Do Monetary Incentives Crowd Out Intrinsic Motivation? A Field Test in the Workplace.*

Constança Esteves-Sorenson† Rosario Macera‡ Robert Broce§

August 2014

Abstract

Economists have long been intrigued by an influential strand of psychology positing that monetary payments may crowd out an agent’s intrinsic interest in a task and thus harm performance. We present the first explicit test in economics for whether this crowding out occurs in the workplace. Further, we investigate whether the evidence for crowding out in psychology is consistent with a different hypothesis—retaliation or disappointment from expected-but-unfulfilled payments—formalized in a novel model of reference-dependent reciprocity. Our field experiment, building on the two-session canonical test in psychology for crowding out, invites intrinsically motivated volunteers for a blind tasting, a common task in market research. After dealing with the typical confounds in the canonical test—ambiguity in the identification of an intrinsically motivating task, satiation and fatigue—we find evidence not for crowding out, but rather for the retaliation or disappointment hypothesis.

JEL Codes: C93, D03, J33, M52, M55.

*We thank Kaitlyn Croteau, Cathy Le, Tiffany Lin, Karen O’Brien and Ignacio Osorio for excellent research assistance. We also thank Jason Abaluck, Daylian Cain, Jason Dana, Kelsey Jack, Lisa Kahn, Botond Köszegi, Ian Larkin, Ulrike Malmendier, Stephan Meier, Pedro Rey-Biel, Olav Sorenson, Amy Wrzesniewski and seminar participants at Pontificia Universidad Católica de Chile, Universidad de Chile, University of Warwick, Yale University and the 36th Annual NBER Summer Institute in Personnel Economics for comments. 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.

‡rosario.macera@uc.cl, Pontificia Universidad Católica de Chile.

§brocer1@southernct.edu, Southern Connecticut State University.
1 Introduction

The effectiveness of monetary payments lies at the crux of economics. An influential strand of research in psychology, however, has argued that there is one important case where payments may undermine performance: when an agent is intrinsically motivated, that is, “when he receives no apparent rewards except the activity itself” (Deci, (1971) page 105). In this case, payments may feel controlling (e.g., Deci, Koestner, and Ryan (1999)) or lead to the perception that the task is not interesting (e.g., Lepper, Greene, and Nisbett (1973)), crowding out the agent’s intrinsic interest in it and thus lowering effort.

Economists, intrigued by the idea that monetary incentives may crowd out agents’ inherent enjoyment in a task and thus undermine performance in the workplace, have asked for conclusive evidence of this phenomenon. Gibbons (1998, page 130), for example, argued that “Field experiments […] would be especially useful” on the issue of whether economic incentives dampen employees’ intrinsic motivation. Rebitzer and Taylor (2011, page 765) recently renewed this call when reviewing incentives in labor markets: “Although […] the evidence is not yet conclusive, we are intrigued by the notion that extrinsic rewards can undermine intrinsic motives.”\(^1\)

We answer this call by testing whether payments harm performance in an intrinsically motivating workplace task and, further, by investigating whether the evidence of crowding out in psychology could be consistent with a different mechanism: retaliation or disappointment. This alternative hypothesis is motivated by the design of the canonical test for crowding out in psychology. In this two-period test, a reward is introduced unexpectedly in the first period and withdrawn in the second. Then, whenever there is a shortfall in outcomes (e.g., time spent on the task)—relative to a control group, which receives no payment—this is taken as evidence of crowding out (see Deci and Ryan (1985) for an extensive review of all studies using this two-period design). The rationale is that the first-period payment undermined subjects’ perception of the task leading to a worse second-period outcome than would have occurred in the absence of a first-period reward. By using a novel model of reference-dependent reciprocity, we demonstrate that this performance shortfall could instead be due to the unexpected first-period reward creating an expectation of a second-period payment, which, when unfulfilled, leads workers to retaliate or to be disappointed, decreasing their performance.

\(^1\)For another review of the evidence on the interaction of monetary incentives and crowding out, see Kamenica (2012).
also addresses the three major confounds in the standard two-period test in psychology: ambiguity in the identification of intrinsically motivated subjects, satiation and fatigue. We find no evidence of crowding out, but instead find support for the retaliation or disappointment hypothesis. Our results therefore suggest that prior evidence for the crowding out of intrinsic motivation by payments could have been driven instead by unmet payment expectations.

Our field experiment starts by replicating the two-period structure of the crowding-out test in psychology in a workplace setting. Students on two campuses were invited to blind-taste and rate cookies, a common task in market research. Volunteers took part in two sessions, each one week apart, and received thank-you cookies at the end of the second session for participating. To allow for significant variation in outcomes, subjects were nearly unconstrained in the number of cookies they could taste and in the time they could spend tasting: up to 70 cookies per three-hour session.

To test for crowding out and to distinguish it from the retaliation or disappointment hypothesis, we randomly assigned volunteers to three groups. Those in the Control tasted and rated cookies in both sessions, only receiving the thank-you cookies at the end, as promised. Those in the Surprising treatment were surprised with a $0.75 piece rate per cookie evaluated in the first session (but received no information about a second-session payment), while at the beginning of the second session they were informed that the piece rate would not be granted. This treatment mirrors that in the canonical test of crowding out in psychology, whereby a shortfall in second-period outcomes in the Surprising treatment versus the Control, when the payment is withdrawn, is ascribed to crowding out. We argue, however, that this shortfall could also be due to retaliation or disappointment stemming from the withdrawal of the expected reward. To separate these two hypotheses, we introduce a treatment where unmet payment expectations do not affect outcomes, as the introduction and the withdrawal of the payment are disclosed in advance: Volunteers in the Anticipated treatment are told one week before the first session that they will receive the $0.75 piece rate in the first session but not in the second. This treatment therefore isolates the role of crowding out by payments in lowering performance.

Our design also improves on the canonical test in psychology by dealing with its three typical confounds. First, a necessary condition for payments to harm performance under the crowding out hypothesis is that subjects are intrinsically motivated. Offering no monetary reward for participating in the tasting, but rather rewarding participants with additional cookies with no resale value, ensured
that only those with an intrinsic liking for sampling cookies selected into the tasting. Selection into the task not only solves the thorny issue of assessing whether subjects inherently enjoy the activity (Frey and Jegen (2001)) but also does so while not relying on subjective criteria, such as an arbitrary task enjoyment rating.

Second, separating the tasting sessions by one week and allowing subjects to sample each cookie ensured that any productivity shortfalls in the second session would not be due to satiation or fatigue. Namely, if the increased consumption induced by the piece rate led subjects to feel full in the first session, despite their being allowed to taste a small portion of each cookie, they would have a week to recover. Similarly, if they felt tired from filling out evaluation forms in the first session, they would also have a week to rest. This temporal separation between sessions thus addressed the two additional confounds of satiation and fatigue, which are common in the canonical two-period test in psychology, in particular with adult subjects, where the non-reward session immediately follows the reward period (Deci, Koestner, and Ryan (1999), page 650). Because of this temporal proximity, outcome shortfalls during the non-reward session could stem from satiation or fatigue due to the increased effort in the first session induced by the payment, instead of from crowding out.

Our main result is that we find no evidence for crowding out in the workplace: a decline in outcomes following the withdrawal of the payment can be ascribed instead to retaliation or disappointment. When the piece rate was introduced unexpectedly in the first session and removed in the second session (SURPRISING treatment), subjects spent less time on the task than those in the CONTROL during the second session. This shortfall in the time spent on the activity during the non-reward period is consistent with the crowding-out findings in psychology. However, when expectations about payments did not play a role, that is, when subjects knew in advance that they would not receive a piece rate in session two (ANTICIPATED treatment), they spent the same amount of time on the task as those in the CONTROL during the non-reward session. Thus the shortfall in the time spent on the task after the withdrawal of a payment found in the canonical two-period test in psychology could be caused by retaliation or disappointment following unmet payment expectations rather than by the crowding out of intrinsic motivation. We found a similar pattern in the quality of the evaluations: Subjects in the SURPRISING treatment lowered the quality of their evaluations below that in the CONTROL during the non-reward period. In contrast, during the same period, those in the ANTICIPATED treatment did not.
This paper is the first direct test in economics of whether payments harm intrinsic motivation and thus performance in non-prosocial tasks, the norm in the workplace. Intrinsically motivating non-prosocial tasks are those where the agent derives utility from the task itself, such as a researcher enjoying exploring ideas or a professional photographer enjoying taking pictures. Monetary payments have been thought to undermine effort in these tasks because they may, for example, increase the perceived expected cost of performing them (Bénabou and Tirole (2003)). Our test thus differs from others in economics, which have focused on testing whether payments undermine prosocial behavior, such as donating to charities or giving blood. Payments may undermine outcomes in these prosocial tasks—tasks which benefit others than the agent and the principal—because they may spoil the agent’s signal of being prosocial (Bénabou and Tirole (2006)).

Further, even though there are several studies in psychology on whether payments harm performance in intrinsically motivating non-prosocial tasks, such as puzzle-solving in the seminal crowding out study by Deci (1971), the evidence is inconclusive. Tests on non-prosocial tasks with adults have been mostly conducted in laboratory settings (except for two field studies, one in Deci (1971) and another in Boal and Cummings (1981)). And though this body of research finds suggestive evidence for crowding out, there is no uniform agreement on the robustness of its findings due to the three confounds addressed by our design (see the discussion between Deci, Koestner, and Ryan (1999) and Cameron, Banko, and Pierce (2001) on the role of ambiguity in the definition of intrinsically motivated subjects, fatigue and satiation in the robustness of the evidence for crowding out).

Importantly, our paper proposes and tests a novel alternative explanation for the apparent evidence supporting the crowding-out hypothesis: retaliation or disappointment from expected but unfulfilled payment. Using a novel model, which combines the theory on social and expectation-based reference-dependent preferences, we demonstrate that unmet payment expectations could decrease performance.

---

2We discuss the most relevant literature in psychology and in economics, such as Deci (1971) and Gneezy and Rustichini (2000), in the following Section 2.1.

3The evidence that payments undermine outcomes in prosocial tasks is mixed. Ariely, Bracha, and Meier (2009) found supportive evidence that monetary incentives induced effort contributions to a charity only when these incentives could not be publicly observed. And though Mellström and Johannesson (2008) found that payments 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 (forthcoming) also find suggestive evidence that financial incentives do not crowd out intrinsic motivation among volunteers for a task with a strong prosocial component: the sale of female condoms for HIV prevention. Chetty, Saez, and Sandór (2014) also find suggestive evidence that incentives do not crowd out prosocial behavior in the provision of referee reports. For a review of the effect of incentives in non-employments settings, such as education and contributions to public goods, see Gneezy, Meier, and Rey-Biel (2011).
and generate the outcome shortfalls that have been attributed to the crowding out in the canonical two-period test.\textsuperscript{4} Our field experiment, moreover, suggests that this mechanism—rather than the crowding out of intrinsic motivation—accounts for the decline in performance following the withdrawal of a reward.

2 Field Experiment Design

This section describes the field experiment and how it tests for crowding out and the alternative hypothesis of retaliation or disappointment due to unmet wage expectations. To provide context for the design, we start by describing the leading evidence in psychology for crowding out in non-prosocial tasks.

2.1 Prior Tests of Crowding Out in Non-Prosocial Tasks

In the most cited paper of this literature, Deci (1971) investigated whether payments undermined intrinsic motivation, and thus performance, with two tests. In the first, students were asked to solve puzzles in three sessions to fulfill a class requirement. This task was selected because it was presumed that subjects would be intrinsically motivated to do it. Subjects were randomly assigned to a control and a treatment group of 12 subjects each. In each session they solved puzzles in the presence of the experimenter for about one hour. Those in the treatment group were surprised with a $1 piece rate per puzzle in the second session (the reward session). At the beginning of the next session, however, they were informed that the piece rate would be withdrawn due to lack of funds (the non-reward session). Those in the control did not receive any payment in any session. Intrinsic motivation was measured as the time subjects spent working on puzzles during an eight-minute window in which the experimenter left the room (the “free-choice” window). He found that when subjects were given the unexpected piece rate, they spent more time than the control solving puzzles during the free-choice window. However, after this payment was withdrawn, subjects spent less time than the control solving puzzles during the free-choice window, a statistically significant difference at the 10% level in a one-tailed test. Whether students enjoyed the task was verified post-experiment when they rated it as highly enjoyable on a nine-point scale. However, there were no differences in enjoyment ratings across groups or sessions.

The influential Deci (1971) paper also describes a field experiment, one of the few in this literature,\textsuperscript{4} Other models, by contrast, have explored other mechanisms by which incentives could undermine performance in employment settings: In Bénaïm and Tirole (2003) incentives undermine effort by increasing the agent’s perceived expected cost of effort or by decreasing his perception of his own ability; in Sliwka (2007), incentives signal the principal’s negative appraisal of the agent’s type; and in Ellingsen and Johannesson (2008) incentives signal that the principal is not worth impressing with effort.
providing additional evidence of a shortfall in outcomes following the withdrawal of the payment. The eight-student staff of a college newspaper were randomly assigned to a control and a treatment group. Four subjects in the treatment were offered $0.50 per headline written during three weeks. At the end of this period they were informed that funds had been exhausted and thus they would no longer be paid. The four subjects in the control were never paid. Intrinsic motivation was measured as the number of minutes spent per headline: the faster subjects worked, the more intrinsically motivated they were. Deci observed that in the three weeks after the piece rate was withdrawn, subjects in the treatment were slower than those in the control (with the caveat that the control attritted to two subjects).

In both these tests, the shortfall in outcomes versus the control during the non-reward period has been interpreted as evidence of the perverse effect of payments on intrinsic motivation. The two leading explanations for this phenomenon have been cognitive evaluation theory and the overjustification hypothesis. The first proposes that rewards can be interpreted as unpleasant controllers of behavior thus undermining a person’s “intrinsic motivation, which refers to doing something because it is inherently interesting or enjoyable” (Ryan and Deci (2000), page 55). The second postulates that if one is paid for an inherently interesting activity “the person might come to 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), page 130). This latter paper, the second most cited in this literature, uses this hypothesis to rationalize why nursery school children who were surprised with a prize for drawing spent less time on this activity during the subsequent non-reward period than those children who were never rewarded. Namely, when the reward was introduced, children attributed their interest in drawing to the reward and not the activity per se. Thus, when the reward was withdrawn and they had to rely solely on their intrinsic motivation to perform the task, they engaged less with it.

Tests for crowding out based on these two mechanisms necessitate a reward period followed by a non-reward one. Tests for overjustification require an initial period in which subjects are paid so that they misattribute their interest in the task to the reward, followed by a non-reward period where the consequences of the misattribution become apparent. Tests of cognitive evaluation theory also require a two-period design, as argued by Deci and Ryan ((1985), page 184), as in the first period there is a trade-off between the displeasing nature of the reward and its incentive effect, and therefore one cannot disentangle which dominates. The idea that the incentive effect of the payment can coexist
with crowding out is highlighted in our model and in Bénabou and Tirole (2003). Therefore, studies of crowding out have focussed on a two-period, between-subjects design (except for the three-period setup in Deci (1971)), where the control is unpaid in both periods and the treatment receives an unexpected reward in period one which is withdrawn in period two.

We argue, however, that this two-period design, though appropriate for testing the two theories above, might introduce the confound of retaliation or disappointment following unmet payment expectations. Namely, when expectations about future payments are unfulfilled, agents may decrease effort to retaliate (e.g., Bewley (1999)). This is supported by the evidence in, for example, Mas (2006), showing that police officers’ performance declines when their wages fall short of those expected from arbitration proceedings. Alternatively, agents may decrease effort temporarily because they are disappointed. Depressed mood states, such as disappointment, have been shown to undermine effort (e.g., Ellis, Thomas, and Rodriguez (1984), Hartlage, Alloy, Vázquez, and Dykman (1993)) by, for example, diverting cognitive resources from the task at hand to negative thoughts and worries (Baddeley (1972), Callaway and Dembo (1958), Easterbrook (1959)). Therefore, our test, which is based on the two-period canonical design for crowding out, aims to disentangle the role of retaliation or disappointment from the crowding out of the intrinsic interest in the task.

The standard two-period test, which may conflate crowding out with retaliation, must also ensure that subjects are intrinsically motivated to perform the task. This is crucial because if the task is not intrinsically motivating then payments will not undermine performance since “there is little or no intrinsic motivation to crowd out” (Deci, Koestner, and Ryan (1999), page 633). Whether subjects are intrinsically motivated in many of these tests has been a source of contention in psychology, as failures to replicate the crowding out have led to a debate on whether subjects were intrinsically motivated in the first place (see the discussion between Cameron, Banko, and Pierce (2001) and Deci, Koestner, and Ryan (1999)). This is partly because the typical measures used to assess intrinsic motivation for a task have relied on reasonable, but arbitrary, cut-offs in enjoyment ratings in a scale or in the time spent on the task prior to the start of the experiment. For instance, it is unclear whether those who rated their enjoyment as a five on a nine-point scale or spent four minutes on the task are intrinsically motivated, while those who rated it as a four or spent three minutes on it are not. Our test thus relies on a novel and non-arbitrary criterion to assess inherent interest in a task: selection into it.
Since a two-period design and an intrinsically motivating task are necessary conditions, research in economics showing that small payments may undermine performance in non-prosocial tasks has not ascribed these results to crowding out. Gneezy and Rustichini (2000), for example, recruited students to answer questions from an IQ test for a fixed fee for a single period. Once in the laboratory, those surprised with a low per-question piece rate underperformed versus the control who received no payment. They stated that the crowding out explanation is “less appropriate in our single-stage set up” (page 803), as they only had the reward period, which allowed misattribution to occur, but no subsequent non-reward session to assess whether it undermined performance. Further, it was unclear whether subjects were inherently interested in the activity, since intrinsic motivation was not measured and students were paid to participate. In Hossain and Li (2014) it was similarly unclear whether the activity was intrinsically motivating in their two-period setting. Using a clever framing of a data-entry task, they found that paying students to perform it did not crowd out their willingness to participate in subsequent data-entry work when it was framed simply as additional work (instead of as a favor to the principal). Since they did not measure whether students were inherently interested in entering data (and indeed were hired for a payment), it is possible that payments did not undermine subsequent participation as the task was viewed as “boring” and thus had little or no intrinsic motivation to crowd out.

2.2 The Field Experiment

Our design extends the canonical two-period test in psychology to investigate crowding out and whether it may be confounded with retaliation or disappointment following unmet payment expectations. Our novel model in Section 3 complements our design by explaining both of these hypotheses.

The field experiment comprised one leg on campus A in January 2012 and three additional legs on campuses A and B in April, June and July of 2012, respectively. Using two campuses allowed us to gather a larger sample and to assess whether the results would hold across two separate environments.

(1) Task and recruitment of intrinsically motivated subjects. Our experiment builds on a blind tasting, a common workplace task occurring often in market research for wines, sodas, cheese, cookies, and many other goods. We chose blind-tasting of cookies for two reasons. First, it is an engaging activity that naturally attracts volunteers. Second, it is a non-prosocial activity as subjects participate in it because of the utility they derive from the activity itself and not because of the benefits it provides to others (as with prosocial tasks).
The blind cookie-tasting was advertised on two campuses in Connecticut through flyers and electronic mailing lists. 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 blind tasting, the principal who commissioned the study was unknown to subjects so as not to bias their evaluations.

Selection into the study revealed whether subjects were inherently interested in sampling cookies since by enrolling they only gained additional thank-you cookies. These were offered for ecological validity as, in most tastings, participants receive a thank-you gift (e.g., gift-certificates). However, they had a $0 resale value due to their perishability. This allayed concerns that subjects volunteered for the tasting not because they liked sampling cookies but rather for the resale value of the thank-you cookies.

This revealed-preference approach to whether subjects are intrinsically motivated for an activity solves the difficult issue in field research of identifying intrinsically motivated agents. As Frey and Jegen (2001, page 591) state: “In spite of the seemingly simple definition offered by Deci, it is difficult, if not impossible, to determine which parts of an employee’s motivation to perform his or her job are intrinsic, and which are not”. Moreover, it has the advantage of not relying on subjective criteria for ascertaining intrinsic motivation, such as arbitrary cut-offs on enjoyment scales, as discussed in Section 2.1.

(2) Task implementation and how it deals with fatigue, satiation and other confounds. In each weekly session, subjects filled out an evaluation form after tasting each cookie. 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?”) and Flavor (e.g., “Does it have a minty

---

6Further, it is hard to disentangle crowding out of intrinsic motivation from other factors in several workplace settings. For example, Gubler, Larkin, and Pierce (2013) documented that an award program implemented to decrease plant-level absenteeism undermined productivity for the most punctual workers. They discuss the difficulty of ascribing this behavior to crowding out, as it could also be due to peer or manager pressure or cultural norms.
flavor?”) and Overall Rating (“What is the overall rating of this cookie?”). At the start of session one, subjects signed a consent form ensuring that, for example, they were informed about the potential for allergies (e.g., some cookies could contain nuts) and answered a short demographic questionnaire.

We view the filling out of each evaluation as the effort cost associated with the intrinsically interesting activity of sampling each cookie—e.g., examining it, smelling it, tasting it. The fact that even inherently enjoyable activities entail a cost mirrors the classic tests in, for example, Deci (1971), where subjects enjoyed solving puzzles or writing headlines, but incurred the mental effort cost of doing so.

To maximize the variability in outcomes, subjects were virtually unconstrained in the number of cookies they could taste, how much of each they tasted, and the amount of time spent tasting. They could sample up to 70 cookies per session (a binding constraint for only 17% of sessions) and eat as much or as little of each cookie as they would like (as conveyed by the research assistant and in the evaluation form). Finally, subjects could spend at most three hours at the tasting site due to scheduling limitations (this constraint was binding for just one subject).

However, to prevent satiation and fatigue, two potential confounds with crowding out, the two cookie-tasting sessions were scheduled one week apart and subjects were allowed to only taste each cookie. Namely, if despite being allowed to take a small bite of each cookie, subjects became satiated in the first session, they still had one week to recover. If they were fatigued from filling out evaluation forms, they also had one week to rest. Avoiding satiation and fatigue is important because they can cause a second-period effort shortfall following a first-period effort increase induced by the reward. Whether crowding out has been confounded by satiation or fatigue has been a topic of debate in psychology (see the discussion between Cameron and Pierce (1994) and Deci, Koestner, and Ryan (1999)) since in all the studies of crowding out with adults the reward session is followed immediately by the non-reward session (Deci, Koestner, and Ryan (1999), page 650). Finally, since we worried that declining marginal utility from eating cookies could also be confounded with crowding out, we offered a different set of 70 cookies in the first and second sessions.

We also ensured that differences in outcomes across the three conditions could not be due to unobserved differences in cookies or in research assistants. First, within a leg, campus and session, all subjects across the three groups where given the same set of 70 cookies to taste. Thus, differences in  

---

7See Appendix B for the cookie evaluation sheet.  
8See Appendix C for the protocol.
outcomes are not due to a different set of cookies across conditions. Second, all subjects interacted with the same research assistant who was blind to the research hypothesis. Thus differences in outcomes are not due to different research assistants or to demand effects.

Finally, subjects also worked alone to avoid peer effects, another potential confound (e.g., Falk and Ichino (2006), Mas and Moretti (2009)).

(3) Treatments and how they disentangle crowding out from retaliation or disappointment. Volunteers for the cookie tasting were randomly assigned without their knowledge to three treatments: CONTROL, SURPRISING and ANTICIPATED. Those in the CONTROL performed as agreed upon recruitment. They came to their assigned rooms in the two sessions, tasted and evaluated cookies, and received the thank-you cookies at the end of session two. The CONTROL thus establishes the baseline outcomes in the absence of a monetary compensation.

Those in the SURPRISING treatment were surprised with a piece rate of $0.75 per cookie evaluated, at the start of session one, immediately before starting the tasting. There was no mention of a payment in session two. One week later, at the start of session two, they were informed that they would not be paid the piece rate. This setup replicates that in the canonical two-period design in psychology where subjects receive a surprise payment in the first session, which is withdrawn in the subsequent session.

However, as discussed, observed shortfalls in outcomes in this treatment versus the CONTROL during the non-reward period can be due not only to crowding out—e.g., the previous period’s piece rate per cookie sampled and evaluated tarnishes the enjoyment of sampling cookies—but also to retaliation or disappointment from expected-but-unfulfilled payments. Our next treatment isolates the effect of payments on crowding out as expectations do not play a role.

Volunteers randomly assigned in the ANTICIPATED treatment received a telephone call or email one week before the first session, informing them that they would be paid a $0.75 piece rate per cookie in the first session but not in the second. Importantly, the piece rate was offered after recruiting had finished and all subjects had been randomly assigned to the three conditions, so their participation in the study was not due to the piece rate. Thus, this treatment only differs from the SURPRISING one in that subjects received advance notice of the payment structure and were thus unsurprised by the introduction and withdrawal of the reward. It therefore isolates the role of crowding out in the undersupply of effort during the non-reward period as expectations do not affect outcomes. In particular, both cognitive evaluation
theory and the overjustification hypothesis predict that the expectation of a contingent payment in session one should reinforce the undermining of intrinsic motivation as the reward becomes more salient, thus feeling more controlling or leading to more misattribution (Deci, Koestner, and Ryan (1999), page 630). See Appendix C with the protocol for each treatment.

We now formalize the retaliation or disappointment hypothesis before discussing our test’s results.

3 The Retaliation or Disappointment Hypothesis

This section formalizes the retaliation or disappointment hypothesis. Following Deci (1971) and Bénabou and Tirole (2003), we model intrinsic interest in the activity as consumption utility. We then combine a reduced-form crowding-out mechanism, where payments can erode consumption utility (as in Deci (1971)), with the displeasure resulting from an expected payment which is unfulfilled. We model this disutility from unmet payment expectations by assuming agents have reference-dependent reciprocal preferences where their reference point corresponds to recent outcomes (Kőszegi and Rabin (2006)).

The model shows that retaliation or disappointment can cause the temporary outcome shortfalls observed in prior crowding out research without undermining the agent’s intrinsic interest in the task.

There are three periods. In period zero, the principal offers the agent a non-contingent, non-monetary reward, $T$, to be granted in period two for performing a task in periods one and two (the thank-you cookies in our design). The agent accepts or rejects this offer. Let $w_t \in \{0, w\}$, $w > 0$ represent the period-$t$ piece rate. Further, let $V_t$ represent the period-$t$ marginal intrinsic utility from sampling one cookie (e.g., examining and tasting it) and $e_t \geq 0$ the effort required to fill out the evaluation form.

To capture in a reduced form the negative relationship between payments and intrinsic motivation in the context of the task, let $V_t, t = 1, 2$ evolve as $V_t = V_{t-1} - \beta w_t$ if $V_0 > 0$ and $V_t = 0$ if $V_0 = 0$, where $V_0 \geq 0$ is the agent’s endowment of marginal intrinsic utility and $\beta \geq 0$ is the crowding-out parameter.

The agent is intrinsically motivated if $V_0 > 0$ and monetary payments diminish intrinsic motivation if $V_0 > 0$ and $\beta > 0$. Whenever $V_0 > 0$ and $\beta = 0$ there is no crowding out. Notice that $V_t = V_{t-1} - \beta w_t$

---

9Though 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) and Gill and Prowse (2012).

10This definition of intrinsic motivation as an underlying preference reflecting the agent’s intrinsic interest in the task is that used in the psychology literature since the seminal paper by Deci (1971). This definition is also embraced by Bénabou and Tirole (2003). Recent theoretical literature, however, as also defined intrinsic motivation as actions exerted without a financial reward (see Kőszegi (forthcoming)).
rules out satiation: since $V_t$ does not depend on the number of cookies the agent tastes, only payments can decrease the agent’s intrinsic utility. We assume the consumption utility from payments is linear and the cost of effort is a positive, strictly increasing and convex function $c(e_t)$, $c(0) = 0$. Notice that since the cost of effort is separable across periods, first-period fatigue does not carry over to the next. Finally, we model $T$ as $T = \alpha V_0 1(t = 2)$, where $\alpha > 0$ and $1(t = 2)$ is an indicator function taking value 1 if $t = 2$. This functional form shows that the cookie tin is only valuable as long as the agent has an intrinsic valuation for cookies: if $V_0 = 0$ then $T = 0$.

(1) Agent’s Preferences. (1.1) Period-$t$ consumption utility. We assume agents experience consumption utility from wages and intrinsic motivation, effort and the thank-you cookies granted at the end,

$$(V_t + w_t)e_t - c(e_t) + T$$  \hfill (1)

(1.2) Period-$t$ expectation-based reference-dependent utility. We assume agents experience reference-dependent social preferences, where the reference point corresponds to their recent expectations about outcomes (Kőszegi and Rabin, (2006, 2007)). Let $(\tilde{V}_t, \tilde{w}_t, \tilde{e}_t)$ represent the expectation made in period $t - 1$ for $(V_t, w_t, e_t)$. Total period-$t$ expectation-based reference-dependent utility is,

$$\eta \mu (w_t - \tilde{w}_t) \mu (e_t - \tilde{e}_t) + \eta \mu (c(\tilde{e}_t) - c(e_t))$$  \hfill (2)

The function $\mu$ compares actual with expected outcomes following the intuition that the true carriers of utility are departures of the consumption level from a reference point (Kahneman and Tversky (1979)). As usual in the literature, $\mu(x)$ is a strictly increasing, piece-wise linear function with a slope of one if $x > 0$ and a slope of $\lambda > 1$ if $x < 0$, where $\lambda$ represents the loss aversion parameter showing how loses hurt more than same-sized gains please, and $\mu(0) = 0$. The parameter $\eta$ represents the importance of the reference-dependent domain relative to consumption utility.

The first term in equation (2) corresponds to reference-dependent reciprocity, the novel component of the utility function. If the agent receives a wage higher than expected ($\mu (w - \tilde{w}) > 0$) he has an incentive to exert more effort than expected to reciprocate ($\mu (w - \tilde{w}) \mu (e - \tilde{e}) > 0$). To the contrary, if the agent receives a wage lower than expected ($\mu (w - \tilde{w}) < 0$) then he has an incentive to exert less effort than expected ($\mu (w - \tilde{w}) \mu (e - \tilde{e}) > 0$).\footnote{The functional form in this reciprocity term extends the literature of reciprocity in levels after Rabin (1993), Charness and Rabin (2002), Gächter and Falk (2002), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006).} This can be due to retaliation because the agent derives utility from punishing the principal; or disappointment, as the agent shifts cognitive resources from the
task to negative thoughts or worries, as discussed previously.

(1.3) **Total period-\(t\) utility flow.** Corresponds to the summation of equations (1) and (2).

(2) **Equilibrium.** In environments with no surprises—as in the CONTROL and ANTICIPATED treatments—reference-dependent agents behave as consumption-utility maximizers: because agents form plans rationally, expectations are always met and reference-dependent utility is zero. Whenever information arrives (as in the SURPRISING treatment), this new information causes departures from expectations. Because agents choose effort immediately after being informed of the first-period payment, however, we assume expectations do not adapt and agents maximize utility conditional to the reference. Furthermore, because the agent does not foresee any information arrival, his period-zero and period-one plans are computed as consumption utility maximizers.

(3) **Timeline.** In period zero agents accept or reject the offer \(w_1 = w_2 = 0\) and \(T\). At the end of this period those in the ANTICIPATED treatment are informed of the new payment structure \(w_1 = w > 0\) and \(w_2 = 0\), while no new information is given to agents in the CONTROL and SURPRISING treatments. In period one agents exert effort and they receive \(w_1 \geq 0\) (\(w_1 = 0\) for the CONTROL, \(w_1 = w\) for the SURPRISING and ANTICIPATED treatments). At the end of the first period agents update their wage and effort expectations for period two. In the second period all treatments receive \(w_2 = 0\) and agents exert effort for the second time. See Figure D.1 for a visualization of the timeline.

(4) **Selection of Intrinsically Motivated Subjects.** Participation in the task reveals intrinsic motivation. To see this, consider the period-zero participation constraint where the agent does not expect to be paid (\(\tilde{w}_1 = \tilde{w}_2 = 0\)) and does not expect any departure from expectations. Period-zero indirect utility is,

\[
U_0 = (\tilde{V}_1 + \tilde{w}_1)\tilde{e}_1 - c(\tilde{e}_1) + T + \left[(\tilde{V}_2 + \tilde{w}_2)\tilde{e}_2 - c(\tilde{e}_2) + T\right]
\]

\[
= V_0(\alpha + \tilde{e}_1 + \tilde{e}_2) - c(\tilde{e}_1) - c(\tilde{e}_2)
\]

(3)

where \(\tilde{e}_1\) and \(\tilde{e}_1\) are the optimal period-zero effort plans solving \(V_0 = c'(e)\). Equation (3) shows that if \(V_0 = 0\) (the agent derives no pleasure from sampling cookies and thus \(T = 0\)) then \(U_0 = -c(\tilde{e}_1) - c(\tilde{e}_2)\)

---

12 Notice that this specification does not include reciprocity in levels and reference-dependent utility from payments and intrinsic motivation. This allows us to simplify the exposition while preserving the main forces behind our predictions. In the full model in Appendix F we show that our predictions are robust to adding these terms.

13 The standard utility maximization that takes place in this case differs from the equilibria proposed in Köszegi and Rabin (2006, 2007), which apply to environments of uncertainty rather than environments of full surprise.

14 There is no systematic research on how long expectations take to adjust. Frederick and Loewenstein (1999) suggest that adaptation varies across domains, others that adaptation occurs immediately (e.g., Gill and Prowse (2012)), within hours (e.g., Card and Dahl (2011)).
and the agent would reject the offer in favor of doing nothing since \(-c(\tilde{e}_1) - c(\tilde{e}_2) < 0\). Furthermore, since the thank-you cookies T are only valuable if \(V_0 > 0\) and equation (3) is increasing in T, the higher the agents’ intrinsic utility for sampling cookies, the more likely they are to select into the study.

Let \(e^g_t\), \(t \in \{1, 2\}\), and \(g \in \{c, a, s\}\) denote the period-\(t\) behavior of the CONTROL, ANTICIPATED and SURPRISING groups, respectively. From now on, whenever positive, we assume \(V_0\) is large enough.\(^{15}\)

(5) Period-One Behavior. (5.1) Subjects in the CONTROL group expected no payment and received no payment \((\tilde{w}_1 = w_1 = 0)\). Subjects thus solve \(e^c_1 = \arg\max_e (V_1 + w_1)e - c(e) = V_0 - c(e)\) with first order condition (f.o.c.) \(V_0 = c'(e^c_1) \equiv c'(\bar{e})\), where \(\bar{e}\) denotes the benchmark effort.

(5.2) Subjects in the ANTICIPATED treatment expected a piece rate, which is actually granted \((\tilde{w}_1 = w_1 = w)\). Subjects thus solve \(e^a_1 = \arg\max_e (V_1 + w_1)e - c(e) = (V_0 + (1 - \beta)w)e - c(e)\) with f.o.c.,

\[
V_0 + (1 - \beta)w = c'(e^a_1) \quad (4)
\]

(5.3) Subjects in the SURPRISING treatment expected no payment, but they are surprised by a piece rate \((\tilde{w}_1 = 0\text{ but } w_1 = w)\). Using \(V_1 = V_0 - \beta w_1\) and \(\bar{e} = \bar{e}\) in period one, subjects solve \(e^s_1 = \arg\max_e (V_0 + (1 - \beta)w)e - c(e) + \eta w \mu (e - \bar{e}) + \eta \mu (c(\bar{e}) - c(e))\) with f.o.c.,

\[
(V_0 + (1 - \beta)w) + \eta w \mu' = [1 + \eta \mu'] c'(e^s_1) \quad (5)
\]

where \(\mu'_k \equiv \mu'(e^s_1 - \bar{e})\) and \(\mu'_e \equiv \mu'(c(\bar{e}) - c(e^s_1))\). This first order condition shows that in the SURPRISING treatment there are two reference-dependent forces: a reference-dependent marginal benefit \((\eta w \mu'_k)\) and a reference-dependent marginal cost \((\eta \mu'_e c'(e^s_1))\). The marginal benefit, which is proportional to \(w\), arises from reference-dependent reciprocity: because agents were not expecting \(w\), they have an incentive to set effort above the expected one to reciprocate the principal’s kindness of offering a payment. The marginal cost, which does not depend on \(w\), arises from reference-dependent effort: because the agent has deviated from her plans, she experiences utility from unmet expectations.

Proposition 1 compares the first-period behavior of the SURPRISING and ANTICIPATED treatments against that of the CONTROL during this reward period. All proofs in Appendix E.

**Proposition 1** (First-Period Effect of the Piece Rate)

(i) Suppose \(\beta \leq 1\). For any \(w\), \(e^c_1 \geq \bar{e}\). Moreover, if \(w\) is big enough, \(e^c_1 \geq \bar{e}\).

(ii) Suppose \(\beta > 1\). For any \(w\), \(e^c_1 < \bar{e}\). Moreover, if \(w\) is small enough, \(e^c_1 < \bar{e}\).

---

\(^{15}\)To ensure interior solutions, we assume \(V_0 > \max\{(1 - \beta)w, \beta w + \lambda^2 w \eta\}\) for \(\beta \geq 0\). See equations (4) and (7).
Proposition 1 shows that the first-period effort in the SURPRISING and the ANTICIPATED treatments can be greater or smaller than that of the CONTROL depending on $\beta$. Intuitively, $\beta$ determines the size of the crowding out effect, which in equilibrium is compared to the standard incentive effect of the piece rate. Consider the case of the ANTICIPATED treatment. Whenever $\beta \leq 1$ the incentive effect of the piece rate outweighs that of crowding out and thus the net effect of the wage corresponds to a marginal benefit $((1 - \beta)w \geq 0)$. To the contrary, whenever $\beta > 1$ the incentive effect of the piece rate is outweighed by that of the crowding out and thus the net effect of the wage is a marginal cost $((1 - \beta)w < 0)$. Thus, the wage reduces incentives and subjects exert less effort than the CONTROL. The case of the SURPRISING treatment builds on the same intuition. The only difference is in the sufficient restriction over $w$, ensuring that the equilibrium is unique. This idea that the net first-period effect of the piece rate can be positive or negative is consistent with Bénabou and Tirole (2003).

(6) Period-Two Behavior. (6.1) Because $T$—the only extrinsic reward—does not depend on the piece rate, agents in the CONTROL group solve the same problem as that in the first period.

(6.2) Agents in the ANTICIPATED treatment did not expect a piece rate and do not receive one ($\tilde{w}_2 = w_2 = 0$). They thus solve $e_2^a = \arg\max_e (V_2 + w_2)e - c(e) + \alpha V_0 = (V_0 - \beta w)e - c(e) + \alpha V_0$ with f.o.c.,

$$V_0 - \beta w = c'(e_2^a) \quad (6)$$

Equation (6) shows the negative effect of $w$ over the second-period effort, arising from a pure crowding-out mechanism: whenever $\beta > 0$, the marginal utility is lower than that of the CONTROL.

(6.3) Because of the surprising first-period payment, subjects in the SURPRISING treatment expected a piece rate in period two but they do not receive it ($\tilde{w}_2 = 0 < w_2 = w$). Using $V_2 = V_1 = V_0 - \beta w$ and the fact that the agent expected to exert effort $\tilde{e}_2$ in period two, subjects solve $e_2^s = \arg\max_e (V_0 - \beta w)e - c(e) + \alpha V_0 + \eta \mu(-w)\mu(e - \tilde{e}_2) + \eta \mu(c(\tilde{e}_2) - c(e))$ with f.o.c,

$$(V_0 - \beta w) - \lambda w \eta \mu'_k = [1 + \eta \mu'_c] c'(e_2^s) \quad (7)$$

where $\mu'_k \equiv \mu'(e_2^s - \tilde{e}_2)$ and $\mu'_c \equiv \mu'(c(\tilde{e}_2) - c(e_2^s))$. The most important thing to notice from equation (7) is that—in contrast to period one—reference-dependent reciprocity is now a marginal cost, which is proportional to loss aversion, $-\lambda w \eta \mu'_k < 0$. Intuitively, because the agent’s expectation of receiving a payment was not fulfilled, he optimally decreased effort below what he had planned to retaliate or because he is disappointed. Proposition 2 formalizes this hypothesis.
Proposition 2 (The Retaliation or Disappointment Hypothesis)

(i) Suppose $\beta = 0$. Then $e_2^s < \bar{e}$ and $e_2^a = \bar{e}$.

(ii) Suppose $\beta > 0$. Then $e_2^s < \bar{e}$ and $e_2^a < \bar{e}$.

Proposition 2 part (i) shows that in the SURPRISING treatment the negative surprise of not receiving $w$ in the second period can cause an effort shortfall relative to the CONTROL even if there is no crowding out. The intuition for why unmet payment expectations generate this behaviorally equivalent response to the crowding out is that the unexpected payment in period one generated the expectation of a payment in period two. Once agents realize this expectation will not be met, they retaliate or become disappointed, decreasing effort relative to the CONTROL. In contrast, those in the ANTICIPATED treatment are not surprised by the payment withdrawal and thus they do not have an incentive to reduce effort. Therefore, they exert the same amount of effort as those in the CONTROL. Part (ii) shows that whenever there is crowding out, effort in the ANTICIPATED treatment also falls below that of the CONTROL. Appendix F shows that Proposition 2 is robust to the case where agents in the SURPRISING treatment think that the second-period payment is possible but not certain.

4 Results and Discussion

This section presents the results of the field experiment. We find, just as in the canonical two-period test for crowding out, that subjects surprised by the introduction of the reward undersupply time on the task after this reward is withdrawn. Namely, when subjects receive the unexpected piece rate in the first session they spend less time tasting cookies than the CONTROL when this piece rate is withdrawn in the second session. Further, they provide faster but lower quality evaluations. However, this appears to be caused by unmet expectations instead of crowding out: when subjects expect the piece rate to be granted and removed, they behave similarly to the CONTROL during the non-reward session. Payments could thus affect subsequent outcomes, not through crowding out, but by generating expectations of future payments which, when unmet, trigger retaliation or disappointment.

We discuss the results in light of the literature on crowding out and our model. Importantly, though we describe the outcomes for the reward session, these do not constitute a test for crowding out as it is unclear whether the incentive effect of the piece rate dominates crowding out or vice versa, as outlined by our model and consistent with both the psychology and the economics literature (e.g., Deci and Ryan (1985), Bénabou and Tirole (2003)). Therefore, we largely base our conclusions on the outcomes for the
subsequent non-reward session, as discussed above in the context of the need for a two-period test.

### 4.1 Sample and Descriptive Statistics

**Sample.** We recruited 91 participants across four legs on two campuses, A and B. Once recruited, they were randomly assigned across three conditions: 37 in the **CONTROL**, 27 in the **SURPRISING** and 27 in the **ANTICIPATED**. Most subjects (76) came from campus A, where the facilities could accommodate more people (see Appendix Table A.2 with a breakdown of the participants per campus and treatment).

**Attrition.** Of the 91 subjects who completed the first session, 10 attritted in session two. This attrition is not likely to bias our results for two reasons. First, it was not substantial, at only 11% (three subjects in the **CONTROL**, one in the **SURPRISING** and six in the **ANTICIPATED** treatment). Second, differences in attrition rates between the treatments and the **CONTROL** were not statistically significant, as documented in Table A.1.

**Summary statistics across all subjects.** Table 1 shows that tasters engaged significantly with the task, though there was substantial variability. Across the 172 person-sessions, tasters sampled an average of 35 cookies, with a minimum of four and a maximum of 70. The average time tasting cookies was 81 minutes, with a minimum of 12 and a maximum of 182. Subjects spent an average of 2.6 minutes per evaluation. Most cookies (70%) were not fully eaten indicating that subjects regulated their consumption.

**Disaggregated statistics per treatment.** Disaggregated summary statistics by treatment and session in Table 2 suggest that the piece rate raised performance in the first session. Panel A, column (1) shows that subjects in the **SURPRISING** and **ANTICIPATED** treatments spent more time tasting cookies than those in the **CONTROL**: 102 and 112 minutes versus 79, respectively. They also evaluated more cookies: 48 and 51 versus 30, respectively. And they also spent less time completing each evaluation: 2.2 and 2.3 minutes versus 3.2, respectively.

During the subsequent non-reward period, however, only subjects for whom the withdrawal of the piece rate was unexpected spent less time on the task than the **CONTROL**. Panel B, column (1), documents that those in the **SURPRISING** treatment only spent 56 minutes tasting cookies during the non-reward session versus 62 in the **CONTROL**. In contrast, those in the **ANTICIPATED** treatment, for whom the withdrawal of the piece rate was expected, spent 15 minutes more than the **CONTROL**. Thus the raw data, unadjusted for unobservable campus and leg differences in performance, hints that subjects only spend less time on the task when the reward is removed unexpectedly.
Panel B, column (1), also documents that though the surprise withdrawal of the payment led subjects in the SURPRISING treatment to spend less time tasting cookies than those in the CONTROL and ANTICIPATED conditions, they evaluated a similar number of cookies: 29 versus 26 and 32, respectively.

As a result, only those surprised by the withdrawal of the payment complete their evaluations faster than the CONTROL during the non-reward period, as they spend less time tasting but evaluate a similar number of cookies. Panel B, column (1), highlights that those in the SURPRISING treatment spent 2.0 minutes per evaluation versus 2.8 and 2.7 in the CONTROL and ANTICIPATED conditions, respectively. In Section 4.3, we find that the faster rate of completing evaluations by subjects in the SURPRISING treatment appears to result from their exerting less effort on rating each cookie.

Though these means on the three main measures of interest—time spent on the task, number of evaluations, and minutes per evaluation—result from an aggregation of subjects’ behavior across campuses and legs, the pattern they convey holds, in general, within campuses (see table A.2) and legs (see table A.3). Therefore, the average differences between the treatments and the CONTROL will not change substantially when we estimate them within campus, leg, and session, and then pool them. We now turn to this analysis, as we control for unobserved factors in each campus, leg and session that could bias our results.

4.2 Empirical Method and Results on Main Measures

**Empirical method.** Within each leg, and for each of the campuses, we randomized volunteers into each condition. For example, for leg two, we randomized those in campus A into the three conditions and did the same for campus B. Therefore, we estimate the outcomes for subject $i$, in the $t1$ (CONTROL), $t2$ (SURPRISING), $t3$ (ANTICIPATED) conditions in campus $c$, leg $l$, and session $s$ as follows:

$$outcome_{i,t,c,l,s} = \alpha_{1,1} + \alpha_{1,2}t_1s_2 + \sum_{\tau=2}^{3} \sum_{j=1}^{2} \beta_{\tau,j}t_\tau s_j + \lambda_c \times \lambda_l \times \lambda_s + \epsilon_{i,t,c,l,s} \tag{1}$$

The interaction of campus, leg and session, $(\lambda_c \times \lambda_l \times \lambda_s)$, captures unobservable time-invariant, campus, leg and session determinants of outcomes. Campus fixed effects control, for example, for unobserved heterogeneity in the level of health consciousness per campus, which can influence the sensitivity to payments within campus and thus the differences in outcomes between the treatments and the CONTROL (e.g., students in one campus not increasing consumption in response to the piece rate). Leg fixed effects control, for example, for the unobserved temperature level during a leg, which can also affect the response...
to incentives within leg and thus differences in outcomes between the treatments and the control (e.g., cookies may be less appealing in a leg occurring in a hot month). The interaction of campus, leg, and session effects conservatively addresses whether these unobservables could affect the differences between the treatments and the CONTROL differentially within each leg, campus and session (e.g., whether a hot summer leg affects outcomes differentially by campus in the first session). Importantly, this interaction also allows us to estimate differences between the treatments and the CONTROL, within campus, leg and session, and then to pool them.

The causal parameters of interest are the $\beta_{t,s}$, which pool the differences in outcomes between the treatments and the CONTROL for each session, within a campus and leg. For example, $\beta_{2,1}$ identifies the difference between treatment two and the CONTROL in session one, by pooling this difference across each campus and leg. The parameter $\alpha_{1,1}$ is the outcome for the baseline category—the CONTROL in session one—which cannot be separately identified from the interaction of the fixed effects, as usual.

Besides analyzing time spent on the task, number of evaluations, and minutes per evaluation with this specification, we also examine the proportion of partially tasted cookies. Because there is serial correlation in outcomes for each subject across sessions, we conservatively cluster the standard errors at the subject level (Bertrand, Duflo, and Mullainathan (2004)).

**Result 1** *Time spent on the task. During the reward session, the piece rate increased the time subjects spent tasting cookies versus the CONTROL. In the subsequent non-reward session, however, subjects only spent less time on the task than the CONTROL when the withdrawal of the piece rate was unexpected.*

We start by discussing time spent on the task as this is, by far, the most common metric by which crowding out has been measured in psychology. This literature has also relied on other measures, albeit more rarely, such as output (e.g., number of puzzles solved) and productivity (e.g., number of headlines per minute as in Deci (1971)), so we discuss them next. Subjective self-reports of enjoyment or of interest in the activity have also been used and we discuss them at the end.

We replicate Deci’s (1971) and similar findings in the crowding-out literature: the surprising introduction of the reward and its withdrawal in the subsequent period caused subjects to spend more time on the task than the CONTROL during the reward period, but less time during the non-reward period. However, this seems to be due to unmet payment expectations rather than crowding out.
Namely, panel A in Table 3 shows that the introduction of the piece rate in the first session for the Surprising and Anticipated treatments led participants to spend more time tasting. Column (1), displaying the raw difference in means, documents an excess of 23 and 32 minutes, respectively, versus the Control (column (1)). Adjusting these magnitudes successively within campus (column (3)), campus and leg (column (5)), and campus, leg and session (column (7)) does not change them meaningfully, yielding an excess of 16 and 32 minutes, respectively (where the latter is statistically significant at the 1% level). The stability in the estimates is unsurprising given that the pattern of results is generally consistent across legs, campuses and sessions. Thus, we focus the remainder of the discussion on the results of specification (1)—in columns (7) and (8)—as it correctly estimates the average differences between the treatments and the Control, within campus and leg, and then pools them.

The subsequent unexpected withdrawal of the piece rate, however, led to a shortfall in the time spent tasting cookies, as column (8) documents. Namely, those in the Surprising treatment—which mirrors the canonical treatment in the crowding-out research in psychology—spent 12 minutes less than the Control tasting cookies during the non-reward period. This result is marginally significant at the 10% level, in a one-tailed test, just as in Deci (1971). One-tailed tests are not uncommon in this literature, since crowding out makes the one-directional prediction of a decrease in performance versus the Control during the non-reward period. This time shortfall could be evidence for crowding out as pointed out by its proponents, but also of disappointed payment expectations, as we conjectured. The behavior of subjects when the removal of the piece rate is expected helps us answer this question.

When the withdrawal of the piece rate was expected there was no shortfall in the time spent on the task, as column (8) also shows. Those in the Anticipated treatment spent 12 minutes more on the task during the non-reward period than the Control, not less as predicted by crowding out, though this estimate is not statistically significant. Had the payment in the first period felt controlling or led subjects to misattribute their performance to the reward—as proposed by cognitive evaluation theory and the overjustification hypothesis, respectively—we should have observed a shortfall in the time tasting during the non-reward period for these subjects as well.

The fact that intrinsically motivated subjects only undersupply time during the non-reward period when the removal of the reward is unexpected suggests that the standard evidence for the crowding-out could be ascribed to unmet payment expectations. We further buttress this premise with the subsequent
findings, which use two-sided tests.

**Result 2 Number of cookies tasted and evaluated.** During the reward session, the piece rate increased the number of cookies tasted and evaluated versus the Control. During the non-reward session, its withdrawal caused no shortfalls in output.

Table 3, panel B, documents that the piece rate increased performance in the first session as it increased the number of cookies subjects tasted and evaluated. Column (7) shows that subjects in the Surprising and Anticipated treatments completed 16 and 22 more evaluations, respectively, than those in the Control, where both these estimates are statistically significant the 1% level. This result parallels Result 1 where the reward also increased the time spent on the task in the first session.

The withdrawal of the piece rate in the subsequent period, however, did not lead to lower output than the Control. During the non-reward period subjects in the Surprising and Anticipated treatments completed one and five more evaluations, respectively, than those in the Control, though these estimates are not statistically significant.

Therefore, though subjects for whom the reward was removed unexpectedly spent 12 fewer minutes than the Control at the tasting site during the non-reward period, they evaluated a similar number of cookies (one more than the Control). This suggests that these subjects increased the speed at which they filled out evaluations, upon the removal of the piece rate, as we document next.

**Result 3 Minutes per evaluation.** During the reward session, the piece rate decreased the number of minutes spent filling out each evaluation versus the Control. In the subsequent non-reward session, however, subjects only spent less time filling out each evaluation than the Control when the withdrawal of the piece rate was unexpected.

Piece rates also increased the speed at which subjects filled out each evaluation. Table 3, Panel, C, column (7) documents that the average number of minutes spent per evaluation by subjects in the Surprising and Anticipated treatments was 0.95 and 0.86 lower than in the Control, respectively (differences which are statistically significant at the 1% level). This increase in the speed during the reward session did not sacrifice the quality of the evaluations, as we show in Section 4.3.

During the subsequent non-reward session, however, only subjects surprised by the withdrawal of the payment produced faster than the Control, as shown in column (8). When the expected payment was
unrealized, tasters spent, on average, 0.80 minutes less filling out each evaluation, which is significant at the 5% level. In contrast, when they expected the withdrawal of the payment, they spent approximately as much time filling out each evaluation as the Control during the non-reward period (-0.12 minutes per evaluation, which is not statistically significant). This finding for the Surprising treatment is a logical consequence of Results 1 and 2: when subjects’ payment expectations were unmet they spent less time tasting cookies, but the number of cookies tasted was essentially the same as in the Control.

This decrease in the number of minutes filling out each evaluation during the non-reward period for the Surprising treatment cannot be ascribed to learning or to subjects’ partially eating a larger fraction of cookies, but rather to a decrease in effort, as we document in Section 4.3. First, had subjects in the Surprising treatment learned to rate cookies faster during the reward session and had this learning carried over to the subsequent non-reward session, we should have observed the same pattern for the Anticipated treatment. This is because those in the Anticipated treatment tasted more cookies, spent more time at the tasting site, and spent a similar amount of time completing each evaluation as those in the Surprising treatment during the reward session. However, in the subsequent non-reward session those in the Anticipated treatment spent the same amount of time as the Control completing each evaluation whereas those in the Surprising treatment spent less time. Therefore, had learning led to higher speed in the evaluations for the Surprising treatment during the non-reward session, it should have led to a higher speed of evaluations for the Anticipated treatment as well, which we do not observe.

Second, the faster rate of filling out evaluations for the Surprising treatment during the non-reward period also cannot be reconciled with subjects partially tasting each cookie, instead of fully eating it, so as to economize on time digesting. Table 4, column (7) shows that the fraction of partially eaten cookies per subject (cookies not fully eaten as a proportion of the total cookies tasted) in the Surprising treatment was 0.18 higher than the Control during the reward period, and statistically significant at the 5% level. However, during the non-reward period there is no statistically significant difference between this treatment and the Control. A secondary finding, is that those in the Anticipated treatment show no statistical difference versus the Control in either period. This suggests that during the reward period the fraction of partially eaten cookies in the Surprising treatment increased because partial tasting allowed subjects to raise the number of evaluations above what they had planned in the
absence of the piece rate. But there is no statistically significant difference across the three treatments during the subsequent non-reward sessions.

Figure 2 summarizes Results 1 through 3, depicting the previously described differences between the Surprising and Anticipated treatments and the Control, adjusting for campus, leg and session time-invariant unobservables. We now document that the faster rate of completing evaluations by those surprised by the withdrawal of the payment appears to be due to subjects' investing less effort in each evaluation.

4.3 Results on Effort Spent Per Evaluation

Effort measure. To investigate the amount of effort tasters devoted to filling out evaluations, we use a high dispersion in ratings within cookie as a proxy for low effort. This measure stems from the extensive literature on survey implementation in psychology and in education, which has found that choosing at random is a common strategy to economize on effort (e.g., Krosnick, (1991, 1999)), which also correlates with faster response times (e.g., Meijer (2003), Wise and Kong (2005)). Other standard proxies for low effort, such as the proportion of correct answers, are not naturally applicable because of the subjective nature of our task. Other easily detectable measures, such as missing ratings in a category (e.g., a missing rating for Appearance), occurred rarely in our data, possibly due to this type of error being easy to spot by the research assistant who counted the evaluations at the end of each session.

Sample. To investigate how the different treatments affect the ratings dispersion for a given cookie, we restrict the sample of cookies in two ways. First, we excluded those for which the same identification number (e.g., cookie number 1) could have more than one flavor. This is the case, for example, with cookie assortments, where the same identification number could refer to a chocolate or to a vanilla cookie. It is important to exclude these cookies to ensure that differences in dispersion are not due to differences in the actual cookie. Second, we only included cookies that were sampled in the three treatments in a given campus, leg and session. This way we compare that cookie’s dispersion across any of the three conditions in that campus, leg and session. Although all subjects within a campus, leg and session received the same set of 70 cookies to taste, they did not end up tasting exactly the same cookies. Subjects tasted different numbers of cookies and, within those, some were more likely to be tasted as they were randomly assigned to tasting trays closer to the subject.

Because of these two requirements, the analysis drops 5 and 12 subjects in sessions one and two,
respectively. For example, a subject who tasted 17 cookies in session two, of which 10 were from cookie assortments (thus excluded from the analysis) while the remaining 7 were not tasted by subjects in the two other conditions in his campus, leg and session, is dropped from the analysis. Appendix Table A.4 documents, however, that results (1) through (3) using the full and the restricted samples are similar. This allays the concern that the dispersion analysis on this smaller sample, which comprises 86 subjects in session one (and a total of 2,020 cookie-person observations) and 69 subjects in session two (and a total of 1,044 cookie-person observations) is not generalizable to the full sample.

**Empirical method.** We estimated the dispersion in the ratings of cookie \( k \), for subject \( i \), in the \( t1 \) (Control), \( t2 \) (Surprising), \( t3 \) (Anticipated) conditions in campus \( c \), leg \( l \), and session \( s \) as follows:

\[
\text{dispersion}_{k,i,t,c,l,s} = \alpha_{1,1} + \alpha_{1,2}t_1s_2 + \sum_{\tau=2}^{3} \sum_{j=1}^{2} \beta_{\tau,j}t_\tau s_j + \lambda_c \times \lambda_l \times \lambda_s \times \lambda_k + \epsilon_{k,i,t,c,l,s} \quad (2)
\]

We chose the standard deviation in the ratings of a cookie as the measure of dispersion.\(^{16}\) 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 them ran from 1 (Excellent) to 5 (Poor). Therefore, a cookie rated as 1, 3, 4, 2, 5, 5, 3, respectively, has a standard deviation of 1.49 whereas a cookie rated as a 3, 3, 3, 3, 3, 3 has a standard deviation of zero.

The interaction of campus, leg, session and cookie fixed effects \((\lambda_c \times \lambda_l \times \lambda_s \times \lambda_k)\) captures unobserved time-invariant, campus, leg, session and cookie determinants of standard deviation. Campus fixed effects control, for example, for whether subjects in one campus are less conscientious than in the other, thus showing a higher propensity to respond at random, which affects the differences in dispersion between the treatments and the Control within campuses. Leg fixed effects address, for example, whether during a hot summer leg, some cookies may have a higher tendency to melt, leading to more dispersion in their ratings for this leg than in others (e.g., their Appearance rating would be worse than in other legs while its Flavor rating would remain unchanged). Cookie fixed effects controls for unobserved time-invariant differences in cookie characteristics that could lead to differences in dispersion, such as a cookie having a good appearance but a bad flavor. The interaction addresses whether these unobservables could differentially affect outcomes within campus, leg, session and cookie (e.g., whether a hot summer leg affects the dispersion of a given cookie more in one campus than in the other). Further, similarly to specification (1), the interaction of these fixed effects allows us to estimate differences between the

---

\(^{16}\) The use of other measures such as the variance and range yielded qualitatively similar results.
treatments and the CONTROL, within a campus, leg, session and cookie and to pool them.

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

Finally, we cluster the standard errors by individual to address the potential correlation in cookie ratings’ dispersion for a subject within and across sessions (Bertrand, Duflo, and Mullainathan (2004)).

**Result 4** Standard deviation of the ratings within an evaluation. During the non-reward session, the piece rate did not increase ratings’ dispersion. In the subsequent non-reward session, however, the ratings’ dispersion was higher than that of the CONTROL only when the withdrawal of the piece rate was unexpected.

Table 5 shows that for the same cookie tasted across the three conditions, within a given leg, campus, and session, the standard deviation of the ratings during the reward period is no different between the treatments and the CONTROL. This finding is consistent when comparing the raw differences in dispersion between the treatments and the CONTROL (column (1)) and these differences within the same campus, leg and session (column (3)), and within campus, leg, session and cookie (column (5)). This suggests that though the piece rate increased the number of evaluations per minute during the reward session (Result 3), it did not damage the ratings dispersion, our measure of quality. This is congruent with previous findings where piece rates increase productivity without lowering quality (e.g., Bandiera, Barankay, and Rasul (2005)).

However, column (6) documents that during the subsequent non-reward period, only subjects surprised by the withdrawal of the payment increased the dispersion in ratings for that cookie relative to the CONTROL. The standard deviation for that same cookie, in that campus, leg and session, is 0.10 higher than that in the CONTROL and significant at the 5% level. This suggests that the unexpected withdrawal of the piece rate led subjects to complete each evaluation faster, as documented in Result 3, but at the expense of increasing the dispersion in their ratings. In contrast, the dispersion in ratings
for that same cookie when subjects expected the reward to be withdrawn (ANTICIPATED treatment) is not higher than that in the CONTROL. It is slightly smaller, but not statistically significant.

Results (3) and (4) together indicate that during the non-reward period, subjects surprised by the withdrawal of the payment (SURPRISING treatment) both completed evaluations faster than the CONTROL and increased the dispersion in their ratings for a fixed cookie. In contrast, during the same period, those who expected the piece rate to be removed (ANTICIPATED treatment), spent the same time as the CONTROL filling out each evaluation and did not have more dispersed ratings. This difference between the treatments in minutes spent per evaluation and the dispersion (quality) of the ratings cannot be ascribed to learning during the previous reward period, as both these groups showed similar productivity (minutes per evaluation) and quality profiles (dispersion) during the non-reward period.

These evidence thus suggests that only those surprised by the withdrawal of the payment decreased effort, suggesting that unmet payment expectations rather than crowding out led to this outcome. Namely, as described in our model, those surprised by the withdrawal of the reward reduced effort during the non-reward period, reflected in the more random and thus less careful evaluations. In contrast, those expecting the withdrawal of the payment (ANTICIPATED treatment) did not. Under crowding out, however, we should have observed a higher dispersion (lower quality) in the evaluations in this treatment relative to the CONTROL during the non-reward period.

*Is speed inversely correlated with dispersion in our setting?* Though studies in psychology and education may have found that the less time subjects spend responding to questions the higher the dispersion in answers, we might worry that it does not occur in our setting. Therefore, to complete the analysis, we now document that, in line with these literatures, we also find this negative relationship: the time spent filling out each evaluation is inversely correlated with the dispersion in its ratings.

*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 4.3, as follows:

$$
\text{sd}_{k,i,c,l,s} = \theta_1 + \theta_2 m + \sum_{r=1}^{70} \beta_r d_r + \psi_i \times \psi_s + \lambda_c \times \lambda_l \times \lambda_s \times \lambda_k + \epsilon_{i,k,s,t,c}
$$

(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 started eating the next cookie.\(^{17}\) 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 $\theta_2$, 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 $\sum_{r=1}^{70} \beta_r d_r$ corresponds to dummies that control for the order in which the cookie was tasted (e.g., subjects both spending less time and increasing 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, which takes long 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 6 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 0.0024. Column (2) adds controls for cookie order, which both increase 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 0.0018. Column (3) adds controls for the time-invariant unobserved ability of the person tasting the cookie. This increases the fit of the model to 28% and yields an estimate of -0.0022. Column (4) adds the experience of the taster, increasing the magnitude of this estimate to -0.0027, while increasing the fit of the model to 29%. 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 in a given campus, leg and session, the ability and experience of the rater of that cookie, and the order in which 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 -0.0042. Despite in several evaluation sheets the end time was missing or matched the start time.
the large number of controls, this estimate is statistically significant at the 10% level.

Having documented that the surprise withdrawal of the piece rate appears to lead to less time and effort in each evaluation versus the CONTROL, we now investigate whether it also led subjects to perceive the cookies more unfavorably.

4.4 Results on the Overall Perception of Cookies

One final question of interest is whether the introduction and withdrawal of the piece rate affected the overall perception of the pleasantness of the cookies. This measure may correlate with the enjoyment ratings for the task, a subjective measure that is occasionally used in psychology to assess crowding out. Specifically, a shortfall in the self-reported enjoyment for the task for the treatment during the non-reward period relative to the CONTROL has also been interpreted as evidence of crowding out. However, this evidence has been sometimes at odds with the evidence from others measures. For example, as described in Section 2.1, though subjects solving puzzles in Deci (1971) spent less time on this activity than the CONTROL upon the removal of the piece rate, their enjoyment ratings for the task remained the same (Deci (1971), page 110). Although we did not collect enjoyment ratings, we can investigate whether the overall rating for cookies declined during the non-reward period for the treatments.

Result 5 Overall rating for a cookie. The piece rate and its withdrawal did not affect the overall rating of the cookies in either treatment or period.

Table 7 documents that there is no difference in the overall rating of a cookie across the different conditions in either session. We use the same sample as in Section 4.3, as well as the same specification (2), with the exception that we change the dependent variable to the overall rating of the cookie. Columns (1) and (2) show that the unadjusted mean rating of the cookies for the CONTROL in sessions one and two is 3.06 and 2.85, respectively, and that the ratings for the SURPRISING and ANTICIPATED treatments do not differ statistically from these estimates. Columns (3) and (4) compare the ratings across the three conditions, for the same cookie tasted within a campus, leg and session, and show no statistical difference versus the CONTROL in either period. These results are thus consistent with those in Deci (1971), who found no differences in enjoyment ratings across groups or sessions.
5 Conclusion

Despite suggestive evidence in psychology that payments may harm performance by crowding out intrinsic motivation, whether this holds in the workplace has been an open question. For example, Kreps ((1997), page 360) stated that “strong empirical support of the stylized fact is hard to find” and Pendergast ((1999), page 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.”

We do not find evidence of crowding out but rather for an alternative hypothesis: retaliation or disappointment. We revisit the canonical two-period test in psychology, where a payment is introduced unexpectedly in session one and withdrawn in session two and show that the adverse effects on effort (time spent on the task and output quality) of introducing a reward appear to be due to its unexpected withdrawal, rather than to crowding out of intrinsic motivation. This suggests that extrinsic incentives do not appear to substitute intrinsic ones in the workplace, favoring the standard intuition that, at least for the workplace, monetary payments do not have perverse effects on intrinsic motivation. Payments, however, may be deleterious if they are expected but not met. This latter result dovetails with research in other settings suggesting that unmet payment expectations lessen effort (e.g., Bewley (1998), Mas (2006), and Kube, Maréchal, and Puppe (2013)).

References


### Figures

Figure 1: Room Layout for Cookie Tasting
Figure 2: Average Number of Cookies Tasted/Evaluated, Minutes Spent Tasting Cookies and Minutes Spent per Evaluation—per Treatment and Session

Notes: This figure presents a visual summary of the results in Table 3, columns (7) and (8). The upper left panel depicts the difference in the average number of minutes spent tasting cookies in session one between the SURPRISING and ANTICIPATED treatments and the CONTROL, adjusting for campus, leg and session time-invariant unobservables. The upper center and right panels depict the same information, but for the average number of cookies tasted and minutes spent filling out each evaluation, respectively. The bottom panel depicts the same information as the upper panel, but for session two. Standard error bars were omitted for clarity.
<table>
<thead>
<tr>
<th></th>
<th>Number of Subjects X Sessions</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cookies tasted and evaluated</td>
<td>172</td>
<td>35</td>
<td>19</td>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>172</td>
<td>81</td>
<td>39</td>
<td>12</td>
<td>182</td>
</tr>
<tr>
<td>Evaluations per minute</td>
<td>172</td>
<td>2.6</td>
<td>1.1</td>
<td>1.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Proportion of partially eaten cookies</td>
<td>172</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 2: Disaggregated Summary Statistics by Treatment and Session

<table>
<thead>
<tr>
<th></th>
<th>Mean (1)</th>
<th>SD (2)</th>
<th>Min (3)</th>
<th>Max (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Session One</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T1- Control (N=37)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>79</td>
<td>37</td>
<td>19</td>
<td>178</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>30</td>
<td>19</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>3.2</td>
<td>1.2</td>
<td>1.4</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>T2-Surprising (N=27)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>102</td>
<td>37</td>
<td>43</td>
<td>177</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>48</td>
<td>19</td>
<td>17</td>
<td>70</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>2.2</td>
<td>0.5</td>
<td>1.5</td>
<td>3.2</td>
</tr>
<tr>
<td><strong>T3-Anticipated (N=27)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>112</td>
<td>36</td>
<td>47</td>
<td>179</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>51</td>
<td>19</td>
<td>23</td>
<td>70</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>2.3</td>
<td>0.6</td>
<td>1.4</td>
<td>4.0</td>
</tr>
<tr>
<td><strong>Panel B: Session Two</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T1- Control (N=34)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>62</td>
<td>33</td>
<td>12</td>
<td>173</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>26</td>
<td>15</td>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>2.8</td>
<td>1.3</td>
<td>1.2</td>
<td>7.3</td>
</tr>
<tr>
<td><strong>T2-Surprising (N=26)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>56</td>
<td>21</td>
<td>22</td>
<td>100</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>29</td>
<td>12</td>
<td>10</td>
<td>53</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>2.0</td>
<td>0.5</td>
<td>1.2</td>
<td>3.6</td>
</tr>
<tr>
<td><strong>T3-Anticipated (N=21)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time tasting cookies (minutes)</td>
<td>77</td>
<td>38</td>
<td>26</td>
<td>182</td>
</tr>
<tr>
<td>Cookies tasted and evaluated</td>
<td>32</td>
<td>15</td>
<td>9</td>
<td>62</td>
</tr>
<tr>
<td>Minutes per evaluation</td>
<td>2.7</td>
<td>1.3</td>
<td>1.3</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Notes: Column (1) documents the average time tasting cookies at the tasting site (in minutes), the number of cookies tasted and evaluated and the minutes per evaluation, in session ones and two. Columns (2) through (4) show the standard deviation, minimum and maximum, respectively, of these measures.
### Table 3: Average Cookies Tasted, Minutes Spent Tasting cookies and Minute per Evaluation

**PANEL A:**

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Minutes Spent Tasting Cookies per Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted Session</td>
</tr>
<tr>
<td></td>
<td>One</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>79  (6)**</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td>23 (9)**</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>32 (9)***</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>3.15 (0.20)***</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td>-0.93 (0.22)***</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>-0.85 (0.24)***</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>-0.85 (0.24)***</td>
</tr>
</tbody>
</table>

**PANEL B:**

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Number of Cookies Tasted and Evaluated per Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>30</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td>18 (5)***</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>21</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>3.15 (0.20)***</td>
</tr>
</tbody>
</table>

**PANEL C:**

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Minutes per Evaluation per Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>3.15</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td>-0.93 (0.22)***</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>-0.85 (0.24)***</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>-0.85 (0.24)***</td>
</tr>
</tbody>
</table>

Number of subjects (clusters) 91 81 91 81 91 81 91 81

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. The significance levels for Panel A (minutes spent tasting cookies per subject) are for a one-sided test: $H_0: \beta_i = 0$ versus $H_a: \beta_i < 0$ for ease of comparison with Deci (1971). The significance levels on all other tests are for two-tailed tests. The results for columns (7) and (8) are from specification 1, with the exception that the baseline category for column (8) is the outcome for the CONTROL in session two.
Table 4: Fraction of Partially Sampled Cookies across Treatments

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>Within campus</th>
<th>Within campusXleg</th>
<th>Within campusXlegXsession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Session</td>
<td>Session</td>
<td>Session</td>
<td>Session</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>0.71</td>
<td>0.64</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.05)***</td>
<td>(0.06)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>0.15</td>
<td>0.10</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(0.06)**</td>
<td>(0.07)</td>
<td>(0.06)**</td>
<td>(0.07)**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>0.09</td>
<td>0.08</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of subjects (clusters)</td>
<td>91</td>
<td>81</td>
<td>91</td>
<td>81</td>
</tr>
</tbody>
</table>

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level; **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed. The results for columns (7) and (8) are from specification 1, with the exception that the baseline category for column (8) is the outcome for the CONTROL in session two.
Table 5: Standard Deviation of Cookie Ratings across Treatments

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Standard Deviation of Ratings in an Cookie Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted</td>
</tr>
<tr>
<td></td>
<td>Session</td>
</tr>
<tr>
<td></td>
<td>One</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>0.74</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td>(0.04)***</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
</tr>
</tbody>
</table>

| Number of cookieXsubject observations | 2,020 | 1,044 | 2,020 | 1,044 | 2,020 | 1,044 |
| Number of subjects (clusters) | 86 | 69 | 86 | 69 | 86 | 69 |

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed. The results for columns (7) and (8) are from specification 2, with the exception that the baseline category for column (8) is the outcome for the CONTROL in session two.
Table 6: Correlation between the Standard Deviation and Time spent per Evaluation

<table>
<thead>
<tr>
<th>Minutes spent in evaluation</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0024</td>
<td>-0.0018</td>
<td>-0.0022</td>
<td>-0.0027</td>
<td>-0.0042</td>
</tr>
<tr>
<td></td>
<td>(0.0034)</td>
<td>(0.0024)</td>
<td>(0.0024)</td>
<td>(0.0027)</td>
<td>(0.0025)*</td>
</tr>
</tbody>
</table>

**Controls:**
- Dummies for Order Cookie was tasted
- Subject fixed effects
- SubjectXsession fixed effects
- CampusXlegXsessionXcookie fixed effects

<table>
<thead>
<tr>
<th>Controls:</th>
<th>-</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dummies for Order Cookie was tasted</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Subject fixed effects</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SubjectXsession fixed effects</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CampusXlegXsessionXcookie fixed effects</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-squared</th>
<th>0.00</th>
<th>0.24</th>
<th>0.28</th>
<th>0.29</th>
<th>0.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cookieXsubject observations</td>
<td>2,961</td>
<td>2,961</td>
<td>2,961</td>
<td>2,961</td>
<td>2,961</td>
</tr>
<tr>
<td>Number of subjectXweek (observations)</td>
<td>155</td>
<td>155</td>
<td>155</td>
<td>155</td>
<td>155</td>
</tr>
</tbody>
</table>

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level;**Significant at the 5% level;***Significant at the 1% level. All tests are two-tailed. This sample has 103 fewer observations (2,961 versus 3,064 in Table 5) due to some cookies having unreadable times at which their started being tasted.
Table 7: Overall Rating of Cookies across Treatments

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th></th>
<th></th>
<th>Within</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Session</td>
<td></td>
<td>campusXlegXsessionXcookie</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>One</td>
<td>Two</td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>3.06</td>
<td>2.85</td>
<td>(0.11)**</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Diff. vs. Control group</td>
<td></td>
<td></td>
<td></td>
<td>(2) T2 - Surprising</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>(0.15)***</td>
<td>(0.13)***</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
</tr>
<tr>
<td></td>
<td>-0.08</td>
<td>-0.01</td>
<td>-0.11</td>
<td>-0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.18)</td>
<td>(0.15)</td>
<td>(0.19)</td>
<td></td>
</tr>
</tbody>
</table>

Number of cookieXsubject observations 2,018 1,044 2,018 1,044
Number of subjects (clusters) 86 69 86 69

Notes. Standard errors are in parentheses and are clustered by subject. *Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. All tests are two-tailed. Two subjects did not give the overall rating for a cookie, resulting in 2,018 cookie-subject observations instead of 2,020 in Table 5. The results for columns (7) and (8) are from specification 2, with the exception that the baseline category for column (8) is the outcome for the CONTROL in session two.
A Additional Figures and Tables

Table A.1: No Differential Attrition Between the Treatments and the CONTROL

<table>
<thead>
<tr>
<th>Diff. vs. Control group</th>
<th>Unadjusted Within CampusXLeg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>T2 - Surprising</td>
<td>-0.83</td>
</tr>
<tr>
<td></td>
<td>(1.19)</td>
</tr>
<tr>
<td>T3 - Anticipated</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.43</td>
</tr>
<tr>
<td></td>
<td>(0.61)***</td>
</tr>
<tr>
<td>CampusXLeg Fixed Effects</td>
<td>-</td>
</tr>
<tr>
<td>Pseudo R-squared</td>
<td>0.08</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>91</td>
</tr>
</tbody>
</table>

Notes. Logistic regression testing whether there are different attrition rates in the treatments relative to the CONTROL. Column (1) shows raw attrition, unadjusted for time-invariant campus, leg and session heterogeneity. The coefficient of -2.43 for the constant implies a baseline likelihood of attrition of 8.8% (exp(-2.43)) for the CONTROL which is statistically different from zero. Column (2) controls for unobserved time-invariant campus and leg determinants of attrition (e.g., a campus may have had a leg where the second session occurred close to an examination week leading to differential attrition in the second session versus that in another campus in another leg). Coefficients for SURPRISING and ANTICIPATED treatments in columns (1) and (2) indicate the log-odds of attrition versus the CONTROL. None is statistically significant, indicating that none of the treatments showed a higher propensity for attrition in session two than the Control. Robust standard errors. ***Significant at the 1% level. Using a linear regression model for this analysis leads to similar results.
Table A.2: Summary Statistics per Campus, Session and Treatment on the Number of Cookies Evaluated, Time Tasting Cookies and Minutes per Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Panel A: Campus A</th>
<th>Panel B: Campus B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (1)</td>
<td>Surprising (3)</td>
</tr>
<tr>
<td></td>
<td>Session One (2)</td>
<td>Session One (4)</td>
</tr>
<tr>
<td></td>
<td>Session Two</td>
<td>Session Two</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>29 26</td>
<td>24 23</td>
</tr>
<tr>
<td>Mean time evaluating</td>
<td>83 61</td>
<td>99 56</td>
</tr>
<tr>
<td>cookies (minutes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean cookie evaluations</td>
<td>31 26</td>
<td>47 29</td>
</tr>
<tr>
<td>Mean minutes per</td>
<td>3.2 2.8</td>
<td>2.2 2.0</td>
</tr>
<tr>
<td>evaluation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                          |                  |                  |                  |
| Total number of subjects | 37 34 27 26 27 21 |                  |                  |
Table A.3: Summary Statistics per Leg on the Number of Cookies Evaluated, Time Tasting Cookies and Minutes per Evaluation

<table>
<thead>
<tr>
<th>Treatments</th>
<th>Control</th>
<th>Surprising</th>
<th>Anticipated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Session One</td>
<td>Session Two</td>
<td>Session One</td>
</tr>
<tr>
<td>Leg 1</td>
<td>Number of subjects</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Mean time evaluating cookies (minutes)</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Mean cookie evaluations</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Mean minutes per evaluation</td>
<td>3.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Leg 2</td>
<td>Number of subjects</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Mean time evaluating cookies (minutes)</td>
<td>82</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Mean cookie evaluations</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Mean minutes per evaluation</td>
<td>3.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Leg 3</td>
<td>Number of subjects</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Mean time evaluating cookies (minutes)</td>
<td>75</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Mean cookie evaluations</td>
<td>30</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Mean minutes per evaluation</td>
<td>3.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Leg 4</td>
<td>Number of subjects</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Mean time evaluating cookies (minutes)</td>
<td>92</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Mean cookie evaluations</td>
<td>33</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Mean minutes per evaluation</td>
<td>3.2</td>
<td>2.0</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>37</td>
<td>34</td>
<td>27</td>
</tr>
</tbody>
</table>
### Table A.4: Side-by-side Comparison of Main Outcomes in Full and Restricted Sample

<table>
<thead>
<tr>
<th>Dependent Variables:</th>
<th>Minutes per Subject</th>
<th>Evaluations per Subject</th>
<th>Minutes per Evaluation per Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within campusXlegXsession</td>
<td></td>
<td>Within campusXlegXsession</td>
</tr>
<tr>
<td></td>
<td>Full Sample</td>
<td>Restricted Sample</td>
<td>Full Sample</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
<td>One</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>(1) T1 - Control</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(2) T2 - Surprising</td>
<td>16</td>
<td>-12</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(8)*</td>
<td>(11)</td>
</tr>
<tr>
<td>(3) T3 - Anticipated</td>
<td>32</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(9)**</td>
<td>(9)</td>
<td>(9)</td>
</tr>
<tr>
<td>Subjects (clusters)</td>
<td>91</td>
<td>81</td>
<td>86</td>
</tr>
</tbody>
</table>

Notes. Standard errors are in parentheses and are clustered by subject. Significant at the 10% level, **Significant at the 5% level; ***Significant at the 1% level. The significance levels for the “Minutes per Subject” panel are for a one-sided test: $H_0 : \beta_i = 0$ versus $H_a : \beta_i < 0$ for ease of comparison with Deci (1971). The significance levels on all other tests are for two-tailed tests.
B The Cookie Evaluation Form

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.

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the flavor develop quickly?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does one particular flavor develop too quickly?</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Overall, how does it rate on “flavor developing nicely”?  

### Flavor.

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does it have a chocolaty flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a buttery flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a peanutty flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a strawberry-like flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have an almondly flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a ginger-like flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a minty flavor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does it have a cheesecake-like flavor?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
C 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.”

**C.1 Protocol for the 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

52
“Thank you for your participation! Did you enjoy the experience? [Research assistant listens the answer]. Here is your Godiva thank-you cookie tin.”

C.2 Protocol for the Surprising Treatment

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 cents 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.”

C.3 Protocol for the Anticipated Treatment

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 sure subjects had received this information.

1. The phone call and/or the email to inform subjects in advance about the payment scheme. The phone call 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 the information that “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.”

### D Time Line

![Timeline Diagram](image-url)

Figure D.1: Timeline

### E Proofs and Extra Proposition

#### Proof of Proposition 1

(i) Suppose $\beta \leq 1$.

**Anticipated Treatment.** Straight from the first order condition in equation (4) whenever $\beta \leq 1$, which implies $(1 - \beta)w \geq 0$.

**Surprising Treatment.** Start assuming $w$ is big enough, in particular, $w \geq \frac{\nu_0\eta}{0.5 - \beta + \eta} > 0$, where the last
inequality follows from the assumption that $\beta \leq 1$. This condition over $w$ can thus be written as,

$$V_0 + (1 - \beta)w + \eta w \geq V_0(1 + \eta \lambda)$$

Because in equilibrium $\mu'_k \equiv \mu(e_1^s - \bar{e}) = 1$ and $\mu'_e \equiv \mu(c(\bar{e}) - c(e_1^s)) = \lambda$, using the equation above and the first-order conditions for the CONTROL and SURPRISING treatments in the first period we have,

$$c'(e_1^s) = \frac{(V_0 + (1 - \beta)w + \eta w \mu'_k)}{[1 + \eta \mu'_e]} \geq V_0 = c'(\bar{e}).$$

(ii) Suppose $\beta > 1$.

ANTICIPATED Treatment. Straight from the first order condition in equation (4) whenever $\beta > 1$, which implies $(1 - \beta)w < 0$.

SURPRISING Treatment. Start assuming $w$ is small enough, in particular, $w \leq \frac{V_0 \eta}{(1 - \beta) + \eta \lambda}$. We consider two cases. First, $1 < \beta < 1 + \eta \lambda$ and thus $\frac{V_0 \eta}{(1 - \beta) + \eta \lambda} > 0$. The condition over $w$ can be then be written as

$$V_0 + (1 - \beta)w + \eta w \lambda \leq V_0(1 + \eta)$$

Because in equilibrium $\mu'_k \equiv \mu(e_1^s - \bar{e}) = \lambda$ and $\mu'_e \equiv \mu(c(\bar{e}) - c(e_1^s)) = 1$, using the equation above and the first-order conditions for the CONTROL and SURPRISING treatments in the first period we have,

$$c'(e_1^s) = \frac{(V_0 + (1 - \beta)w + \eta w \mu'_k)}{[1 + \eta \mu'_e]} \leq V_0 = c'(\bar{e}).$$

Second, suppose $\beta > 1 + \eta \lambda$. In this case, from the first-order conditions for the CONTROL and SURPRISING treatments in the first period we have that the sufficient condition for $e_1^s < \bar{e}$ is,

$$[(1 - \beta) + \eta \mu'_e]w \leq V_0 \eta \mu'_e$$

which holds for all $w$ as the LHS is negative because $\beta > 1 + \eta \lambda$ and the RHS side is positive because $\mu$ is strictly increasing. ∗

**Proof of Proposition 2**

(i) Suppose $\beta = 0$.

ANTICIPATED Treatment. $e_2^s = \bar{e}$ follows straight from comparing the first-order conditions for the CONTROL and ANTICIPATED treatments in the second period, which are equal whenever $\beta = 0$.

SURPRISING Treatment. To see that $e_2^s < \bar{e}$ notice that because $\mu$ is strictly increasing it must be the case that,

$$-\lambda \eta \mu'_k \nu < V_0 \eta \mu'_e$$
where $\mu'_k$ and $\mu'_e$ are defined as in equation (7) for any plan $\tilde{e}_2$. Adding and subtracting conveniently, this inequality can be expressed as:

$$c'(\tilde{e}_2) = \frac{V_0}{1 + \eta \mu'_e} - \frac{\lambda w \eta \mu'_k}{1 + \eta \mu'_e} < V_0 = c'(\bar{e})$$  \hspace{1cm} (4)

**(ii)** Suppose $\beta > 0$.

**ANTICIPATED** Treatment follows straight from from comparing the first-order conditions for the CONTROL and ANTICIPATED treatments in the second period where $0 < V_0 - \beta w < V_0$.

**SURPRISING** Treatment. To see that $e^*_2 < \bar{e}$ notice that because $\mu$ is strictly increasing it must be the case that,

$$-\beta w - \lambda w \eta \mu'_k < V_0 \eta \mu'_e$$

where $\mu'_k$ and $\mu'_e$ are defined as in equation (7) for any plan $\tilde{e}_2$. Adding and subtracting conveniently, this inequality can be expressed as:

$$c'(e^*_2) = \frac{(V_0 - \beta w)}{1 + \eta \mu'_e} - \frac{\lambda w \eta \mu'_k}{1 + \eta \mu'_e} < V_0 - \beta w = c'(\bar{e})$$  \hspace{1cm} (5)

which concludes the proof. $\star$

**Preamble to Proposition 3**

We now consider the case where agents in the SURPRISING treatment make a probabilistic assessment about the period-two payment. We show that in this case a sufficient condition for agents in the SURPRISING treatment to exert less effort than the CONTROL is that the agent is sufficiently loss averse.

For simplicity we assume there are two levels of effort, $e \in \{0,1\}$ where $c(1) = c < V_0$ and $c(0) = 0$. The rest of the set up follows that in Section 3. To keep utility specifications tractable, we omit reciprocity in levels and reference-dependent utility in payments as in the main text.

Following our experimental design, the timing is as follows. Period zero mimics that in Section 3. In period one subjects in the SURPRISING treatment are told about the piece rate and immediately after they exert effort. At the end of the period the agent makes plans about the effort he will exert in period two. To form these plans he assigns probability $0 < \pi < 1$ to receiving the piece rate at the beginning of the second session. There are, therefore, four possible contingent effort plans: work for any realization of $w_2 \in \{0,w\}$, work if $w_2 = w$ and not work if $w_2 = 0$, do not work if $w_2 = w$ and work if $w_2 = 0$ and never work for any realization of $w_2 \in \{0,w\}$.

Following Kőszegi and Rabin (2006), the agent’s reference point corresponds to his rational expec-
tations about outcomes. In our set up, this reference point corresponds to one of the agent’s contingent effort plans described above. The fact the these plans will constitute both the reference and the actual action, however, posses a circularity problem: preferences must be used to determine the reference plan but at the same time the reference plan determines the shape of the preference structure because agents have reference-dependent preferences.

To solve the circularity problem Kőszegi and Rabin (2006, 2007) propose to use a Personal Equilibrium or PE. This solution concept takes advantage of the rational expectations assumption: agents can only form expectations they are willing to follow once these expectations constitute the reference point. An action therefore constitutes a PE if, once planned and implemented, it provides the highest utility relative to any possible deviation from the plan. For instance, in our framework, if the agent chooses to exert effort for any realization of \( w_2 \), then implementing this plan must give him higher utility than deviating to low effort—having planned to work—in case \( w_2 = 0 \) is observed at the beginning of period two. Whenever a given effort plan obeys this rational expectation condition, the plan it is said to be credible and it constitutes a personal equilibrium.

There could be, however, several plans that are credible and thus multiple personal equilibria exists. For instance, both, work for any wage and work only if \( w_2 = w \) can be credible, in the sense that the agent will actually exert effort for any \( w_2 \) and \( w_2 = w \), respectively, once the respective plan constitutes the reference point. In this situation the agent chooses the credible effort plan that gives him the highest total utility. Such plan constitutes what Kőszegi and Rabin call a Preferred Personal Equilibrium or PPE.\(^\text{18}\)

Personal Equilibrium is not the only possible solution concept. Kőszegi and Rabin (2007) also propose a Choice-Acclimating Personal Equilibrium or CPE. As PPE, CPE proposes that the reference point corresponds to the agent’s expectations about outcomes (one of the effort plans described above). Contrary to the PPE, however, CPE does not make sure that the equilibrium effort plan is credible—by comparing the equilibrium utility with the utility of all possible deviations from the equilibrium plan—but rather it assumes it: given that agents have rational expectations CPE imposes that agents will implement the effort plan that constitutes his reference and then he will chose among these plans the

\(^{18}\)In the model presented in Section 3, forming expectations and choosing actual actions given expectations are simpler problems. First, because plans were made under certainty—and therefore they are computed through a straight consumption utility maximization—and second because departures from expectations occur right before effort was exerted, providing the agent no time to update his expectations.
one that provides the highest total utility. For instance, to check whether “work” for any piece-rate realization is a CPE, the agent computes the utility of working for any piece rate (conditional on this being the plan) and compares it against the utility that planning and implementing any other contingent effort plans provides.\textsuperscript{19}

PPE and CPE are adequate in different temporal contexts. In particular, PPE is adequate whenever plan formation and its implementation are simultaneous or take place within a short period of time. Meanwhile, CPE is adequate whenever there is a time lag between the plan formation stage and their implementation. The crucial assumption in a CPE, therefore, is that this time lag will allow the agent to commit to his effort plans and thus implement them once planned. PPE, to the contrary, does not assume commitment (but rather imposes credibility) and thus it allows for short time distances between planning and implementation.\textsuperscript{20}

Because in our experimental design there is a week between the payment news and no information arrives between the end of period one—when period-two effort plans are made—and the second-period effort decision, we use the CPE as a solution concept.

We now present the effort predictions for period two. We show that effort in the SURPRISING treatment will still be lower to that of the CONTROL in the case the agent forms stochastic expectations about his second-period payment.\textsuperscript{21}

**Proposition 3** There exists an upper bound $\bar{\lambda}$ such that if $\lambda > \bar{\lambda}$ then $e_2^s = 0 < \bar{e} = 1$ whenever $w_2 = 0$.

**Proof** First recall from Section 3 that $\bar{e}$ solves $V_0 = c'(\bar{e})$. For agents in the SURPRISING treatments, there are four possible second-period effort plans depending on whether there is a piece rate or not. First work for any realization of $w_2$, not work for any realization of $w_2$, work if $w_2 = w$ (and thus not work if $w_2 = w$) and work if $w_2 = 0$ (and thus not work if $w_2 = w$). We show that first and last are not equilibriums and thus, if an equilibrium exists, it must be the case that $e_2^s = 0$.

\textsuperscript{19}The CPE is actually a very similar concept to the solution concepts proposed by Bell (1985) and Loomes and Sugden (1986). The difference is only computational, in that in these papers the agent’s reference point is assumed to be the lottery’s certainty equivalent, meanwhile in Kőszegi and Rabin agents compare each possible scenario with each possible realization of the reference point.

\textsuperscript{20}This is the reason why Kőszegi and Rabin (2006) say that in a PPE the reference point corresponds to the agent’s recent rational expectations.

\textsuperscript{21}Period-one behavior is not interesting for this section purposes, as it is not affected by probabilistic expectations. Because of the assumption that agents exert effort immediately after receive in the payment news, agents do not update the (fixed) expectations they formed in period zero.
The utility if the agent plans and actually works for any realization of \( w_2 \) corresponds to

\[
U = \pi[V_2 + w - c + \eta\pi (\mu(w - w)\mu(0 - 1) + \mu(c - c)) + \eta(1 - \pi) (\mu(w - 0)\mu(0 - 1) + \mu(c - c))]
\]

\[
+ (1 - \pi)[V_2 - c + \eta\pi (\mu(0 - w)\mu(0 - 1) + \mu(c - c)) + \eta(1 - \pi) (\mu(0 - 0)\mu(0 - 1) + \mu(c - c))]
\]

\[
= V_0 - \beta w + \pi w - c
\]

(6)

We start showing that working for any realization of the wage is not a PE because deviating to not work for \( w_2 = 0 \) is profitable whenever the agent is sufficiently loss averse. To see this notice that deviating towards \( e = 0 \) if \( w_2 = 0 \) and \( e = 1 \) if \( w_2 = w \) having planned to exert \( e = 0 \) if \( w_2 = 0 \) or \( w_2 = w \) involves gains and losses. On the one hand he losses the expected amount of \((1 - \pi)(V_0 - \beta w - c)\) from a decrease in expected consumption utility. This, however, increases his expected reference-dependent utility from two sources. First, he creates an expected gain because he is now able to reciprocate the principal’s denial of the wage. This expected gain corresponds to \((1 - \pi)\eta\mu(0 - w)\mu(0 - 1)\) (relative to \((1 - \pi)\eta\mu(0 - w)\mu(1 - 1) = 0\) in case he sticks to the plan). Second, he also creates a pleasant surprise of not having to exert effort having planned to, expected gain which amounts to \((1 - \pi)\eta\mu(c - 0)\) (relative to \((1 - \pi)\eta\mu(c - c) = 0\) if he sticks to the plan). Then, deviating to no work if \( w_2 = 0 \) is a profitable deviation if:

\[
\lambda^2 > \frac{V_0 - \beta w - (1 + \eta)c}{\eta w}
\]

(7)

We now show that working whenever \( w_2 = 0 \) and not working whenever \( w_2 = w \) is not a PE because deviating to not work for any \( w_2 \) is a profitable deviation whenever the agent is sufficiently loss averse. Doing a similar calculation to that in equation (6), we have that the utility if the agent plans to work only if there is no wage and no work if there is wage and actually implements this plan, corresponds to

\[
U = \pi[0 + \eta\pi (\mu(w - w)\mu(0 - 0) + \mu(0 - 0)) + \eta(1 - \pi) (\mu(w - 0)\mu(0 - 1) + \mu(c - 0))]
\]

\[
+ (1 - \pi)[V_2 - c + \eta\pi (\mu(0 - w)\mu(1 - 0) + \mu(0 - c)) + \eta(1 - \pi) (\mu(0 - 0)\mu(1 - 1) + \mu(c - c))]
\]

\[
= (1 - \pi)[V_0 - \beta w - c] - \eta\pi(1 - \pi)[2\lambda w + (\lambda - 1)c]
\]

(8)

As before, there are expected gains and losses to deviating to not work for any \( w_2 \). On the one hand, the expected losses amount to \((1 - \pi)(V_0 - \beta w - c)\) from decrease in consumption utility. On the other hand, there are three gains. First, the agent saves himself the disappointment of not being able to reciprocate the lack of \( w \), turning the loss \((1 - \pi)\pi\mu(0 - w)\mu(1 - 0)\) into \((1 - \pi)\pi\mu(0 - w)\mu(0 - 0) = 0\). Second,
he also saves himself the disappointment of having to exert effort in the scenario where he planned not to, turning the loss \((1 - \pi)\pi \mu(0 - c)\) into \((1 - \pi)\pi \mu(0 - w)\mu(0 - 0) = 0\). Finally, he also creates a pleasant surprise by not exerting effort having planned to, turning the zero \((1 - \pi)^2 \eta \mu(c - c)\) into a gain \((1 - \pi)^2 \eta \mu(c - 0) > 0\). Then, deviating to no work for any \(w_2\) is a profitable deviation if:

\[(1 - \pi)\pi \eta \lambda(w + c) + (1 - \pi)^2 \eta c > (1 - \pi)(V_0 - \beta w - c)\]

Thus, the condition over \(\lambda\) corresponds to

\[\lambda > \frac{(V_0 - \beta w - c) - (1 - \pi)\eta c}{\pi \eta (w + c)}\]

where using \(V_0 > c\) an upper bound can be written as

\[\lambda > \frac{V_0(1 + \pi \eta) - \beta w - (1 + \eta)c}{\pi \eta (w + c)}\] (9)

The result follows from choosing the \(\bar{\lambda}\) as the maximum lower bound determined by equations (7) and (9).

\section*{F Extended Model}

In this section we show that Section 3 results are robust to the inclusion of reciprocity in levels and a reference-dependent term for utility from payments and intrinsic motivation. The rest of the set up follows that in the body of the paper. We start formalizing reciprocity in levels and then we reintroduce consumption utility and reference-dependent utility.

(1) Pure Reciprocity. To model reciprocity in levels, we assume there exists kindness functions \(K_w(w - \bar{w})\) and \(K_e(e - \bar{e})\) that capture the utility the agent gets from better-than-fair actions. Let \((\bar{w}, \bar{e})\) represent these fair actions. Following the literature on reciprocity (Rabin (1993), Charness and Rabin (2002), Gächter and Falk (2002), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006)), period-\(t\) reciprocal consumption utility has the following multiplicative form

\[\gamma K_w(w_t - \bar{w}) K_e(e_t - \bar{e})\] (10)

where and \(\gamma \geq 0\) is a preference parameter representing the importance of the reciprocity term relative to consumption utility form wages and effort. Intuitively, whenever the principal grants a better than fair wage \((w > \bar{w})\) the agent gets a marginal benefit from exerting a higher-than-fair effort \((e > \bar{e})\). Equivalently, if the principal grants a worse-than-fair wage \((w < \bar{w})\) it is marginal benefit to exert lower-than-fair effort to reciprocate the principal’s unkindness. For simplicity we assume both kindness
functions are linear and equation (10) reduces to
\[ \gamma K_w(w_t - \bar{w})K_e(e_t - \bar{e}) = \gamma w_t(e - \bar{e}). \]

Finally, to define fair action, we use the period-zero agreement. In particular, we argue that in cases where a contract governs the relationship between the parties, the fair actions are determined by the initial contract. Therefore, from our experimental design we assume \( \bar{w} = 0 \) and \( \bar{e} \) solves \( V_0 = c'(\bar{e}). \)

(2) Total Consumption Utility. Period-\( t \) total consumption utility corresponds to the summation of consumption utility from wages and intrinsic motivation, pure reciprocity, effort and the cookies granted at the of the period as a thank-you gift.

\[ (V_t + w_t)e_t + \gamma w_t(e_t - \bar{e}) - c(e_t) + T \] (11)

where we have used the linearity of the kindness functions. \( V_t \) and \( T \) were defined in Section 3.

(3) Expectation-Based Reference-Dependent Utility. As before, agents experience reference-dependent utility in all consumption domains: intrinsic utility, wages, reciprocity and effort. Let \( (\tilde{V}_t, \tilde{w}_t, \tilde{e}_t) \) represent the expectation made in period \( t - 1 \) for \( (V_t, w_t, e_t) \). Total period-\( t \) expectation-based reference-dependent utility corresponds to:

\[ \eta_w \mu \left( (V_t + w_t)e_t - (\tilde{V}_t + \tilde{w}_t)\tilde{e}_t \right) + \eta_k \gamma \mu (w_t - \tilde{w}_t)(e_t - \tilde{e}_t) + \eta_e \mu (c(\tilde{e}_t) - c(e_t)) \] (12)

where \( \mu \) function is the value function as described in equation (2) and the parameters \( \eta = (\eta_w, \eta_k, \eta_e) \) represent the importance of each reference-dependent domain relative to consumption utility. Notice we do not add a reference-dependent term for \( T \) because it is fixed and thus expectations are always met.

The main difference between equation (12) and equation (2) in Section 3 is the extra term representing reference-dependent utility in payments and intrinsic motivation (the first term in equation (12)). The second term replicates that in Section 3 and it represents reference-dependent reciprocal utility. In equation (12), however, we have made explicit the \( \gamma \) parameter which multiplies the function. This parameter representing the importance of pure reciprocity (see equation (11) above) multiplies reference-dependent reciprocity to capture the intuition that the importance of reference-dependent preferences is proportional to the strength of consumption utility (Kőszegi and Rabin (2006)). Namely, if \( \gamma = 0 \) then the agent does not experience reference-dependent reciprocity either.

\^{22}The notation of the fair effort preserves that in Section 3, where \( \bar{e} \) represents the effort exerted by the CONTROL group.
(4) **Total Utility.** Total period-\(t\) utility, \(U_t\), corresponds to the summation of equations (11) and (12).

(5) **Equilibrium.** Identical to that in Section 3.

Let \(e^g_t, t \in \{1, 2\} \) and \(g \in \{c, a, s\}\) denote the period behavior of the CONTROL, ANTICIPATED, and SURPRISING groups respectively.

(6) **Period-One Behavior.**

(6.1) For Subjects in the CONTROL group \(w_1 = \tilde{w}_1 = 0\). Thus they solve

\[
e^c_1 = \underset{e}{\text{argmax}} (V_1 + w_1)e - c(e) = V_0 - c(\bar{e}) \quad \Rightarrow \quad V_0 = c'(e^c_1) \equiv c'(\bar{e})
\]

Notice these agents do not experience reciprocity in levels because their are paid the contracted wage. Further, they do not experience any reference-dependent utility as all expectations are met.

(6.2) For subjects in the ANTICIPATED treatment, \(w_1 = \tilde{w}_1 = w\). Thus they solve

\[
e^a_1 = \underset{e}{\text{argmax}} (V_1 + w_1)e + \gamma w_1(e - \bar{e}) - c(e) + T = (V_0 + (1 - \beta)w)e + \gamma w(e - \bar{e}) - c(e)
\]

\[
\Rightarrow \quad V_0 + (1 - \beta)w + \gamma w = c'(e^a_1)
\]

Notice agents in the ANTICIPATED treatment do experience reciprocity in levels because their are paid higher than the contracted wage. They don’t, however, experience any reference-dependent utility as all expectations are met.

(6.3) For subjects in the SURPRISING treatment, \(w_1 = w\) but \(\tilde{w}_1 = 0\). Thus they solve

\[
e^s_1 = \underset{e}{\text{argmax}} (V_1 + w_1)e + \gamma w_1(e - \bar{e}) - c(e)
\]

\[
+ \eta_w \mu_s ((V_1 + w_1)e - (\tilde{V}_1 + \tilde{w}_1)e) + \eta_k \mu_k (w_1 - \tilde{w}_1) \mu(e - \bar{e}) + \eta_{c0} \mu_c (c(\bar{e}) - c(e))
\]

because in period zero the agent expected to exert effort \(\bar{e} = \bar{e}\) in period one. Using \(V_1 = V_0 - \beta w_1\), \(\tilde{V}_1 = V_0 - \beta \tilde{w}_1 = V_0 \) and \(T = 0\), we can rewrite the agent’s problem as

\[
e^s_1 = \underset{e}{\text{argmax}} (V_0 + (1 - \beta)w)e + \gamma w(e - \bar{e}) - c(e)
\]

\[
+ \eta_w \mu_s ((V_0 + (1 - \beta)w)e - V_0 \bar{e}) + \eta_k \mu_k (e - \bar{e}) + \eta_{c0} \mu_c (c(\bar{e}) - c(e))
\]

The f.o.c. corresponds to

\[
[1 + \eta_w \mu'_w](V_0 + (1 - \beta)w) + [\gamma + \eta_k \mu'_k]w = [1 + \eta_{c0} \mu'_c] c'(e^s_1)
\]

where \(\mu'_w \equiv \mu'((V_0 + (1 - \beta)w)e^s_1 - V_0 \bar{e})\), \(\mu'_k \equiv \mu'(e^s_1 - \bar{e})\) and \(\mu'_c \equiv \mu'(c(\bar{e}) - c(e^s_1))\)
We focus here on the two new terms: reciprocity in levels and reference-dependence in payments (the intuition behind the other terms is equal to that in equation (5)). Briefly, both terms in this first period correspond to a marginal benefit: agents reciprocate the excess wage above the contracted wage and the pleasant surprise of having more money than expected.

Proposition 4 proves the equivalent to Proposition 1 using the extended version of the preference structure.

**Proposition 4 (First-Period Effect of the piece rate)**

Assume \( \lambda > \eta_w \lambda_e \).

(i) Suppose \( \beta \leq 1 \). For any \( w \), \( e_1^a \geq \bar{e} \). Moreover, if \( w \) is big enough, then \( e_1^s \geq \bar{e} \).

(ii) Suppose \( \beta > 1 + \gamma \). For any \( w \), \( e_1^a < \bar{e} \). Moreover, if \( w \) and \( \beta \) are big enough, then \( e_1^s < \bar{e} \).

**Proof** (i) Suppose \( \beta \leq 1 \).

**Anticipated treatment.** \( e_1^a \geq \bar{e} \) follows straight from comparing equations (13) and (14) whenever \( \beta \leq 1 \).

**Surprising treatment.** To see that \( e_1^s \geq \bar{e} \), start assuming \( w \) is big enough, in particular, \( w > \frac{V_0(\eta_e - \eta_w)}{(1-\beta)(1+\eta_w) + \gamma + \eta_k} \), where the lower bound is positive because of assumption \( \lambda > \eta_e \lambda \). Doing some algebra, this condition can be rewritten as

\[
[1 + \eta_w](V_0 + (1 - \beta)w) + [\gamma + \eta_k]w \geq V_0(1 + \eta_e \lambda)
\]

Because in equilibrium \( \mu'_w = \mu'_k = 1 \) and \( \mu'_e = \lambda \), using the equation above and the first-order conditions in equations (13) and (15) we have

\[
c'(e_1^s) = \frac{[1 + \eta_w \mu'_w](V_0 + (1 - \beta)w) + [\gamma + \eta_k \mu'_k]w}{[1 + \eta_e \mu'_e]} \geq V_0 = c'(\bar{e}).
\]

(ii) Suppose \( \beta > 1 + \gamma \).

**Anticipated treatment.** \( e_1^a < \bar{e} \) follows straight from comparing equations (13) and (14) whenever \( \beta > 1 + \gamma \).

**Surprising treatment.** To see that \( e_1^s \geq \bar{e} \), start assuming \( \beta \) and \( w \) are big enough. In particular, assume that \( \beta > \frac{(1 + \eta_w \lambda + \gamma + \eta_k \lambda)}{(1 + \eta_e \lambda)} > 1 \) and that \( w > \frac{-V_0(\eta_e - \eta_w)}{((1-\beta)(1+\eta_w \lambda) + \gamma + \eta_k \lambda)} \).

There are two cases to consider depending on the size of \( \lambda \) and thus whether the lower bound for the wage is positive or negative.
**Case 1**: \(\lambda \eta_w > \eta_e > \frac{\eta_w}{\lambda}\). In this case \(V_0(\eta_e - \eta_w, \lambda) > 0\). Using this, the restriction over \(\beta\) and the lower bound for the wage we have that,

\[
[(1 - \beta)(1 + \eta_w \lambda) + \gamma + \eta_k \lambda]w < V_0(\eta_e - \lambda \eta_w) < 0
\]  

(16)

Because in equilibrium we have \(\mu'_w = \mu'_k = \lambda\) and \(\mu'_e = 1\), adding and subtracting conveniently in equation (16) we can show that,

\[
c'(e^1) = \frac{[1 + \eta_w \lambda]V_0 + [(1 - \beta)(1 + \eta_w \lambda) + \gamma + \eta_k \lambda]w}{[1 + \eta_e \mu'_e]} < V_0 = c'(\bar{e}).
\]

where the equalities follow the first-order conditions in equations (13) and (15).

**Case 2**: \(\eta_e > \lambda \eta_w > \frac{\eta_w}{\lambda}\) and thus \(V_0(\lambda \eta_w - \eta_e) < 0\). Using this and the restriction over \(\beta\) we have that,

\[
[(1 - \beta)(1 + \eta_w \lambda) + \gamma + \eta_k \lambda]w < 0 < V_0(\eta_e - \lambda \eta_w)
\]  

(17)

which holds for all \(w\). The rest of the proof follows that in Case 1. *

Two caveats are important from Proposition 4. First, because of the new forces introduced in the extended model, the number of parameter combinations that lead to different effort predictions is considerably greater than that in the reduced model. Proposition 4 thus only presents a few possible set of sufficient conditions. Second, even though the conditions presented in part \((i)\) mimic those in Proposition 1, those presented in path \((ii)\) do not. In particular, in Proposition 1 part \((ii)\) we require the wage to be big enough rather than low enough. Despite this, we notice that in this extended model it is also possible to find an equilibrium for a sufficiently low wage. This equilibrium, however, is harder to identify and thus we have chosen to describe a simpler set of parameters. This comes at no cost, since equivalently to Section 3, the purpose of Proposition 4 is to show that effort in the treatment conditions can be greater as well as smaller relative to the Control.

(7) Period-Two Behavior.

(7.1) Subjects in the Control group solve the same problem as that in period one, because the utility from the period-two fixed reward—the thank-you cookies—does not change with the piece rate and \(w_2 = \tilde{w}_2 = 0\).

(7.2) For subjects in the Anticipated treatment, \(w_2 = \tilde{w}_2 = 0\) (just as in the Control group). Thus
they solve
\[ e^a_2 = \arg\max_e (V_2 + w_2)e + \gamma w_2(e - \bar{e}) - c(e) + T \]
\[ = \arg\max_e (V_0 - \beta w)e - c(e) + \alpha V_0 \Rightarrow V_0 - \beta w = c'(e^a_2) \] (18)

where there is no reciprocity in level as agents are paid the market wage and all reference-dependent terms are zero because the agent does not expect any departure from expectations.

(7.3) For subjects in the Surprising treatment \( w_2 = 0 \) but \( \tilde{w}_2 = w \) and thus they solve
\[ \max_e (V_2 + w_2)e + \gamma w_2(e - \bar{e}) - c(e) + T \]
\[ + \eta w \mu((V_2 + w_2)e - (\tilde{V}_2 + \tilde{w}_2)e) + \eta k w_1 \mu(e - \tilde{e}_2) + \eta e \mu(c(\tilde{e}_2) - c(e)) \]

because the agent expected to exert effort \( \tilde{e}_2 \) in period two. Using \( V_2 = V_1 = V_0 - \beta w \) and \( \tilde{V}_2 = V_1 - \beta \tilde{w}_2 = V_0 - 2\beta w \), we have
\[ \max_e (V_0 - \beta w)e - c(e) + \alpha V_0 \]
\[ + \eta w \mu((V_0 - \beta w)e - (V_0 + (1 - 2\beta)w)\tilde{e}_2) + \eta k w_1 \mu(e - \tilde{e}_2) + \eta e \mu(c(\tilde{e}_2) - c(e)) \]

The FOC corresponds to
\[ [1 + \eta w \mu'_w](V_0 - \beta w) - \lambda w \eta k \mu'_k = [1 + \eta e \mu'_e] c'(e^a_2) \] (19)

where \( \mu'_w = \mu'((V_0 - \beta w)e - (V_0 + (1 - 2\beta)w)\tilde{e}_2) \), \( \mu'_k = \mu'(e^a_s - \tilde{e}_2) \) and \( \mu'_e = \mu'(c(\tilde{e}_2) - c(e^a_s)) \)

We focus here on the two new terms: reciprocity in levels and reference-dependence in payments (the intuition behind the other terms is equal to that in equation (7)). Briefly, reciprocity in levels is zero in period two: subjects in the Surprising treatment are paid the contracted wage, meanwhile reference-dependence in payment is once more a marginal benefit as increasing effort decreases the loss the agent experiences from getting a payment lower than expected.

Proposition 5 proves that the retaliation of disappointment hypothesis holds using the extended version of the preference structure.

**Proposition 5** *(Retaliation or disappointment Hypothesis)*
Suppose \( \beta = 0 \). If the wage is big enough or if the agent is sufficiently loss averse, then \( e^*_2 < e^a_1 = \bar{e} \).

**Proof** ANTICIPATED treatment. \( e^a_1 = \bar{e} \) follows straight from comparing equations (13) and (18) whenever \( \beta = 0 \).
Surprising Treatment. To see that $e_2^2 < \bar{e}$ recall that for any plan for the second period $\bar{e}_2$ we defined 

$$\mu'_w \equiv \mu'(V_0 - \beta w)e_2^s - (V_0 + (1 - 2\beta)w)\bar{e}_2), \mu'_k \equiv \mu'(e_2^s - \bar{e}_2) \text{ and } \mu'_e \equiv \mu'(c(\bar{e}_2) - c(e_2^s)).$$

We consider two cases. Case 1: $\eta e \mu'_e \leq \eta w \mu'_w$. In this case the following inequality holds,

$$-\lambda w \eta_k \mu'_k < 0 \leq V_0(\eta e \mu'_e - \eta w \mu'_w)$$

Adding and subtracting conveniently, this inequality can be expressed as,

$$c'(e_2^s) = \frac{V_0}{1 + \eta \mu'_e} - \frac{\lambda w \eta \mu'_k}{1 + \eta \mu'_e} < V_0 = c'(\bar{e}) \quad (20)$$

Case 2: $\eta e \mu'_e < \eta w \mu'_w$. In this case the following inequality holds,

$$-\lambda w \eta_k \mu'_k < V_0(\eta e \mu'_e - \eta w \mu'_w) < 0$$

and thus for equation (20) to hold, we need,

$$\lambda w > \frac{V_0(\eta e \mu'_e - \eta w \mu'_w)}{\eta_k \mu'_k}$$

Finally, Proposition 6 shows that if agents have classical preferences no difference between treatments should be observed.

**Proposition 6** *(Behavior Under Classical Preferences)*

Suppose $\beta = \gamma = 0$ and $\eta = 0$. Then, $e_2^s = e_2^a = \bar{e}$.

**Proof** Obvious from equations $V_0 = c'(\bar{e})$, (18) and (19).

Proposition 6 highlights the agent’s behavior whenever their preferences follow the classical model: no crowding out and no social preferences and thus no reference-dependent retaliation. In this case, we should observe all treatment behave equally. It is important to notice, however, that the classical model is not the only one that make this prediction. In fact, agents with $\beta > 0$ and $\eta > 0$ can also display this behavior if marginal benefits and costs balance out perfectly. We consider such event as a low-probability one and favor the classical model, that is, the most parsimonious model.