the Importance of Context," *Journal of Personality and Social Psychology*, 57(5), 819–829.

Sarafino, E. (1984): "Intrinsic Motivation and Delay of Gratification in Preschoolers - the Variables of Reward Salience and Length of Expected Delay," *British Journal of Developmental Psychology*, 2(JUN), 149–156.

Shanab, M., D. Peterson, S. Dargahi, and P. Deroian (1981): "The Effects of Positive and Negative Verbal Feedback on the Intrinsic Motivation of Male and Female Subjects," *Journal of Social Psychology*, 115(2), 195–205.

Shapria, Z. (1976): "Expectancy Determinants of Instrincially Motivated Behavior," *Journal of Personality and Social Psychology*, 34(6), 1235–1244.

Smith, T., and T. Pittman (1978): "Reward, Distraction, and Over-Justification Effect," *Journal of Personality and Social Psychology*, 36(5), 565–572.

Sorensen, R., and M. Maehr (1976): "Toward Experimental-Analysis of Continuing Motivation," *Journal of Educational Research*, 69(9), 319–322.

Staw, B., B. Calder, R. Hess, and L. Sandelands (1980): "Intrinsic Motivation and Norms about Payment," *Journal of Personality*, 48(1), 1–14.

Swann, W., and T. Pittman (1977): "Initiating Play Activity of Children - Moderating Influence of Verbal Cues on Intrinsic Motivation," *Child Development*, 48(3), 1128–1132.

Taub, S., and S. Dollinger (1975): "Reward and Purpose as Incentives for Children Differing in Locus of Control Expectancies," *Journal of Personality*, 43(2), 179–195.

Thompson, E., S. Chaiken, and J. Hazlewood (1993): "Need for Cognition and Desire for Control as Moderators of Extrinsic Reward Effects - A Person X Situation Approach to the Study of Intrinsic Motivation," *Journal of Personality and Social Psychology*, 64(6), 987–999.

Tripathi, K., and A. Agarwal (1988): "Effect of Reward Contingency on Intrinsic Motivation," *Journal of General Psychology*, 115(3), 241–246.

Vallerand, R. (1983): "The Effect of Differential Amounts of Positive Verbal Feedback on the Intrinsic Motivation of Male Hockey Players," *Journal of Sport Psychology*, 5(1), 100–107.

Vallerand, R., and G. Reid (1984): "On the Causal Effects of Perceived Competence On Intrinsic Motivation - a Test of Cognitive Evaluation Theory," *Journal of Sport Psychology*, 6(1), 94–102.

Vasta, R., D. Andrews, A. Mclaughlin, L. Stirpe, and C. Comfort (1978): "Reinforcement Effects on Intrinsic Interest - Classroom Analog," *Journal of School of Psychology*, 16(2), 161–166.

Vasta, R., and L. Stirpe (1979): "Reinforcement Effects on Three Measures of Children's Interest in Math," *Behavior Modification*.

Weiner, M. (1980): "The Effect of Incentive and Control Over Outcomes Upon Intrinsic Motivation and Performance," *Journal of Social Psychology*, 112(2), 247–254.

Weiner, M., and A. Mander (1978): "The Effects of Reward and Perception of Competency upon Intrinsic Motivation," *Motivation and Emotion*.

Wicker, F., G. Brown, J. Wiehe, and W. Shim (1990): "Moods, Goals, and Measures of Intrinsic Motivation," *Journal of Psychology*, 124(1), 75–86.

Williams, B. (1980): "Reinforcement, Behavior Constraint, and the Over-Justification Effect," *Journal of Personality and Social Psychology*, 39(4), 599–614.

Wimperis, B., and J. Farr (1979): "Effects of Task Content and Reward Contingency Upon Task-Performance and Satisfiaction," *Journal of Applied Social Psychology*, 9(3), 229–249.

Zinser, O., J. Young, and P. King (1982): "The Influence of Verbal Reward on Intrinsic Motivation in Children," *Journal of General Psychology*, 106(1), 85–91.

F Standard, Crowding-Out and Unmet-Wage Expectations Models

This section outlines the predictions of a standard economics model and of a crowding-out-of-enjoyment model. Further, it also describes the predictions of an unmet-wage expectations model. The original model is in Macera and te Velde (2016). These three models are nested in a single framework below. We capture the standard model by assuming that agents derive intrinsic utility from tasting (or from tasting and evaluating) each cookie, they like being paid the piece rate, but have disutility over effort. We conceptualize crowding out by assuming that monetary payments may erode agents' intrinsic utility for tasting (or tasting and evaluating). Finally, we model the displeasure from not receiving an expected payment by assuming that agents have reference-dependent reciprocal preferences in which their reference point corresponds to expectations about future outcomes (Kőszegi and Rabin, 2006).[36]

In the exposition that follows, the "standard model" is the typical model of economic behavior, with no crowding out and no reference-dependent preferences; the "crowding-out model" has crowding out but no reference-dependent preferences; and the "unmet-wage expectations model" has no crowding out but does have reference-dependent preferences.

Effort is a unidimensional measure representing either output or productivity. Each of these performance metrics matters to principals. Further, for these two outcomes, the two main models of interest—standard and crowding-out—yield distinctive predictions for the second session when the piece rate is withdrawn. Principals also care about quits, but the

[36]Though there are several theories that use expectations as reference points ((e.g., Bell, 1985; Loomes and Sugden, 1986; Gul, 1991; Shalev, 2000)), we use Kőszegi and Rabin's framework because it is a portable model and has received empirical support. See, Pope and Schweitzer (2011); Crawford and Meng (2011); Abeler, Falk, Goette, and Huffman (2011); Ericson and Fuster (2011); Gill and Prowse (2012).

increase in quits in session two predicted by crowding out is also consistent with an income effect in the standard model. Since this does not allow us to distinguish between the two models, we do not formalize predictions on this metric.[37]

F.1 Setup

The principal and agent interact over three periods. In period zero, the principal offers the agent a non-monetary reward T (the thank-you cookies), which is not contingent on performance, to carry out the task in the first and second periods (i.e., the first and second week). This reward is paid at the end of period two.

Agents derive intrinsic utility from tasting each cookie (e.g., inspecting it and taking a bite) or from tasting and evaluating each cookie (both inspecting it and taking a bite and completing the evaluation form). We thus let V_t represent the period-t marginal intrinsic utility from tasting (or tasting and evaluating) one cookie. Agents may be paid a piece rate: the period t piece rate may be zero (the agents are not paid) or positive (if the agents are paid). The period t piece rate is thus represented by $w_t \in \{0, w\}$.

Effort at each t period is represented by $e_t \geqslant 0$. This effort entails a cost. If agents like tasting and dislike evaluating then effort costs can entail completing the evaluation form and the cognitive effort of providing accurate ratings. If agents like both tasting and evaluating (completing the evaluation form) then the effort cost, can include, for example, the cognitive effort of providing accurate ratings and the physical cost of holding the pen. The idea is that even this enjoyable activity should entail some effort costs otherwise effort could be unbounded. The cost of effort $c(e_t)$, $c(0) = 0$ is, as usual, a positive, strictly increasing and convex function and separable across periods. That is, the first-period (week) fatigue does

[37] Although a multitasking model in which agents also care about quality would be closer to our experimental analysis, we felt this would add complexity to the model, with little gain. Our model, in its current form, already captures the main forces and predictions underpinning the three models.

not affect cost of effort in the following period (week), as subjects have one period (week) to rest.

F.2 Preferences

We assume that intrinsic enjoyment V_t, $t = 1, 2$ evolves as $V_t = V_{t-1} - \beta w_t$ if $V_0 > 0$ and $V_t = 0$ if $V_0 = 0$, where $V_0 \geqslant 0$ is the agent's initial stock of marginal intrinsic utility for tasting (or tasting and evaluating) each cookie and $\beta \geqslant 0$ is the crowding-out parameter.

This setup implies three things. First, if the agent has an endowment of intrinsic enjoyment in tasting, or in both tasting and evaluating ($V_0 > 0$) but the crowding out parameter $\beta = 0$, there is no crowding out, and thus a standard model: agents derive utility from each cookie tasted and from the piece rate and disutility from effort. Second, if $V_0 > 0$ and β is positive then there is crowding out: pay erodes the agents' endowment of intrinsic utility. Third, if the agent has no intrinsic interest in the task ($V_0 = 0$) then there is no intrinsic interest to undermine (in line with Deci, Koestner, and Ryan, 1999) and thus intrinsic enjoyment continues to be zero at each period ($V_t = 0$) even if the agent is paid.[38,39]

Importantly, the thank-you cookies T are only valuable as long as the agent enjoys tasting (or tasting and evaluating cookies) as explained the main text: $T = \alpha V_0 1(t = 2)$, where $\alpha > 0$ and $1(t = 2)$ is an indicator function taking value 1 if $t = 2$. Thus if $V_0 = 0$ (there is no enjoyment for tasting or for tasting and evaluating) then agents do not derive utility from the thank-you cookies ($T = 0$).[40]

[38] This conceptualization of intrinsic interest as a preference is in line with Deci (1971) and Bénabou and Tirole (2003). Recent work in economics has defined intrinsic motivation as actions taken without a financial reward (see Kőszegi, 2014), thus nesting our particular case: agents engaging in an action because of their intrinsic enjoyment for it.

[39] There is no declining marginal utility for tasting or tasting and evaluating in this model, for simplicity. This choice also approximates the setup in our test, which separates sessions by one week to curb the role of satiation.

[40] Recall in the main text that the thank-you cookies were perishable and hard to resell.

Period-t consumption utility. We assume agents experience consumption utility from tasting (or tasting and evaluating each cookie), the piece rate, if offered, effort and the thank-you cookies.

$$(V_t + w_t)e_t - c(e_t) + T \tag{4}$$

Period-t expectation-based reference-dependent utility. To capture agents' disutility arising from expecting payments from the principal, which are not made, we assume that agents have reference-dependent reciprocal preferences where the reference point corresponds to their recent expectations about wages and effort (Kőszegi and Rabin, 2006, 2007). Thus we let $(\widetilde{w}_t, \widetilde{e}_t)$ represent the expectation made in period $t-1$ for (w_t, e_t). As a result, the total period-t expectation-based reference-dependent utility is

$$\eta\mu\big(w_t - \widetilde{w}_t\big)\mu\big(e_t - \tilde{e}_t\big) + \eta\mu\left(c(\tilde{e}_t) - c(e_t)\right) \tag{5}$$

The function μ compares actual with expected outcomes, operationalizing the idea that consumption level departures from the reference point (Kahneman and Tversky, 1979) affect agents' utility. As usual, $\mu(x)$ is a strictly increasing, piece-wise linear function, where $\mu(0) = 0$, with a slope of one if $x > 0$ and a slope of $\lambda > 1$ if $x < 0$. The parameter λ represents the loss aversion parameter and this formulation captures the idea that losses hurt more than same-sized gains please. The parameter η represents the importance of the reference-dependent domain relative to the consumption utility domain in the agents' utility. For example, if $\eta = 0$ there is not gain/loss utility from departures from expectations and we return to simpler case of equation (4).

The first term in equation (5) corresponds to reference-dependent reciprocity: if the agent

receives a piece rate that is higher than expected ($\mu(w - \widetilde{w}) > 0$), he has an incentive to exert more effort than expected ($\mu(w - \widetilde{w})\mu(e - \tilde{e}) > 0$). In contrast, if the agent receives a piece rate lower than expected ($\mu(w - \widetilde{w}) < 0$) then he has an incentive to exert less effort than expected ($\mu(w - \widetilde{w})\mu(e - \tilde{e}) > 0$).[41]

Total period-t utility flow. Corresponds to the summation of equations (4) and (5).

F.3 Timeline for the three periods

In period zero (the recruiting period), agents are not offered any piece rate, just the thank-you cookies ($w_1 = w_2 = 0$ and T) and they decide to accept or reject participating in the task. At the end of the recruiting period those in ANTICIPATED are informed they will receive a piece rate in period one but not in period two ($w_1 = w > 0$ and $w_2 = 0$), while no new information is given to agents in CONTROL and UNANTICIPATED. In period one agents exert effort and those in ANTICIPATED and UNANTICIPATED receive the piece rate ($w_1 = 0$ for CONTROL and $w_1 = w$ for UNANTICIPATED and ANTICIPATED). At the end of the first period agents update their wage and effort expectations for period two, given the information available to them. In the second period no subject receives the piece rate in any condition ($w_2 = 0$), and agents exert effort again.

F.4 Agent Behavior in Period Zero: Decision to Participate in the Tasting

Participating in the blind tasting in the absence of a monetary reward revealed an agent's interest in tasting (or in tasting and evaluating) cookies. We can see this through the agents' participation constraint. In period zero (the recruiting period) and across the three conditions the agent does not expect to be paid ($\widetilde{w}_1 = \widetilde{w}_2 = 0$) and does not expect any

[41] See Rabin (1993); Charness and Rabin (2002); Gächter and Falk (2002); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006).

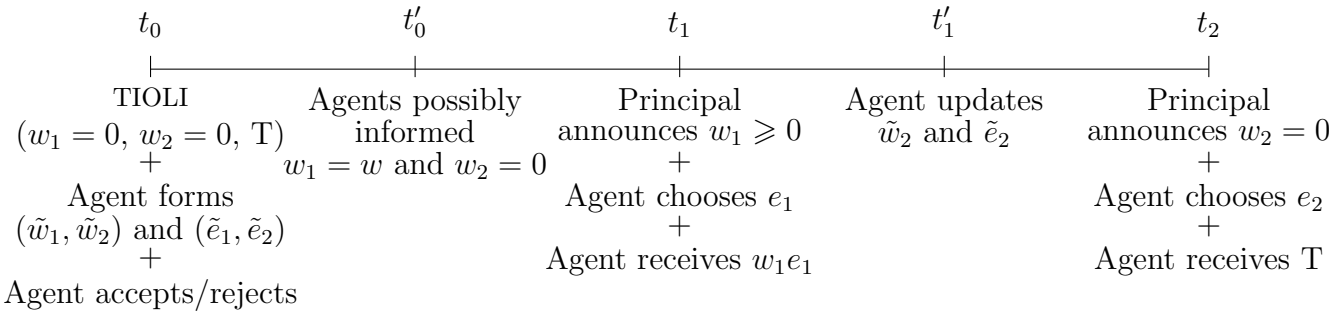


Figure F.1: Timeline

surprises relative to expectations. Thus, period-zero indirect utility is,

$$\begin{aligned} U_0 &= (\widetilde{V}_1 + \widetilde{w}_1)\widetilde{e}_1 - c(\widetilde{e}_1) + T + \left[(\widetilde{V}_2 + \widetilde{w}_2)\widetilde{e}_2 - c(\widetilde{e}_2) + T\right] \\ &= V_0(\alpha + \widetilde{e}_1 + \widetilde{e}_2) - c(\widetilde{e}_1) - c(\widetilde{e}_2) \end{aligned} \tag{6}$$

where $\widetilde{e}_1$ and $\widetilde{e}_1$ are the optimal period-zero effort plans when the agent solves the standard problem $V_0 = c'(e)$ (the solution to the first-order condition with respect to effort in equation (4) under no piece rate—$w_t = 0$—given that agents are not told about the piece rate during recruitment). Equation (6) shows that if $V_0 = 0$ (the agent derives no pleasure from sampling cookies and thus has no utility for the thank-you cookies $T = 0$) then the agent only experiences costs of participating but no benefits ($-c(\widetilde{e}_1) - c(\widetilde{e}_2) < 0$). As a result the agent would reject the offer in favor of doing nothing. Further, notice that beyond the thank-you cookies T being only valuable if $V_0 > 0$, equation (6) shows that the higher the agent's intrinsic utility for sampling cookies (the higher the V_0), the more likely they are to participate in the tasting.

F.5 Agent Behavior in Period (Week) One Under the Three Models

Let e_t^g, $t \in \{1, 2\}$, and $g \in \{c, a, u\}$ denote the period-t behavior of Control, Anticipated and Unanticipated, respectively. We assume V_0 is large enough to ensure interior

solutions.[42]

Period-One Behavior. Subjects solve $e_1 = \text{argmax}_e(V_0 + (1-\beta)w)e - c(e) + \eta w\mu\big(e - \bar{e}\big) + \eta\mu\left(c(\bar{e}) - c(e)\right)$ with first order condition (f.o.c.),

$$(V_0 + (1-\beta)w) + \eta w\mu'_k = \left[1 + \eta\mu'_e\right] c'(e_1) \tag{7}$$

where $\mu'_k \equiv \mu'\big(e_1 - \bar{e}\big)$ and $\mu'_e \equiv \mu'\left(c(\bar{e}) - c(e_1)\right)$. This first-order condition highlights that there exist a reference-dependent marginal benefit and reference-dependent marginal cost. The reference-dependent marginal benefit ($\eta w\mu'_k$), proportional to w, arises from reference-dependent reciprocity: because agents were not expecting w but get it, they have an incentive to set effort above the expected level to, for example, reciprocate the principal's increased wage. The reference-dependent marginal cost, arises from reference-dependent effort: the agent's having to expend more effort than he had expected ($\eta\mu'_e c'(e_1^u)$).

Behavior of Control **during period one**. Subjects in Control receive no incentive pay, so they experience no crowding out. They also expect no incentive pay and receive none so they also experience no departures from expectations ($\widetilde{w}_1 = w_1 = 0$). As a result, their f.o.c. from 7 simplifies to $V_0 = c'(e_1)$ with optimal effort $V_0 = c'(e_1^c) \equiv c'(\bar{e})$, where $\bar{e}$ denotes the benchmark effort.

Behavior of the treatments during period one. Proposition 1 compares the first-period behavior of Unanticipated and Anticipated with that of Control during the reward period for the three models. All proofs are in Appendix G.

Proposition 1 *(First-Period Effect of the Piece Rate Under a Standard, Crowding-Out and Unmet-Wage-Expectations Model)*

[42] In particular, we assume $V_0 > \max\{(1-\beta)w, \beta w + \lambda^2 w\eta\}$ for $\beta \geqslant 0$. See equations (7) when $\eta = 0$ and (8).

(1) Standard model ($\beta = 0$ and $\eta = 0$): For any w, $e_1^u = e_1^a > \bar{e}_1$.

(2) Crowding-out Model ($\beta > 0$ and $\eta = 0$)

(2.i) Suppose $\beta \leqslant 1$. For any w, $e_1^u = e_1^a \geqslant \bar{e}_1$.

(2.ii) Suppose $\beta > 1$. For any w, $e_1^u = e_1^a < \bar{e}$.

(3) Unmet-Wage Expectations Model ($\beta = 0$ and $\eta > 0$). For any w, $e_1^a > \bar{e}_1$. Moreover, if w is big enough, $e_1^u > \bar{e}_1$.

Proposition 1, part (1) shows that, under a standard model, the first-period effort in both treatments (UNANTICIPATED and ANTICIPATED) should be higher than that in CONTROL: agents work more when they are paid the piece rate than when they are not.

Part (2), shows that, under a crowding-out model effort in either treatment can be greater or smaller than that of CONTROL depending on β. Recall that β determines the size of the crowding-out effect, which is compared to the standard incentive effect of the piece rate. Whenever $\beta \leqslant 1$ the incentive effect of the piece rate outweighs that of crowding out, leading subjects to exert more effort than those in CONTROL. In contrast, whenever $\beta > 1$, the incentive effect of the piece rate is outweighed by that of the crowding out, leading subjects to exert less effort than those in CONTROL. The idea that the net first-period effect of the piece rate can be positive or negative is consistent with Deci, Koestner, and Ryan (1999) and Bénabou and Tirole (2003).

Part (3) shows that, under an unmet wage expectations model, behavior in ANTICIPATED conforms to a standard model as there are no departures from expectations (agents are expecting the introduction and withdrawal of the reward). Thus their effort should be higher than that in CONTROL. Effort in UNANTICIPATED can be larger than that in ANTICIPATED if the wage w is high enough to so that its incentive effect and the pleasant surprise of this payment compensate agents for the unpleasant surprise of having to exert more effort than

expected.

F.6 Agent Behavior in Period (Week) Two Under the Three Models

Behavior of the CONTROL **during period two**. Subjects in CONTROL solve the same problem as that in the first period because receiving T (the thank-you cookies) do not depend on the piece rate or on effort.

Behavior of the treatments during period two. Since $V_2 = V_1 = V_0 - \beta w$, and given the second-period effort plan $\tilde{e}_2$, subjects solve $e_2^u = \text{argmax}_e(V_0 - \beta w)e - c(e) + \alpha V_0 + \eta\mu(-w)\mu\big(e - \tilde{e}_2\big) + \eta\mu\left(c(\tilde{e}_2) - c(e)\right)$ with f.o.c,

$$(V_0 - \beta w) - \lambda w \eta \mu'_k = \left[1 + \eta\mu'_e\right] c'(e_2^u) \tag{8}$$

where $\mu'_k \equiv \mu'\big(e_2^u - \tilde{e}_2\big)$ and $\mu'_e \equiv \mu'\left(c(\tilde{e}_2) - c(e_2^u)\right)$.

The negative term $(-\lambda w \eta \mu'_k < 0)$ captures the intuition that because the agent's expectation of receiving a payment was not fulfilled, he has an incentive to decrease effort below what he had planned to exert, to, for example, retaliate. Proposition 2 formalizes this hypothesis.

Proposition 2 *(Second-Period Effort in the Absence of the Piece Rate Under a Standard, Crowding-Out and Unmet-Wage-Expectations Models)*
(1) Standard model ($\beta = 0$ and $\eta = 0$): $e_2^u = e_2^a = \bar{e}_2$
(2) Crowding-out Model ($\beta > 0$ and $\eta = 0$): $e_2^u = e_2^a < \bar{e}_2$.
(3) Unmet-Wage Expectations Model ($\beta = 0$ and $\eta > 0$): $e_2^a = \bar{e}_2$ and $e_2^u < \bar{e}_2$

Proposition 2, part (1) shows that under a standard model, the second-period effort in both (UNANTICIPATED and ANTICIPATED) in the absence of the piece rate should be the same as the unpaid CONTROL.

Part (2), shows that, under a crowding-out model effort in either treatment should be lower than that in CONTROL: the first-period piece rate eroded interest in the task, resulting in a decline in effort in period two.

Part (3) shows that, under an unmet wage expectations model, behavior in ANTICIPATED conforms to a standard model as there are no departures from expectations (agents are expecting the introduction and withdrawal of the reward). Thus their effort should be similar to that in CONTROL. Effort in UNANTICIPATED will be lower due to unmet wage expectations, which might induce lower effort due to, for example, retaliation.

G Proofs

Proof of Proposition 1

(1) Suppose $\beta = 0$ and $\eta = 0$. Then first order condition in equation 7 becomes $(V_0 + w) = c'(e_1) > c'(\bar{e}_1)$.

(2.*i*) Suppose $\beta \leqslant 1$ and $\eta = 0$. Straight from the first-order condition in equation (7): if $\beta \leqslant 1$ this implies $(1-\beta)w \geqslant 0$ for any w and thus that $V_0 + (1-\beta)w = c'(e_1) > V_0 = c'(\bar{e}_1)$.

(2.*ii*) Suppose $\beta > 1$ and $\eta = 0$. Straight from the first-order condition in equation (7): if $\beta > 1$ this implies $(1-\beta)w < 0$ for any w and thus that $V_0 + (1-\beta)w = c'(e_1) < V_0 = c'(\bar{e}_1)$.

(3) Suppose that $\beta = 0$ and $\eta > 0$. From the first order condition in equation 7

$$c'(e_1) = \frac{(V_0 + (1-\beta)w) + \eta w \mu'_k}{[1 + \eta \mu'_e]}$$

Assume w is sufficiently large, in particular, $w \geqslant \frac{V_0 \eta \lambda}{1+\eta} > 0$. This condition over w can thus be written as,

$$V_0 + w + \eta w \geqslant V_0(1 + \eta\lambda)$$

by adding V_0 to both sides. Because in equilibrium $\mu'_k \equiv \mu(e_1^u - \bar{e}_1) = 1$ and $\mu'_e \equiv \mu(c(\bar{e}_1) -$

$c(e_1^u)) = \lambda$, using the f.o.c and the equation above we have

$$c'(e_1) = \frac{(V_0 + w) + \eta w \mu_k'}{[1 + \eta \mu_e']} \geqslant V_0 = c'(\bar{e}_1)$$

Proof of Proposition 2

(1) Suppose $\beta = 0$ and $\eta = 0$. From the first order condition in equation 8, $V_0 = c'(e_2) = c'(\bar{e}_2)$.

(2) Suppose $\beta > 0$ and $\eta = 0$. From the first order condition in equation 8, $(V_0 - \beta w) = c'(e_2) < c'(\bar{e}_2)$.

(3) Suppose $\beta = 0$ and $\eta > 0$. Agents who expected the withdrawal of the reward (ANTICIPATED) experience no gain/loss utility from expectations. So they solve the same problem as that in part (1). For agents who did not expect the withdrawal of the piece rate (UNANTICIPATED) condition, $e_2 < \bar{e}_2$. We see this by noticing that because μ is strictly increasing it must be the case that,

$$-\lambda w \eta \mu_k' < V_0 \eta \mu_e'$$

where μ_k' and μ_e' are defined as in equation (8). Adding and subtracting convenient terms, we can express this inequality as:

$$c'(e_2) = \frac{V_0}{[1 + \eta \mu_e']} - \frac{\lambda w \eta \mu_k'}{[1 + \eta \mu_e']} < V_0 = c'(\bar{e}_2) \tag{9}$$

H The Cookie Evaluation Form

Cookie Number

Cookie Tasting Score Sheet

Initials: _____
Subject ID: _____

Starting Time: _____
Ending Time: _____

Excellent=1, Very Good=2, Good=3, Fair=4, Poor=5

Appearance.

Does it look chewy?	Yes	No
Does it look fresh?	Yes	No
Does it look hard?	Yes	No
Does it look rich?	Yes	No
Does it look home-baked?	Yes	No
Does it look colorful?	Yes	No

Overall, how does it rate on "appearing desirable"? _____

Aroma.

Does it have a strong smell?	Yes	No
Does it smell home-baked?	Yes	No

Overall, how does it rate on "having an attractive aroma"? _____

Snap.

Does it have a clean snap?	Yes	No
Does it break easily?	Yes	No
Is it hard?	Yes	No

Overall, how does it rate on "breaking nicely"? _____

Texture.

Does it crumble?	Yes	No
Is it crunchy?	Yes	No
Is it chalky?	Yes	No
Does it melt on your mouth?	Yes	No

Overall, how does it rate on "having a nice texture"? _____

Cookie Tasting Score Sheet

Start.

Does the flavor develop quickly?	Yes	No
Does one particular flavor develop too quickly?	Yes	No

Overall, how does it rate on "flavor developing nicely"? _____

Flavor.

Does it have a chocolaty flavor?	Yes	No
Does it have a buttery flavor?	Yes	No
Does it have a peanuty flavor?	Yes	No
Does it have a strawberry-like flavor?	Yes	No
Does it have an almondy flavor?	Yes	No
Does it have a ginger-like flavor?	Yes	No
Does it have a minty flavor?	Yes	No
Does it have a cheesecake-like flavor?	Yes	No
Does it have a strong flavor?	Yes	No
Does it have a natural flavor?	Yes	No

Overall, how does it rate on "having a nice flavor?" _____

Which is the overall rating of this cookie? _____

* You don't have to sample all the cookies: there is no predetermined amount of cookies to taste. Sample as many cookies as you wish.

* You don't have to eat the whole cookie: you can eat as much of each cookie as you want. Leave any leftovers in the corresponding cups.

I Protocol

Students interested in the blind-cookie tasting contacted the research assistant via the telephone number or email in the campus flyer.

(1) *Recruiting wording when interested subjects contacted the research assistant.*

Thank you for your interest in this cookie-tasting study. This is a research study on cookie preferences, which involves tasting and evaluating several brands of cookies. The tasting takes place on two separate sessions, one week apart, at *[campus address of the tasting facilities]*. In each of these sessions you will need to rate cookies along a few dimensions, such as taste and aroma. You can taste cookies for as long as you like but for no more than three hours. At the end of the second session, as a thank-you gift for having participated in the cookie-tasting study, you will receive a large Godiva luxury cookie tin.

We are currently still recruiting participants, but the target start date is the *[dates]*.

Before continuing, I will ask you a few questions

1. Are you a *[University Name]* student?
2. Are you 18 or older?
3. Do you know anyone else who is participating in this activity or has participated in the past? *[If yes, research assistant asks who]*.

Thank you.

Would you be willing to participate?

[If subject agrees]. What times and days work?

It is important that you pick the same day and time slot in both weeks. Please

choose time slots of three hours for ease of scheduling and to comply with room availability restrictions. You are not required, however, to taste cookies for 3 hours as I mentioned above.

Finally, can I have your:

1. Name
2. Phone number
3. Email

Thank you. I will be in touch and let you know when we will be starting the study. In the meantime, please feel free to ask me any questions."

I.1 Protocol for CONTROL

(C1) **The reminder/confirmation email the day before the start of the first session**.

Hello *[Name of the volunteer]*,

Thank you for participating in the cookie experience.

You are scheduled for:

Session 1: *[date]* at *[time]*

Session 2: *[date]* at *[time]*

Both tasting sessions take place on *[campus address of the tasting facilities]*. There will be a cookie sign that will help you find the room. If you have trouble finding the entrance, please give me a call to *(XXX) XXXX-XXXX*.

Remember that at the end of the second session you will receive a Godiva luxury cookie tin as a thank-you gift for your participation.

We are looking forward to see you!

[Name of the research assistant]

Project Coordinator?

(C2) **Protocol for the first tasting session**.

(C2.1) *Welcoming wording while walking the subject to the room*

Welcome! Thank you for participating in this blind cookie tasting experience. We have wonderful cookies for you to taste in this session and the next. You will receive a Godiva luxury cookie tin at the end of the second session.

(C2.2) *Wording once in the room*

Before starting this session please fill in this small questionnaire *[Research assistant hands in the demographic questionnaire]*. You also have to read and sign this consent form *[Research assistant hands in the consent form]*. Please note that the ingredient list for each cookie is available in case you have any food-allergy concerns. If you are diabetic, please be aware that cookies contain sugar. If you have any questions, please let me know. *[Research assistant waits for the subject to finish completing the forms]*.

(C2.3) *Wording explaining the task*

These are the cookies. There are 70 of them. You can evaluate as many as you want. Importantly, there is neither a fixed time to evaluate each sample nor a fixed amount of each cookie to taste. Evaluate as many cookies as you feel like. The only restriction is that the rooms is only available for three hours.

To evaluate a cookie, you need to taste the cookie and rate it on this evaluation sheet *[The research assistants shows the evaluation form]*. You do not have to

eat the whole cookie. Just as much of it as you want. Please, leave any leftovers in their corresponding paper cups.

Each cookie is rated with a number between 1 and 5 along each of the following dimensions: appearance, aroma, snap, start, texture, start and flavor. To identify the cookies, you have to write down the cookie number on this box at the top of the evaluation sheet. Also, on each sheet you have to write down your name and initials and the starting and finishing time. There is water if you need to clear your palate and there are napkins as well.

(C2.4) *Wording instructing subjects on how to leave the room*

Once you are done with your tasting, please text me your room number or call me and I will come to pick up your evaluations and check you out. If you need to leave the room temporarily or have any questions, please let me know. That is all. Do you have any questions? Happy tasting!

(C2.5) *Wording used as farewell once subjects finished the task*

Thank you. Did you enjoy the experience? *[Research assistant listens the answer]*. Remember that your next evaluation session will be exactly one week from today at this same time and in the same location, where you will be given a different set of cookies to evaluate. It is important that you repeat your experience so we can gain a better understanding of your preferences. I look forward to seeing you next week! *[Research assistant walks the subject out to the exit]*.

(C3) **Protocol for the second tasting session**.

(C3.1) *Welcoming wording while walking the subject to the room*

Welcome! We have wonderful cookies for you to taste in this second session.

(C3.2) *Wording once in the room*

Before starting this session please fill in this small demographic questionnaire *[Research assistant hands in the questionnaire].* Just answer questions 4 (level of hungriness and 5 (time of last meal). As in the previous session remember that the ingredient list for each cookie is available in case you have any food-allergy concerns. If you are diabetic, please be aware that cookies contain sugar. If you have any questions, please let me know. *[Research assistant waits for the subject to finish the questionnaire].*

(C3.3) *Wording explaining the task*

These are the new cookies. As in the previous session, there are 70 of them. Remember, you can evaluate as many as you want. Just as in the previous session, there is neither a fixed time to evaluate each cookie nor a fixed amount of each cookie to sample. Evaluate as many cookies as you feel like. The only restriction is that the room will only be available for three hours. The evaluation sheets are the same as before. To evaluate a cookie, you need to taste the cookie and write down your ratings on this evaluation sheet. You do not have to eat the whole cookie. Just sample as much of it as you would like. Each cookie is rated with a number between 1 and 5 along each of the following dimensions: appearance, aroma, snap, start, texture, start and flavor. To identify the cookies, you have to write down the cookie number on this box at the top of the evaluation sheet. Also, on each sheet you have to write down your name and initials and the starting and finishing time. There is water if you need to clear your palate and there are napkins as well.

(C3.4) *Wording to instruct subjects on how to leave the room.* Same as (C2.4).

(C3.5) *Wording used as farewell once subjects finished the task*

> Thank you for your participation! Did you enjoy the experience? *[Research assistant listens the answer].* Here is your Godiva thank-you cookie tin.

I.2 Protocol for Unanticipated

Same as the protocol for the Control with two exceptions:

(1) At the end of the wording in (C2.3), in the first session, the research assistant states: "Finally, we have a surprise for you. We will be able to pay you 75 per each cookie you evaluate. Once you are done with your tasting, please contact me and I will come to pick up your evaluations. Then I will count how many cookies did you evaluate and I will give you the money before checking you out. That is all. Do you have any questions? If you have any questions, please let me know. Happy tasting!" (2) At the end of the wording in (C3.3), in the second session, the research assistant states: Finally, there is no monetary payment per cookie tasted in this session.

I.3 Protocol for Anticipated

Same as the Control with the exception that volunteers were informed in advance that they would be paid in the first session but not in the second. This was accomplished via a phone call and an email in case the research assistant could not reach the subject that day and had left a message in voice mail. In this latter case, the research assistant followed-up the following day to make subjects had received this information.

(1) The phone call and/or the email to inform subjects in advance about the payment scheme. The phone called occurred only after all volunteers had been recruited and randomly assigned to the treatments.

Hello *[Name of the volunteer]*,

I am the project coordinator of the blind cookie tasting experience for which you are participant. I am calling/emailing you to inform you that you will receive 75 cents per each cookie you evaluate in your first tasting session. However, there will be no payment in the second tasting session. Are you fine with receiving this payment in the first session and no payment in the second session? *[After subject agrees]* Thanks. We will send you a reminder email before your first tasting session with you of your schedule, the tasting facilities and of the payment scheme. We are looking forward seeing you!

(2) The reminder email in (1) added this "As we discussed over the phone, you will receive 75 cents per each cookie you evaluate in your first session only. There will be no payment in the second tasting session."

(3) The wording in (C2.3) for the first session added, at the end: "Finally, let me remind you that in this session we will pay you 75 cents per each cookie you evaluate. Once you are done with your tasting, please contact me and I will pick up your evaluations and check you out. Then I will count how many cookies did you evaluate and I will give you the money before checking you out. That is all. Do you have any questions? If you have any questions, please let me know. Happy tasting!"

(4) Wording in (C2.5) for the first session added, at the end: "Finally, remember that there will be no monetary payment for the next session."

(5) Wording in (C3.3) for the second session added at the end: "Finally, remember that there will be no monetary payment for this session."