Gift Exchange in the Workplace: Addressing the Conflicting Evidence with a Careful Test

Constança Esteves–Sorenson*

March 2017

Abstract

Tests of gift exchange, wherein agents receive excess wages which are non-contingent on performance in one-shot settings, have yielded contradictory evidence: they sometimes find effort boosts, consistent with gift exchange whereas they sometimes find no effort increases, consistent with a standard model. We identify eight confounds that could have led to the mixed evidence—agent disutility from being viewed as selfish, small samples, insufficient wage raises, an effort ceiling, fatigue, selection of abler workers, reemployment concerns and peer effects—and run a comprehensive test addressing them. Our test consisted of a field experiment hiring workers for a data entry job, followed by laboratory games assessing their prosocial behavior. After addressing these confounds, we find that behavior during the field test was consistent with a standard model: workers did not repay fixed wage raises with an effort boost but they did raise effort in response to a piece rate. The piece rate scheme was also more efficient: the effort boost came at lower expense than paying fixed wage raises. Further, workers who behaved prosocially in laboratory games did not behave prosocially in the field. (JEL: D03, J41, M52)

Keywords: Incentives, Gift Exchange, Field Experiments, Laboratory Experiments.

*constanca.esteves-sorenson@yale.edu, Yale University. This paper was previously circulated with the title "Revisiting Gift Exchange: Theoretical Considerations and a Field Test." We thank Rosario Macera for her work during the initial stages of this project. We thank Ben Arsenault, James Floman, Julia Levinson, Tiffany Lin, Andrew Pearlmutter and Beau Wittmer for excellent research assistance. We also thank Dan Benjamin, Gary Charness, Stefano DellaVigna, Florian Ederer, Uri Gneezy, Mitch Hoffman, Lisa Kahn, Botond Kőszegi, Kory Kroft, Ulrike Malmendier, Michel Maréchal, Mushfiq Mubarak, Ted O'Donoghue, Matthew Rabin, Frédéric Schneider, Olav Sorenson, Josh Tashoff and Roberto Weber for helpful discussions and suggestions at different stages of this project, and seminar participants at Cornell University, Stanford University's Institute for Theoretical Economics, University of California Berkeley, University of Warwick, Yale University, and Universidad Católica de Chile for helpful comments. This research was partially funded by Whitebox grants in Behavioral Economics and the Institution for Social and Policy Studies, both at Yale University.

1 Introduction

Gift exchange has important implications for the provision of incentives. Advanced by Akerlof (1982), it arose as one of three main theories for why it may be efficient for firms to pay above-market wages: (i) such wages attract and retain higher-ability workers (Weiss (1980)); (ii) they motivate workers to raise effort so as to not be fired and lose the wage premium yielded by the repeated employment relationship (e.g., Shapiro and Stiglitz (1984)); or, (iii) as suggested by gift exchange, they prompt agents to boost effort because agents develop sentiment for the firm, thus acquiring utility from reciprocating the firm's gift of above-market wages with the gift of above-minimal effort.[1,2]

Gift exchange thus holds the important promise that workers will exert excess effort (to repay the firm's excess wage) even in the absence of performance-contingent rewards, such as performance pay, reemployment, promotion, or a good market reputation. It could thus diminish not only the need for costly monitoring technologies and performance evaluations to motivate workers, but also the use of repeated interactions, as agents would exert high effort even if offered no chance of reemployment.

Given its importance, gift exchange has been extensively tested, with mixed results: the evidence is sometimes consistent with gift exchange and sometimes with a standard principal-agent model. Typical studies of gift exchange have tested whether paying workers higher non-performance-contingent wages boosts their effort in one-shot settings. Whereas a standard model would predict no effort increases, effort boosts in these tests have been seen as evidence for gift exchange: that workers repaid higher wages with higher effort. In this context, laboratory tests have yielded large pay-effort elasticities: for example, Fehr, Kirchsteiger, and Riedl (1993), the most cited laboratory gift-exchange test found that a 140% wage raise led to a 300% effort increase (a 2.14 elasticity). Yet, other studies, particularly field tests, have found low, statistically insignificant elasticities: for example, a 33% wage raise in Kube, Maréchal, and Puppe (2013) yielded a statistically insignificant effort change of -0.3% (a -0.01 elasticity), in line with a standard model.

The conflicting gift-exchange elasticities have led to a broad debate, to which we contribute by (i) identifying the factors that could have led to the mixed evidence (confounds) and (ii) by running a test dealing with them. We identify eight factors that could have led to the conflicting results: agent disutility from being perceived as selfish; small samples;

[1] Akerlof (1982, pages 543-544) states: "As a consequence of sentiment for the firm, the workers acquire utility for an exchange of 'gifts.' ... On the worker's side, the 'gift' given is work in excess of the minimum work standard; and on the firm's side the "gift" given is wages in excess of what [workers] could receive if they left their current jobs."

[2] Though these efficiency wages' theories aimed to explain involuntary unemployment, we focus on their implications for the literature on the provision of incentives (e.g., Gibbons and Waldman (1999)).

insufficient wage raises; an effort ceiling; fatigue; selection of abler workers; reemployment concerns; and peer effects. We then implemented a field test, free (as much as possible) of these confounds, and in which workers were unaware they were part of a study. After the field test, we ran laboratory games to assess workers' prosociality.

After addressing these factors that could have led to the mixed evidence, the field test results are more consistent with a standard model: non-performance-contingent (i.e., fixed) wage raises induced no effort increases whereas a piece rate did. Further, the piece rate boosted effort despite entailing a lower expense than the fixed wage raises. Last, workers who behaved prosocially in the laboratory did not do so in the field.

The field test consisted of recruiting 194 students for a one-time data entry job—the most used worker-job combination in field tests of gift exchange—for the going market wage, and randomly assigning them to four conditions, varying additional pay. Students from two universities were hired for \$12 per hour to digitize an academic library for six hours split into three two-hour weekly shifts. After hiring, 47 were randomly assigned to the Control, where they got no raises; 70 to a 67%Raise condition, where they got a raise to \$20 per hour prior to starting the job; 45 to a 50%-100%Raise condition, where they got a raise to \$18 per hour prior to starting the job and a further surprise raise to \$24 per hour before the third shift; and 32 to a PieceRate condition, where instead of the fixed wage raise, workers got an additional per record piece rate before starting the job, representing an average 21% extra pay. All fixed wage raises were framed as gifts.

The field test dealt with the above mentioned confounds in gift-exchange studies. First, our between-subjects design shrouded individual workers' actions and thus prevented agents' disutility from being perceived as selfish from yielding effort boosts that could be confounded with gift exchange (a confound identified by Levitt and List (2007)). Between-subjects tests, such as ours, compare the effort of a group receiving a wage raise before the job (treatment) and a group not receiving it (control). Thus each worker in the treatment only labors under the wage raise. As a result, his baseline effort (effort in the absence of the raise) is unknown to others (e.g., the principal and/or experimenters) and, therefore, it is also unknown whether he is behaving selfishly by not increasing effort under the raise. By enabling each worker in the treatment to behave selfishly without detection, between-subjects designs ensure that effort increases are due to gift exchange (the intrinsic taste for repaying the wage) and not due to the disutility from being perceived as selfish. In contrast, in the two other tests of gift exchange—laboratory and within-subject field tests—effort in the absence of a raise is observable and thus selfishness can be easily detected. In laboratory studies, each worker's minimum effort is known to other players and/or the experimenter, while in within-subject field tests, each worker's

effort is observed before and after the raise. Thus, in either case, whether the worker increased effort after the raise is observable, giving him a separate reason, unrelated to gift exchange, to boost effort: to avoid the moral cost of being perceived as selfish.

Second, our large sample addressed the dual issues posed by small samples: (1) low power for detecting statistically significant gift-exchange elasticities, leading to support of a standard model; or, alternatively, (2) the potential to yield large statistically significant gift-exchange elasticities, as only elasticities which are large, by chance, are able to be statistically significant despite the large standard errors arising from the small sample.

Third, we used substantial raises of 50%, 67% and 100%, matching or exceeding those in other field tests, to allay concerns that behavior in line with a standard model could be due to raises being too small to elicit effort. We also compared effort under these raises to that under the piece rate, where the average pay increase was smaller, at 21%.

Fourth, the piece rate scheme also helped ascertain whether an effort ceiling caused behavior consistent with a standard model—workers wanted to boost effort to reciprocate the wage but the task is too onerous—by observing if effort increased with the piece rate.

Fifth, we divided the six-hour task into three two-hour weekly shifts to minimize the role of fatigue in curbing gift exchange. If workers engaged in gift exchange during a two-hour shift they had one week to rest, so as to continue exerting high effort. Further, splitting the job over three weeks, allowed workers to reflect on how to boost productivity from shift to shift, thus giving gift exchange more scope to increase performance.

Sixth, hiring at the market rate minimized the role of selection of high-productivity workers in possibly generating evidence consistent with a standard model. Above-market wages might attract abler workers (Weiss (1980)), whose output is close to the highest feasible, preventing them from lifting it substantially to repay the raise.

Seventh, the one-time job discouraged workers from increasing effort to be rehired at the higher wage, thus preventing reemployment concerns from being confounded with gift exchange (e.g., Shapiro and Stiglitz (1984)). Eighth, workers labored alone and received no peer information to preclude peer effects confounds from biasing effort estimates.

In the second part of our study, the same workers were invited to play Sequential Prisoner's Dilemma (SPD) games after the conclusion of the field test, as during it they had to be unaware they were part of a study. These SPD games measure prosocial behavior in the laboratory (e.g., Burks, Carpenter, and Goette (2009), Schneider and Weber (2012)). We tested whether workers who behaved prosocially in these games, where their individual actions could be observed, also behaved prosocially in the field (by raising effort in response to wage raises) where their prosocial actions could not be observed due to the between-subjects design.

We find that workers behaved as in a standard model during the field test: large raises of 50% and 67%, and a further raise to 100% yielded statistically insignificant effort changes of -4% to 4% (elasticities of -0.08 to 0.06) in the specification most favorable to gift exchange. However, the cheaper piece-rate scheme, peaking at only 30% extra pay by the third shift, boosted effort by a conservative 18% (an elasticity of 0.60). The piece rate was thus substantially more efficient at boosting effort. Further, workers who behaved prosocially in the laboratory games did not do so on the job, exhibiting similar, statistically insignificant, negative to small positive effort changes of -3% to 7%.

This paper is, to our knowledge, the first to identify and describe how one or a combination of confounds could have led to the conflicting gift-exchange evidence and to address them in a single study: a paired field-laboratory test. Though others have tried to tackle some of these factors, such as an effort ceiling (Kube, Maréchal, and Puppe (2013)) and small samples (e.g., Cohn, Fehr, and Goette (2014)) none has jointly dealt comprehensively with all of the above-mentioned factors, as shown in Section 2.

Beyond this contribution, our test also introduced two novel features—the piece rate and splitting the task over three weeks—not present in prior tests. First, hiring workers at a flat wage and then comparing their behavior under a fixed wage raise and the piece rate allowed us not only to address confounds (e.g., an effort ceiling), but also to assess which of the two schemes more efficiently boosted effort. Second, separating the task into three periods gave workers time to rest and assess how to improve productivity.[3]

Our findings further the debate on whether gift exchange is a powerful motivator. Prior positive elasticities, at times quite large, suggested that principals, especially those employing agents whose performance cannot be easily monitored or assessed, may not need to use costly monitoring and/or performance evaluation schemes as well as long-term relationships to elicit effort (see Gibbons (2005) for a review): workers would boost effort to repay the excess wage even in the absence of these tools. Our results, which are more consistent with a standard model, together with the finding that the piece rate boosted effort more efficiently, suggest that gift exchange is not as powerful a motivator as thought, though it still may operate in other contexts (discussed in Section 6).[4]

[3]Our test differs from that in Kube, Maréchal, and Puppe (2013) who also tested whether an effort ceiling could explain the absence of gift exchange. They found that workers hired via a piece-rate contract outproduced those hired with a fixed-wage contract and subsequently given a 33% fixed raise. This evidence, though suggestive of no effort ceiling, had been viewed as inconclusive as piece-rate contracts attract higher-productivity workers than fixed-wage contracts (Lazear (2000)). The higher output of the piece-rate hires could thus have been due to their higher productivity versus the fixed-wage hires rather than due to the absence of an effort ceiling for the fixed-wage hires. We ensured that workers receiving fixed wage raises and the piece rate were similar as both selected into the same fixed-wage contract.

[4]Our results are also consistent with those in two recent working papers—DellaVigna, List, Malmendier, and Rao (2016) and Muralidharal, Pradhan, Ree, and Rogers (2016)—discussed at the end.

2 Prior Gift-Exchange Tests in the Workplace and Potential Confounds

To motivate the design of our test, we describe the most-cited field and laboratory studies with gift-exchange tests in the workplace. They all test whether workers boost effort in response to wage raises that are not conditional on performance and in one-shot interactions. These two features ensure that, under a standard principal-agent model, there should be no increase in effort as pay is not contingent on performance and reemployment incentives are absent (e.g., Shapiro and Stiglitz (1984)). Thus, when workers have raised effort in tests with these two features, this has been viewed as evidence for gift exchange.

We show that similar wage raises in these tests have sometimes yielded effort responses, some quite large, consistent with gift exchange whereas sometimes they have yielded no effort responses (elasticities close to zero or statistically indistinguishable to zero) in line with a standard model. We describe how the disparate pay-effort elasticities in these tests—ranging from a statistically insignificant -7.6 to 2.9— could stem from unanticipated confounds introduced by features of these tests, and briefly note how we addressed them. Table 1, Panels I and II, summarizes field and laboratory tests, respectively; Panel III surveys real-effort laboratory tests that have accompanied field tests. We now describe their potential confounds, also summarized in Table 2.

(1) **Moral costs of selfish actions, observability and the advantage of between-subjects designs**. Much of the evidence for gift exchange has arisen in the laboratory, starting with the influential work by Fehr, Kirchsteiger, and Riedl (1993), introducing the Gift-Exchange Game and showing that fixed wage raises of 140% over the market clearing level in one-shot games increased workers' effort beyond the minimum by 300%, an elasticity of 2.14 (Table 1, Panel II, row (1)), constituting a clear departure from the standard model. The authors created four labor markets in the laboratory, each with 13-15 students randomly assigned to the role of employer or employee. Employers offered wages for three minutes, which workers accepted or rejected. Once hired, workers chose effort from the same cost-of-effort table—where the minimal effort of 0.1 units equated to a cost of effort of zero, for example—and each worker's effort choice was observable by the principal and the experimenters. There was no worker-level heterogeneity in cost of effort as all workers had the same table. Each one-shot game, from the wage bids to the workers' effort choices, lasted ten minutes. Subjects played twelve one-shot games with an anonymous trading partner to curb reputation confounds.

An extensive experimental literature based on tests with similar features to those in Fehr, Kirchsteiger, and Riedl (1993) has since shown large elasticities for gift-exchange games: excess wages of over 100% have induced excess effort of over 200% (Table 1, Panel

II, columns (3) and (4)). For more details on these tests, see Appendix D.

These large laboratory elasticities could, however be inflated by subjects' disutility of projecting a selfish image. This hypothesis was first proposed by Levitt and List (2007), who posited that agents' actions resulted from a trade-off between wealth and the moral costs of behaving selfishly and that these moral costs increased in the observability and scrutiny of agents' actions by others. Consistent with this, in a review of several laboratory and field studies, they found that the lower the observability and scrutiny of agents' actions, the more selfishly they behaved. Further, the lower the stakes in the game—i.e., the cheaper it was to behave unselfishly—the higher the prosocial behavior.

Beyond possibly inducing behavior consistent with gift-exchange in the laboratory, disutility from being perceived as selfish may also induce such behavior in within-subject field tests. In field tasks, individual cost of effort is heterogeneous and unobservable: the task is harder for some workers than for others. Thus, some studies test for gift exchange by observing each worker's effort before and after a raise, a within-subject test. Though this approach is reasonable, agents' behavior may reflect the disutility of being perceived as selfish: the worker knows that the principal knows his baseline effort before the raise and thus whether the worker is behaving selfishly by not increasing effort after the raise. Thus, the worker may lift his effort for a reason unrelated to gift exchange (the intrinsic desire to repay the wage): to avoid the moral costs of being perceived as selfish.

Between-subjects tests of gift exchange prevent selfish image concerns from inflating effort responses. Between-subjects tests compare a treatment group that receives the raise with a control group that does not. Thus, the baseline effort of each "treated" worker—effort in the absence of the raise—is unknown, and thus whether he exceeded his baseline under the raise is also unknown. Thus, each worker can refrain from boosting effort in response to the raise without incurring the cost of projecting a selfish image.

The disutility of projecting a selfish image may thus partially explain why between-subjects field tests document lower effort responses to the same wage raise than laboratory tests. For example, in the between-subjects field experiment in Gneezy and List (2006), the most cited gift-exchange field test, a 67% increase in the fixed wage increased output by 27% in the first 1.5 hours (an elasticity of 0.4), which was significant at the 5% level in a one-tailed test (Table 1, Panel I, row (1)). Effort waned thereafter, however, resulting in a statistically insignificant increase of 2% over the six hours of the task (an elasticity of 0.03). In contrast, the same 67% raise in the laboratory Bilateral Gift-Exchange Game in Fehr, Kirchler, Weichbold, and Gächter (1998) yielded a much larger effort increase of 50%-100% and thus a substantially larger elasticity of 0.74-1.5.[5]

[5]See Appendix D, Panel II, bullet (3.5) showing the effort response to a 67% raise. Table 1, Panel II,

The disutility of projecting a selfish image may also partially explain why between-subjects tests often yield lower effort responses, often consistent with a standard model, than within-subject tests. For example, Kube, Maréchal, and Puppe (2013) found, in their between-subjects test, that a 33% raise did not increase output, but rather decreased it slightly by a statistically insignificant -0.3% (an elasticity of -0.01; Table 1, Panel I, row (7)). In contrast, the within-subject test in Cohn, Fehr, and Goette (2014), where workers' productivity was observed with and without a smaller wage raise of 23%, showed a statistically significant 3% increase in productivity, with an upper bound of 14% for a subsample (elasticities of 0.13 and 0.61, respectively, in Panel I, row 8).

We used a between-subjects test to address the confound of selfish image concerns. To further investigate this confound we asked workers to play laboratory SPD games after the field experiment. As gift-exchange games are also SPD games, we tested whether prosocial actions by each worker in our SPD games, which could be observed (e.g., by the experimental team), corresponded to those in the field, where they could not.

(2) **Small samples.** Small samples can yield evidence at times consistent with a standard model and at times consistent with gift exchange. For example, Cohn, Fehr, and Goette (2014) argue that the smaller and thus lower-power sample in Kube, Maréchal, and Puppe (2012) could account for evidence consistent with a standard model: that a 19% wage raise yielded a statistically insignificant 5% effort boost in a sample of 69 workers (a statistically insignificant elasticity of 0.26; Table 1, Panel I, row (6)).[6]

Small samples can, alternatively, lead to large gift exchange pay-effort elasticities, as only estimates that are high by chance manage to be statistically significant despite the large standard errors induced by the small sample.[7] Tests with few subjects, often with 13 or fewer in an experimental condition (see Table 1, column (3) in Panel I and column (2) in Panels II and III) could lead to inflated short-term gift exchange elasticities. For example, in a 30-minute real-effort laboratory experiment, Hennig-Schmidt, Sadrieh, and Rockenbach (2010) found that a 10% wage raise supplemented by surplus information (the value of the task to the principal) for a 19-student treatment yielded an increase in effort by 29% vis-à-vis the 10-student control, who received no raise or surplus information (Table 1, Panel III, row (2)). This very large elasticity of 2.9, if statistically significant, could overstate the true elasticity, due to the very small control sample.[8]

row (3) summarizes this game: many of the raises exceeded 67%, resulting in an average raise of 215%.

[6]Kube, Maréchal, and Puppe (2012) also show that workers responded to in-kind gifts. Though important, these incentives are outside this review's scope, which focuses on the provision of monetary incentives.

[7]See Button, Ioannidis, Mokrysz, and et al. (2013) on the statistical properties of estimates from small samples.

[8]The statistical significance of this result was not reported.

To prevent these two issues we used one of the largest samples in field tests of gift exchange: our CONTROL, 67%RAISE and 50%-100%RAISE start at 47, 70 and 45 workers, respectively. Appendix Section B shows that we had ex-ante 98% power to reject the null hypothesis, at the 5% level, that a 67% raise would not increase effort in favor of the one-sided alternative that it would do so by 20%.

(3) **Size of fixed wage raise**. Behavior consistent with a standard model could also be due to the wage raise being too small to elicit effort responses. For example, though Gneezy and List (2006) found that a 67% raise boosted effort by 27% in the first 1.5 hours of their 6-hour data entry task, Kube, Maréchal, and Puppe (2013) showed that a smaller, 33% raise did not increase effort in the first 1.5 hours on a similar job (rather it lowered it by a statistically insignificant 10%).

To prevent small raises from curbing effort increases, we offered 50%, 67% and 100% wage raises, matching or exceeding those in prior field studies: raises on data entry jobs with student workers, such as ours, have ranged from 19% to 67% and those in other field tests have ranged from 10% to 100% (Table 1).

(4) **Effort ceiling**. Behavior consistent with a standard model could have been due to an effort ceiling: workers wanted to reciprocate the wage but the task was so onerous that they quickly reached an infinite marginal cost of effort, preventing any effort boosts. For example, perhaps the 33% raise in Kube, Maréchal, and Puppe (2013) elicited no extra effort (a statistically insignificant elasticity of -0.01) because the task was very onerous, whereas the same 33% wage raise in Gilchrist, Luca, and Malhotra (2015) led to an effort increase of 18% (an elasticity of 0.55). An effort ceiling curbing effort responses could also explain why gift exchange in laboratory games is generally much larger than in field experiments: workers' cost-of-effort function in the laboratory, which is designed by experimenters, may minimize bounded effort responses.

To test whether an effort ceiling could curb gift exchange, we ran a piece-rate treatment to assess whether it was feasible to increase effort in our task: we offered a subsample an additional per record piece rate instead of a fixed wage raise.

(5) **Fatigue**. Fatigue could have also deflated past gift-exchange estimates, as workers became tired of exerting higher effort in response to a raise. For example, Gneezy and List (2006) showed that the bulk of the effort response to the 67% raise in their 6-hour data entry task occurred in the first 1.5 hours; effort over the 6 hours only increased by a statistically insignificant 2% (an elasticity of 0.03). Similarly, in a companion field test with a fund-raising task, they found that most of the response to the 100% fixed wage raise occurred in the first 3 hours of the 6-hour job, waning thereafter. The authors tested if this waning was due to fatigue by inviting workers to return the next day, after

resting. However, only 4 control and 9 treatment subjects returned, yielding low power to detect a difference. Similarly, Bellemare and Shearer (2009), showed that the bulk of effort increases occurred on the day their workers got a raise (Table 1, Panel I, row (3)).

To prevent fatigue from dampening effort and leading agents to behave, over time, as in a standard model, we split our task into three two-hour shifts exactly one week apart. Thus, workers who became tired on one shift from reciprocating the raise had one week to recover. Further, workers could think about how to improve productivity over the one-week interval, giving them further scope to reciprocate.

(6) **Selection of higher-productivity workers**. Selection of higher-productivity workers might have led to behavior consistent with a standard model, as these workers' baseline output may be close to the upper bound for a task, preventing them from lifting it substantially after a raise. Hiring at above-market wages may attract these workers if reservation wages and ability are positively correlated (Weiss (1980), Bewley (1999)).

Selection of abler workers could thus be another reason why Kube, Maréchal, and Puppe (2013) found that a 33% wage raise did not increase effort but Cohn, Fehr, and Goette (2014) found that a 23% raise did. The former hired at 15 euros per hour, almost twice the market wage of 8 euros, potentially attracting a high proportion of higher-output workers (who would not have applied for the job at the lower market wage). Thus, workers receiving the 33% raise in the treatment group might have had little room to increase effort over that of similar high-productivity workers in the control. Workers in the latter study, however, were hired at the market wage, so they may have had more leeway to increase productivity.[9]

To prevent such selection from curbing gift exchange, we hired at the market wage; this also has the advantage of more closely following the gift-exchange hypothesis, whereby a natural proxy for the workers' outside option is the market wage.

(7) **Reemployment confounds**. In contrast, effort increases ascribed to gift exchange could have been due to reemployment confounds. For example, Bellemare and Shearer (2009), showed that tree planters who returned the following year responded more to the 37% fixed raise than those who did not. Also, contrary to Kube, Maréchal, and Puppe (2013), where a surprise 33% raise did not increase performance, the same surprise raise in Gilchrist, Luca, and Malhotra (2015) increased performance by 26% among frequent oDesk workers; yet, it did not do so among first-timers (Panel I, row (9)). One reason why only frequent workers boosted effort could have been reputation concerns. In

[9]Cohn, Fehr, and Goette (2014) hired at the market wage as they argued that hiring at higher wages might curb effort responses to ensuing raises via another mechanism other than selection of high-productivity workers: that agents *do not* reciprocate raises to wages viewed as fair, as in the fair wage-effort hypothesis by Akerlof and Yellen (1990).

Table 1: Overview of Studies of Gift Exchange in the Workplace

	PANEL I: Field Studies (In chronological order)	Task (1)	Design (2)	Sample Sizes (3)	% Wage Increase (4)	% Effort Response (5)	Elasticity (6)
(1)	Gneezy and List (2006)	Data entry (6 consecutive hours in one day).	Between subjects	9 student workers in wage-raise ("Gift") treatment; 10 in control ("noGift").	67% ($8 per-hour raise relative to $12 per-hour base).	2% vs. control (whole 6 hours; not significant); 27% vs. control (first 1.5 hours; significant at 5%); 11% vs. control (first 3 hours; not significant); All one-tailed tests.	0.03 (whole 6 hours); 0.40 (first 1.5 hours); 0.16 (first 3 hours).
(2)		Door-to-door fundrasing (3 hours pre-lunch and 3 hours post-lunch).	Between subjects	13 student workers in wage-raise ("Gift") treatment; 10 in control ("noGift").	100% ($10 per-hour raise relative to $10 per-hour base).	38% vs. control (whole 6 hours; significant at 10%); 72% vs. control (first 3 hours; significant at 5%); 6% vs. control (second 3 hours; not significant); All one-tailed tests.	0.38 (whole 6 hours); 0.72 (first 3 hours); 0.06 (second 3 hours).
(3)	Bellemare and Shearer (2009)	Tree planting (7 days spread over two weeks).	Within subjects	18 workers of a tree-planting firm.	37 % ($80 one-day raise relative to $215 average daily earnings).	11% -14% (significant at 5%)[(1)].	0.30-0.38
(4)	Hennig-Schmidt, Rockenbach and Sadrieh (2010)	Data entry (2 hours; 1-hour shift per month; wage raise in second hour).	Within and between subjects	25 student workers in wage raise ("F10") treatment; 24 in control ("F0").	10% (2 DM per-hour raise relative to 20 DM per-hour base).	-76% decrease in effort across shifts vs. control (not significant).	-7.6
(5)			Within and between subjects	23 student workers in wage raise ("F40 peer") treatment; 24 in control ("F0").	40% (8 DM per-hour raise relative to 20 DM per-hour base) + Peers' wage information.	-28% decrease in effort across shifts vs. control (not significant).	-0.69
(6)	Kube, Maréchal and Puppe (2012)	Data entry (3 consecutive hours in one day).	Between subjects	34 student workers in wage raise ("Money") treatment; 35 in control ("Baseline").	19% (€7 raise relative to €36 pay, from the €12 per-hour base).	5% vs. control (not significant).	0.26
(7)	Kube, Maréchal and Puppe (2013)	Data entry (6 consecutive hours in one day).	Between subjects	22 student workers in wage raise ("PayRaise") treatment; 25 in control ("Baseline").	33% (€5 per-hour raise relative to €15 per-hour base announced as a "projected" wage).	-0.3% vs. control (whole 6 hours; not significant); -10%, 1%, 0.2% and 7% vs. control (1st, 2nd, 3rd and 4th 1.5 hours, respectively; none significant).	-0.01 (whole 6 hours); -0.33, 0.03, 0.01 and 0.21 (1st, 2nd, 3rd and 4th 1.5 hours, respectively).
(8)	Cohn, Fehr and Goette (2014)	Newspaper distribution (average of 6.5 three-hour shifts in 4 weeks).	Within and between subjects	196 workers of a promotion agency commissioned to distribute newspapers.	23% (5 ChF per-hour raise relative to 22 ChF per-hour base).	3% (full sample; significant at 5%); 14% (reciprocal subjects who felt underpaid by 5 ChF; significant at 5%).	0.13 (full sample); 0.61 (reciprocal subjects who felt underpaid by 5 ChF (upper bound)).
(9)	Gilchrist, Luca and Malhotra (2015)	CAPTCHA entry in oDesk (4 hours spread as desired across 7 days).	Between subjects	58 oDesk workers in wage raise ("3+1") treatment; 110 in control ("3").	33% ($1 per-hour raise relative to $3 per hour base).	18% vs. control (full sample; significant at 5%); 26% vs. control frequent workers; significant at 5%); 3% (first-time workers; not significant).	0.55 (full sample); 0.79 (frequent workers); 0.09 (first-time workers).

Notes: The most cited published field studies. All tests are two-tailed unless otherwise stated. "vs." is an abbreviation for "vis-à-vis". "Not significant" means not significant at either the 5% or 10% levels.[(1)]Estimate of the percentage increase in effort as the fixed-effects model only identifies level (instead of percentage) changes in effort. "Within and between-subjects" describe tests where workers' effort was observed pre- and post-raise for a given wage sequence in one condition and pre- and post-raise for a different wage sequence in another condition. "Significant at 5%" and "Significant at 10%" means significant at the 5% and 10% levels, respectively. Appendix D elaborates on the studies on this table.

Table 1: Continued

	PANEL II: Most Cited Laboratory Studies (In chronological order)	Task	Sample Sizes	% Wage Increase (experimental currency)	% Effort Response (experimental currency)	Elasticity
		(1)	(2)	(3)	(4)	(5)
(1)	Fehr, Kirchsteiger and Riedl (1993)	Employers make public wage bids. Upon accepting, workers choose effort.	Four 2-hour sessions, each with 8 to 9 student workers and 5 to 6 student employers.	140%: 72 units average wage over 30 units minimum (equilibrium) wage.	300%: 0.4 average effort over 0.1 minimum (equilibrium) effort.	2.14
(2)	Fehr, Kirchsteiger and Riedl (1998)	Buyers make price offers. Upon accepting, sellers choose a quality level.	Approx. two 3-hour sessions, each with 9 to 12 student sellers and 6 to 8 student buyers ("Reciprocity" treatment).	147%: 74 units average wage over 30 units minimum (equilibrium) wage.	250%: Approx. 0.35 average effort over 0.1 minimum (equilibrium) effort.	1.70
(3)	Fehr, Kirchler, Weichbold and Gächter (1998)	Employer-Worker random matching. Employer makes wage offer. Upon accepting, worker chooses effort.	Four 2-hour sessions with Austrian soldiers as subjects, each with 10 workers and 10 employers ("Bilateral GE" treatment).	215%: Approx. 63 units average wage over 20 units minimum (equilibrium) wage.	260%: Approx. 0.36 average effort over 0.1 minimum (equilibrium) effort.	1.21
(4)		Employers make public wage bids. Upon accepting, workers choose effort.	Four 2-hour sessions with Austrian soldiers as subjects, each with approximately 9 to 12 workers and 6 to 8 employers ("GE Market" treatment).	195%: Approx. 59 units average wage over 20 units minimum (equilibrium) wage.	290%: Approx. 0.40 average effort over 0.1 minimum (equilibrium) effort.	1.54
(5)	Gächter and Falk (2002)	Employers make public wage bids. Upon accepting, workers choose effort.	Three 2-hour sessions, each with 10 student workers and 10 student employers ("One-Shot" treatment).	248%: Approx. 73 units average wage over 21 units minimum (equilibrium) wage.	310%: 0.41 average effort over 0.1 minimum (equilibrium) effort.	1.25
(6)	Brown, Falk and Fehr (2004)	Employers make public or private wage bids. Upon accepting, workers choose effort.	Four 1.5-hour sessions, each with 10 student workers and 7 student employers ("Incomplete Contract Random" treatment).	380%: Approx. 24 units average wage over 5 units minimum (equilibrium) wage.	230%: Approx. 3.3 average effort over 1 minimum (equilibrium) effort.	0.61
	PANEL III: Companion Real-Effort Laboratory Experiments	Task	Sample Sizes	% Wage Increase	% Effort Response	Elasticity
(1)	Hennig-Schmidt, Rockenbach and Sadrieh (2010)	Folding letters and enveloping them in two 15-minute shifts separated by a 5-minute break; wage raise in second shift.	10 students in wage raise ("L10") treatment; 10 in control ("L0").	10%: €0.25 per 15 minutes over €2.5 per 15 minutes base wage.	-1% decrease in effort across shifts vs. control (not significant).	-0.10
(2)			19 students in wage raise ("L10 surplus") treatment; 10 in control ("L0").	10%: €0.25 per 15 minutes over €2.5 per 15 minutes base wage + Surplus information.	29% increase in effort across shifts vs. control (significance not reported).	2.90

Notes: Only the most cited studies of gift exchange in labor markets were included in Panel II. These are either laboratory tests focusing solely on gift exchange or that contain a gift-exchange treatment. The percentage wage increase in column (3) is the percentage wage increase versus the equilibrium wage when workers have money maximizing preferences (standard i.e., selfish preferences). The percentage effort increase in column (4) is the increase in effort versus the minimum effort (equilibrium effort) when workers have pure money maximizing preferences (standard i.e. selfish preferences). These studies also analyze the wage-effort relationship by regressing effort against a wage offer where the positive coefficient on the wage (signifying that increases in the wage are positively correlated with increases in effort) is always statistically significant at least at the 5% level using a two-tailed test. "Approx." is an abbreviation for "Approximately". Appendix D elaborates on the studies on this table.

Table 2: Summary of Potential Confounds in Studies of Gift Exchange in the Workplace

	PANEL I: Field Studies	Scope for Moral Cost confound? (Has a within-subjects design)	Small sample? (Conditions with fewer than 20 subjects)	%Wage raise large enough to yield a significant effort increase?	Test for effort ceiling?	Potential fatigue? (All work in one or adjacent days)	Potential selection? (Workers hired above the average market wage)	Potential reemployment concerns?	Potential peer effects?	Laboratory test assessing heterogeneity in prosocial preferences?
	(In chronological order)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	Gneezy and List (2006): Data-entry task	No	Yes	67% raise led only to a short-term effort increase	No	Yes	Unclear (market wage not declared).	No	No	No
(2)	Gneezy and List (2006): Fundrasing task	No	Yes	100% raise led to a 38% effort increase, but bulk (72%) in short term.	No	Inconclusive (extra 3-hour shift next day has a small sample: 9 subjects in treatment and 4 in control).	Unclear (market wage not declared).	No	No	No
(3)	Bellemare and Shearer (2009)	Yes	Yes	37% raise led to 11%-14% effort increase.	No	Yes	Unclear (market wage not declared).	Yes (on-going relationship for some workers).	No	No
(4)	Hennig-Schmidt, Rockenbach and Sadrieh (2010): "F10" treatment	Yes	No	10% raise led to no effort increase.	No	No	Yes	No	No	No
(5)	Hennig-Schmidt, Rockenbach and Sadrieh (2010): "F40 peer" treatment	Yes	No	40% raise led to no effort increase.	No	No	Yes	No	Yes	No
(6)	Kube, Maréchal and Puppe (2012)	No	No	19% raise led to no effort increase.	Indirect test (gift in kind increased effort)[(1)]	Yes	Unclear (market wage not declared).	No	No	No
(7)	Kube, Maréchal and Puppe (2013)	No	No	33% raise led to no effort increase.	Inconclusive test (subjects recruited using the piece rate, inducing selection).	Yes	Yes ("projected" wage used in recruitment is above market wage).	No	No	No
(8)	Cohn, Fehr and Goette (2014)	Yes	No	23% raise led to 3% effort increase.	No	No	No	No	No	Yes (Moonlighting game).
(9)	Gilchrist, Luca and Malhotra (2015)	No	No	33% raise led to no effort increase for first-time workers, but a 26% increase for frequent ones.	No	Unclear (4 hours spread across 7 days as desired).	Unclear (subjects hired at their oDesk reservation wage of $2-$3 per hour).	Yes (signaling to future employers via publicly viewable reviews).	No	No

Notes: Column (9) shows whether the field study also contains a laboratory test assessing heterogeneity in prosocial preferences by workers in the field experiment. Even though this is not a confound per se, it is an important feature with the goal of finding a relationship between laboratory and field behavior, similarly to that in our test.[(1)] This paper provides an indirect test of an effort ceiling in that it shows that gifts in kind were able to increase effort beyond the observed statistically insignificant 5% estimated with the wage raise.

Table 2: Continued

	PANEL II: Most Cited Laboratory Studies (In chronological order)	Scope for Moral Cost Confound? (1)	Small sample? (Conditions with fewer than 20 subjects) (2)	%Wage raise large enough to yield a significant effort increase? (3)	Test for effort ceiling? (4)	Potential fatigue? (5)	Potential selection? (6)	Potential reemployment concerns? (7)	Potential peer effects? (8)
(1)	Fehr, Kirchsteiger and Riedl (1993)	Yes	No	140% raise led to a 300% effort increase.	n/a	No[1]	No[2]	No[3]	No
(2)	Fehr, Kirchsteiger and Riedl (1998)	Yes	No	147% raise led to a 250% effort increase.	n/a	No[1]	No[2]	No[3]	No
(3)	Fehr, Kirchler, Weichbold and Gächter (1998): "Bilateral GE" treatment	Yes	No	215% raise led to a 260% effort increase.	n/a	No[1]	No[2]	No[3]	No
(4)	Fehr, Kirchler, Weichbold and Gächter (1998): "GE Market" treatment	Yes	No	195% raise led to a 300% effort increase.	n/a	No[1]	No[2]	No[3]	No
(5)	Gächter and Falk (2002)	Yes	No	248% raise led to a 310% effort increase.	n/a	No[1]	No[2]	No[3]	No
(6)	Brown, Falk and Fehr (2004)	Yes	No	380% raise led to a 230% effort increase.	n/a	No[1]	No[2]	No[3]	No
	PANEL III: Companion Real-Effort Laboratory Experiments								
(1)	Hennig-Schmidt, Rockenbach and Sadrieh (2010): "L10" treatment	Yes	Yes	10% raise led to no effort increase vs. control.	No	Yes	No	No	No
(2)	Hennig-Schmidt, Rockenbach and Sadrieh (2010): "L10 surplus" treatment	Yes	Yes	10% raise + Surplus information led to a 29% effort increase vs. control (significance not reported).	No	Yes	No	No	No

Notes: Only the most cited laboratory studies focusing on gift exchange or with a gift exchange treatment pertaining to labor markets were included in Panel II. "n/a" and "vs." stand for "not applicable" and "vis-à-vis", respectively. [1] In the laboratory, effort is defined by the experimenter in an cost-of-effort table, which is fixed across rounds and sessions; therefore, effort in one round does not affect the marginal cost of effort in subsequent rounds. [2] The cost-of-effort table, written by the experimenter, does not vary across subjects. [3] Reemployment concerns are minimized by having subjects play a series of one-shot rounds.

oDesk, workers' job histories, including the number of jobs, hourly wages and employer ratings, are available for future employers. Thus, frequent workers could have increased effort after the raise not only to reciprocate but also to earn a good rating for the higher-paid job, a valuable signal of their higher marginal product to future employers. Reputation would naturally have a higher value for frequent workers—whose heavy usage of oDesk suggests it is a substantial source of current and future earnings—than for first-timers.[10]

To avoid reemployment concerns, which could be confounded with gift exchange, our workers were reminded during the job that it was a one-time engagement.

(8) **Peer effects**. Some tests may have found evidence consistent with a standard model due to peer effects. For example, Hennig-Schmidt, Sadrieh, and Rockenbach (2010) found that offering peer wage information together with a 40% wage raise *decreased* effort by a statistically insignificant 28% (Table 1, Panel I, row (5)). Wage-raise subjects could have curbed their effort to punish the principal for behaving unfairly towards similar workers (giving them a much lower raise), in line with Fehr and Fischbacher (2004), who show that subjects punish unfair actions towards a third party.[11]

To avoid potential peer effects in our test, our subjects worked in isolation and received no information about other workers, such as their wages.

3 Research Design

This section describes our two-part study and elaborates on how it addresses the confounds discussed previously. The first part of the study, the field experiment, was run between fall 2011 and fall 2012 in two legs at university A (a national university) and then in five further legs both there and also at university B (a local state university). Hiring workers from two campuses allowed not only for a larger sample but also for assessing whether there would be similar patterns across two different samples.

The second part was a laboratory test in winter 2013, after the conclusion of the field experiment. We invited workers to play SPD games and asked their consent to use their responses for research, as is standard in laboratory tests. Because we were required to ask for consent to use workers' field experiment data for research before they participated in the SPD games, we only implemented the games after the last wave of the field experiment had concluded. The workers therefore remained unaware had been partaking in a study while on the job enhancing the findings' external validity.[12]

[10]This paper contains another treatment, in which workers select into a \$4/hour contract. This is outside the scope of this review, which focuses on tests of gift exchange: whether workers reciprocate fixed wage raises with higher effort.

[11]Cohn, Fehr, Herrmann, and Schneider (2014) show that this punishment may not always occur.

[12]Asking for consent to use the field experiment data before the implementation of the SPD games was a Human Subjects requirement. No worker forbade the use of any data.

3.1 The Field Experiment

Recruitment and task. We hired 194 students from the two universities to create a bibliographic database for a department at one of the universities. Workers entered data on academic articles (e.g., title, authors, journal, year, volume, issue and pages) using bibliographic software. Campus flyers advertised the job as a one-time sequence of three weekly two-hour shifts. The hiring wage was \$12 per hour, the standard wage for data entry.[13] Subjects worked alone and did not know that all characters they entered—including spaces between words, backspaces, etc.—were recorded.

Between-subjects design and the treatments. After hiring in each leg, workers at a given campus were randomly assigned to four main conditions (e.g., in leg 3, those at campus A were randomly assigned to each condition and those at campus B were similarly, separately, assigned). The 47 workers in the CONTROL performed the task and received \$72 at the end of the third shift.

The 70 workers in the 67%RAISE condition received a 67% raise to \$20 after being hired but before starting their first shift. This condition results from the aggregation of three subtreatments—67%SURPRISERAISE, 67%ANTICIPATEDRAISE, and 67%PROMISEDRAISE—varying the timing of the information about the raise (immediately before or one week before the first shift) and when the raise was paid (at the start or at the end of each shift). Given that the distribution of outcomes for these three subconditions was not statistically different (see Section 4.1), we aggregated them into a single condition 67%RAISE.[14]

The 45 workers in the 50%-100%RAISE condition received a 50% raise to \$18 per hour after being hired but before they started their first shift; a subsample of these workers (23) got an additional surprise raise of 100%, to \$24 per hour in the third shift.[15]

[13]Source for market wages (06/2012): student pay scale (Campus A) and PayScale.com, Salary.com (Campus B).

[14]Specifically, the 67%SURPRISERAISE and 67%ANTICIPATEDRAISE subtreatments tested whether the pleasant surprise of a wage raise increased reciprocal effort. To this end, those in the 67%SURPRISERAISE received the news of the wage raise immediately before starting work on the first shift whereas those in the 67%ANTICIPATEDRAISE received this news one week in advance of the first shift (but after recruiting), so that the wage raise would not be a surprise. In both these treatments we paid the raise at the beginning of each shift to convey that it was not conditional on performance, as we worried that if it were given at the end, workers might erroneously perceive the raise as conditional on performance, thus inflating gift exchange estimates. The 67%PROMISEDRAISE variation tested for this confound: whether wage raises promised at the beginning of the first shift, but paid at the end of the contract—instead of being paid upfront as in the previous two treatments—could be misconstrued as contingent on performance, thus artificially enhancing effort. Our concerns were unfounded: we found that paying the raise upfront at the beginning of each shift or promising it at the beginning of the task and paying it at the end yielded the same outcome, a result also found in other studies (e.g., Kube, Maréchal, and Puppe (2013)).

[15]We ran one additional treatment, in which the remaining subset of the workers in this condition received a wage cut in the third shift, instead of a wage increase, returning to the \$12 per hour contract wage (thus receiving a \$6 per hour wage cut). We do not report these results as they are outside the

Treatments	Shift One	Shift Two	Shift Three
CONTROL	$2 \times \$12$	$2 \times \$12$	$2 \times \$12$
67%	$2 \times \$20$	$2 \times \$20$	$2 \times \$20$
50%-100%	$2 \times \$18$	$2 \times \$18$	$2 \times \$24$
PIECERATE	$2 \times \$12$ +Piece Rate	$2 \times \$12$ +Piece Rate	$2 \times \$12$ +Piece Rate

Table 3: Compensation per two-hour weekly shift

The 32 workers in the PIECERATE condition were offered a per record piece rate for the duration of the contract, instead of a fixed wage raise, before they started their first shift. Importantly, the piece rate was offered *after* hiring at the fixed \$12 per hour market wage had occurred, as an add-on to this wage, and thus it was not used to recruit workers. We therefore avoided selection of higher-ability workers into piece-rate contracts (Lazear (2000)), ensuring that workers in this treatment were similar to those in the fixed-wage-raise treatments. The piece-rate scheme was piecewise convex, where x is the number of records per shift: $\$0 \times x$ if $x < 70$; $\$0.05 \times x$ if $70 \leq x \leq 110$; $\$0.10 \times x$ if $110 < x \leq 140$; and $\$0.20 \times x$ if $x > 140$. These workers collected their piece-rate earnings at the end of each shift. The timing for this condition and all others, per campus, is in Figure A.1.

The fixed wage raises, offered after recruitment at the market wage but before the start of the job, were given in a gift envelope embossed with the phrase "A Gift for You" at the start of each weekly shift whereas the contract wage of \$72 was paid at the end of the third shift, upon the job's completion. Table 3 shows the total compensation per shift for the different conditions. See Appendix E for the protocol for the treatments.

Beyond having subjects' unaware that they were part of a study and that software tracked their input, strengthening the external validity of our findings, our test deals with the several confounds identified above, as we now describe.

(1) **The disutility from being perceived as selfish**. Because effort boosts could be due to the disutility of being perceived as selfish, we implemented a between-subjects design. Each worker in the fixed-wage-raise conditions knew that the principal had not observed how much he/she could produce in the absence of the raise, as each worker's effort was only observed after the raise (and we tested for gift exchange by comparing the outcomes for the group of workers in each treatment against those in the CONTROL). An exception was the last shift (shift three) in the 50%-100%RAISE, where a subsample got the additional raise to 100% of the contract wage. These workers knew that the principal had observed how much they could produce with a 50% raise, so they might boost effort after the additional raise, if the cost of doing so were smaller than the benefit of not

scope of this paper, which focuses on effort responses to wage increases. They are available upon request.

projecting a selfish image (as posited by the Levitt and List (2007) model).

(2) **Small samples.** To curb the generation of conflicting evidence by small samples, we had one of the largest between-subjects samples in a gift-exchange study. In particular, this sample had 98% power to reject the null of no gift exchange in favor of the one-sided alternative that a 67% raise would increase effort, as shown in the simple power calibration in Appendix B. Further, if workers increased effort in any statistically significant way in the PieceRate, which had fewer subjects, but failed to do so in the wage-raise treatments, then lack of gift exchange would very unlikely be due to low power.

(3) **Insufficient wage raises.** Because behavior consistent with a standard model could be due to insufficient wage raises, we offered 50%, 67%, and 100% raises, matching or exceeding those in prior data entry field tests with students, the most used task-worker combination in field tests of gift exchange. To further assess whether insufficient extra pay was the reason for no effort boosts, we compared effort under the fixed wages raises to that under the piece rate, representing a much lower average 21% extra compensation.

(4) **Effort ceiling**. To assess whether an effort ceiling could curb gift exchange, we ran the PieceRate condition testing whether workers hired at the \$12 per hour wage and later offered the per record piece rate (instead of a fixed raise) increased effort. Thus if workers raised effort with the piece rate but did not with fixed wage raises, then the absence of gift exchange would likely not be due to an effort ceiling.

(5) **Fatigue.** To minimize the role of fatigue in dampening effort increases (thus biasing behavior towards a standard model), we split the work into three two-hour shifts, each exactly one week apart, to give workers time to rest. This also gave them time to assess how to continue boosting productivity to reciprocate the wage.

(6) **Selection of higher-productivity workers.** We minimized the role of selection of higher-productivity workers in biasing results in favor a standard model, by hiring workers at the market wage of \$12 per hour.

(7) **Reemployment concerns.** To ensure gift exchange was not confounded with reemployment concerns, upon recruitment and throughout the six hours, we emphasized that the job was a one-time engagement.

(8) **Peer effects.** Subjects worked in isolation and were not given information about the participation or wages of other workers to avoid peer effects on performance (e.g., Mas and Moretti (2009)).

Boosting the potential for gift exchange and dealing with other confounds. Beyond dealing with these confounds, the design enhanced ex ante the potential for gift exchange by framing the wage raise as a voluntary kind action by the principal not conditional on agents' performance: it was offered in an envelope embossed with the

phrase "A Gift for You" at the start of each shift. Though framing the raise as a gift is not a necessary condition for gift exchange—see Akerlof (1982), Fehr, Kirchsteiger, and Riedl (1993), Gneezy and List (2006), for example—Charness (2004) showed that the perception of a principal's volition boosts reciprocal effort. Thus evidence consistent with a standard model would not be due to ambiguity concerning the principal's volition and kindness.

We also avoided differences in research assistants or demand effects from biasing results by having all subjects interact with the same research assistant who was blind to the research hypothesis. Thus differences in treatments' outcomes cannot be due to different assistants or demand effects.

3.2 Post-Field Experiment Laboratory Games

To investigate whether each workers' prosocial behavior in the laboratory, which could be observed by others (e.g., the experimental team), correlated with that in the field test, where individual prosocial actions could not be observed by the principal or any third party due to our between-subjects design, we invited workers to play laboratory games in the semester after the conclusion of the field experiment. The invitation included a consent form to participate and asked subjects to play three one-shot SPD games, as in Burks, Carpenter, and Goette (2009) and Schneider and Weber (2012).[16] They received a $10 participation fee and any gains from the games, which could amount to up to $15. In the first SPD game, subjects chose their action without knowing the opponent's play. In the second and third games, they chose their action after the first mover cooperated and defected, respectively. The stakes were equal in the three games, following those in the SPD and trust games in Clark and Sefton (2001), and Charness and Rabin (2002), respectively. Each player was randomly and anonymously paired with another from his/her university and this pairing determined the payoffs. Subjects played practice rounds before the actual games to ensure they understood them. See Appendix E.2 for details on the protocol.[17]

The three SPD games offer a less ambiguous taxonomy of agents' prosocial type (to what extent they voluntarily benefit others) than do other games, such as gift-exchange

[16]These games and other similar ones (e.g., trust games) have also been used to test whether prosocial behavior in the laboratory correlates with that in the field (e.g., Karlan (2005), Benz and Meier (2008), Burks, Carpenter, and Goette (2009), Baran, Sapienza, and Zingales (2010), and Carpenter and Seki (2010)). For a review of laboratory games assessing prosocial behavior see Levitt and List (2007).

[17]Our SPD games use the strategy method (where second movers make conditional decisions) instead of the direct-response method (where the second player makes a unique choice after observing the first-player move). This allows for the collection of a broader set of players' choices without sacrificing their validity, as treatment effects found with the strategy method are invariably observed with the direct-response method (Brandts and Charness (2011)).

games, where the action space is continuous (any effort choice in a given range). In our SPD games, each worker's triplet of binary cooperate/defect choices corresponds to one of eight types, of which three are our focus. On one extreme are the *altruists*, who cooperate no matter what (play Cooperate, Cooperate, Cooperate) followed by *conditional cooperators*, who cooperate as first movers, but only cooperate as second movers if the first player cooperates (play Cooperate, Cooperate, Defect). At the opposite extreme are the *selfish*, who defect no matter what (play Defect, Defect, Defect).[18] In contrast, in games with continuous action spaces, classification of workers by prosocial types would be more ambiguous, requiring more arbitrary cut-offs (e.g., defining those in the top decile or quartile of reciprocal effort as prosocial).

The importance of this classification is that if behavior in the games reflects an underlying prosocial preference, then the most prosocial workers in the laboratory, the *altruists* and *conditional cooperators*, should increase effort in response to wage raises in the field experiment, if feasible. Namely, *conditional cooperators*, the most numerous prosocial type in these games, reciprocate the cooperative behavior of the first mover. Thus, they should reciprocate the cooperative behavior of the first mover in the field experiment—the principal, who increased the wage—by increasing effort. Further, the wage raise should also lift effort among the remainder of prosocial workers, the *altruists*, as they always cooperate as second movers. However, if reciprocal behavior in the games reflects, to a large extent, agents' desire to avoid projecting a selfish image, then there should be little or no correlation between the behavior in the laboratory, where selfish actions by each worker can be observed, and in the between-subjects field test, where they cannot.

Importantly, because of their natural correspondence to our field test and to gift-exchange games, SPD games enabled us to test the correlation between laboratory and field behavior without alerting subjects to our purpose, which could bias their responses. Gift-exchange games are SPD games in which the principal cooperates/does not cooperate by offering/not offering above-market wages, and having observed this, workers cooperate/defect by increasing/not increasing effort beyond their minimum. Our field experiment is setup like a gift-exchange game and thus like a SPD game: the principal cooperates by offering a raise above the $12-market wage and workers cooperate/defect by increasing/not increasing effort over the baseline: the effort observed in the Control. We thus inferred our workers' prosocial behavior without implementing a gift-exchange game, which given its similarity to our field experiment, could have signaled our intent.

[18] This terminology is adapted from Schneider and Weber (2012), who call their conditional cooperators "optimistic conditional cooperators" and their selfish players "pessimistic selfish."

4 Field Experiment Results

This section discusses the results of the field experiment and the subsequent laboratory SPD games. After tackling the confounds that could have biased prior elasticities, we find that the evidence in our field experiment is most consistent with a standard model: large fixed wage raises of 50% and 67%, and a further raise of 100%, elicited statistically insignificant negative or small positive effort responses, yielding elasticities of -0.08 to 0.06, in the specification most favorable to gift exchange. The additional piece-rate pay, though entailing a maximal extra expenditure of only 30%, boosted effort by up to 18%-19%, yielding a much larger elasticity of up to 0.63.

4.1 Sample and Effort Measure

Sample. We hired 194 participants at two campuses, A and B. Slightly more than half (57%) came from A, where we got clearance to run the study earlier. Workers within each campus and leg were randomly assigned, without their knowledge, to each condition. We started with 47 students in the Control, 70 in the 67%Raise, 45 in the 50%-100%Raise and 32 in the PieceRate. The 67%Raise pooled three subtreatments giving workers a 67% raise—67%SurpriseRaise, 67%AnticipatedRaise and 67%PromisedRaise—as their outcomes were not statistically different (see Appendix Table A.1).[19] Of the 45 workers in the 50%-100%Raise, a random subsample of 23 got a further surprise raise to 100% of the contract wage in the third shift.[20]

Effort measure. We used the number of characters inputted per subject, instead of the number of records, as the former more closely approximates effort: some records may have longer titles and more coauthors, and so require more characters, for example. Further, number of characters is the measure used in other studies (e.g., Kube, Maréchal, and Puppe (2012)). Nonetheless, Appendix Table A.3 shows that the final results using characters as an outcome are similar to those using number of records (a noisier measure of effort) or number of correct words inputted (a measure of quality).

[19]The distributions are statistically indistinguishable between the three treatments within each campus, leg and shift, with only a minor exception: when including two campus B outliers in the 67%AnticipatedRaise, whose effort was twice the average effort across these three treatments in campus B. Excluding these two outliers from the Kruskall-Wallis test renders the distributions of the three treatments at campus B statistically indistinguishable. Including them, in contrast, leads to the rejection that these three distributions are the same for campus B in shifts one (at the 10% level) and three (at the 4% level). Yet we include the two outliers in the analysis because they increase the average effort of the 67%Raise relative to the Control, thus biasing our estimates in favor of gift exchange. Despite this, we observed that the 67%Raise yielded no increases, as we document later.

[20]Before we started the final experiment with 194 subjects, we ran a small pilot test on 15 subjects varying, among others, the wording in the treatments and the types of gift envelopes, with the goal of settling on the most natural design possible. Because these data are not comparable with that of the final experiment we do not include it in the analysis.

4.2 Descriptive Statistics

Disaggregated overall statistics. Table 4 shows that workers in the 67%Raise and 50%-100%Raise exerted lower effort than those in the Control. However, those in the PieceRate increased effort. Workers in the Control entered an average of 17,591 characters over 131 worker-shifts (column (1)). However, those in the 67%Raise and 50%-100%Raise entered -1% and -2% fewer characters than those in the Control over their respective 207 and 111 worker-shifts (column (2)). The finding that wage raises induce slight effort decreases is consistent with that of previous tests of gift exchange (see Table 1, Panel I). In contrast, those in the PieceRate inputted 15% more characters than those in the Control over their 90 worker-shifts (column (2)). This effort increase results from only an average extra 21% compensation (an average per worker piece-rate expenditure of $5.1 per two-hour shift in addition to the base wage of $24).

Disaggregated statistics per campus. The finding that the piece rate elicited more effort but the fixed wage raises did not holds not only overall but also across campuses. Table 5, Panel A, column (1), shows that the average number of characters inputted by the Control at campus A, an elite university, was 21,382. Fixed wage raises, however, did not increase effort, which was lower by -5% and -9% than in the Control in the 67%Raise and 50%-100%Raise, respectively (column (2)). However, the piece rate increased effort by 26%. This differential response to wage raises and the piece rate is similar for campus B, a local university. Students in the Control inputted 13,240 characters, displaying lower baseline productivity than those at campus A (Panel B, column (1)). Though the baseline performance for students at the state school was lower than that of those at the elite school, a reasonable outcome, all responded in the same differential way to fixed raises and the piece rate: the 67% and 50%-100% raises had a mild and mixed impact on effort (+2% and -2% over the Control, respectively) whereas the piece rate raised effort by 10% (column (2)).[21]

Disaggregated statistics per shift. The piece rate not only elicited more effort than non-performance-contingent wage raises across campuses, but also within shifts. Figure 1 shows that the piece rate increased the number of characters inputted vis-à-vis the Control in every shift, despite the much smaller average additional expenditure per worker-shift of 10%, 24% and 30% in shifts one, two and three, respectively (that is, an average per worker piece-rate expenditure of $2.4, $5.8 and $7.3 in shifts one, two and three, respectively, in addition to the $24 per shift base pay). In contrast, fixed wage raises of 50%, 67% , and 100% over the $24 per shift base pay, in general, did not.

[21]See Appendix Table A.2 for additional and more detailed summary statistics for the whole sample and per campus.

Although the pattern of no response to fixed wage raises, in contrast to piece rates, holds across campus and shifts, it stems from raw average differences across conditions. Given that we randomly assigned workers to the several conditions within campus and leg, it is useful to control for unobserved time-invariant factors at each campus and leg that might affect performance. Further, it is also important to analyze how effort progresses over time: whether effort increases from one shift to the next as workers rest and have an opportunity to reflect on how to improve their productivity for the following shift, and whether the additional fixed raise in the third shift to 100% in the 50%-100%Raise elicited extra effort. We now describe this analysis.

4.3 Empirical Method and Results

This section documents that after controlling for unobserved campus, leg and shift time-invariant heterogeneity, we still observe that fixed wage raises did not increase effort over time whereas the piece rate did, by up to 19%.

Empirical method I: Regression. To compare our elasticities with those in previous research, we used the natural log of characters as our dependent variable. Also, because within each campus and within each leg, we randomized workers to each condition (e.g., we randomly assigned workers in campus A to each condition in leg 4, and similarly randomly assigned workers in campus B to each condition in leg 4), we estimate differences between the conditions within each leg, campus, and shift and then pool them. Thus, we estimate the natural log of the characters entered by a subject i, in t_1 (Control), t_2 (67%Raise), t_3 (50%-100%Raise), t_4 (PieceRate) in campus c, leg l, and shift s as follows:

$$\ln(\text{characters})_{i,t,s,c,l} = \alpha_{1,1} + \alpha_{1,2} t_1 s_2 + \alpha_{1,3} t_1 s_3 + \sum_{\tau=2}^{4} \sum_{j=1}^{3} \beta_{\tau,j} t_\tau s_j + \psi_c \times \psi_l \times \psi_s + \epsilon_{i,s,t,c,l} \quad (1)$$

The interaction of campus, leg and shift fixed effects ($\psi_c \times \psi_l \times \psi_s$) captures unobservable time-invariant, campus, leg, and shift determinants of outcomes. Campus fixed effects control, for example, for unobserved propensities to respond differently to incentives between campuses, which could affect the difference between the treatments and the Control within each campus. For example, students at one campus may be more productive (as is the case with campus A) or may have more taste for reciprocity or more ability to increase effort than students at the other. Leg fixed effects control, for example, for unobserved different conditions across legs (e.g., students may be more tired in a leg occurring closer to final exams than in another occurring early in the term), which could influence both their baseline performance and how they respond to incentives within each

leg. Shift fixed effects control, for example, for unobserved learning across shifts.

The interaction of campus, leg, and shift conservatively addresses whether these unobservables could occur differentially within each leg, campus, and shift. For example, tiredness close to final exams at a given campus and leg could undermine both effort at a shift and learning across shifts.

The causal parameters of interest are the $\beta_{t,s}$ on the interaction of the treatment and shift dummy variables. They pool the percentage differences in characters between the treatments and the CONTROL within a campus, leg, and shift. For example, $\beta_{2,1}$ identifies the percentage difference in characters between t_2 (67%RAISE) and the CONTROL for shift one by pooling all these differences within each campus and leg. The parameter $\alpha_{1,1}$ is the natural log of characters for the baseline category, the CONTROL in shift one, which is not separately identified from the fixed effects, as usual.

To account for serial correlation in worker effort across shifts (namely, in $\epsilon_{i,s,t,c,l}$), we cluster the standard errors by worker (Bertrand, Duflo, and Mullainathan (2004)).

Further, all tests comparing the treatments to the CONTROL are one-tailed following, for example, Gneezy and List (2006), the most influential gift-exchange field study. One-tailed tests are used because gift exchange makes the one-sided prediction of an increase in effort in response to raises. Further, the sample size was partially determined by our power calculations (Appendix B), which yielded a reasonable minimum size needed to reject the null in a one-tailed test. We also use one-tailed tests for the PIECERATE, so as to apply the same standard of statistical significance applied for the wage-raise treatments. Further, reasonable piece rates have been shown to not decrease effort.[22]

Empirical method II: Multiple Hypothesis Testing (MHT). We also run the procedure in List, Shaikh, and Xu (2016) designed to reduce the chance of false positives. It corrects the p-values of a single hypothesis test to account for researchers' testing of multiple hypotheses but with substantial gains in power versus classical methods, such as Bonferroni (1935). In the analyses that follow, we will show p-values both from the regression method above and from the MHT method.

Unadjusted estimates per shift. The results in Table 6 show, as above, that fixed wage raises did not increase effort; the evidence suggests, however, that piece rates not only increased effort, but accelerated it over time. During the first shift, the percentage difference in effort between the wage-raise treatments and the CONTROL was small at 2%, whereas it was 9% for the PIECERATE (column (1)). In the second shift, the difference between the wage-raise treatments and the CONTROL became negative at -3%, whereas

[22]Gneezy and Rustichini (2000) found that very small piece rates depressed effort whereas reasonable piece rates increased effort, as expected.

it increased to a conservative 12% in the PieceRate (column (2)). In the third shift, the difference between 67%Raise and the Control was still negative at -1%, whereas there was no difference between the 50%-100%Raise and the Control (column (3)). However, the difference between the PieceRate and the Control rose to a conservative 15%, hinting that the piece rate accelerated learning, whereas fixed wage raises did not.[23,24] These estimates are statistically insignificant under the regression method, however, due to the large standard errors induced by the model's low explanatory power (R^2 of 0.02), which did not include unobserved campus differences in productivity, for example (workers in campus A had higher productivity than those in campus B). We add these controls, as they reduce both the standard errors and the bias in the estimates.

These estimates are also statistically insignificant under the MHT method, which often yields larger p-values than our regression method (the square brackets show the p-values from the regression method followed by the p-values of the MHT method). Whether the p-values under the MHT method exceed those of the regression method depends, to an extent, on serial correlation of outcomes across sessions. The regression method incorporates serial correlation—thus boosting p-values and reducing the chance of false positives—but no correction for multiple hypothesis testing. The MHT method incorporates multiple hypothesis testing which boosts the p-values but no correction, at this time, for serial correlation.

Estimates per shift adjusted for campus fixed effects. Including campus fixed effects is critical given the previous evidence that workers at campus A tend to be more productive than those at campus B. Campus fixed effects ensures that differences between the treatments and the Control are estimated within campus (treatments vis-à-vis Control in campus A and treatments vis-à-vis Control in campus B) and pooled. Relying solely on the previous raw means to draw conclusions would be incorrect, as these raw means simply average effort across all workers irrespective of their campus membership and their distribution across the experimental conditions. Thus, we would be, for example, comparing the performance of the lower-productivity workers at campus B in the treatments with that of higher-productivity workers at campus A in the Control.

Columns (4)-(6) show that adding campus fixed effects both reduces the bias of the

[23]This is consistent with anecdotal evidence volunteered by subjects in the PieceRate treatment to the research assistant: They spent their time between the weekly shifts mulling over how to increase performance to earn higher earnings on the following shift.

[24]The percentage increases for the PieceRate of 12% and 15% in shifts two and three, respectively, are actually conservative lower bounds for the percentage increases for these shifts, which are $e^{(0.12)} - 1 = 0.13$ and $e^{(0.15)} - 1 = 0.16$, respectively. This is because though the natural log specification approximates well small percentage changes (thus the estimates for the 67%Raise and 50%-100%Raise correspond to actual estimated percentage changes, such as $e^{(0.02)} - 1 = 0.02$ and $e^{(-0.03)} - 1 = -0.03$), it underestimates percentage increases when these are large.

estimates and increases the fit of the model considerably (R^2 of 0.35), as campus membership explains a great deal of worker performance. The better fit of the model reduces the standard errors substantially for most estimates, which become more precise.

Yet, the pattern they convey is similar to that described above. In the first shift, workers' effort in the 67%Raise and 50-100%Raise was respectively 1% and -4% lower than that in the Control, though these estimates are not statistically significant (column (4)). In contrast, effort in the PieceRate was 12% higher than that of the Control in the first shift and statistically significant at the 10% level. In the second shift, those in the 67%Raise and 50%-100%Raise undersupplied effort by -6% and -9%, respectively, compared to those in the Control, though these magnitudes are, again, statistically insignificant (column (5)). In contrast, those in the PieceRate further increased effort vis-à-vis the Control by 14%, which is statistically significant at the 5% level. In the third shift, those in the 67%Raise and 50-100%Raise still supplied slightly less effort than those the Control at -3% and -6%, respectively, though these estimates are statistically insignificant (column (6)). In contrast, those in the PieceRate further boosted effort versus the Control by 19%, significant with a p-value slightly over 1% level.

The null results on the fixed-wage-raise conditions are still null under the MHT procedure. The PieceRate loses some statistical significance with a p-value of 4% in the third shift (versus slightly over 1% with regression).

Yet it could be that, beyond time-invariant campus unobservables, unobserved leg and shift differences biased the results. We thus add the full set of controls for time-invariant campus, leg, and shift unobservable heterogeneity, described in specification 1.

Estimates per shift adjusted for campus, leg and shift fixed effects. Including these factors increases the model fit only slightly (the R^2 increases to 0.38). This reduction in the error variance is not enough to offset the loss of power caused by the large increase in the number of fixed effects (from 2—one for each campus—to close to 40). This decreases the precision of some estimates, such as those for the PieceRate, as their standard errors increase by at least 12%, yet it favorably raises gift-exchange estimates, which reach a high of 4% in one shift. We now describe these magnitudes.

In the first two-hour shift, workers in the 67%Raise and 50%-100%Raise increased effort by a respective 4% (the best estimate for gift exchange) and 2% compared to those in the Control, though these estimates are statistically insignificant (column (7)). This translates into very small and statistically insignificant elasticities for the first two hours of the task—0.06 and 0.04, respectively—which are within the range of those in other fields tests in the literature (see Table 1). Workers in the PieceRate increased effort by 7% vis-à-vis the Control with an additional expenditure of 10% (an average per

worker piece-rate payment of $2.4 in shift one in addition to the $24 base pay). This point estimate of 7%, yielding an elasticity of 0.70, is larger than that for the wage-raise treatments, though not statistically significant.

In the second shift, however, workers in the 67%Raise and 50-100%Raise slowed down, exerting -2% and -4% less effort, respectively, than workers in the Control, though these magnitudes are not statistically significant (column (8)). This translates into statistically insignificant elasticities of -0.03 and -0.08, respectively, consistent with prior negative elasticities in this literature (see Table 1). In contrast, workers in the PieceRate further increased effort relative to the Control, to 10%, though this estimate is not statistically significant due to the larger standard error of 0.10—which is more than 25% higher than that for the wage-raise treatments—induced by the smaller sample and the larger number of fixed effects. Nonetheless, this increase in 10% in effort was associated with only a 24% increase in compensation (an average per worker piece-rate pay of $5.8 in shift two in addition to the $24 base pay), much smaller than the fixed wage raises, resulting in a pay-effort elasticity of 0.41.

In the third shift, those in the 67%Raise and 50%-100%Raise continued exerting less effort than those in the Control, inputting -4% and -3% fewer characters (elasticities of -0.06 and -0.03, respectively) though these magnitudes are statistically insignificant (column (9)). However, workers in the PieceRate further boosted effort relative to the Control to 18%, which is statistically significant at the 10% level (instead of at the 1.5% level, as in the 19% estimate with only campus fixed effects) despite the larger standard errors induced by the numerous campus, leg and shift fixed effects. This further increase in effort to 18% corresponded to a 30% increase in compensation (an average piece-rate payment of $7.3 per worker in addition to the $24 base pay) yielding a pay-effort elasticity of 0.60 (and of 0.63 for the 19% increase in effort in the previous specification). These 18% and 19% estimates are conservative lower bounds on the percentage increases, corresponding actually to 20% and 21% increases, respectively (see footnote 24). Importantly, these effort boosts result from an extra expenditure peaking at 30% of the contract wage, which was much smaller than that in the fixed wage raise conditions, indicating that the piece rate was more efficient at increasing effort. The MHT procedure yields similar conclusions.

The wage increase to 100% in the last shift of the 50%-100%Raise (the only case in our field test in which an individual worker's effort was observed pre- and post-raise) appears to have had little effect on effort: it seems to have only arrested the decline in effort, though the evidence is inconclusive. Table 6, column (9) shows that those in the 50%-100%Raise increased effort in the third shift by 1% (to -3% vis-à-vis the Control)

after having declined by 6% (from +2% to -4% vis-à-vis the CONTROL), in shifts one and two (columns (7) and (8), respectively). In contrast, those in the 67%RAISE, where their baseline effort was unknown, and thus each could refrain from raising effort without fear of being viewed as selfish, showed a steady decline in effort versus the CONTROL: from 4%, to -2% and further to -4% in shifts one, two and three, respectively (columns (7)-(9)). These estimates are, however, statistically insignificant.

4.4 Robustness: Assessing Performance via Output Quality and Quits

Worker behavior was also more consistent with a standard model on other performance measures, such as output quality and quits. Table A.3, columns (7)-(9), shows that the number of correct words inputted in the wage raise conditions was, in general, slightly *lower* than that in the CONTROL, though the difference was statistically insignificant. In contrast, the number of correct words inputted by subjects in the PIECERATE was higher than that in the CONTROL, but also statistically insignificant. Thus, those in the PIECERATE did not produce lower-quality output, a concern in a multitasking model (e.g., Holmström and Milgrom (1991); Baker (1992)), perhaps because the piece rate was not a strong enough incentive to shift effort from the unrewarded performance dimension (quality of records) to the rewarded dimension (number of records).

The results are similar under the MHT procedure, even when the the p-values are adjusted not only for the testing of multiple conditions (fixed-wage-raise and PIECERATE conditions) but also for the testing of multiple outcomes (natural log of characters entered, natural log of records entered and natural log of correct words).

Further, the piece rate seems to have also motivated workers to persist on the task, though this effect is only marginally significant at the 10% level. Of the starting 194 workers, 15 did not complete all three shifts, of which most (7) were in the CONTROL. The remaining 8 were split as follows: 2 in the 67%RAISE, 2 in the 50%-100%RAISE and 4 in the PIECERATE. A regression analysis of sample attrition in Appendix Section C shows that there was no statistically differential attrition between workers in the wage-raise conditions and those in the CONTROL within each campus and leg. In contrast, 11% fewer workers in the PIECERATE missed shifts two and three vis-à-vis the CONTROL, within each campus and leg, though this effect is only marginally significant. This persistence accords with anecdotal evidence that subjects viewed the kinked piece-rate scheme as a challenge, returning each week with the goal of surpassing prior performance and reaching the next kink in earnings.

5 Correlation Between Laboratory and Field Test Behavior

This section shows that workers who behaved more prosocially in SPD games did not behave particularly prosocially during the field test. The best gift-exchange estimates for this select subsample show statistically insignificant effort increases of 7%, -3%, and -3% in shifts one, two, and three, respectively, which are not substantially different from those in the overall sample.

These findings suggest that prosocial behavior in the laboratory, observed and scrutinized by others, did not translate into prosocial behavior in the field—an effort boost after the raise—where the between-subjects design enabled selfish actions without detection.

5.1 Main Results

SPD games participants and classification into prosocial types. Table 7, column (2) shows that a substantial portion of workers (71% of the 194 workers, or 138) participated in the SPD games after the conclusion of all the field experiment's waves. This high participation rate did not vary substantially across conditions: 70% of the workers from the wage-raise treatments participated, and 69% and 74% from the PieceRate and Control, respectively (column (3)).[25]

The two opposite extremes in prosociality concentrated the majority of respondents at 74%: 41% behaved as *Prosocial* and 33% as *Selfish* (Row 1, columns (7) and (9)). The *Prosocial* type adds the two types who behaved the most prosocially in these games: the 10 *altruists*, who always cooperated (CCC), and the 46 *conditional cooperators*, who always cooperated except when the first mover defected (CCD) (columns (4) and (5)). At the opposite extreme were the 45 *Selfish* players who always defected (DDD) (row (1), column (8)). This distribution of types lies within that of other laboratory tests. For instance, a review by Fehr and Fischbacher (2002, page C6) pointed out that 40% or more of responders in gift-exchange games, which are SPD games, behaved prosocially. And, for example, Cohn, Fehr, and Goette (2014) classified 35% of workers as selfish in the post-experiment game described in Section 2.

We now analyze whether the subsample of workers who behaved most prosocially in the games and were in the wage-raise conditions during the field experiment engaged in gift exchange by exerting higher effort than the Control.

Sample. Table 7, row (2), documents that of the 81 workers in the 67%Raise and 50-100%Raise participating in the SPD games, 34 (42%) behaved as *Prosocial*. We thus test whether this subset of 34 workers—who received the fixed wage raises, participated in the SPD games, and behaved most prosocially in them—exerted more effort than those in

[25]For a further breakdown of participation by wage raise treatment, see Appendix Table A.4.

the CONTROL. The idea is that the average effort increase vis-à-vis the CONTROL in the wage-raise treatments for this subset of workers should be higher than that previously documented for all workers in the wage-raise treatments in Table 6, columns (7)-(9), because the full sample in the wage-raise conditions contained both the most prosocial types and the remaining types with weaker prosocial preferences (e.g., as selfish), which may have diluted average effort responses to raises.

Empirical methods and results. We use specification 1 but restrict the workers in the wage-raise and piece-rate treatments to only the *Prosocial* ones. Also, we aggregate the *Prosocial* workers in the two wage-raise treatments into one group—the 34 discussed above—so as to have a large enough sample size from which to draw conclusions. Thus, instead of comparing the effort of all workers in the treatments to those in the CONTROL as in section 4.3, we compare only the effort of the most prosocial workers in the treatments to that in the CONTROL. As before, we control for time-invariant unobserved campus, leg, and shift heterogeneity. We also conduct this analysis using the MHT procedure.

Table 8, columns (4)-(6), shows that workers who behaved the most prosocially did not particularly increase effort in response to the wage raise during the field experiment. Their effort was 7% higher than the CONTROL's in shift one, though statistically insignificant (column (4)). Further, effort subsequently slowed down to -3% less than the CONTROL's in shifts two and three though these estimates are also statistically insignificant (columns (5) and (6)). These estimates' statistical insignificance, despite the large wage raises, and their similarity to those using the whole sample of workers in the wage-raise treatments (columns (1)-(3)), imply that this subset of workers were as non-responsive to wage raises as the whole sample.[26] The MHT procedure yields similar conclusions.

5.2 Robustness Checks

Did *Prosocial* workers face an effort ceiling? It could be that *Prosocial* workers in the games were those who faced an effort ceiling during the field experiment and this is why they failed to increase effort following the raise. We tested this hypothesis by assessing whether *Prosocial* workers in the PIECERATE were able to increase effort. Table 7, row (3), shows that of the 22 workers in the PIECERATE who participated in the SPD games (a 69% response rate), 10 (45%) behaved as *Prosocial*. Table 8, columns (4)-(6) shows that these workers increased effort substantially, by 15%-24% in comparison

[26]These results change little if we exclude the two *altruists* from the 47 workers in the CONTROL. These two workers could have exerted higher effort despite no wage raise as they cooperate no matter the action of the first mover (giving a raise or not), inflating effort in the CONTROL and thus deflating gift exchange. Their exclusion leads to slightly lower gift exchange estimates (a per shift decline of about 1%), instead of higher, as the two *altruists* exerted lower effort than the CONTROL's average. Thus the low elasticities above are nonetheless biased in favor of gift exchange.

to the CONTROL, an effect statistically significant at the 10% level, though this sample is too small to allow for conclusions (in fact under the MHT procedure these results lose statistical significance with p-values of 11%-15%). This pattern suggests, nonetheless, that an effort ceiling did not prevent *Prosocial* workers receiving the fixed wage raises from increasing effort.

Could the lack-of-effort response by *Prosocial* workers be due to lower-effort workers selecting disproportionately into the SPD games? It could also be that participants in the SPD games tended to be workers who exerted low effort during the field experiment. This could have caused the finding of no effort increases by *Prosocial* workers in the wage-raise treatments vis-à-vis the CONTROL, because average effort in the CONTROL could have been inflated by including both game participants and nonparticipants. Restricting the CONTROL to only the 34 workers who played the games, instead of the original 47, when estimating specification 1 allows for effort responses by *Prosocial* workers in the fixed wage-raise treatments to be higher than that of a potentially lower-effort CONTROL. Table 8, columns (7)-(9) documents, however, that the estimates are slightly smaller than those obtained previously, with effort in the wage-raise treatments ending at a statistically insignificant -4% relative to the CONTROL in the third shift. Thus, the lack of effort response by *Prosocial* workers in the wage raise conditions holds conditional of participation in the games. The MHT procedure yields similar results. For a detailed sample decomposition per shift and condition for the estimation in Table 8, see Appendix Table A.5.

6 Discussion and Conclusion

We identify several factors that could have led tests of gift exchange to yield evidence sometimes consistent with gift exchange and sometime consistent with a standard principal-agent model—agent disutility from being perceived as selfish, small samples, insufficient wage raises, an effort ceiling, fatigue, selection of high-productivity workers, reemployment concerns, and peer effects—and implement a field test jointly dealing with all these confounds. Further, we paired this field test with a laboratory one, to compare workers' behavior in the laboratory and in the field. We find that after dealing with all these confounds, our field test results are most consistent with a standard model: workers did not increase effort in response to fixed wage raises but did do so in response to a piece rate scheme. Further, the piece rate was more efficient at boosting effort. Last, workers who behaved prosocially in the laboratory did not do so in the field.

We dealt with several factors that could have curbed gift exchange and thus biased the evidence in favor a standard model. We had a large sample to enhance the chance

of detecting a gift exchange effort increase; we offered wage raises matching or exceeding others in other field tests; we verified that an effort ceiling was not curbing effort boosts; we gave subjects enough time in between shifts to rest and figure out how to boost effort; we avoided the selection of higher-productivity workers; and we dealt with peer effects. Despite this, and despite framing the wage raise as a voluntary kind action by the principal, the evidence is most consistent with a standard model. To our knowledge no other field test has dealt jointly with all these curbs to gift exchange (see Table 2 in Section 2).

Why is our field evidence more consistent with a standard model, despite dealing with all these factors, whereas some prior tests have yielded evidence consistent with gift exchange? The most likely explanation is that our test *jointly* dealt successfully with three important confounds to gift exchange in prior tests: small sample sizes, the disutility of being perceived as selfish, and reemployment concerns.

By using a large sample, we reduced the likelihood of finding large gift-exchange elasticities which arise by chance in small samples. For example, we find smaller elasticities, even in the very short run than those in the most cited gift-exchange test—Gneezy and List (2006)—which had a substantially smaller sample than ours. They found, across their two studies, that their gift-exchange elasticities are short-lived (a statistically significant 0.40 for the first 1.5 hours in one study and 0.72 for the first 3 hours in another study, as shown in row 1 of Table 1). These elasticities waned thereafter, potentially due to fatigue or habituation to the raise, leading them to conclude that gift exchange was not a powerful motivator in the real world. In our study, the pay-effort elasticities, even for the first two-hour shift are smaller—ranging from -0.08 to 0.06—and statistically insignificant. The finding that short-term elasticities in sizable samples are lower than those in smaller ones is also consistent with recent work by DellaVigna, List, Malmendier, and Rao (2016): they found, in a very large sample, that workers' increase in effort following a non-performance-contingent wage raise peaked at 3% for the first 20 minutes (an elasticity of about 0.15), decaying thereafter. Thus, in their large sample, they also find little evidence that workers reciprocate non-performance-contingent wage raises with higher effort. This finding is also consistent with that in Muralidharal, Pradhan, Ree, and Rogers (2016) who found that doubling the fixed wages of a large sample of teachers in Indonesia did not boost performance.[27]

Beyond having a larger sample, we also addressed selfish image concerns. In a review of laboratory studies on social preferences and their external validity for real-world situations, Levitt and List (2007) posit a model in which agents' actions result from a trade-off

[27]Hoffman and Lyons (2015) also show that higher wages do not boost performance for politicians.

between wealth and the moral costs of behaving selfishly. The model predicts that the higher the observability of agents' actions and the higher the scrutiny (the higher the probability they are noticed, if observable), the higher the moral cost for agents and thus the more they will behave prosocially. Further, the lower the stakes of the game (e.g., the lower the costs of exerting the prosocial action), the stronger the prosocial behavior, as it is cheaper to behave prosocially. They present evidence consistent with this view.

The facts that our between-subjects field test yielded evidence consistent with the standard model and that we find little correlation between prosocial behavior in the field and laboratory are consistent with this "moral costs" reasoning. The between-subjects test shrouded agents' selfish behavior, reducing the moral costs of behaving selfishly and thus facilitating behavior consistent with a standard model. The fact that agents who behaved prosocially in the laboratory SPD games did not behave as such during the field experiment is also consistent with this "moral costs" view. In our laboratory test, workers knew their actions could be observed and scrutinized (as is the case with typical laboratory tests), whereas these actions could not be observed in the field due to the between-subjects design. Thus, agents who behaved prosocially in the games might have done so to avoid the disutility from being viewed as selfish. This behavior might have been further enabled by the modest stakes in our game (following those in the literature) allowing the cost of the prosocial action (payoff losses from behaving prosocially) to be more easily outweighed by the benefits of not projecting a selfish image. Relatedly, the absence of substantial effort boosts in the last shift of the 50%-100%Raise, the only instance in the field test where moral costs could have boosted effort, might have been due to this cost-benefit trade-off not being favorable: the cost of the prosocial action (the cost of raising effort in the last shift of 50%-100%Raise) might have loomed as large or larger than the benefit of not projecting a selfish image, yielding little effort change.

Last, we minimized reemployment confounds. We reminded workers that ours was a one-time job and curbed any expectation that worker performance would be appraised for the benefit of future employers, minimizing market reputation incentives.

Why is it important to deal with confounds to gift exchange in order to assess which mechanism underpins effort boosts? For example, why should we care whether effort results from gift exchange or from moral costs? Because different theories pose significant implications for incentive design. If the mechanism through which higher wages boost effort is gift exchange, then there is little need for monitoring, for example: workers would boost effort, even in the absence of monitoring, due to their intrinsic desire to repay the wage. If the mechanism is moral costs instead, than workers would only boost effort if monitored, due to their disutility from being viewed as selfish. This is relevant

considering that firms devote significant resources to monitoring (e.g., Katz (1986)).

Yet, the fact that our field test results are consistent with a standard model does not mean that there is no scope for social preferences in labor markets, in particular towards an employer. Our workers' prosocial behavior in the laboratory and its lack of correlation with that in the field is consistent with workers caring about the image they project to others and therefore with social preferences. Also, in recent work, DellaVigna, List, Malmendier, and Rao (2016) document that workers have social preferences towards the employer: they find evidence consistent with a warm-glow model of social preferences (e.g., workers supplying effort due to norms in the workplace and/or because they value providing meaningful work).[28] However, like us, they also find little evidence that workers reciprocate non-performance-contingent wage raises with higher effort.

Also, our field evidence does not mean that there is no scope for gift exchange as a motivator: it is possible that it may flourish in settings not captured in this or prior tests. One such setting is one where both (i) workers interact with a firm for long periods (months or years) instead of minutes or hours as in prior tests, allowing them to develop affect for the firm that leads them to reciprocate higher wages, and (ii) firms screen for the types who will engage in gift exchange: those who will develop, when given enough time, affect for the firm and thus acquire utility for repaying higher wages. Esteves-Sorenson, Pohl, and Freitas (2015) study an environment with these two characteristics.[29] These and other environments are areas for future research.

References

Akerlof, G. (1982): "Labor Contracts as Partial Gift Exchange," *The Quarterly Journal of Economics*, pp. 543–569.

Akerlof, G., and J. Yellen (1990): "The Fair-Wage Hypothesis and Unemployment," *Quarterly Journal of Economics*, 105(2), 255–83.

Baker, G. (1992): "Incentive contracts and performance measurement," *Journal of Political Economy*, pp. 598–614.

Baran, N., P. Sapienza, and L. Zingales (2010): "Can We Infer Social Preferences from the Lab? Evidence from the Trust Game," *NBER Working Paper 15654*.

Bellemare, C., and B. Shearer (2009): "Gift Giving and Worker Productivity: Evidence from a Firm-Level Experiment," *Games and Economic Behavior*, 67(1), 233–244.

[28] See also, for example, Benjamin (2015a, 2015b) on other social preferences in labor markets.

[29] Firms also may use other tools to gather information to screen prospective workers, beyond temporary contracts, such as personality tests (see Englmaier, Kolaska, and Leider (2015)).

Benjamin, D. J. (2015a): "Distributional Preferences, Reciprocity-Like Behavior, and Efficiency in Bilateral Exchange.," *American Economic Journal: Microeconomics*, 7(1), 70–98.

——— (2015b): "A Theory of Fairness in Labour Markets," *Japanese Economic Review*, 66(2), 182–225.

Benz, M., and S. Meier (2008): "Do People Behave in Experiments as in the Field? Evidence from Donations," *Experimental Economics*, 11(3), 268–281.

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

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

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

Brandts, J., and G. Charness (2011): "The Strategy Versus the Direct-Response Method: a First Survey of Experimental Comparisons," *Experimental Economics*, 14(3), 375–398.

Brown, M., A. Falk, and E. Fehr (2004): "Relational Contracts and the Nature of Market Interactions," *Econometrica*, 72(3), 747–780.

Burks, S., J. Carpenter, and L. Goette (2009): "Performance Pay and Worker Cooperation: Evidence from an Artefactual Field Experiment," *Journal of Economic Behavior and Organization*, 70, 458–469.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, and et al. (2013): "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, 14(5), 365–376.

Carpenter, J., and E. Seki (2010): "Do Social Preferences Increase Productivity? Field Experimental Evidence From Fishermen in Toyama Bay," *Economic Inquiry*, 49(2), 612–630.

Charness, G. (2004): "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, 22(3), 665–688.

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

Clark, K., and M. Sefton (2001): "The Sequential Prisoner's Dilemma: Evidence on Reciprocation," *The Economic Journal*, 111(468), 51–68.

Cohen, J. (1988): *Statistical Power Analysis for The Behavioral Sciences*. Psychology Press, New York, NY, second edition edn.

Cohn, A., E. Fehr, and L. Goette (2014): "Fair Wages and Effort: Evidence from a Field Experiment," *Management Science*, Articles in Advance, 1–18.

Cohn, A., E. Fehr, B. Herrmann, and F. Schneider (2014): "Social Comparison and Effort Provision: Evidence from a Field Experiment," *Journal of the European Economic Association*, 12(4), 877–898.

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

Englmaier, F., T. Kolaska, and S. Leider (2015): "Reciprocity in Organisations - Evidence from the WERS," Working Paper.

Esteves-Sorenson, C., V. Pohl, and E. Freitas (2015): "Efficiency Wages and Its Mechanisms: Empirical Evidence," *Working Paper*.

Fehr, E., and U. Fischbacher (2002): "Why Social Preferences Matter-The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives," *Economic Journal*, pp. 1–33.

Fehr, E., and U. Fischbacher (2004): "Third-Party Punishment And Social Norms," *Evolution and Human Behavior*, 25(2), 63–87.

Fehr, E., E. Kirchler, A. Weichbold, and S. Gächter (1998): "When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets," *Journal of Labor Economics*, 16(2), 324–351.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *The Quarterly Journal of Economics*, 108(2), 437–459.

Fehr, E., G. Kirchsteiger, and A. Riedl (1998): "Gift Exchange and Reciprocity in Competitive Experimental Markets," *European Economic Review*, 42(1), 1–34.

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

Gibbons, R. (2005): "Incentives Between Firms (And Within)," *Management Science*, 51(1), 2 – 17.

Gibbons, R., and M. Waldman (1999): "Careers in Organizations: Theory and Evidence," *in "Handbook of Labor Economics." eds: Ashenfelter, O. and Card, David. (Elsevier)*, 3.

Gilchrist, D., M. Luca, and D. Malhotra (2015): "When 3+1>4: Gift Structure and Reciprocity in the Field," *Management Science*, Forthcoming.

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

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

Hennig-Schmidt, H., A. Sadrieh, and B. Rockenbach (2010): "In Search of Workers' Real Effort Reciprocity – A Field and a Laboratory Field Experiment," *Journal of the European Economic Association*, 8(4), 817–837.

Hoffman, M., and E. Lyons (2015): "Do Higher Salaries Lead to Higher Performance? Evidence from State Politicians," .

Holmström, B., and P. Milgrom (1991): "Multitask Principal-agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7(2), 24–52.

Karlan, D. (2005): "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions," *American Economic Review*, 95(5), 1688–1699.

Katz, L. (1986): "Efficiency Wage Theories: A Partial Evaluation," *NBER Macroeconomics Annual*, 1, 235–276.

Kube, S., M. Maréchal, and C. Puppe (2012): "The Currency of Reciprocity: Gift Exchange in the Workplace," *American Economic Review*, 102(4), 1644–1662.

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

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

Levitt, S., and J. List (2007): "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 21(2), 153–174.

List, J., A. Shaikh, and Y. Xu (2016): "Multiple Hypothesis Testing in Experimental Economics," *NBER Working Paper*.

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

Muralidharal, K., M. Pradhan, J. d. Ree, and H. Rogers (2016): "Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia," Working Paper.

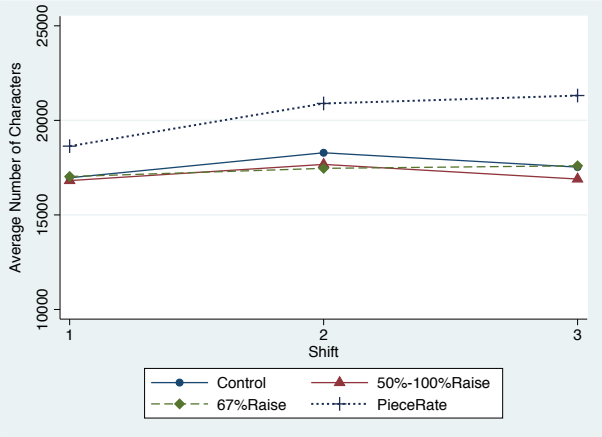
Schneider, F., and R. Weber (2012): "Long Term Commitment and Cooperation," *Unpublished Manuscript, Univeristy of Zurich*.

Shapiro, C., and J. Stiglitz (1984): "Equilibrium Unemployment as a Worker Discipline Device," *The American Economic Review*, 74(3), 433–444.

Weiss, A. (1980): "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, 88(3), 526–538.

Figures and Tables

Figure 1: Raw Average of Characters Inputted per Shift Across the Four Conditions



Note: PieceRate workers received an average per worker piece-rate payment of \$2.4, \$5.8 and \$7.3 in shifts one, two and three, respectively, in addition to the \$24 per shift base pay. This represents an additional average per worker compensation of 10%, 24% and 30% in shifts one, two and three, respectively, which is smaller than that arising from the wage-raise treatments: 67% in all shifts for the 67%Raise and 50% for shifts one and two, and 100% for shift three, in the 50%-100%Raise.

Table 4: Average Number of Characters Inputted per Condition and Worker Sample

	Number of characters inputted per subject			
	Average across all worker-shifts	Ratio relative to the Control	Total workers in shift one	Total number of worker-shifts
	(1)	(2)	(3)	(4)
Control	17,591	1.00	47	131
67%Raise	17,361	0.99	70	207
50%-100%Raise	17,164	0.98	45	111
PieceRate	20,301	1.15	32	90
			194	539

Notes: Column (1) documents the average number of characters entered across all the worker-shifts in the CONTROL and the treatments. For example, workers in the CONTROL inputted 17,591 characters across all their shifts, whereas those in the 67%RAISE inputted 17,361 characters. Column (2) depicts the ratio of the number of characters inputted versus the CONTROL. For example, those in the 67%RAISE treatment inputted $0.99 = 17,361/17,591$ characters relative to those in the CONTROL. Column (3) documents the number of recruited workers and therefore the number of workers in the first shift for each treatment. The sample for 50%-100%RAISE treatment though starting at 45 workers, only had 23 subjects in the third shift has only as subsample of 23 workers received the surprise wage raise to 100% of the market wage at the beginning of the third shift. Column (4) documents the number of worker-shift observations to account for missing shifts. For example, the 131 worker-shifts in the CONTROL result from 46 workers in shift one (the character-recording software did not record the characters for one worker), 45 in shift two (as two workers attrited in shift two) and 40 in shift three (as an additional 5 workers attrited in shift three). The 207 worker-observations in the 67%RAISE result from 70, 69 and 68 workers in shifts one, two and three, as one worker attrited in shift two and an additional worker attrited in shift three. The 111 worker-shift observations for the 50%-100%RAISE treatment results from 45 and 43 workers in shifts one and two, respectively (two attritted in shift two) and the additional subsample of 23 workers receiving the additional raise up 100%. The 90 worker-shift observations in the PIECERATE treatment result from 31 workers in shift one, 31 in shift two (as the character-recording software did not capture one worker's characters in shift one and another worker's characters in shift two) and 28 in shift three (as four workers attrited in shift three).

Table 5: Average Number of Characters Inputted per Condition per Campus, and Worker Sample per Campus

	Average across all worker shifts	Ratio relative to the Control	Total workers in shift one	Total number of worker-shifts
	(1)	(2)	(3)	(4)
Panel A: Campus A				
Control	21,382	1.00	26	70
67%Raise	20,217	0.95	40	119
50%-100%Raise	19,363	0.91	30	73
PieceRate	26,931	1.26	15	42
			111	304
Panel B: Campus B				
Control	13,240	1.00	21	61
67%Raise	13,498	1.02	30	88
50%-100%Raise	12,977	0.98	15	38
PieceRate	14,500	1.10	17	48
			83	235

Notes: Panel A, column (1) shows the average number of characters inputted across the four conditions for campus A. For example, workers in the CONTROL, in campus A entered an average of 21,382 characters in their shifts whereas those in the 67%RAISE entered 20,217. Column (2) shows the ratio of the number of characters inputted versus the CONTROL. For example, those in the 67%RAISE treatment in campus A inputted $0.95 = 20,217/21,382$ characters relative to those in the CONTROL in campus A. Column (3) documents the number of recruited workers and therefore the number of workers in the first shift for each treatment in campus A. For example, the CONTROL started with 26 workers in shift one. Column (4) represents the total number of worker shift-observations. A decomposition of the number of workers per shift for the whole sample and per campus is in Appendix Table A.2. Panel B shows the same information as Panel A, but for campus B.

Table 6: Characters Inputted Across the Different Treatments Relative to the CONTROL

	Dependent Variable: ln(Number of Characters per Subject)								
	Unadjusted			Within campus			Within campus X leg X shift		
	Shifts			Shifts			Shifts		
	One	Two	Three	One	Two	Three	One	Two	Three
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Difference vs. Control									
67%Raise	0.02	-0.03	-0.01	0.01	-0.06	-0.03	0.04	-0.02	-0.04
	(0.07)	(0.08)	(0.08)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)
	[0.39/0.47]	[0.67/0.56]	[0.56/0.51]	[0.44/0.44]	[0.82/0.82]	[0.70/0.70]	[0.26/0.43]	[0.62/0.59]	[0.71/0.57]
50%-100%Raise	0.02	-0.03	0.00	-0.04	-0.09	-0.06	0.02	-0.04	-0.03
	(0.08)	(0.09)	(0.08)	(0.06)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
	[0.39/0.39]	[0.63/0.62]	[0.50/0.50]	[0.71/0.60]	[0.89/0.81]	[0.79/0.68]	[0.37/0.49]	[0.70/0.53]	[0.67/0.62]
PieceRate	0.09	0.12	0.15	0.12	0.14	0.19	0.07	0.10	0.18
	(0.10)	(0.10)	(0.12)	(0.07)*	(0.08)**	(0.09)**	(0.09)	(0.10)	(0.12)*
	[0.17/0.32]	[0.12/0.26]	[0.10/0.22]	[0.06/0.14]	[0.03/0.08]	[0.01/0.04]	[0.22/0.38]	[0.17/0.30]	[0.08/0.10]
Constant	9.66	9.73	9.70	9.40	9.48	9.45	9.79	9.76	9.74
	(0.06)***	(0.06)***	(0.06)***	(0.06)***	(0.06)***	(0.06)***	(0.07)***	(0.03)***	(0.04)***
Campus fixed effects	-	-	-	Yes	Yes	Yes	-	-	-
CampusXlegXshift fixed effects	-	-	-	-	-	-	Yes	Yes	Yes
R-squared		0.02			0.35			0.38	
Number of workers	192	188	159	192	188	159	192	188	159
Number of workerXshift observations		539			539			539	

Notes: Columns (1)-(3) document the percentage difference between the raw means of the three treatments versus the CONTROL by using specification 1 without the fixed effects. The estimates for shifts two and three are also obtained using specification 1 without the fixed effects, but using the CONTROL in shifts two and three, respectively, as the baseline category, for ease of exposition. We also change the baseline this way for shifts two and three for the remaining analysis in this table. Columns (4)-(6) control for unobserved time-invariant campus heterogeneity by only having campus fixed effects in specification 1, instead of campus, leg and shift fixed effects. Columns (7)-(9) control for unobserved time-invariant determinants of performance within campus, leg and shift, with campus, leg and shift fixed effects, as outlined in specification 1. Standard errors from the regression are in parentheses and are clustered by worker (*Significant at the 10% level, **Significant at the 5% level, ***Significant at the 1% level). All tests are one-tailed. The square brackets contain the [p-values under regression/p-values under MHT]. The MHT procedure in columns (7)-(9) beyond correcting the p-values for testing multiple conditions (fixed-wage-raise and PIECERATE conditions) also corrects the p-values for testing multiple outcomes (ln(characters), ln(records) and ln(correct words)), so that these p-values are congruent with those of the robustness checks in Table A.3.

Table 7: Distribution of Prosocial Types Among Field Experiment Workers—Overall and by Condition

	Workers in Field Experiment	Workers who Responded to Survey		Breakdown Of Respondents by Prosocial Type					
				Prosocial				Selfish	
	Total	Total	Response Rate	Altruists	Conditional Cooperators	Total	Prop. of Respondents	Total	Prop. of Respondents
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) All workers	194	138	0.71	10	46	56	0.41	45	0.33
(2) Wage raise treatments (67%Raise and 50%-100%Raise)	115	81	0.70	4	30	34	0.42	28	0.35
(3) PieceRate	32	22	0.69	4	6	10	0.45	3	0.14
(4) Control	47	35	0.74	2	10	12	0.34	14	0.40

Notes: Column (1) documents the distribution of workers across the whole field experiment. Column (2) documents the number of workers who responded to the survey in total and by condition. We aggregate the 67%Raise and 50%-100%Raise condition into row (2) to garner a larger sample size of at least 30 subjects. For a detailed breakdown of prosocial types by these two conditions, please see appendix table A.4. Column (3) documents the response rates to the survey for the whole sample and by condition. It is thus the ratio of column (2) over column (1). Column (4) documents the number of workers who behaved as *altruists* in the SPD games for the whole sample and by condition. They cooperated as a first mover, cooperated as a second mover if the first mover cooperated, and cooperated as second mover even if the first player defected (CCC). Column (5) documents the number of workers who behaved as *conditional cooperators* in the SPD games for the whole sample and by condition. They cooperated as a first mover, cooperated as a second mover if the first mover cooperated, but did not cooperate as a second mover if the first player defected (CCD). Column (6) documents the number of workers who behaved as *Prosocial* for the whole sample and by condition. Column (7) documents the proportion of prosocial types among all the respondents, both for the whole sample and by condition. It is thus the ratio of column (6) over (2). Column (8) documents the number of workers who behaved as selfish players in the SPD games for the whole sample and by treatment. They defected as a first mover, defected as a second mover even if the first mover cooperated, and defected as a second mover even if the first player defected (DDD). Column (9) documents the proportion of *Selfish* types among all the respondents, both for the whole sample and by treatment. It is thus the ratio of column (8) over (2). The remaining workers were at neither end of this spectrum.

Table 8: Characters Inputted Across the Different Treatments Relative to the CONTROL by Prosocial Type—within Campus, Leg and Shift

Dependent Variable: ln(Characters inputted by subjects)									
				Prosocial in Treatments					
Sample:	Total sample			Prosocial Workers in Treatments Versus Full Control			Prosocial Workers in Treatments Versus Survey Respondents in Control		
	Shifts			Shifts			Shifts		
	One	Two	Three	One	Two	Three	One	Two	Three
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Difference vs. Control									
Wage Raise treatments	0.04	-0.03	-0.04	0.07	-0.03	-0.03	0.05	-0.04	-0.04
(67%Raise and	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.10)	(0.10)	(0.10)	(0.11)
50%-100%Raise)	[0.29/0.24]	[0.67/0.70]	[0.71/0.72]	[0.20/0.16]	[0.62/0.63]	[0.60/0.60]	[0.31/0.27]	[0.65/0.66]	[0.66/0.66]
PieceRate	0.07	0.10	0.18	0.15	0.19	0.24	0.11	0.15	0.22
	(0.09)	(0.10)	(0.12)*	(0.11)*	(0.14)*	(0.17)*	(0.12)	(0.16)	(0.19)
	[0.22/0.27]	[0.17/0.17]	[0.08/0.05]	[0.09/0.15]	[0.08/0.14]	[0.08/0.11]	[0.20/0.28]	[0.17/0.26]	[0.13/0.17]
Constant	9.79	9.76	9.74	9.73	9.76	9.64	9.76	9.78	9.67
	(0.07)***	(0.03)***	(0.04)***	(0.05)***	(0.05)***	(0.06)***	(0.06)***	(0.06)***	(0.07)***
CampusXlegXshift fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-Squared		0.38			0.44			0.40	
Number of subjectsXsession		539			250			218	

Notes: Columns (1)-(3) document the percentage difference between the combined wage-raise treatments (67%RAISE and 50%-100%RAISE) and the PIECERATE versus the CONTROL for the whole sample for each of the three shifts. Columns (4)-(6) document these percentage differences for the subsample of workers in the treatments who behaved most prosocially in the SPD games (*Prosocial* workers), relative to the whole CONTROL. Columns (7)-(9) document, as a robustness check, the same percentage differences for subsample of the *Prosocial* workers in the treatments relative the subsample of workers in the CONTROL who responded to the survey. Standard errors from the regression are in parentheses and clustered by worker (*Significant at the 10% level, **Significant at the 5% level, ***Significant at the 1% level). All tests are one-tailed. The square brackets contain the [p-values under regression/p-values under MHT].

A Additional Figures and Tables

Figure A.1: Timing of field experiment and post-field experiment survey

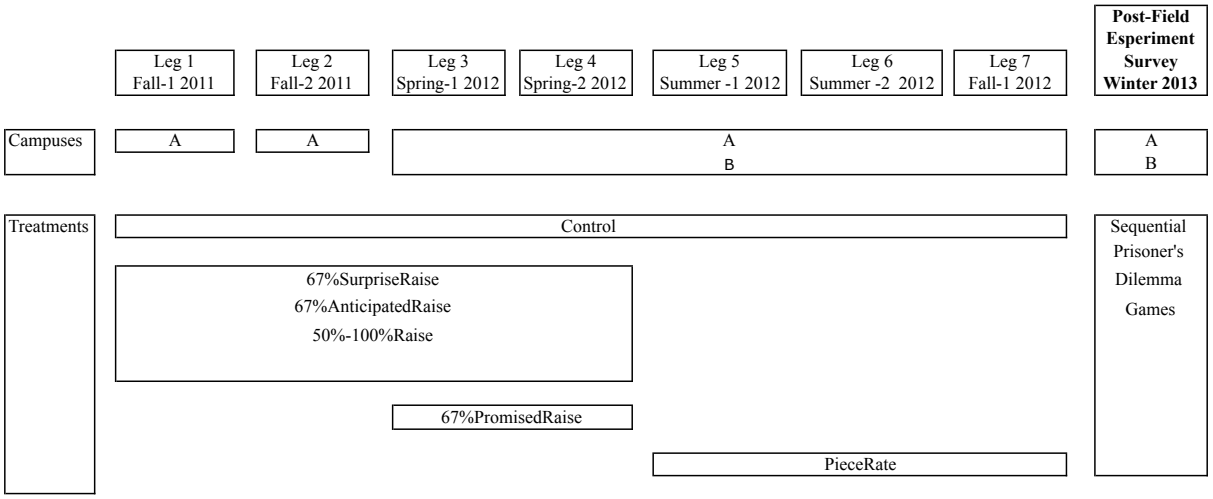


Table A.1: Kruskall-Wallis Test for Equality of Distributions in the 67%SurpriseRaise, 67%AnticipatedRaise and 67%PromisedRaise Treatments

	Shift One (p-value) (1)	Shift Two (p-value) (2)	Shift Three (p-value) (3)
Panel A: Campus A			
(1) 67%SurpriseRaise=67%AnticipatedRaise= =67%PromisedRaise	0.32 (N_1=12, N_2=18, N_3=10)	0.27 (N_1=12, N_2=18, N_3=10)	0.27 (N_1=12, N_2=17, N_3=10)
Panel B: Campus B			
Without the two high productivity outliers in the 67%AnticipatedRaise			
(2) 67%SurpriseRaise=67%AnticipatedRaise= =67%PromisedRaise	0.21 (N_1=11, N_2=5, N_3=12)	0.41 (N_1=11, N_2=5, N_3=11)	0.14 (N_1=11, N_2=5, N_3=11)
With the two high productivity outliers in the 67\%AnticipatedRaise			
(3) 67%SurpriseRaise=67%AnticipatedRaise= =67%PromisedRaise	0.10 (N_1=11, N_2=7, N_3=12)	0.17 (N_1=11, N_2=7, N_3=11)	0.04 (N_1=11, N_2=7, N_3=11)

Notes: This table presents the results of the Kruskall-Wallis test for whether the 67%SurpriseRaise, 67%AnticipatedRaise and 67%PromisedRaise samples were drawn from the same population (against the alternative that they were not), within a given campus and shift, per leg. Panel A, row (1) tests for whether the samples for the 67%SurpriseRaise, 67%AnticipatedRaise, 67%PromisedRaise in campus A, at each leg, were drawn from the same population (e.g., whether the 67%SurpriseRaise sample in leg 1 is drawn from the same population as that in legs 2, 3 and 4 and these samples are no different from the ones drawn for the 67%AnticipatedRaise and 67%PromisedRaise treatments in the same legs for campus A). The p-values of 0.32, 0.27 and 0.27 in columns (1), (2) and (3), respectively, indicate that we cannot reject that these samples were drawn from the same population. The samples sizes are below the p-values, where N_1, N_2 and N_3 are the sample sizes in the 67%SurpriseRaise, 67%AnticipatedRaise and 67%PromisedRaise, respectively, in campus A. Row (2) performs the same analysis for campus B, without its two very high-productivity outliers in the 67%AnticipatedRaise treatment (two workers inputted twice as many characters than the average worker across the three treatments in this campus). The p-values of 0.21, 0.41 and 0.14 in shifts one through three, documented in columns (1) through (3), respectively, show that we cannot reject that the samples were drawn from the same population. Given that we include these high-productivity outliers in the 67%Raise condition, as the effort increase in this condition is one of the focuses of the analysis, we show in Row (3), for completeness, how the test for equality of distributions changes when we include these two outliers. As expected, their inclusion changes the results of the test: we can reject the samples are drawn from the same distributions in shift one marginally at the 10% level and in shift three at the 4% level. The outcome measure for this test is the number of characters inputted, which more closely approximates effort, as we argue in Section 4.1.

Table A.2: Summary Statistics–Whole sample and Per Campus

	Average across all worker shifts	Shift One	Shift Two	Shift Three	SD	Min	Max
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Whole sample							
Control	17,591	16,965	18,286	17,528	6,917	6,298	37,722
N (worker-shifts)	131	46	45	40			
67%Raise	17,361	17,030	17,463	17,597	6,560	4,983	37,828
N (worker-shifts)	207	70	69	68			
50%-100%Raise	17,164	16,816	17,670	16,897	5,587	3,792	30,935
N (worker-shifts)	111	45	43	23			
PieceRate	20,301	18,797	20,893	21,312	8,841	6,455	42,200
N (worker-shifts)	90	31	31	28			
Panel B: Campus A							
Control	21,382	20,593	22,564	20,970	6,295	9,335	37,722
N (worker-shifts)	70	25	24	21			
67%Raise	20,217	19,542	20,344	20,780	6,340	7,800	37,828
N (worker-shifts)	119	40	40	39			
50%-100%Raise	19,343	18,789	20,444	18,397	5,141	11,475	30,935
N (worker-shifts)	73	30	28	15			
PieceRate	26,931	24,197	27,306	29,878	6,972	15,622	42,200
N (worker-shifts)	42	15	15	12			
Panel C: Campus B							
Control	13,240	12,644	13,398	13,723	4,700	6,298	26,183
N (worker-shifts)	61	21	21	19			
67%Raise	13,498	13,682	13,489	13,318	4,594	4,983	28,955
N (worker-shifts)	88	30	29	29			
50%-100%Raise	12,977	12,869	12,494	14,084	3,741	3,792	19,239
N (worker-shifts)	38	15	15	8			
PieceRate	14,500	13,734	14,880	14,887	5,655	6,455	28,526
N (worker-shifts)	48	16	16	16			

Notes: Panel A, column (1) shows the average characters inputted across by all workers in all shifts in each condition, for the whole sample, with the number of worker observations below. For example, workers in the CONTROL inputted an average of 17,591 characters across their total of 131 shifts. Column (2) shows the average number of characters inputted for shift one, with the number of workers per shift below. For example, the average characters inputted by the observed 46 workers in the CONTROL in the shift one was 16,965. Columns (3) and (4) depict the same information but for shifts two and three, respectively. Column (5) shows the standard deviation of characters across all worker-shifts per condition. For example, 6,917 was the standard deviation in the number of characters inputted by workers in the CONTROL across the 131 worker-shifts. Columns (6) and (7) represent the minimum and maximum of characters across all worker-shifts. For example, the 6,298 and 37,722 are the minimum and maximum number of characters entered in any shift across the 131 worker-shifts. The 131 worker-shifts in the CONTROL result from 46 workers in shift one (the character-recording software did not record the characters for one worker), 45 in shift two (as two workers attrited in shift two) and 40 in shift three (as an additional 5 workers attrited in shift three). The 207 worker-observations in the 67%RAISE result from 70, 69 and 68 workers in shifts one, two and three, as one worker attrited in shift two and an additional worker attrited in shift three. The 111 worker-shift observations for the 50%-100%RAISE treatment results from 45 and 43 workers in shifts one and two, respectively (two attritted in shift two) and the additional subsample of 23 workers receiving the additional raise up 100%. The 90 worker-shift observations in the PIECERATE treatment result from 31 workers in shift one, 31 in shift two (as the character-recording software did not capture one worker's characters in shift one and another worker's characters in shift two) and 28 in shift three, (as four workers attrited in shift three). Panels B and C depict the same information as above, but for campuses A and B, respectively.

Table A.3: Differences Between the Treatments and the CONTROL on Characters, Records and Correct Words Inputted

Dependent Variable	ln(charaters inputted per subject)			ln(records inputted per subject)			ln(correct words inputted per subject)		
	Within campus X leg X shift			Within campus X leg X shift			Within campus X leg X shift		
	Shifts			Shifts			Shifts		
	One	Two	Three	One	Two	Three	One	Two	Three
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Results using specification (1)									
Difference vs. Control									
67%Raise	0.04	-0.02	-0.04	0.02	-0.05	-0.08	0.06	0.00	-0.07
	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.09)	(0.09)	(0.09)
	[0.26/0.43]	[0.62/0.59]	[0.71/0.57]	[0.37/0.49]	[0.73/0.55]	[0.84/0.70]	[0.26/0.44]	[0.50/0.49]	[0.78/0.64]
50-100%Raise	0.02	-0.04	-0.03	0.00	-0.03	-0.04	-0.01	-0.04	-0.02
	(0.07)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)
	[0.37/0.49]	[0.70/0.53]	[0.67/0.62]	[0.49/0.49]	[0.65/0.55]	[0.68/0.64]	[0.56/0.53]	[0.65/0.52]	[0.58/0.60]
PieceRate	0.07	0.10	0.18	0.03	0.07	0.18	0.02	0.04	0.11
	(0.09)	(0.10)	(0.12)*	(0.10)	(0.11)	(0.13)*	(0.10)	(0.12)	(0.11)
	[0.22/0.38]	[0.17/0.30]	[0.08/0.10]	[0.38/0.48]	[0.27/0.43]	[0.08/0.11]	[0.44/0.50]	[0.36/0.48]	[0.15/0.27]
Constant	9.79	9.76	9.74	4.37	4.35	4.35	7.54	7.56	7.54
	(0.07)***	(0.03)***	(0.04)***	(0.04)***	(0.04)***	(0.04)***	(0.07)***	(0.04)***	(0.04)***
CampusXlegXshift fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared		0.38			0.48			0.40	
Number of subjects	192	188	159	194	189	159	175	168	154
Number of subjectXsession observations		539			542			497	
Panel B: Unadjusted Control Averages									
Raw Average number of inputted by workers in the Control	16,965	18,286	17,526	75	86	85	1,943	2,098	2,064

Notes: Panel A, columns (1)-(3) replicate the prior analysis in Table 6, columns (7)-(9), from specification 1. Columns (4)-(6) replicate this analysis but where the outcome variable is the natural log of the number of records inputted by a subject. This analysis yields similar though sometimes smaller and slightly less precise estimates than that with ln(characters) as the dependent variable, as records are a noisier measure of effort (two records may differ substantially in the amount of characters required to enter them, e.g., their titles' length may differ). The sample for the records analysis has three more observations than the one for characters (542 versus 539) because the software in three instances recorded the number of records entered but not the characters. Columns (7)-(9) replicate the analysis in specification 1, but where the outcome variable is the natural log of correct words inputted per subject. The number of observations for this measure is slightly smaller at 497 as these data was unavailable across all conditions in one of our seven legs. Panel B contains the raw average of the characters, records and correct words for reference. Standard errors from the regression are in parentheses and clustered by worker (*Significant at the 10% level, **Significant at the 5% level, ***Significant at the 1% level). All tests are one-tailed. The square brackets contain the [p-values under regression/p-values under MHT]. The MHT procedure in this table corrects the p-values not only for the test of multiple conditions (fixed-wage-raise and PIECERATE conditions) but also multiple outcomes (ln(characters), ln(records) and ln(correct words)).

Table A.4: Distribution of Prosocial Types Among Field Experiment Workers in Wage Raise Treatments

	Workers in Field Experiment	Workers who Responded to Survey		Breakdown Of Respondents by Prosocial Type					
				Prosocial				Selfish	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Wage Raise Treatments	Total	Total Respondents	Response Rate	Altruists	Conditional Cooperators	Total	Prop. of Respondents	Total	Prop. of Respondents
(1) 67% Raise	70	47	0.67	2	20	22	0.47	14	0.30
(2) 50%-100%Raise	45	34	0.76	2	9	11	0.32	14	0.41

Notes: Column (1) documents the distribution of workers across the two treatments. Column (2) documents the number of workers who responded to the survey in total and by treatment. Column (3) documents the response rates to the survey per treatment. It is thus the ratio of column (2) over column (1). Column (4) documents the number of workers who behaved as *altruists* in the SPD games, by always cooperating, by treatment. Column (5) documents the number of workers who behaved as *conditional cooperators* in the SPD games, by always cooperating except when the first mover defected, by treatment. Column (6) documents the number of workers who behaved as *Prosocial* by treatment. Column (7) documents the proportion of prosocial types among all the respondents, by treatment. It is thus the ratio of column (6) over (2). Column (8) documents the number of workers who behaved as selfish players in the SPD games, by always defecting, by treatment. Column (9) documents the proportion of *Selfish* types among all the respondents by treatment. It is thus the ratio of column (8) over (2). The remainder workers' were at neither end of these spectrum.

Table A.5: Sample Breakdown Per Condition and Shift for Table 8

				Prosocial in Treatments					
	Total sample			Prosocial Workers in Treatments Versus Full Control			Prosocial Workers in Treatments Versus Survey Respondents in Control		
	Shifts			Shifts			Shifts		
	One	Two	Three	One	Two	Three	One	Two	Three
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total Observations	539			250			218		
Control	46	45	40	46	45	40	34	34	30
Combined wage raise treatments	115	112	91	34	33	24	34	33	24
67%Raise	70	69	68	22	21	20	22	21	20
50%-100%Raise	45	43	23	12	12	4	12	12	4
PieceRate	31	31	28	10	9	10	10	9	10

Notes: Columns (1)-(3) document the sample breakdown per condition and shift for the analysis in columns (1)-(3) in Table 8. Columns (4)-(6) breakdown the sample for the analysis in columns (4)-(6) in Table 8. One worker observation was missing in shift one for the CONTROL resulting in 46 workers observations (instead of 47) as the character-recording software did not record the characters for one worker in the CONTROL. Similarly the 9 workers in the PIECERATE in the second shift, instead of 10, result from the character-recording software not recording the characters for one worker in this treatment in the second shift. The 4 observations for the 50%-100%RAISE treatment result from only 4 workers among the 23-worker subsample in this treatments in shift three behaving as *Prosocial*. Columns (7)-(9) breakdown the sample for the analysis in columns (7)-(9) in Table 8. The sample for the CONTROL is the sample of those who were in the CONTROL and responded to the survey.

B Online Appendix - Power Calculations

This section outlines our ex-ante power calculations ensuring that our sample would be large enough to reject the null of no gift exchange, at the 5% level, in favor of the one-sided alternative that a 67% fixed wage raise would elicit 20% extra effort. Based on the study most similar to ours—Gneezy and List (2006), which documented an increase in effort by 27% for the first 1.5 hours of their task, statistically significant at the 5% level in a one-tailed test—we show, through a simple calibration that we achieved 98% power to detect a 20% increase in effort in response to a 67% raise. Thus the reason we cannot document effort responses to fixed wage raises is not due to larger standard errors than those in this prior study—they were, in fact, smaller—but rather due smaller effect sizes ranging from -4% to 4%.

Nonetheless, because we worried that the short-term effect size of 27% in Gneezy and List (2006), on which we based our power calculations, could be potentially inflated due to the risk of overstatement in small samples, we also used the PieceRate condition to help us assess the power of our gift exchange test. As documented in the main text, the PieceRate detected statistically significant effort increases despite its having a smaller sample than wage-raise treatments, which did not detect any statistically significant effort increases across any of the specifications. Thus the absence of gift exchange is unlikely due to insufficient sample sizes. We now describe the power calculations.

B.1 Ex-Ante Power Calculations Based on the Most Similar Study

The goal was to estimate the sample size necessary to achieve at least 80% power to reject, at the 5% level, the null of no effort increase in favor of the one-sided alternative of an effort increase of 20% following a 67% fixed wage raise. We used the conventional methodology of selecting an effect size and variance estimates of the closest study to ours, which is the data-entry study in Gneezy and List (2006). Our study uses an analogous task to their book digitization task: entering data on academic articles (title, authors etc.). It also has a similar sample (U.S. undergraduates), going market wage ($12 per hour), and wage increase (of $8 per hour to $20 per hour).

We thus use the effect size and variances for the first 1.5 hours in Gneezy and List (2006) to calibrate the sample size required for each two-hour shift. They found a 27% productivity increase for the first 1.5 hours which was statistically significant at the 5% level in a one-sided test. We calibrate our sample conservatively to find a 20% effort increase per two-hour shift, which is statistically significant at the 5% level in a one-sided test. In order to convert the productivity increases into elasticities, we convert the productivity (number of records) data in their study to natural logs.

Mean and Variance Estimates. The parameters in the Gneezy and List (2006) data-entry study were as follows (where the subscripts t and c indicate the treatment and control groups, respectively):

- For the Control group: a sample size $N_c = 10$ and a sample variance $\sigma_c^2 = (0.23)^2$.
- For the Gift treatment (an \$8 hourly wage raise to \$20 per hour): a sample size $N_t = 9$ and a sample variance $\sigma_t^2 = (0.34)^2$.
- Alternative hypothesis. Let μ_s^a designate the population mean, where the superscript a indicates the alternative hypothesis for $s = t, c$, the treatment and control, respectively. Thus, the effect size—the increase in effort under the alternative hypothesis—corresponds to:[30]

  $$H_a: \quad \mu_t^a - \mu_c^a = 0.2$$

  where the alternative hypothesis is naturally one-sided following gift exchange's theoretical prediction that above-market wages increase, but never decrease effort. A two-sided alternative hypothesis would not only have been theoretically inaccurate, but would also have required different power calculations resulting in a larger ex-ante sample size.

Null Hypothesis. The null hypothesis of no effort responses to fixed wage raises corresponds to

$$H_0 : \mu_t - \mu_c = 0$$

Power Calculations. We can now compute the sample size to achieve at least 80% power to reject the null of no effort increase in favor of the one-sided alternative of a 20% increase, at the 5% level.

Given our large sample size, the estimator for the difference in population means $\bar{X}_t - \bar{X}_c$ is asymptotically normally distributed through a straightforward application of the Central Limit Theorem. Therefore, we compute:

[30] The effect size is defined as a specific non-zero value in the population for our alternative hypothesis. This effect size definition was, for example, popularized by Cohen (1988), page 10.

$$
\begin{aligned}
0.8 &= P(\text{reject } H_0 \text{at the 5\% level } |H_a \text{is true}) \\
&= P(\bar{X}_t - \bar{X}_c \text{ is in the rejection region } |H_a \text{ is true}) \\
&= P\left(\frac{\bar{X}_t - \bar{X}_c}{\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}} > 1.65 | H_a \text{ is true} \right) \\
&= P\left(\bar{X}_t - \bar{X}_c > 1.65 \sqrt{\frac{\hat{\sigma}_2^t}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}} | H_a \text{ is true} \right) \\
&= P\left(\frac{\bar{X}_t - \bar{X}_c - (\mu_t^a - \mu_c^a)}{\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}} > \frac{1.65\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}} - (\mu_t^a - \mu_c^a)}{\sqrt{\frac{\hat{\sigma}_2^t}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}} \right) \\
&= P\left(Z > \frac{1.65\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}} - (\mu_t^a - \mu_c^a)}{\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}} \right) \\
&= P\left(Z > \frac{1.65\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}} - 0.2}{\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}} \right)
\end{aligned}
$$

After plugging the parameters shown above, power depends on the sample sizes for the control and treatment groups according to the following equation :

$$
0.8 = P\left(Z > \frac{1.65\sqrt{\frac{0.23^2}{N_t} + \frac{0.34^2}{N_c}} - 0.2}{\sqrt{\frac{0.23^2}{N_t} + \frac{0.34^2}{N_c}}} \right)
$$

For example, a sample size of 26 subjects for the treatment and control groups ($N_c = N_t = 26$), would achieve approximately 80% power in rejecting the null that wage raises do not increase effort in favor of the one-sided alternative that they increase it by 20%.

In our specific case, we exceed the sample size of 26 for both our control and treatment groups. Our control contains 47 subjects whereas the 67%Raise treatment contains 70, achieving 98% power. The standard error in our 67%Raise, in any two-hour shift, is lower, at 0.07, as shown in Table 6, columns (7)-(9), compared to the benchmark standard error of 0.13 in Gneezy and List (2006), resulting from the samples of 10 and 9 subjects in the control and treatment groups, respectively.[31] This suggests that the lack of detection

[31]To compute this standard error we just plug 10 and 9 into N_c and N_t respectively, in $\sqrt{\frac{\hat{\sigma}_t^2}{N_t} + \frac{\hat{\sigma}_c^2}{N_c}}$.

of effort in response to fixed wage raises is no due larger standard errors in the estimates from the field experiment.

C Online Appendix - Attrition Analysis

This section documents the results of a linear regression model ascertaining whether attrition in the second or third shifts differs between the treatments and the CONTROL. We estimate the parameters in the specifications that follow using a linear probability model.

Empirical method. We estimate whether subject i, in t_1 (CONTROL), t_2 (67%RAISE), t_3 (50-100%Raise), t_4 (PIECERATE) in campus c, leg l, and shift s attritted as follows:

$$\text{attrit}_{i,t,s,c,l} = \theta_1 + \theta_2 t_2 + \theta_3 t_3 + \theta_4 t_4 + \psi_c \times \psi_l \times \psi_s + \epsilon_{i,s,t,c,l} \tag{2}$$

The variable *attrit* is binary, taking the value one if the worker attrited in a given shift and zero otherwise. The interaction $\psi_c \times \psi_l \times \psi_s$, controls, as discussed previously in the context of specification 1, for unobserved time-invariant campus, leg and shift determinants of attrition. These unobservables can affect the difference in attrition between the treatments and the CONTROL within a given campus, leg and shift. For example, a given leg, in a given campus, shift three may occur closer to finals, leading to less attrition in the wage raise or piece rate treatments, where subjects receive additional payments beyond the contract wage, and the CONTROL, where they do not. Further, this interaction allows us to estimate differences between the treatments and the control, within campus, leg and shift, in line with the previously described random assignment scheme, and then to pool them.

The causal parameters of interest are the θ_i. They pool the percentage differences in attrition between the treatments and the CONTROL within a campus, leg and shift, for shifts two and three. For example, θ_2 identifies percentage difference in attrition between treatment two (67%RAISE) in shifts two and three and the CONTROL in shifts two and three, by pooling all these differences within each campus, leg and shift. The parameter θ_1 estimates the average percentage of attriters the baseline category—the outcome for the CONTROL in both shifts two and three—which cannot be separately identified from the fixed effects, as usual.

Due to serial correlation in the attrition of each worker across shifts (serial correlation in $\epsilon_{i,s,t,c,l}$), we cluster the standard errors at the subject level (Bertrand, Duflo, and Mullainathan (2004)).

Results. Table 6, column (1) documents the rate of attrition without controlling for unobserved time-invariant campus, leg and shift factors. It thus transforms into percent-

ages the attrition rates in Table A.2, with the summary statistics for the whole sample and per campus, in columns (2)-(4). For example, the estimate of 10% for the constant in Table 6, column (1), row (4), documents the baseline attrition for the Control. It corresponds exactly to attrition documented Table A.2, columns (2)-(4) for the Control: of the potential 94 worker-shift observations across shifts two and three (from the 47 workers in shift one, we should have had 47 workers in shift two and 47 in shift three), 2 workers missed the second and third shifts (4 missed shifts) and an additional 5 workers missed the third shift (and additional 5 missed shifts). These 9 missed shifts represent the documented rate of attrition of 10% (9/94).

Column (1), rows (1), (2) and (3) document that the unadjusted differences in attrition between the treatments and the Control are not statistically significant, except for the 67%Raise treatment. Attrition in the 67%Raise is 7% smaller than that in the Control, though only marginally significant at the 10% level. As before, this 7% estimate also accords with the attrition documented in Table A.2, columns (2)-(4). Of the 140 potential worker-shifts in the 67%Raise in the second and third shifts (from the 70 workers in shift one, we should have had 70 workers in shifts two and 70 in shift three), one worker missed the second and third shift and an additional worker missed the third shift, resulting in 3 missed shifts. This amounts to an attrition rate of 2.1% (3/140), which is approximately 7% lower than that in the Control, as documented. However, none of these estimates control for time-invariant campus, leg an shift determinants of attrition.

Column (2), with estimates including campus, leg and shift fixed effects, documents that there are no statistically significant differences in attrition between the treatment and the Control, at the 5% level, when correctly conducting the analysis within campus and leg, following the random assignment scheme. Namely, attrition in the 67%Raise becomes positive but statistically insignificant and attrition in the PieceRate though negative, is only marginally significant.

The reason the lower attrition in the 67%Raise becomes positive and not statistically significant with the campus, leg and shift fixed effects, when it was negative and marginally significant when comparing the raw means in column (1), is that using campus, leg and shift fixed effects answers the question "When the 67%Raise treatment was run within a given campus and leg, was the attrition in shifts two and three in this condition higher or lower than that in the Control, in that same campus, leg and shifts?". In contrast, the raw means estimation in column (1) compares attrition in the 67%Raise to that in the Control, which includes students from other campuses and legs where the 67%Raise was not run (e.g., legs 1 and 2 in campus B and legs 5 through 7 on both

campuses A and B, as in Figure A.1.). Thus, for the legs and campuses where we ran the 67%Raise condition, attrition was 3% higher than that in the Control in those same campuses and legs, though not statistically significant.

In contrast, in the campus and legs where we ran the PieceRate, attrition was 11% lower than that in the Control in those campuses and legs, though only marginally significant.

Table 6: Differential Attrition Between the CONTROL and the Treatments

Dependent variable: =1 if Subject Attrited in Shift Two or Three		
Sample:	Full Worker Sample	
	Unadjusted	Within campusXlegXshift
	(1)	(2)
Diff. vs. Control group		
(1) 67%Raise	-0.07	0.03
	(0.04)*	(0.02)
(2) 50%-100%Raise	-0.07	0.03
	(0.04)	(0.02)
(3) PieceRate	-0.02	-0.11
	(0.05)	(0.07)*
(4) Constant	0.10	(0.06)
	(0.04)***	(0.02)***
Campus X leg X shift fixed effects	-	Yes
R-squared	0.01	0.18
Number of subjectsXsession observations	366	366

Notes: Number of observations: 194 observations in shift two and 172 observations in shift three, totaling 366 worker-shift observations. The sample declines by 22 observations from shift two to three (from 194 to 172) as we only implemented the 100% wage raise on the subsample of 23 workers in the 50%-100%RAISE treatment for the third shift, instead of the full sample of 45 workers, which naturally reduced the sample size by 22 workers. Standard errors clustered by individual. **Significant at the 5% level, *Significant at the 10% level. All tests are two-tailed, as we did not have a specific hypothesis for whether workers should attrit more or less in the treatments than in the CONTROL.

D Sources for the Overview of the Evidence on Gift Exchange in the Workplace

This Appendix presents the sources and computations for Tables 1 and 2. All studies considered use changes in effort in response to different wage raises. To compare effort responses across studies, we calculate the associated elasticities by dividing the % wage variation by the % effort increase. Effort responses are significant estimates at the 5% level and using two-tailed tests, unless otherwise stated.

PANEL I: Field Studies

(1) **Gneezy and List (2006)**

Gneezy and List (2006) contains two task. In the first, students were hired for \$12 per hour for a one-time job of digitizing the holdings of a library for six hours and randomly assigned them to two groups. The 10 in the control received the agreed wage whereas the 9 in the treatment received a surprise 67% raise to \$20 per hour. Output was the number of records inputted (e.g., the author's name, title of book etc.). In the second, 23 students were hired at \$10 per hour to raise funds for a charity, 13 (treatment) received a 100% raise to \$20 per hour and 10 (control) did not. Those in the treatment raised 72% more funds in the first 3 hours than those in the control (Table 1, Panel I, row (2)). However, effort waned thereafter, translating into an average increase of 38% over the 6 hours, which was significant at the 10% level in a one-tailed test. The authors tested if this waning could be due to fatigue by inviting workers to raise more funds the next day, after resting. However, only 4 and 9 subjects in the control and treatment, respectively, returned the next day, yielding low power for detecting a difference.

1.1) *Sample sizes* for the "Gift" treatment and the control ("noGift") are in Table I, column participant number, page 1371 (data entry task) and in Table V, column participant number, page 1376 (fundraising task).

1.2) *Wage increases*

- *data entry task.* Raise from \$12 to \$20 dollars in Section 2.A, second paragraph in page 1368. It represents a 8*100/12=67% increase.
- *Fundraising task.* Raise from \$10 to \$20 dollars in Section 2.B, first paragraph in page 1370. It represents a 10*100/10=100% increase.

1.3) *Effort responses* to wage increases and their significance levels for each task were calculated as follows.

- *data entry task.* Productivity is measured by the number of books logged. Productivity differences between the "Gift" treatment and the control for the 90, 180, 270 and 360 minutes are presented in Table I, page 1371,
  - The overall effort response over the six hours is reported in Table I, column 360 minutes, row Average, page 1371. The difference between the average number of records imputed by the 9 subjects in the "Gift" treatment over the 6 hours interval (40.3 records) and the same number for the 10 subjects in the control (39.6 records) leads to a difference of 40.3-39.6=0.7 records. This represents a (0.7*100)/39.6=2% increase, which is not statistically significant using a one-sided Wilcoxon test (see page 1372).
  - The effort response for the first 90-minute interval is reported in Table I, column 90 minutes, row Average, page 1371. The difference between the "Gift" treatment and the control corresponds to 51.7-40.7=11, which represents an statistically significant increase at the 5% level of (11*100)/40.7=27% using a one-sided Wilcoxon test and a t-test (see pages 1370-1372).
  - The effort response for the first three hours is reported in Table I, column 180 minutes, row Average, page 1371. The difference between the average number of records imputed by the 9 subjects in the "Gift" treatment over the first three hours (44.9 records) and the same number for the 10 subjects in the control (40.5 records) leads to a difference of 44.9-40.5=4.4 records. This represents a (4.4*100)/40.5=11% increase, which is not statistically significant using a one-sided Wilcoxon test (see second paragraph in page 1372).

1.4) *Elasticities.* The overall elasticity for the whole six hours corresponds to 2%/67%=0.03. For the first 90 minutes it corresponds to 27%/67%=0.40. For the first three hours the elasticity corresponds to 11%/67%=0.16.

- *Fundraising task.* Productivity is measured by the earnings raised. Productivity differences between the "Gift" treatment and the control for the overall six hours and by three-hour intervals are displayed in Table III, page 1374 with their significance levels. Averages by 90-minute intervals are not reported.
  - The overall effect corresponds to the difference between the average earnings by the 13 subjects in the "Gift" treatment over the 6 hours interval ($9.013) and the same number for the 10 subjects in the control ($6.516 dollars) (see Table III, column Gift and NoGift, respectively, row Entire day per hour). The difference 9.013-6.516=2.496 represents

a (2.496*100)/6.516=38% increase, which is statistically significant at the 10% level using a one-sided Wilcoxon test (see Table III, column Difference, row Entire day per hour).

- The effect for the first three-hour window corresponds to the difference between the earnings by the 13 subjects in the "Gift" treatment over the first three hours ($11.00) and the same number for the 10 subjects in the control ($6.40) (see Table III, column Gift and NoGift, respectively, row Pre Lunch per hour). The difference 11.00-6.40=4.6 represents a (4.6*100)/6.40=72% increase, which is statistically significant using a one-sided Wilcoxon test (see Table III, column Difference, row Pre Lunch per hour). For the second-three hours the difference between treatment and control of 7.026-6.633=0.392 represents a (0.392*100)/6.633=6% increase, which is not statistically significant using a one-sided Wilcoxon test (see Table III, column Difference, row Post Lunch per hour).

1.4) *Elasticities*. The overall elasticity corresponds to 38%/100%=0.38; for the first and second hours it corresponds to 72%/100%=0.72 and 6%/100%=0.06, respectively.

(2) **Bellemare and Shearer (2009)**

In a seven-day field experiment, 18 tree planters received a one-time lump-sum of $80 on the second day they planted. This amounted to a 37% raise over the average daily earnings per worker of $215 and increased effort by 11%-14% (Table 1, Panel I, row (3)).

2.1) *Sample size* of 18 workers is in first paragraph of Section 3, page 253.

2.2) *Wage increase* of $80 dollars in the second day of work (in addition to the $0.20 piece rate for all seven days) is described in Section 3, second paragraph in page 235. The total $215 average daily earnings using the piece rate is described in Section 4, page 236, end of the last paragraph. The gift thus corresponds to an average 80*100/215=37% increase.

2.3) *Effort responses* to the wage increase measured by the increase in the daily average number of trees planted, are presented in Table 2, page 238 with their significance levels. Effort responses change according to whether only data on the experimental block is considered (block for which the workers received the gift) or if productivity of the same workers in neighboring blocks is also included (Table 2, columns I and II, respectively). Fixed effects by planter and block are used in both cases.

* Table 1, column I shows the estimates using daily productivity for the seven-day window considering only productivity in the experimental block. Workers increase productivity by 118 trees on average after receiving the gift, which is statistically at the 1% level. Given that the average baseline effort of all workers pre-gift is unknown (pre-gift effort and worker fixed effects cannot be separately identified), we use the average productivity for the experimental block of 1075.59 tree as a proxy for pre-gift average worker effort when computing the percentage increase in productivity with the gift. Given that raw average of 1075.59 trees incorporates productivity both with and without the gift, the increase in 118 trees lead to an estimate of the lower bound on the average productivity increase, which is 118.31*100/1075.59=11% increase. On the other hand, the upper bound on the increase in productivity arising from the 118 tree increase is 118/(1075.59-118.31)=12%
* Table 1, column II, shows the estimates using daily productivity for the seven-day window considering productivity in the experimental and non-experimental blocks. Workers increased productivity by 132.271 trees on average after receiving the gift, which is statistically at the 1% level. Again, we use the average productivity for the experimental block of 1075.59 tree as a proxy for pre-gift average worker effort when computing the percentage increase in productivity. The average of 1075.59 trees not only incorporates productivity both with and without the gift but it is also higher than that in the non-experimental blocks 971.55 trees. Thus the increase of 132 trees is an estimate of the lower bound on the average productivity increase, which is 132.27*100/1075.59=13% increase. On the other hand, using the experimental block as the baseline, the upper bound on the increase in productivity arising from the 132.27 tree increase is 132.27/(1075.59-132.27)=14%.

2.4) *Elasticities.* The resulting elasticities range from to 11%/37%=0.30 to 14%/37%=0.38.

(3) **Hennig-Schmidt, Sadrieh, and Rockenbach (2010)**

They hired 103 students to transcribe abstracts for 20 DM per hour, for two one-hour sessions, each one month apart. Workers received the agreed hourly wage in the first session, whereas in the second session a random subsample of 23 received a surprise raise of 40% and information that peers were only receiving a 10% raise.In this paper. The increase in productivity from the first to the second hour was 28% *lower* than that for the 24 subjects in the control (Table 1, Panel I, row (5))

3.1) *Sample sizes* for the control ("F0"), "F10" and "F40 peer" treatments of 24, 25 and 23 subjects respectively, are in Table 1, columns F0, F10 and F40 peer, respectively, row Number of typist, in page 821.

3.2) *Wage increases* of $2 Deutsche Marks (DM) for the "F10" treatment and of $8 DM for the "F40 peer" treatment above the baseline of $20 DM per hour are in Table 1, columns F0, F10 and F40 peer, respectively, row Wage 2nd hour in page 821. Because the wage raise only applies to the second hour, the percentage wage raise corresponds to 2*100/20=10% for the "F10" treatment and 8*100/20=40% for the "F40 peer" treatment. In the "F40 peer" treatment, the wage raise is accompanied by information about the wage raise of the "F10" treatment.

3.3) *Effort responses* to the wage increase, measured as the number of correctly typed words per minute, are presented in Table 2, page 822. The change from period one to two in the number of correct imputed words per minute in the "F10" treatment corresponds to 0.152, while the control increased by 0.634 (see Table 2 column F10 and F0, row 2nd minus 1st hour). The difference 0.152-0.634=-0.48, which corresponds to a -0.48*100/0.634=-76% increase is not statistically significant (see Table A.1, column F0 vs. F10, row Output ratio-usable difference in Appendix A, page 832). The change from period one to two in the number of correct imputed words per minute in the "F40 peer" treatment corresponds to 0.459 (see Table 2, column F40 peer, row 2nd minus 1st hour). The difference with the control 0.459-0.634=-0.18 corresponds to a -0.18*100/0.634=-28% increase, which is not significant (formal test is not reported in Table A.1; significance only reported verbally in Section 2.3, page 823, last paragraph).

3.4) *Elasticities.* The elasticity in the "F10" treatment corresponds to -76%10%=-7.6. The elasticity for the "F40peer" corresponds to -28%40%=-0.69.

(4) **Kube, Maréchal, and Puppe (2012)**

They hired students for the one-time job of digitizing library holdings for three hours for 12 euros per hour. The 35 in the control received the agreed wage whereas the 34 in the treatment received a 19% raise.

4.1) *Sample size* of 34 student workers in the "Money" treatment and 35 in the control ("Baseline") stated at the end of Section I, page 1648, second paragraph.

4.2) *Wage increase* of a total of 7 euros from the 12 euros per-hour baseline in Section I, page 1646, first and fourth paragraphs, respectively. See also Appendix

Table A4 in page 1659. Because this is a three-hour task, the gift corresponds to a 7*100/(3*12)=19% increase.

4.3) *Effort responses* to the wage increase, measured by the number of charters entered, are presented in Table I, page 1649 in percentages and with their significance levels. Subjects in the "Money" treatment increased their productivity by 5.2%, which is not significant (see Table I, column baseline, row Money; In Table 1 we round this estimate to 5% to keep all effort estimates without decimals). For reference, productivity levels for the treatment and the control are presented in Appendix Table A2, page 1658. The average productivity of the control, which lumps the productivity of the "Baseline I" and "Baseline II" treatments, corresponds to 7,983.5+8,622.1=16,605.6 (see Table A2, column Characters, rows Average for Baseline I and Baseline II treatments). The average productivity of the "Money" treatment, which lumps the productivity of the "Money" and the "MoneyUpfront" treatments, correspond to 8,462.3+8,989.9=17,452.2 (see Table A2, column Characters, rows Average for Money and MoneyUpfront treatments).

4.4) *Elasticity* corresponds to 5%/19%=0.26.

(5) **Kube, Maréchal, and Puppe (2013)**

In a setup similar to Gneezy and List's (2006), students digitized library holdings for six hours, a one-time job for a projected 15 euros per hour. The 25 subjects in the control received the agreed pay, whereas the 22 in the treatment received a surprise 33% raise to 20 euros per hour.

5.1) *Sample size* of 22 student workers in the "PayRaise" treatment and 25 in the control ("Baseline") stated at the end of Section 2, page 858, third paragraph.

5.2) *Wage increase* of 5 euros per-hour from 15 euros per-hour baseline is at the end of Section 2, page 857, second paragraph. The gift corresponds thus to a 5*100/15=33% increase.

5.3) *Effort responses* to the wage increase, measured by the number of books entered, are presented in percentages in Table 1, page 859 by 90-minutes intervals and in the overall six hours with their significance levels.

* The overall effort response of subjects in the "PayRaise" treatment corresponds to a productivity decrease of by -0.3% (see Table 1, column PayRaise-Baseline, row All quarters), which is not significant (see Table 1, column p> $|z|$, row All quarters). For reference, productivity levels for the treatment and control are presented in the last paragraph of page 859. The average productivity of the control corresponds to 219.4 books en-

tered. The average productivity of the PayRaise treatment corresponds to 218.6 books entered.

* The effort response of the "PayRaise" treatment by 90-minutes intervals correspond to -9.5%, 1%, 0.2% and 6.5% (see Table 1, column PayRaise-Baseline, rows Quarter I, Quarter II, Quarter III and Quarter V, respectively; In Table 1 we round these estimates to -10%, 1%, 0.2% and 7% to keep all effort estimates above one without decimals). None is statistically significant (see Table 1, column p> $|z|$, rows Quarter I, Quarter II, Quarter III and Quarter IV, respectively).

5.4) *Elasticity* for the overall six hours corresponds to -0.3%/33%=-0.01. For each 90-minute intervals, the elasticities correspond to -10%/33%=-0.33, 1%/33%=0.03, 0.2%/33%=0.01 and 7%/33%=0.21 for the first, second, third and fourth intervals respectively.

(6) **Cohn, Fehr, and Goette (2014)**

In the within-subject test in Cohn, Fehr, and Goette (2014) 196 workers distributed newspapers in three-hour shifts for 22 CHF per hour where, unbeknown to them, their wages would alternate between 22 CHF and 27 CHF during the study's four weeks. Effort increased by 3%. They subsequently surveyed workers on the amount by which they had felt underpaid and had them play a laboratory game assessing their reciprocity. Of the 61% who answered, 65% were labeled as reciprocal and 35% as nonreciprocal. Only those who answered, felt underpaid, and were reciprocal, responded to the raise: they increased effort by 2.8% for each CHF of underpayment, yielding an upper bound of 14% for those who felt underpaid by 5 CHF. References to page numbers are omitted since only the online version of the paper is currently available.

6.1) *Sample size* of 196 workers of a promotion agency hired to distribute the newly launched newspaper of a publishing company is in the first paragraph in section 2.4.

6.2) *Wage increase* of 5 Swiss Francs (CHF) per hour from the 22 CHF per-hour baseline is in the first paragraph in section 2.2. The gift corresponds thus to a 5*100/22=23% increase.

6.3) *Effort responses* to the wage increase were measured by the hourly number of newspaper copies distributed.

* The effort response for the full sample of workers comes from Table 6 displaying the coefficient estimates of regressing the logarithm of hourly

number of copies distributed on a treatment dummy variable (1 if received a wage raise; 0 otherwise) plus location and day fixed effects. Table 6, column (1), row CHF27, shows that the parameter associated with the treatment dummy is 0.037, which is significant. Column (2) shows that when adding worker fixed effects this coefficient estimate corresponds to 0.030, which is also significant. Table 1 presents this estimate. Thus, the estimated increase in effort is 3%.

* The effort response for reciprocal workers who felt underpaid at the baseline wage is displayed in column (1) Table 10, where workers were classified as reciprocal and non-reciprocal using a Moonlighting game. The coefficient estimates of regressing the logarithm of hourly number of copies distributed on a treatment dummy variable (1 if received a wage raise; 0 otherwise) and the interaction between the treatment dummy and the difference between the wage a worker considered to be fair and the base wage correspond to 0.000 (not significant) and 0.028 (significant), respectively (See Table 10, column (1), rows Intercept, CHF27 and CHF27×Δ_i). Therefore, the total effort increase for reciprocal workers who felt underpaid by 5 CHF corresponds to 0.000+5*0.028=0.14 or 14%. For reference, neither the treatment dummy nor the interaction term is significant for the non-reciprocal workers.

6.4) *Elasticity* for the overall sample it corresponds to 3%/23%=0.13. The elasticity for the reciprocal subjects who felt 5 CHF underpaid at the base wage corresponds to 14%/23%=0.61%

(7) **Gilchrist, Luca, and Malhotra (2015)**

They hired a random sample of 168 workers (asking a $2-$3 hourly wage) at $3, for a four-hour transcription task to be completed within one week. They gave a subsample of 58 a surprise raise from $3 to $4 per hour and informed them that the job was expected to be one-time.[32] Only the online version of this paper is available so no pages are cited.

7.1) *Sample size* of 58 oDesk workers in the "Wage=3+1" treatment and 110 in "Wage=3" (the baseline category) are shown in Table 1, fourth and fifth columns, fourth row.

[32]This paper contains another treatment, in which workers select into a $4/hour contract. This is outside the scope of this review, which focuses on tests of gift exchange: whether workers reciprocate fixed wage raises with higher effort.

7.2) *Wage increase* of $1 dollar per-hour from baseline $3 dollars per-hour is presented in Figure 2 on the experimental design. The gift corresponds to an 1/3=33% increase.

7.3) *Effort response* to wage increases and their significance levels were calculated as follows. Productivity is measured by the number of completed and correct CAPTCHAs entered in the 4-hour task.

* The effort response for the full sample of workers comes from Table 2, where treatment "Wage=3" corresponds to the baseline category. Column (1) shows the difference between this baseline and the "Wage=3+1" treatment using the number of completed and correct CAPTCHAs as the dependent variable. The baseline productivity of the "Wage=3" treatment is captured by the constant term amounting to 792.1, while the coefficient for "Wage=3+1" corresponds to 146.8, which is significant using robust standard errors (see Table 2, Column (1). The effort increase for the full sample thus corresponds to 146.4*100/(792.1)=18%.
* The effort response for the experienced oDesk workers comes from Table 3, Panel A. The productivity in the "Wage=3+1" treatment for experienced workers is 973 records whereas it is 773.1 for the baseline "Wage=3". This difference in 199.9 records is statistically significant at the 5% level. The effort increase for the experienced sample corresponds thus to 199.9*100/773.1=26%.
* The effort response for the unexperienced oDesk workers comes from Table 3, panel A. The effort increase for the unexperienced sample corresponds thus to (861.8-834.4)*100/(834.4)=3%.

7.4) *Elasticities.* The elasticity for the full sample corresponds to 18%/33%=0.55. For the experienced workers it corresponds to 26%/33%=0.79 and for the unexperienced workers it corresponds to 3%/33%=0.09.

PANEL II: Most Cited Laboratory Studies

The most cited laboratory studies on gift exchange rely on laboratory games with generally similar features: subjects, usually students, were randomly assigned to be employers or employees and respectively given a common-knowledge profit and cost-of-effort function. The employer offered the wage first, in a publicly observable bid or a private offer to a randomly matched worker, and the employee chose effort second, with choices jointly determining payoffs in experimental units. These one-shot interactions lasted a few minutes before subjects were re-paired. We now detail each study.

(1) **Fehr, Kirchsteiger, and Riedl (1993)**

1.1) *Sample size* in Section II, second paragraph in page 440.

1.2) *Average Wage offer* of 72 experimental units above the market wage of 30 units is in the first paragraph of Section V, page 446. Market wage is in Section III, page 443, second paragraph. This represents a (72-30)*100/30= 140% increase.

1.3) *Average effort response* of 0.4 units is in the first paragraph Section V, page 446 (effort range is 0.1, 0,2, ..., 1). The competitive effort level, which corresponds to the minimum effort of 0.1, is in Section III, page 443, second paragraph. This represents a (0.4-0.1)*100/0.1= 300% increase.

1.4) *Elasticity* corresponds to 300%/140%=2.14.

(2) **Fehr, Kirchsteiger, and Riedl (1998)**[33]

2.1) *Sample size* in first paragraph of Section V, pages 9-10.

2.2) *Average Wage offer* of 74 experimental units above the market wage of 30 units is in the first paragraph of Section VI, page 11. Market wage is in Section V, first paragraph in page 11 (denoted by f) . This represents a (74-30)*100/30= 147% increase.

2.3) *Average effort response* is not reported. We estimate it using the experimental data provided in Tables 5 to Table 8 in Appendix B, pages 24 to 32. These tables display the observed wage, the effort, the cost of effort and the id numbers for workers and firms for all shifts and for all periods for the control and reciprocity treatments. The observed wage and effort are displayed in columns "p" and "q", respectively. There is no row or column indicating to which treatment each observation corresponds to, but since in the control condition effort

[33]Fehr, Kirchsteiger, and Riedl (1998) frame the experiment in terms of prices offered by buyers and quality offered by sellers, but argue this framing applies to labor markets, where buyers are employers and sellers are workers. Hence the wages and effort terminology we use here.

was exogenously determined by the experimenter, for this condition effort is filled with a dash in the "q" column. The average effort for the reciprocity treatment corresponds to the raw average of the 213 effort observations in the "q" columns, which are not dashed, across Tables 5 to 8. See section 6, first paragraph in page 11 for a quote that 213 is right number of effort observations for this treatment. This raw average, therefore, pools effort across all employer-employee matches, all rounds and all shifts. From this calculation the average effort response corresponds to 0.35 units (effort range is 0.1, 0,2, ..., 1). The competitive effort level (denoted by q_0), which corresponds to the minimum effort of 0.1, is in Section V, first paragraph in page 11. This represents a (0.35-0.1)*100/0.1= 250% increase.

2.4) *Elasticity* corresponds to 250%/147%=1.70.

(3) **Fehr, Kirchler, Weichbold, and Gächter (1998)**

3.1) *Sample sizes*. Number of shifts is in the first paragraph of Section III, page 333. Number of subjects per shift is in the fist paragraph of Section II.A, page 329 ("Bilateral GE treatment") and in Section II.C, page 331 ("GE Market treatment").

3.2) *Average Wage offers* by treatment are not reported. We estimate them using Figure 2a in page 340.

- *Bilateral GE treatment*. The approximated average wage offer is 63 experimental units above the market wage of 20 units. Market wage is in Table 1, first column, last row, page 329. This represents a (63-20)*100/20= 215% increase.
- *GE Market treatment*. The approximated average wage offer is 59 experimental units above the market wage of 20 units. Market wage is in Table 1, second column, last row, page 329. This represents a (59-20)*100/20= 195% increase.

3.3) *Average effort responses* are not reported. We estimate them using Figure 1 in page 334. Figure 1 shows the average effort by wage intervals. Intervals correspond to 21 to 30 wage units, 31 to 40, 41 to 50, ... , 71 to 80 and more than 80 wage units. Figure 1 also reports the percentage of employer-employee matches in each wage interval. The average effort response is calculated as the weighted average of the average efforts by wage intervals.

- *Bilateral GE treatment*. The approximated average effort response corresponds to 0.36 units (effort range is 0.1, 0,2, ..., 1). Table 7 shows the exact calculation.

Table 7: Calculation of the Average Effort for the Bilateral GE Treatment in Fehr, Kirchler, Weichbold, and Gächter (1998)

Wage Interval (Experimental Currency)	Percentage of Trades Per Wage Interval	Approximated Average Effort Per Wage Interval	Weighted Average Effort Per Wage Interval
(1)	(2)	(3)	(4)
21-30	9	0.16	1.44
31-40	7	0.21	1.47
41-50	17	0.29	4.93
51-60	20	0.35	7
61-70	21	0.43	9.03
71-80	15	0.44	6.6
+ 80	11	0.53	5.83
Total	100	2.41	36.3
	Total Weighted Average Effort		0.36

Notes: Columns (1) shows the wage intervals as shown in the x-axis of Figure 1. Column (2) shows the percentage of trades (employer-employee matches realised) as shown in Figure 1. Column (3) corresponds to an approximation of the average effort of each wage interval, which was estimated visually from Figure 1. Column (4) corresponds to the multiplication of columns (2) and (3). "Total Weighted Average Effort" corresponds to the summation of column (4) divided by 100.

The competitive effort level, which corresponds to the minimum effort of 0.1, is in Table 1, first column, last row, page 329. This represents a (0.36-0.1)*100/0.1= 260% increase.

- *GE Market treatment.* The approximated average effort response corresponds to 0.4 units (effort range is 0.1, 0,2, ..., 1). Table 8 shows the exact calculation.

  The competitive effort level, which corresponds to the minimum effort of 0.1, is in Table 1, second column, last row, page 329. This represents a (0.4-0.1)*100/0.1= 300% increase.

3.4) *Elasticities.* The elasticities correspond to 260%/215%=1.21 ("Bilateral GE treatment") and 300%/195%=1.54 ("GE Market treatment").

3.5) *Effort response and elasticity to a 67% wage increase in the "Bilateral GE treatment".* An increase in 67% in wages corresponds to an increase from the market-clearing wage of 20 to 33. Figure 1 in page 334, documents that for a wage of 33, the effort response raises from 0.1 to 0.15-0.20. This corresponds to a 50% to 100% increase, and a pay-effort elasticity between 50%/67%=0.75 and 100%/67%=1.5.

(4) **Gächter and Falk (2002)**

4.1) *Sample size.* Number of shifts is in the first paragraph of Section IV, page 7.

Table 8: Calculation of the Average Effort for the GE Market Treatment in Fehr, Kirchler, Weichbold, and Gächter (1998)

Wage Interval (Experimental Currency)	Percentage of Trades Per Wage Interval	Approximated Average Effort Per Wage Interval	Weighted Average Effort Per Wage Interval
(1)	(2)	(3)	(4)
21-30	3	0.1	0.3
31-40	7	0.15	1.05
41-50	18	0.3	5.4
51-60	33	0.42	13.86
61-70	26	0.48	12.48
71-80	11	0.5	5.5
+ 80	2	0.55	1.1
Total	100	2.5	39.69
	Total Weighted Average Effort		0.40

Notes: Columns (1) shows the wage intervals as shown in the x-axis of Figure 1. Column (2) shows the percentage of trades (employer-employee matches realised) as shown in Figure 1. Column (3) corresponds to an approximation of the average effort of each wage interval, which was estimated visually from Figure 1. Column (4) corresponds to the multiplication of columns (2) and (3). "Total Weighted Average Effort" corresponds to the summation of column (4) divided by 100.

Number of subjects per shift is in Appendix: Instructions, page 22.

4.2) *Average Wage offer* is not reported. We approximate it using the reported average payoff of the firm, (120-W)*e, which corresponds to 19.4 (see Section IV, first paragraph in page 8). Using the average effort of 0.41 (see below), we have that (120-W)*0.41=19.4, which means that the average wage offer is approximately W=120-(19.4/0.41)=73. The market wage of 21 is in Section IV, first paragraph in page 8 (denoted as w^*). This represents a (73-21)*100/21= 248% increase.[34]

4.3) *Average effort response* of 0.41 units (effort range is 0.1, 0,2, ..., 1) is in Section IV, first paragraph in page 9. The competitive effort level of 0.1 is in Section IV, first paragraph in page 8 (denoted as e^*). This represents a (0.41-0.1)*100/0.1= 310% increase.

4.4) *Elasticity* corresponds to 310%/248%=1.25.

(5) **Brown, Falk, and Fehr (2004)**

5.1) *Sample size*. Number of shifts is in second paragraph in page 755. Number of subjects per shift is second paragraph in Section 4, page 759.

[34]Figure 1, in page 7, however, shows that the average wage, per shift for the "OS" treatment ("One-Shot" treatment) hovers around 61 units, which is below the average 73 units implied by the average payoff for the firm of 19.4 stated in the text. In the case of an average wage of 61, the percentage wage increase is (61-21)/21*100=190%, resulting in an pay-effort elasticity of 310%/190%=1.63. We report the most conservative elasticity of 1.25, calculated below, in the table.

5.2) *Average Wage offer* is not reported. We estimate it using Figure 3, page 763, which shows the average wage offer by period. From this calculation, the average wage offer corresponds to approximately 24 units. The market wage of 5 is in the first paragraph of Section 3, page 755. This represents a (24-5)*100/5= 380% increase.

5.3) *Average effort response* is not reported. We estimate it using Figure 5, page 767, which shows the average effort by period. From this calculation, the average wage offer corresponds to approximately 3.3 units. The minimum effort of 1 is in the first paragraph of Section 3, page 755. This represents a (3.3-1)*100/1= 230% increase.

5.4) *Elasticity* corresponds to 230%/380%=0.61.

PANEL III: Companion Real-Effort Laboratory Experiments

(1) **Hennig-Schmidt, Sadrieh, and Rockenbach (2010)** Students stuffed envelopes for two fifteen-minute sessions, receiving a show-up fee of 1.50 euros and a wage of 2.5 euros per session. All workers received 2.5 euros in session one. In session two, a 19-subject subsample received a 10% raise and surplus information. They raised their output vis-à-vis session one by 12.9 envelopes, whereas those in the control raised it by 10. This 29% magnitude, if statistically significant, could overstate the true elasticity, due to the very small control sample.[35]

3.1) *Sample sizes* for the control ("L0"), "L10" and "L10 surplus" treatments of 10, 10 and 19 subjects respectively, are in Table 3, columns L0, L10 and L10 surplus, respectively, row Number of typists in page 825.

3.2) *Wage increase* of 0.25 euros for both the "L10" and "L10 surplus" treatments above the baseline of 2.5 euros per each 15-minutes shift are in Table 3, columns L10 and L10 surplus, respectively, 2nd work unit wage (15 mins) row in page 825. Because the wage raise only applies to the second hour, the percentage wage raise corresponds to 0.25*100/2.5=10% for both treatments. In the "L10 surplus" treatment, the wage raise is accompanied by information about the employer's surplus as a result of work effort.

3.3) *Effort responses* to the wage increase, measured as the number of filled envelopes, are presented in Table 5, page 826. The change from period one to two in the number of filled envelopes in the "L10" treatment corresponds to 50.4-40.5=9.90, while the control "L0" increased output by 41.1-31.1=10 (see

[35]The statistical significance of this result was not reported.

Table 5, column L0 and L10, row Output quantity work unit 2 minus row Output quantity work unit 1). The difference 9.90-10=-0.10 corresponds to a -0.10*100/10=-1% increase. This difference is not statistically significant (see Section 3.2, third paragraph in page 827). The change from period one to two in the number of filled envelopes in the "L10 surplus" treatment corresponds to 56.2-43.3=12.9. The difference with the control "L0" corresponds to 12.9-10=2.9, which is a 2.9*100/10=29% increase. The significance of this estimate is not reported.

3.4) *Elasticities.* The elasticity in the "L10" treatment corresponds to -1%10%=-0.1. The elasticity for the "L10 surplus" corresponds to 29%10%=2.9.

E Online Appendix - Protocols

E.1 Protocol and Wording for Treatments

Students who answer the campus fliers contact the recruiting assistant by phone or email. The recruiting assistant gathers their contact information and availability. Since this is a natural field experiment, no consent forms are signed to enter the employment relationship. Workers coordinate with the project manager (our research assistant) the time and place to perform the job. Each subject works in a different room, in isolation, where rooms are spread across campus to avoid any contamination.

The description and wording for each treatment is as follows:

1) Control

- Subjects are hired at $12 per hour for the duration of the task—the six hours of work—as advertised. On the first day, subjects are briefly instructed on the very simple bibliographic software before they start working. Immediately after, they execute the agreed two hours of work. On the two subsequent days, subjects work two hours each day as agreed upon recruitment. Finally, subjects are paid the recruiting $12 per hour ($72 total) when they submit their time sheets at the end of the six hours.

2) 67%Raise

In this treatment, workers were offered a 67% wage raise versus the contract wage of $12 per hour: they received a raise to $20 per hour (i.e., an additional $8 per hour) for the duration of the contract.

The 67%Raise treatment aggregates three subtreatments varying the timing of the information about the raise (immediately before or one week before the first shift)

and when the raise was paid (at the start or at the end of each shift): 67%SurpriseRaise, 67%AnticipatedRaise and 67%PromisedRaise. Given that the distribution of outcomes for these three subtreatments was not statistically different we aggregated them into a single condition 67%Raise.

2.A) 67%SurpriseRaise

- Subjects are hired at $12 per hour for the duration of the task—the six hours of work—as advertised. On the first day of work, they are briefly instructed on the very simple bibliographic software, but before they start working the agreed two hours, subjects are offered the envelope with the raise raise of $8 dollars per hour for the first day ($16 dollars total). They are further told that there will be similar gifts for the next two shifts. On the two subsequent days of work subjects are given the raise before they start their two-hour shift. Finally, subjects are paid the recruiting $12 per hour ($72 total) when they submit their time sheets at the end of the six hours.
- Wording:
  - At the beginning of day 1: "We have a thank you gift, in the amount of $8 per hour in addition to the $12 per hour pay. We will give this gift for the hours you work today and we will also give you the same gift on each of the next two shifts."
  - At the beginning of days 2 and 3: "As promised, here is the gift for today".
  - At the end of day 3, upon receiving the time sheets. "Thank you for your work. Here is the $72 payment."

2.B) 67%AnticipatedRaise

- This treatment is exactly like 67%SurpriseRaise, except that the research assistant meets subjects exactly one week in advance of the start of work to give them instructions on the program and shows them the time sheets used to pay them the agreed hiring wage of $72.[36] Further, during this extra shift the research assistant shows workers the envelope with the $8 per hour raise (the $16 in the envelope).[37]
- Wording:

[36] If the meeting was not possible exactly one week in advance, it was scheduled week and 1 day in advance.

[37] The workers in all there treatment also receive time sheets used to pay them the agreed hiring wage of $72, but on the first day they report to work.

* For the extra initial shift: "We have a thank you gift, in the amount of $8 per hour in addition to the $12 per hour pay. We will give this gift at the beginning of your shift of next week and we will give the same gift on each of the next two shifts. (The research assistant only shows the gift, does not give it to subjects. He is instructed to make it natural and to this end he shows the envelope with the gift on top of the time sheets).
* At the beginning of days 1, 2 and 3: "As promised, here is the gift for today" (offered again right before the students start the shift).
* At the end of day 3, upon receiving the time sheets. "Thank you for your work. Here is the $72 payment."

2.C) 67%PromisedRaise

- This treatment is exactly like SurpriseRaise, except that instead of handing in the gifts immediately before subjects work on the task, the gifts are announced in the first shift, but they are only received at the end of the last shift. That is, subjects receive the raise at the end of the third shift ($48) together with the hiring pay of $72.
- Wording:
  * At the beginning of day 1: "We have a gift in the amount of $8 per hour in addition to the $12 per hour pay. We will give this gift for the hours you work today and we will give the same gift on each of the next two shifts. You will receive it at the end of the last shift."
  * For days 2 and 3: Nothing is said.
  * At the end of day 3, upon receiving the time sheets. "Thank you for your work. Here is the $48 gift and the $72 payment."

3) 50%-100%Raise

- In this treatment subjects receive a raise that mounts to $6 per hour in the first and second days of work (i.e., shifts one and two, respectively) to $18 per hour. In the third day before subjects start the task, they are given a further additional raise of $6 per hour, to $24 per hour.
- Wording:
  * At the beginning of days 1 and 2: same wording as in 67%SurpriseRaise treatment, but changing the size of the gift.

* At the beginning of day 3: "We have a further thank-you gift in the amount of $6 per hour in addition to the $12 per hour pay and the gift of $6 per hour in previous shifts. Here are the gifts."
* At the end of day 3, upon receiving the time sheets. "Thank you for your work. Here is the $72 payment."

4) PieceRate

- Subjects are hired at $12 per hour for the duration of the task—the six hours of work—as advertised. On the first day of work, before they start working the agreed two hours, subjects are informed that, in addition to the $12 per hour agreed upon hiring, they will receive a piece rate for each record they enter.
- The piece rate corresponds to $0 \times x$ if $x < 70$; $0.05 \times x$ if $70 \leqslant x \leqslant 110$; $0.10 \times x$ if $110 < x \leqslant 140$; and $0.20 \times x$ if $x > 140$; where x is the number of records entered on the shift. The piece rate is paid in cash at the end of each shift. Subjects are paid the recruiting $12 per hour when they submit their time sheets at the end of the six hours, as it is customary at the host university where experiments were conducted.
- Wording:
  * Wording for the communication with subjects right before the start of the first shift, when they are informed about the piece rate: "In addition to the agreed $12 per hour, you will receive a piece rate in each of the three shifts. The piece rate is as follows (research assistant walks subjects through Table 9 below). The payment for the piece rate will be given to you in cash at the end of each shift".
  * For days 1, 2, 3, when handing in payment for the piece rate at the end of each shift: "You logged XX records during your two-hour shift. This implies that you receive $YY. Here is your payment".

Table 9: Piece Rate Table Shown to Subjects in PieceRate Treatment

In addition to the $12 per hour in each shift, you will receive an amount for each record imputed, per shift, as follows below.

Number of records	**Extra payment per record**	**Total compensation per two-hour shift**
69 or less	$0	$24 for the two-hours of work
Between 70 and 110	$0.05 per record inputted	For example: • If input 70 records, receive an extra 70x0.05=$3.5. (So total compensation per shift is $24+$3.5=$27.5) • If input 110 records, receive an extra 110x0.05=$5.5. (So total compensation per shift is $24+$5.5=$29.5)
Between 111 and 140	$0.1 per record inputted	For example: • If input 111 records, receive an extra 111x0.1=$11.1 (So total compensation per shift is $24+$11.1=$35.1) • If input 140 records, receive an extra 140x0.1=$14 (So total compensation per shift is $24+$14=$38)
141 or more	$0.20 per record inputted	For example: • If input 141 records, receive an extra 141x0.20=$28.2 (So total compensation per shift is $24+$28.2 =$52.2)

E.2 Protocol for the Post-Field Experiment Survey

Next we present the exact protocol used in the survey. Notes to the reader are in corresponding footnotes, which were not part of the protocol.[38]

"Instructions
We will ask you to make decisions on two related situations ("games"). In these games, in addition to your participation fee of $10, you can earn up to $15. To this end you will be (anonymously) paired with another undergraduate student from your university who will be the "other player" (or "partner") in these games.[39]

We will start with a brief training period for you to familiarize yourself with these simple games. You will play exactly the same games you will face in the actual decision period, except that you will not receive the payment corresponding to the outcome of the games during this practice period. After you practice playing each game, you will be asked whether you want to ask the research assistant clarifying questions, play the practice game again or whether you want to play the actual games.[40]

Remember you can contact a research assistant to answer any questions you may have about the survey and the payments anytime between 9am and 5pm. The research assistant is available by (email), (Skype) or (phone).[41]

(New screen)

PRACTICE GAMES

Your will not be paid for the outcomes of these two games

Game 1

First, you have to choose between action A and B. The other player, having observed your choice, will also choose between A and B.

[38]The survey also included a multiple-choice questionnaire and 11 lotteries. We do not dwell on their description since these results would, for the most part, only have been relevant had gift exchange been observed.

[39]To achieve this pairing, we first had a random sample of students from each university play each of the three games. We then paired our workers with a randomly selected subject from this previously surveyed pool as is customary in the literature.

[40]Subjects faced exactly the same choices in these practice games as in the actual games. Furthermore, subjects could contact a research assistant and ask any questions about the games or the survey in general. Finally, they were unconstrained in the number of practice rounds.

[41]The existence of a research assistant who would support subjects in the survey was communicated to the subject in the email that invited them to participate in the survey.

The payment you will receive from your choice depends on your choice **and** on your partner's choice, and it is represented in the following diagram[42]:

	Other player chooses A	Other player chooses B
You choose A	You get \$4 Other player gets \$4	You get \$0 Other player gets \$7.5
You choose B	You get \$7.5 Other player gets \$0	You get \$1 Other player gets \$1

Please examine the payments on the diagram carefully and choose between actions A or B by clicking below:

(button) I choose A
(button) I choose B

Message on screen: "Thank you for your choice"
Message on the next screen (This screen also displays the payment diagram):

"You chose X *(A or B)*.

(1) Suppose your partner plays A

- What would be you payment in this case?

  (If payment is correct, display"Yes, that is correct". If subject is incorrect, display"That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display"Yes, that is correct". If subject is incorrect, display"That is not correct, please try again" and display the same question again)

(2) Suppose your partner plays B

- What would be you payment in this case?

[42]The stakes in the games were as follows: If both subjects cooperated they would both receive \$4; if both defected they would both receive \$1, following Clark and Sefton (2001). The deviation payoff for defecting when the other cooperated was \$7.5, following the Trust game in Charness and Rabin (2002) (pages 861 and 862). Finally, the payoff of cooperating when the other defected was \$0, following Clark and Sefton (2001).

(If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(New screen) Now suppose that you had chosen Y *(Y is A if X was B and Y is B if X was A)*

(3) Suppose your partner plays A

- What would be you payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(4) Suppose your partner plays B

- What would be you payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(New screen) What would you like to do?

a. Ask the research assistant clarifying questions
b. Play this practice game again

c. Continue to the second game of the practice shift

If subject selects a) "You can contact a research assistant (Name) by calling (Phone) or by calling the Skype id [···]".
If subject selects b), repeat the game.
If subject selects c) proceed to Game 2 below.

Game 2

Suppose now that instead of choosing first, you will choose second. That is, your partner will choose between A or B and having observed her/his choice, you will choose between A and B. The payment you will receive is represented in the following diagram, which is the same diagram, with the same amounts, as that in the previous games.

	Other player chooses A	Other player chooses B
You choose A	You get \$4 Other player gets \$4	You get \$0 Other player gets \$7.5
You choose B	You get \$7.5 Other player gets \$0	You get \$1 Other player gets \$1

Please examine the payments on the diagram carefully and choose between actions A or B by clicking below:
If my partner chooses A:

(button) I choose A
(button) I choose B

If my partner chooses B:

(button) I choose A
(button) I choose B

Message on the next screen (This screen also displays the payment diagram):

(1) "You chose X *(A or B)* if your partner chooses A.

- What would be you payment in this case?

  (If payment is correct, display"Yes, that is correct". If subject is incorrect, display"That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(2) "You chose X *(A or B)* if your partner chooses B.

- What would be you payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(3) "Suppose you had chosen Y if your partner had chosen A. *(Y is A if X was B or Y is B if X was A)*

- What would be you payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(4) "Suppose you had chosen Y if your partner had chosen B. *(Y is A if X was B or Y is B if X was A)*

- What would be you payment in this case?

  (If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

- What would be your partner's payment in this case?

(If payment is correct, display "Yes, that is correct". If subject is incorrect, display "That is not correct, please try again" and display the same question again)

(New screen) What would you like to do?

a. Ask the research assistant clarifying questions
b. Play this practice game again
c. Continue to the second game of the practice session

If subject selects a) "You can contact a research assistant (Name) by calling (Phone) or by calling the Skype id [···]".
If subject selects b), return to the initial screen "PRACTICE GAMES"
If the subject selects c), go to new screen with the "ACTUAL GAMES".

(New screen)

ACTUAL GAMES

Now you will make your actual decisions. The resulting monetary outcomes of the games and gambles will be added to your Amazon gift card with your participation fee. To play the game you have now been anonymously paired with another undergraduate student from your university.

Game 1

First, you have to choose between action A and B. The other player, having observed your choice, will also choose between A and B.
The payment you will receive from your choice depends on your choice **and** on your partner's choice, and it is represented in the following diagram:

	Other player chooses A	Other player chooses B
You choose A	You get \$4 Other player gets \$4	You get \$0 Other player gets \$7.5
You choose B	You get \$7.5 Other player gets \$0	You get \$1 Other player gets \$1

Please examine the payments on the diagram carefully and choose between actions A or B by clicking below:

(button) I choose A
(button) I choose B

Message on screen: "Thank you for your choice.You will know the outcome of the game once you play the second game."

Game 2

Suppose now that instead of choosing first, you will choose second. That is, your partner will choose between A or B and having observed her/his choice, you will choose between A and B.

The payment you will receive is represented in the following diagram, which is the same diagram, with the same amounts, as that in the previous games.

	Other player chooses A	Other player chooses B
You choose A	You get \$4 Other player gets \$4	You get \$0 Other player gets \$7.5
You choose B	You get \$7.5 Other player gets \$0	You get \$1 Other player gets \$1

Please examine the payments on the diagram carefully and choose between actions A or B by clicking below:

If my partner chooses A:

(button) I choose A
(button) I choose B

If my partner chooses B:

(button) I choose A
(button) I choose B

(New screen)

OUTCOME OF GAMES AND PAYMENTS

Outcome of Game 1:

You chose X
Your partner chose XX
Therefore, you won XX

Outcome of Game 2:

You chose X
Your partner chose XX
Therefore, you won XX

Closing Window: Farewell Message

"Thank you for participating in this survey. You will receive a payment of $XX in addition to your participation fee. The payment is being processed now. You will receive an email within the next hour with an electronic Amazon gift card. Please contact the research assistant (Name) by calling (Phone) or by calling the Skype id [···] if you have any concerns."