A reassessment of the potential for loss-framed incentive contracts to increase productivity: a meta-analysis and a real-effort experiment

Paul J. Ferraro*[1] and J. Dustin Tracy†[2]

[1] *Carey Business School & Whiting School of Engineering, Johns Hopkins University*
[2] *Economic Science Institute, Chapman University*

August 4, 2021

Abstract

Behavioral scientists have reported substantial increases in worker productivity when incentives are framed as losses rather than gains. Loss-framed incentive contracts have also been reported to be preferred by workers. These claims are challenged by results from our meta-analysis and real-effort experiment. Whereas the summary effect size from loss-framed contracts in laboratory experiments is a 0.4 SD increase in productivity, the summary effect size from field experiments is 0.0 SD. Although this difference may reflect differing labor environments in the laboratory and field, we detect evidence of publication biases among laboratory experiments. In a new laboratory experiment that addresses prior design weaknesses, we estimate an effect size of 0.1 SD. This result, in combination with evidence from the meta-analysis, suggests that the difference between the effect size estimates in published laboratory and field experiments does not stem from the limited external validity of laboratory experiments, but may instead stem from a mix of underpowered laboratory designs and publication biases. Moreover, in our experiment, most workers preferred the gain-framed contract and the increase in average productivity is only detectable in the subgroup of workers (∼20%) who preferred the loss-framed contracts. This result suggests that employers may find using these contracts in real labor environments challenging. Based on the results from our experiment and meta-analysis, we believe that further research is warranted to assess the robustness and magnitude of the impacts from loss-framed contracts before advocating for their adoption by private and public sector actors.

Keywords: framing effects, incentive contracts, meta-analysis, real-effort experiment, and behavioral insights

JEL Codes: C91 J24 J33

*pferrar5@jhu.edu
†tracy@chapman.edu

1 Introduction

In the behavioral sciences, scholars have invoked loss aversion to explain behavioral patterns that appear to contradict traditional economic theories (Kahneman and Tversky, 1979; Thaler and Johnson, 1990; Hardie et al., 1993; Haigh and List, 2005; Jarrow and Zhao, 2006; Looney and Hardin, 2009; Chrisman and Patel, 2012). For a loss-averse individual, the disutility of a loss is larger than the utility of an equivalent gain. Recently, scholars and practitioners working in the private and public sectors have tried to harness loss aversion to induce behavioral changes Convery et al. (2007); Thaler and Sunstein (2008); Jakovcevic et al. (2014); Homonoff (2018).

One application is the loss-framed incentive contract, in which incentives are framed as losses from an earnings benchmark rather than as gains from a zero earnings default (Hossain and List, 2012; Imas et al., 2016). For example, rather than offer a worker \$1 for every unit of output, a firm can offer a loss-framed contract that pays workers \$B for a performance benchmark of B units from which \$1 would be subtracted for every unit short of the benchmark. As long as the worker does not exceed the benchmark, the contracts pay the same for a given level of performance. Given this equivalency, traditional economic theory predicts no difference in expected productivity under the two contracts. The theory of loss aversion, however, assumes workers assess outcomes relative to reference points, which differ under the two contracts. Loss-averse workers will work harder to avoid losses from \$B than to achieve similar gains from \$0 in the gain-framed contract.

Consistent with this prediction, 16 of the 20 experiments that contrast productivity under loss-framed and gain-framed contracts report that loss-framed contracts induce greater productivity (Figure 1). Half of them report effect sizes of about a half standard deviation (SD) or larger. Even after weighting the studies by the precision of their estimates in a meta-analysis, the summary estimated average effect size is still 0.21 SD.

Yet despite the experimental evidence pointing to a substantial productivity impact from a simple change in framing, loss-framed contracts are rare outside behavioral science experiments. If the impacts of loss-framed contracts were only observed in cases in which workers received the benchmark earnings up-front, a plausible explanation is that employers find it costly, financially or socially, to claw back the lost earnings from workers. Yet more than half of the experiments do not endow the workers with the earnings in advance; they simply change the contract wording. We can detect no difference in the estimated effect sizes between studies that advance the earnings and those that do not ($\chi^2 = 1.45$, $df = 1$, $p = 0.23$); see supplementary materials. In other words, with little cost, loss-framed contracts appear to be able to increase effort exerted by workers (or other agents, such as citizens targeted by government programs that aim to incentivize the supply of positive externalities). Indeed, this promise has inspired some behavioral scientists to encourage private and public sector actors to adopt loss-framed contracts. For example, designers of an incentive program in South Africa to encourage citizens to exert more effort to drive carefully framed the incentives as losses ("loss aversion lotteries") because people have a "tendency to be especially troubled by the risk of losing things that already belong to them."[1]

We postulate three reasons why, despite the abundance of evidence for productivity gains associated with loss-framed contracts, there is a paucity of such contracts in the private and public sectors: (1) the estimates that form the evidence base exaggerate the true effects or are not externally valid; (2) workers prefer gain-framed contracts, and thus any productivity gains at the intensive margin may be more than offset by productivity losses at the extensive margin (i.e., employers offering gain-framed contracts attract more workers); and (3) the framing effect does not persist over time (most studies observe behavior for short time periods). In our study, we focus on the first two reasons and find evidence consistent with them both.

In our meta-analysis, we observe that the average estimated effect masks substantial heterogeneity in effect sizes between laboratory and field experiments. Whereas laboratory experiments suggest that loss framing increases productivity by nearly 0.40 SD, the summary effect size for field experiments is zero (to two decimal places) 0.00 SD (Figure 1). The laboratory experiment estimates are also much more variable, and thus the summary effect is much less precisely estimated among the laboratory experiments. Whether the difference between laboratory and field experiments reflects different types of workers, different types of

[1] https://www.ideas42.org/wp-content/uploads/2018/04/Project-Brief_Offering-Rewards-to-Safe-Drivers.pdf

working environments, or publication biases, these patterns imply that real-world organizations may expect more modest impacts from loss-framed contracts than implied by the academic literature.

To shed more light on the performance of loss-framed contracts and worker preferences for these contracts, we designed a real effort laboratory experiment in which workers had an opportunity to select their contract after working under both contract types. In the experiment, workers were randomized to work under a loss-framed contract and then a gain-framed contract, or vice-versa (i.e., within-worker design). Our estimated effect size of 0.12 SD is one-third the summary effect size of prior laboratory experiments and closer to the summary effect size from field experiments.

Furthermore, when given the opportunity to choose the contract framing, only one in five workers chose the loss-framed contract. Three out of four workers chose the gain frame, and the rest reported indifference. This pattern, in which a minority of workers report preferring the loss frame, is consistent with prior claims (Lazear, 1991) and survey studies (Tannenbaum et al., 2013; Evers et al., 2017), but not with two incentivized experiments (Imas et al., 2016; de Quidt, 2018). In Section S.2.1, we show how the designs and preference metrics used by these two studies can mask a majority preference for gain-framed contracts.

Furthermore, we observe that the loss-framed contract effect is only detectable among the minority of workers who prefer the loss-framed contract. The estimated effect among the rest of the workers is indistinguishable from zero, in both practical and statistical senses. If the loss-framed-contract-preferring workers were more productive than the average worker, employers could offer loss-framed contracts to screen for productive workers. However, we find that loss-frame-contract-preferring workers are less productive under the gain-framed contract; a loss-framed contract only brings their average performance up to the average performance of the other workers. Even if the loss-framed contract cannot serve as a screen for high productivity workers, it may serve as a commitment device for low-productivity workers (Imas et al., 2016). To exploit this commitment device, however, an employer would have only two options. It could offer two types of contracts simultaneously within the organization and let workers select their preferred contract. Or it could separate its operations into two units, each offering a different contract. In many contexts, the organizational complexities and fixed costs implied by such strategies could easily dominate the modest performance benefits from loss-framed contracts.

Our experimental results also suggest that the difference between the summary effect size estimates from laboratory and field experiments may not be an artifact of the laboratory context per se, but instead may arise from the mix of underpowered designs and publication bias. Underpowered designs produce highly variable estimated effect sizes. This variability yields an exaggerated picture of the true effect size when combined with a bias against publishing estimates that are statistically insignificant or estimates of an unanticipated sign. Adjusting for this bias through a simple trim-and-fill method yields a summary effect size for laboratory experiments of 0.05 SD, whose confidence interval (CI) includes the summary effect size for field experiments.

In summary, our meta-analysis and experimental results imply that the effects of loss-framed contracts on productivity may be real. Yet these effects are also likely to be, on average, modest and heterogeneous in ways that are difficult to exploit in a cost-effective manner. In the next section, we present the meta-analysis. We present our experimental design and results in Section 3. In Section 4, we discuss our results and outline paths for future research. All code and data for the meta-analysis and experimental analyses are available at `https://osf.io/3nqgd/`.

2 Meta-analysis

2.1 Methods

The meta analysis was performed in R using the package *meta* (Schwarzer et al., 2015).

2.1.1 Inclusion criteria

We sought to identify all studies, published or unpublished, that have the following features: (1) an experimental design in which (a) loss-framed and gain-framed incentive contracts were (b) randomized within or

across workers, providing an unbiased estimator of the average effect of loss-framed contracts rather than gain-framed contracts (this criterion rules out, for example, an early study (Luft, 1994) that allowed workers to choose their contract and a recent study (de Quidt, 2018) that allowed workers to opt out); (2) a real or stated behavioral outcome measure of productivity; and (3) publicly accessible data or sufficient statistical details in the published article to allow for inclusion of the study in a meta-analysis. If an article provided neither data nor sufficient details, we contacted the authors to obtain the data or estimates of the requisite parameters for inclusion in the meta-analysis (see next section for a description of required inputs). Although we did not explicitly exclude un-incentivized experimental designs (i.e., hypothetical choices), all studies that met the inclusion criteria also used incentivized designs.

Using Google Scholar, we searched for publications that met our inclusion criteria by using combinations of the terms "loss framing" or "penalty" and "real effort" or "work". When we found publications that met one or more of the inclusion criteria, we used their bibliographies and Google Scholar's list of citations for the publication to identify additional publications that met the inclusion criteria. In all, 126 publications were identified as warranting closer investigation to see if they met all three criteria. The criteria were applied sequentially and the evaluation stopped as soon as the publication failed to meet an inclusion criterion.

Fifteen publications containing 20 experiments met our inclusion criteria, 14 of which used a between-subject design (median number of subjects is 130). Table-1 reports how many studies were excluded by each criterion. The complete list of 111 excluded publications is available at `https://osf.io/3nqgd/`.

Table 1: Number of publications excluded by each criterion

Criteria	**Count**
1	19
1a	50
1b	3
2	37
3	2

Table 2: Excluded estimates within included publications

Study	**Excluded**	**Reason**
Goldsmith and Dhar (2011)	Exp 1A	The task was nearly impossible and designed to increase time spent on task w/o increasing productivity.
Goldsmith and Dhar (2011)	Exps 2-4	Surveys about how frames are perceived.
Armantier and Boly (2012)	Time	Not a productivity measure.
Imas et al. (2016)	Exp 2	Subjects selected into participation through WTP.
Brooks et al. (2017)	Low-Bar & Extreme	Were designed to demonstrate prepaying for too few (too many) units can harm productivity.
Bulte et al. (2020)	Task 2	Subjects selected desired frame.

Some of the publications in the meta-analysis report multiple effect sizes. Hossain and List (2012), Armantier and Boly (2015), and Imas et al. (2016) report on multiple experiments within the same publication. Levitt et al. (2016), and de Quidt et al. (2017) report multiple treatments. Where we report multiple effect sizes from a study, we distinguish each effect with a term in quotation marks that matches the term used by

the authors to describe the experiment or treatment. Table 2 reports which estimates within the included studies were excluded and why.

2.1.2 Estimator

We standardized the treatment effects reported in each study by the pooled standard deviation. To calculate this standardized mean difference (SMD), we employed Hedges (1981, 1982) method, which corrects for bias in the estimation of standard error (*meta's* default). Thus, all estimated treatment effects are reported in proportions of a standard deviation. To calculate a 95% CI of the standardized mean difference, we multiply the standard error of the SMD by the Z score.

To estimate a summary effect size and its confidence interval, we used an empirical Bayes regression estimator, which employs a random-effects model for both the subgroup effect and differences within the subgroup (Raudenbush and Bryk, 1985). Van der Linden and Goldberg (2020) have shown the empirical Bayes method is superior to models for which random effects are only employed for the within-subgroup differences. In contrast to a fixed-effects estimator, the random-effects estimator does not assume that the (unobserved) true treatment effect is constant, but rather permits it to vary from study to study, i.e. it allows the effect of loss-framed contracts to vary according the particulars of the study, e.g. setting, task contract details. Thus the summary effect we seek to estimate is not a single true treatment effect, but rather the mean of the population of true treatment effects (i.e., the studies in the meta-analysis are assumed to be a random sample from this population). We allow for a distribution of treatment effects because prior studies have argued that loss-framed contracting effects are likely to be heterogeneous. For example, loss-framed contracting may be ineffective when the goal is unattainable (Brooks et al., 2017) or in tasks requiring special knowledge to succeed (Luft, 1994; Goldsmith and Dhar, 2011). For details on the way in which the random-effects estimator weighs observations and estimates standard errors, as well as the assumptions it makes with regard to meta-analysis methods, see Schwarzer et al. (2015).

We estimate a summary effect size for all twenty estimates, as well as effect sizes conditional on whether the studies were laboratory or field experiments. To test the null of homogeneous effects across studies within the overall group or subgroups (laboratory and field), we employ the method proposed by Higgins and Thompson (2002). First, we calculate the squares of the differences between each study estimate and the summary estimate. Then we find their weighted sum, Q, and the probability that Q resulted from a chi-square distribution with (number of studies -1) degrees of freedom. We then calculate S, the sum of the fixed-effects weights minus the sum of the square of the weights divided by the sum. Tau Squared, τ^2, is the difference between Q and the degrees of freedom divided by S. The test across subgroups is analogous, with the subgroups' estimates in place of the studies' estimates. The Q for the test across subgroups is the weighted sum of the squared differences between each subgroup's estimate and the joint estimate. There are only two subgroups, so there is only a single degree of freedom. χ is used in place of τ. I^2 is the sum of squared errors $\tau^2(\chi^2)$ scaled to lie between 0 and 100%. Higher values of I^2 imply more heterogeneity.

2.2 Results

Figure 1 plots the standardized treatment effect estimates, along with their respective confidence intervals. The summary estimated effect from loss-framed contracts is an increase in performance by one-fifth of a standard deviation: 0.21 SD (95% CI [0.06, 0.35]). This estimated effect, however, may mask heterogeneity in effects conditional on the study designs.

An important design attribute is whether the experiment was conducted in a laboratory setting or in the field. Some prior publications have questioned the external validity of laboratory experiments, particularly those that, like some of the publications in Figure 1, use student workers (Levitt and List, 2007; Galizzi and Navarro-Martinez, 2018). One might question whether a framing applied to subjects engaged in small-stakes labor in front of an experimenter in a laboratory might affect behavioral mechanisms in ways that differ from naturally occurring field contexts with employees or contractors.

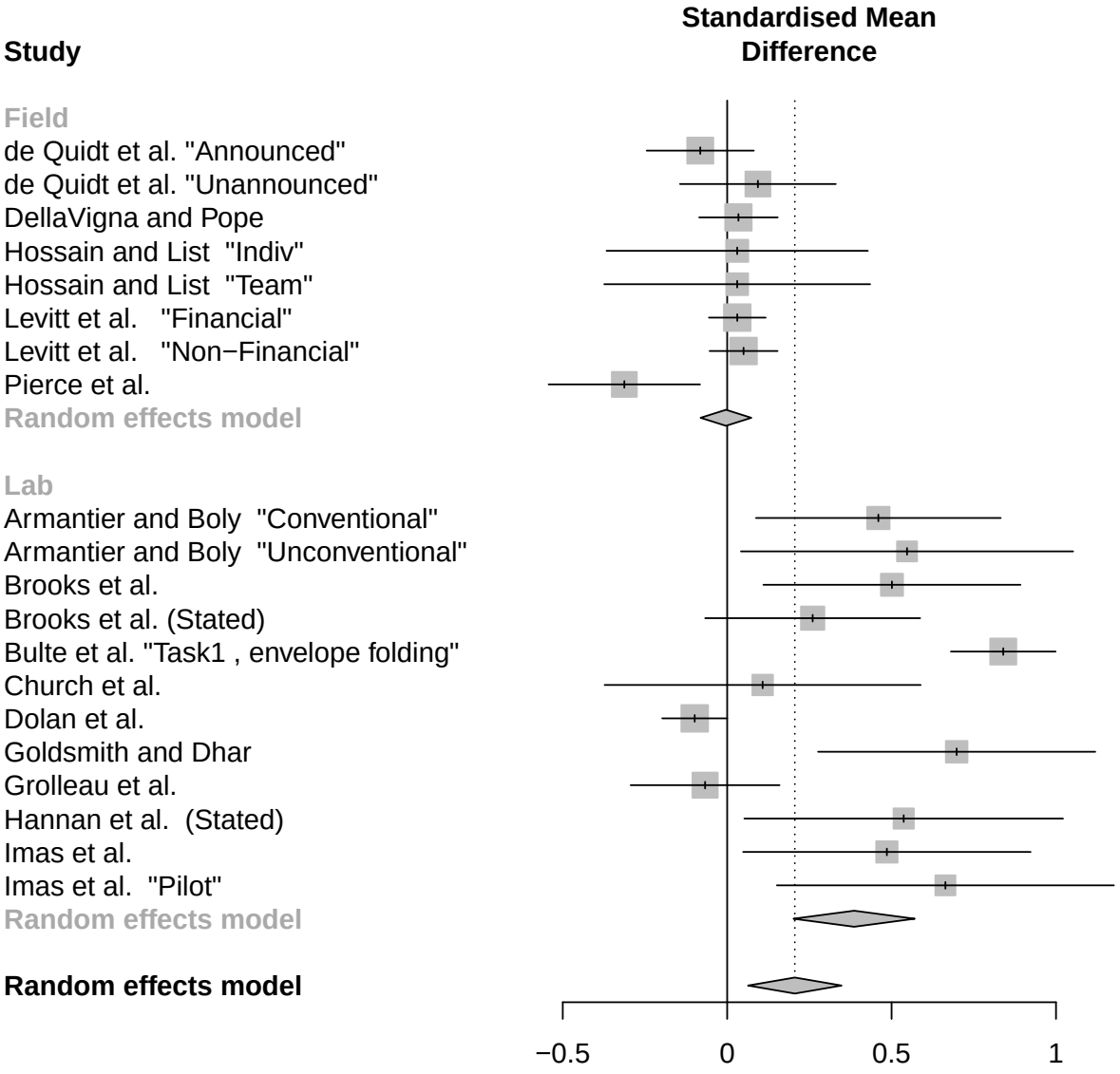


Figure 1: **Meta-analysis of experimental studies estimating the effect of loss-framed contracts on productivity (effort).** The squares represent the mean estimated effects for each experiment, standardized as fractions of the pooled standard deviation. Larger squares imply larger sample sizes. Lines through the squares represent 95% confidence intervals. The centers of the parallelograms represent the estimated summary effect sizes, by type of experiment and overall (dotted vertical line). The width of the parallelograms represent the 95% confidence interval. See Methods for explanation of study designations in quotation marks.

To shed light on this question, Figure 1 also presents estimated treatment effects conditional on whether the studies were field experiments or laboratory experiments. By "field experiment," we mean real effort experiments outside of the laboratory (we do not include so-called "lab-in-the-field" studies that are laboratory experiments that include non-student subjects). In two of the laboratory experiments Hannan et al. (2005) and Brooks et al. (2012), subjects do not engage in real effort but choose "work levels" that result in variable

payoffs (labeled "(Stated)" in Figure 1). Removing these two studies does not change our inferences.

For laboratory experiments, the summary estimated effect size is 0.39 SD (95% CI [0.20, 0.57]). In contrast, the summary effect size for field experiments is -0.00 SD (zero to two decimal places) (95% CI [-0.08, 0.07]). This difference is statistically significant ($\chi^2 = 14.61$, df = 1, $p < 0.01$). Moreover, a measure of variation in effect size estimates across studies (I^2) is substantially larger among the laboratory experiments. In fact, in a test of the null of homogeneity in effect sizes across studies, we can easily reject the null for laboratory experiments ($I^2 = 91\%$, $\tau^2 = 0.071$, $p < 0.01$), but cannot reject it for field experiments ($I^2 = 31\%, \tau^2 = 0.0046$, $p = 0.18$).

One of the laboratory studies is not a peer-reviewed article, but rather was published as part of a government report that includes the results of many experiments (Dolan et al., 2012). Given that one has to look carefully to find the loss-framed contract experiment in the report, it is not surprising that this experiment is not typically cited by later articles in Figure 1. Another laboratory study (Grolleau et al., 2016) was missed in an earlier version of this meta-analysis because the result that loss-framing has a noisy, negative estimated effect on productivity is not apparent from the title, abstract and conclusions, which focus on how loss-framing increases cheating. Thus, when most scholars think about the loss-framed incentive contract literature, these two studies are not typically part of the mix. Removing them from the meta-analysis (see supplementary materials) yields a larger summary estimated effect size for laboratory experiments of 0.55 SD (95% CI [0.41, 0.70]) and much lower measure of heterogeneity ($I^2 = 50\%$, $\tau^2 = 0.0159$, $p = 0.03$). The summary estimated effect size for all studies increases to 0.25 SD (95% CI [0.09, 0.40]).

In addition to differing in the type of workers and their work environments, the laboratory and field experiments also differ in their sample sizes. The median sample size for the laboratory experiments is 98, whereas the median for the field experiments is 436. In Figure 2, the study effect sizes are displayed in a funnel plot that illustrates the relationship between effect sizes and standard errors. In the absence of publication bias or systematic heterogeneity, 95% of the data would be expected to lie within the funnel-shaped lines radiating from the top. An asymmetric distribution reflects the possibility of publication bias or a systematic difference between small and large studies.

Visual inspection of the plot reveals a clear asymmetry, with a shift toward larger estimated treatment effects as the standard error gets larger. To supplement the visual inspection, we run two popular tests that test the null hypothesis of no asymmetry: the regression-based Thompson-Sharp test (Thompson and Sharp, 1999), and the rank-correlation Begg-Mazumdar test (Begg and Mazumdar, 1994). Both reject the null of no asymmetry ($p < 0.05$). We also implement the non-parametric Duvall-Tweedie trim-and-fill method (Duval and Tweedie, 2000a,b) and find evidence of asymmetry (see Supplement 2.1 for details on tests and results).

We know of no theory that predicts that small studies will systematically yield larger estimated treatment effects. In our meta-analysis, however, the small studies are mostly laboratory studies. Thus one explanation for the asymmetry is that the loss-framed contracting is more effective in laboratory experiments, implying that the results in laboratory environments are not generalizable to the field.

Another explanation for the asymmetry is publication bias. There is a widely discussed bias in all of science, including the behavioral sciences, to produce, submit, and publish results that are large in magnitude and statistically significant at conventional thresholds (e.g. (Rosenthal, 1979; Ioannidis, 2005; Duval and Tweedie, 2000a,b; Simmons et al., 2011; Maniadis et al., 2014; Miguel et al., 2014; Simonsohn et al., 2014; Open Science Collaboration, 2015; Baker, 2016)). In other words, the difference between the estimated effects of laboratory and field experiments in Figure 1 may not reflect weak external validity of laboratory experiments, but rather a stronger publication bias among laboratory experiments. Such a bias is plausible: field experiments, on average, are more novel and costly than laboratory experiments, and thus may be more likely to be published even when the treatment effect is small or statistically insignificant. If laboratory experiments are only published when they yield statistically significant results, and such studies tend to be underpowered, then the only estimates that will appear in the published literature will be exaggerated compared to their true values (so called Type M error) — because if the true treatment effects were small, only inflated estimates will pass the statistical significance threshold and be published (Button et al., 2013; Gelman and Carlin, 2014; Ioannidis et al., 2017).

The trim-and-fill method not only detects asymmetry, but also adjusts the summary estimate by "filling"

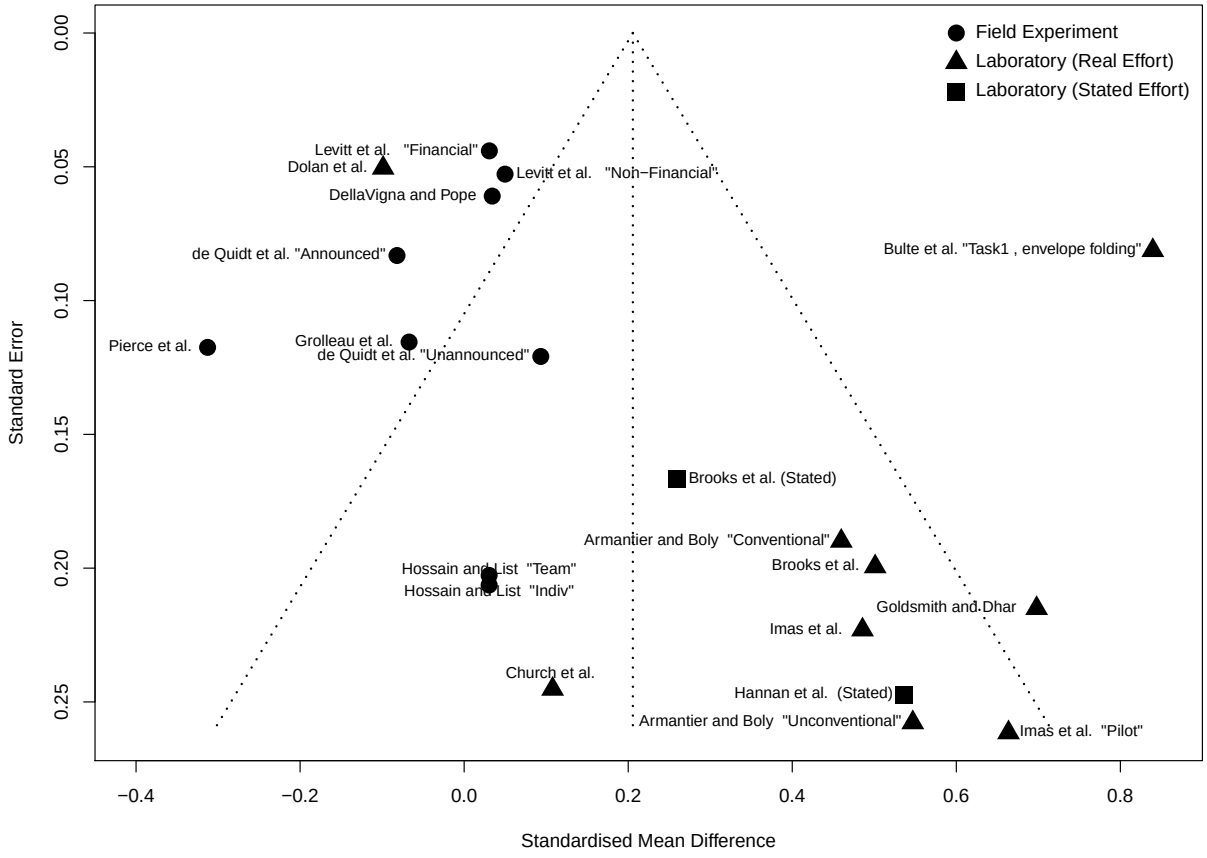


Figure 2: **Funnel plot of estimated effect sizes.** Estimated standardized effect sizes and their standard errors. The dotted vertical line is the summary estimated effect from loss-framed contracts.

in missing publications to generate symmetry (Figure S.2; see Supplement 2.1 for details). Note that in the presence of p-hacking, this method may under-adjust for publication bias (Simonsohn et al., 2014). After this adjustment, the estimated summary treatment effect shrinks over three-quarters and its confidence interval includes negative values: 0.02 SD (95% CI [-0.16, 0.19]) Even if we assume that working in the laboratory is fundamentally different from working in the field or that the workers in laboratory studies are different from workers in field studies - and thus the two types of experiments should not be pooled - the trim-and-fill adjustment applied only to the laboratory studies yields an summary effect size of 0.05 SD (95% CI [-0.22, 0.32])

These adjusted summary effect sizes are nearly identical to the summary estimated effect size for field experiments. To shed more light on why we observe this heterogeneity in effect sizes, and to explore preferences for contract types, we design a new experiment.

3 Experiment

3.1 Methods

Based on the results from the meta-analysis, we sought an experimental design with three features: (1) laboratory control combined with the realism of a real-effort task that included tradeoffs between working and taking breaks; (2) high statistical power; and (3) an opportunity for workers to express their preferences for framing under informed and incentivized conditions. We chose a piece-rate incentive contract because it matches real-world incentive contracts more closely, and because the meta-analysis provided weak evidence that such contracts may yield larger treatment effects (See Supplement S.1.2).

3.1.1 Opportunity costs for real effort

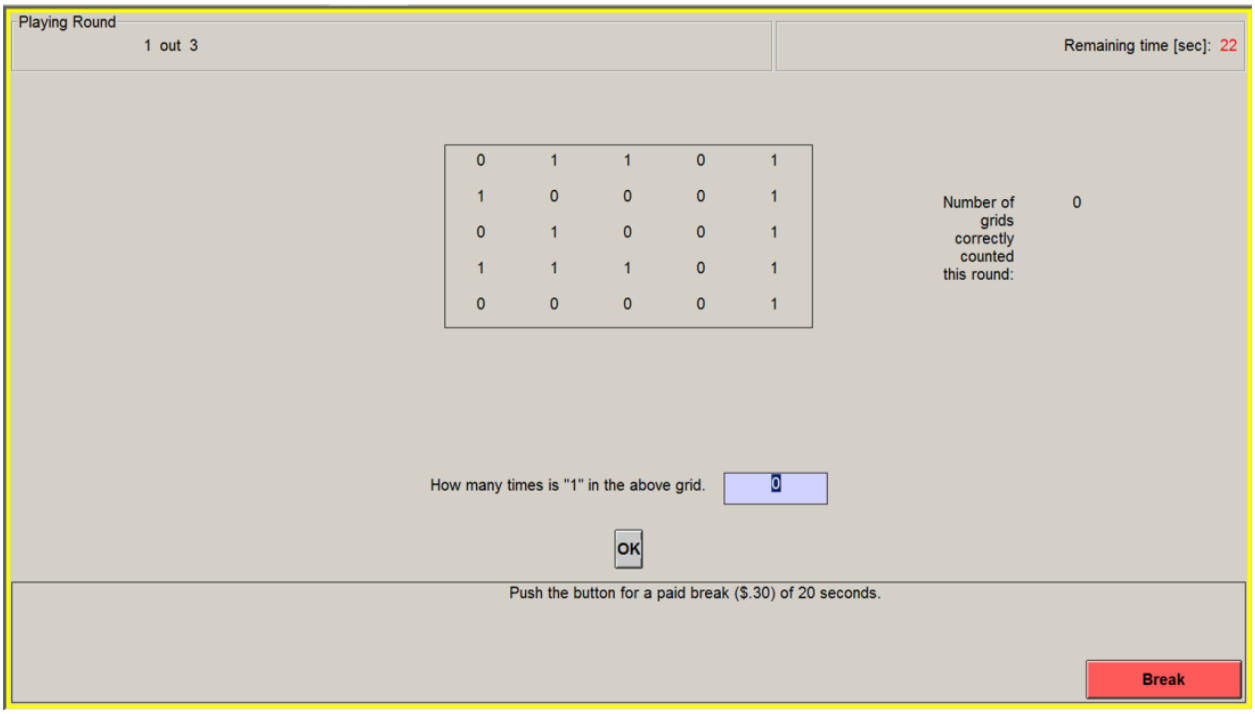


Figure 3: **Screen shot of effort task.**Workers were presented with 25 digits in a 5 x 5 grid. Each digit was either a "0" or a "1," chosen at random. Workers were rewarded for entering the correct number of 1s into a box on the screen, and clicking the "OK" button using a mouse. If the worker entered the correct number, the word "Correct" appeared in green at the top of the screen, the number of correctly answered grids displayed to the right of the grid increased by one, and the grid immediately refreshed. If the entry was incorrect, the word "Incorrect" appeared in red at the top of the screen, and the worker stayed on that grid until the correct number was entered.

Gächter et al. (2016) have questioned whether real-effort tasks actually measure effort. In prior studies, workers may have derived utility from the task itself or simply had no other option for activity during the experiment. The Eckartz (2014) design has two features that make it more likely that behavior in the task reflects effort. First, the task is tedious and thus workers are unlikely to derive utility from the task itself. Second, workers can opt to take paid breaks by pushing a button at the bottom of the screen, as shown in Figure 3. In our experiment, a break lasted 20 seconds and workers received USD $0.30 for each break. The break payment value was chosen to make cash payments easier, and the break length was chosen so that

taking a break was about a third to a half as profitable as we anticipated working to be. The grid, entry box and OK button disappeared from the screen for the duration of the break, so that workers could not work on the task while they were on break. The button to take a break was removed from the screen when there were fewer than 20 seconds left in the round to prevent workers from collecting the 30 cents without losing the 20 seconds of work time. Workers could take as many breaks as they wanted (they did take breaks; see supplementary materials). In Eckhart's study, offering breaks was shown to increase responsiveness to incentive contracts.

In the gain-framed contract, workers were informed they would be paid USD $0.25 for each correct grid, up to 100, at the end of the round. In the loss-framed contract, workers were paid USD $25.00 in cash at the beginning of the round, and informed that, for each grid they were short of 100, USD $0.25 would be collected from them at the end of the round. Immediate cash payments were used to increase saliency and because of a concern that paper losses, as opposed to real losses, could dilute loss aversion (Imas, 2016).

The experiment comprised one practice "round" of work to familiarize workers with the task, followed by three paid rounds (Figure 4). In the first round, workers worked under one of the contract frames, and in the second round they worked under the other contract frame (randomized order). Then, they were asked under which of the two contract frames they would prefer to work in the third round. Workers were also allowed to enter no preference, in which case they were randomly assigned a contract frame. We do not analyze this third round; it served to incentivize the revelation of workers' true preferences for contract framing.

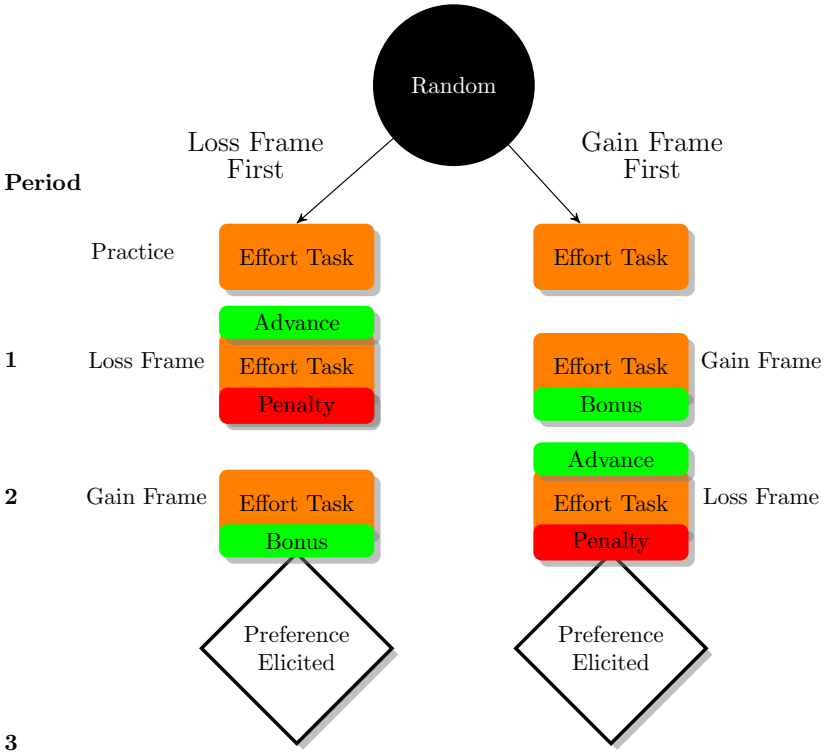


Figure 4: **Experimental design.** "Advance" refers to the payment provided in advance, prior to the effort task under a loss-framed contract. "Penalty" refers to the reduction in the advanced payment, based on performance. "Bonus" refers the end-of-round payment, based on performance.

3.1.2 Power analysis

We had no strong priors about the mean number of grids that workers could complete in the time allotted or its standard deviation, so we waited until the first three sessions were completed and used the data on the 33 subjects (dropping one worker who spent a whole round on break) to estimate the mean in the gain-framed contract (29.66), its standard deviation (5.74), and the intra-worker correlation between grids completed in the gain-framed contract and grids completed in the loss-framed contract (0.66). We used a simple analytical formula for power, with an adjustment for the within-workers design. We set power equal to 80%, the Type 1 error rate to 0.05, and the minimum detectable effect size equal to the lower bound of the 95% CI of the summary effect for the laboratory experiments in an earlier version of the meta-analysis that included fewer studies (0.16 SD). Those parameters yield a required sample size of 255 workers.

3.1.3 Other details

The experiment was conducted using z-Tree software (Fischbacher, 2007). The experiment consisted of one practice round, followed by three paid rounds (Figure 4), each of 4 minutes in length. After the practice rounds and ensuring that all workers understood the task, workers were informed that there would be three paid rounds and the payment scheme (contract) would be explained before each round. In the first two paid rounds, a worker worked one round under a gain-framed contract and the other round under a loss-framed contract. To control for learning effects, the contract order was randomized at the session level. Order was randomized at the session level, rather than at the worker level, because of concerns about within-session spillovers (interference) among workers: the upfront payment under the loss-framed contract might be observed by gain-framed contract workers and that observation of differential treatment across workers could have affected their productivity.

The experimental sessions were conducted in ExCen (the Experimental Economics Center) laboratory at Georgia State University. Each session lasted between 60 and 75 minutes. Workers earned $23.50 on average. In total, 268 undergraduate-student workers participated (see Table 3 for demographics). Cash was handled immediately before and after each round, a process that is labor-intensive. The experimental protocol required the ratio of staff-to-workers to be constant across sessions. Two sessions (16 workers) in early fall 2016 were run with fewer laboratory personnel than the experimental protocol required and thus were eliminated immediately from the study. If these sessions are included, the estimated effect of loss-framed contracting is reduced by 21% (0.7 rather than 0.9 additional grids; see Result 1). This research was approved by the Institutional Review Board of Georgia State University (Protocol: H16459). All participants gave informed consent of their participation in the study.

3.2 Results

Result 1. *The mean number of completed grids was 0.91 higher under the loss-framed contract (paired comparison of means 95% CI [0.18, 1.63]; Table 4)*, a difference that is statistically significant in a paired t-test ($p = 0.01$). Using a covariate-adjusted regression estimator to control for the starting frame and round (Table 6, column 1), the estimated effect is 0.89 (95% CI [.23, 1.55]. This estimated effect implies a standardized effect size of 0.12 SD, which is less than one-third of the summary estimated effect size from prior laboratory experiments (Figure 1) and well outside of its 95% CI. This small estimated treatment effect is not an artifact of the within-subject design: using only the first round data implies an effect size of 0.16 SD. This result suggests that the large difference between the estimated loss-framed contract effects in laboratory and field experiments is more likely an artifact of publication biases and underpowered designs, rather than the low external validity of laboratory experiments.

Note that, although loss aversion is the purported mechanism behind the estimated effects in the studies in the meta-analysis, we take no stand on whether or not our framing effect is driven by loss aversion or some other mechanism. For example, the loss-framed contract may induce greater effort through worker preferences for conformity or pro-sociality: the loss-framed contract communicates an expectation that a

Table 3: Demographic characteristics of workers

Variable		Proportion
Female		0.57
US Citizen		0.91
African		0.06
African-American		0.57
Asian		0.09
Asian-American		0.06
Hispanic/Latino		0.03
Middle-Eastern		0.01
Multiracial		0.06
Other		0.03
White		0.08
Age	Mean	20.92
	SD	3.56

Table 4: Productivity by Frame and Round

Frame	Round	Obs	Mean	SD	Min	Max
Gain Frame	All	268	29.26	7.29	0	50
Loss Frame	All	268	30.17	6.95	2	50

worker ought to achieve the benchmark performance level and workers seek to comply with that expectation (Brooks et al., 2012). Our results neither support nor refute the existence of loss aversion. The study was designed to assess the performance of loss-framed contracts.

Result 2. *When given a choice of working under the gain-framed contract or the loss-framed contract, two-thirds of workers expressed a preference for the gain-framed contract (Table 5).*

Table 5: Frame Preferences

	Preferred Frame			
	Indifferent	Gain	Loss	Total Observations
Number of workers	34	176	58	268
(% of sample)	(13%)	(66%)	(22%)	
Number of workers		176	58	234
after removing indifferent worker				
(% of sample)		(75%)	(25%)	

Result 3. *The productivity-enhancing effect from loss-framed contracts was concentrated in the subgroup of workers who expressed a preference for the loss-framed contract (Table 6, col. 2).* We report treatment effect estimates conditional on whether or not the worker expressed a preference for the loss-framed contract. In the first row is the estimated loss-framed contracting effect conditional on the worker expressing a preference for the gain-framed contract or indifference. The estimated effect is small, 0.18. In contrast, the estimated loss-framed contracting effect for workers who expressed a preference for a loss-framed contract is large (0.18 + 3.28= 3.46, 95% CI [1.98, 4.94]). As a robustness check, in column 3, we re-estimate the regression,

breaking out the small subgroup of workers who expressed indifference. The inferences are identical.

Table 6: Estimated effect of loss-framed contracts on grids completed

	(1) Impact of LF	(2) Impact by Preference	(3) Impact by Preference (w/ Indifference)
Loss Framed	0.89	0.18	0.03
	[0.23,1.55]	[-0.56,0.92]	[-0.83,0.89]
Prefer Loss Frame		-2.60	-2.71
		[-5.04,-0.16]	[-5.17,-0.24]
Prefer LF & Loss Framed		3.28	3.46
		[1.58,4.99]	[1.67,5.24]
Indifferent			-0.55
			[-3.02,1.93]
Indifferent & Loss Framed			0.91
			[-0.88,2.70]
Observations	536	536	536
Number of Subjects	268	268	268

95% CI in brackets, based on heteroskedastic-robust standard errors, clustered by worker. Tables S.1-S.3 in the supplementary materials present results with alternative clustering assumptions for the variance estimator. All regressions also included dummy variables for order effects (=1 if started in loss frame), and for round effects (=1 if second round), whose estimated coefficients are suppressed for clarity.

Results 2 and 3 suggests that an employer that exclusively offers loss-framed contracts in a market in which competitors offer gain-framed contracts would be at a disadvantage unless loss-framed-contract-preferring workers are more productive; in other words, unless employers could offer loss-framed contracts to screen for productive workers. The data, however, do not support a screening function for loss-framed contracts (Result 4).

Result 4. *The subgroup of workers who preferred the loss-framed contract are less productive, on average, in the gain-framed contract than other workers, and the loss-framed contract only served to increase their productivity to match the other workers' productivity.* In Table 6, column 2, the estimated coefficient in the second row reflects the productivity difference, in the gain-framed contract condition, between the 20% of workers who prefer the loss-framed contract and the other workers. The negative value (-2.60) implies that the loss-frame-preferring subgroup is less productive, on average.

Thus the output of the loss-frame-preferring workers under the loss-framed contract is similar to the output of the gain-framed contract preferring workers under the gain-framed contract: $0.18 - 2.60 + 3.28 = 0.86$ (95% CI [-1.23, 2.95]). In other words, the loss-framed contract appears to make the subgroup of loss-frame-preferring workers equally productive as the other workers.

Although the loss-framed-contract does not appear to serve as a screen for high-productivity workers in our context, it may serve as a commitment device for low-productivity workers (Imas et al., 2016). To exploit this commitment device, however, an employer would have to offer two types of framing simultaneously within its organization and let workers select their preferred framing, or separate its operations into two units, each offering a different framing. In many contexts, the organizational complexities and fixed costs implied by such strategies could easily dominate the modest performance benefits from loss-framed contracts.

4 Discussion

We sought to understand preferences for contract framing for two reasons. First, heterogeneity of preferences might shed light on the paucity of explicitly (written) loss-framed contracts in the field. Second, the literature has yielded ambiguous conclusions about worker preferences for contract framing. An older literature suggests that people are hesitant to apply, or work under, contracts that implicitly label people as low performers (Baker et al., 1988) and penalize low performance, rather than reward high performance (Lazear, 1991). In the loss-aversion context, Imas et al. (2016, p. 1271) write that "[s]tandard behavioral models predict a tradeoff in the use of loss contracts: employees will work harder under loss contracts than under gain contracts; but, anticipating loss aversion, they will prefer gain contracts to loss contracts."

Despite these predictions, however, the only two studies with payoff-equivalent contracts and incentivized elicitations of preferences for the contracts conclude that people prefer loss-framed contracts (Imas et al., 2016; de Quidt, 2018).

Imas et al. (2016) first ran a between-workers design (N=83) in which workers completed a slider task under a loss-framed contract or a gain-framed contract. The reward was a t-shirt. Performance was either sufficient to keep/receive the t-shirt or it was not. The authors estimate that the loss-framed contract increased effort by about 0.5 SD. In a second experiment with different workers and a between-workers design (N=85), workers were endowed with \$5, and then presented with the opportunity to participate in the slider task for a t-shirt reward. Workers were offered either a gain-framed or a loss-framed contract, and then participated in the equivalent of a random-price auction in which they expressed their willingness-to-pay (WTP) to work for the t-shirt. If the random price was greater than their stated WTP, the workers waited for others to complete the task. If the price was less than or equal to their stated WTP, the worker engaged in the task. Imas et al. report that the average WTP was \$2.54 for the loss-framed contract and \$1.76 for the gain-framed contract. They conclude, "[s]urprisingly, rather than a preference for the gain contract, we find that people actually prefer loss contracts." (2016, p. 1272). To explain this apparent preference for loss-framed contracts, the authors hypothesize that loss-framed contracts serves as a commitment device.

Although a higher average WTP for the loss-framed contract is one metric of group preferences, it may be sensitive to outliers in small sample sizes and it does not reveal whether more people prefer the loss-framed contract or the game-framed contract. If the margin by which loss-frame-preferring people are willing to pay more for their preferred contract is larger than the margin by which gain-frame-preferring people are willing to pay more for their preferred contract, then it is possible for $Mean(WTP_{LF}) > Mean(WTP_{GF})$ even if most people prefer the gain-framed contract (see more details in supplementary materials). In our design, workers chose directly between a loss-framed and gain-framed contract, and thus there is no ambiguity about the proportion of people who prefer each contract type.

In contrast to Imas et al. (2016), de Quidt (2018) elicits contract choice rather than WTP. Six-hundred and eighty-seven mTurk workers are randomized into two groups. One group chooses between working on a task under a loss-framed contract or doing their next-best alternative activity, which is unknown to the experimenter The other group chooses between working on the same task under a gain-framed contract or doing their next-best alternative activity. The acceptance rate was higher, by 11 percentage points, for the loss-framed contract. Based on that experiment, and some variations in which the salience of the framing is varied, the author concludes (p.523) that there is "no evidence of the predicted distaste for penalties [loss-framed contracts]."

Yet in both the Imas et al. and de Quidt studies, workers express their preferences for contracts under an information asymmetry: they were likely familiar with a gain-framed contract, but they may never have seen, much less worked under, a loss-framed contract. Their choices may thus have included a value for experiencing (sampling) a novel form of incentive contract. We believe that, prior to expressing a preference, workers should be familiar with both contracts, and then given an opportunity to choose which contract they wish to work under in subsequent rounds of the same task. We achieve that level playing field with our design, in which workers experience both contracts prior to choosing directly between a loss-framed and gain-framed contract (and the order in which they experienced the contracts is randomized). Thus, there is no ambiguity about the proportion of people who prefer each contract type.

In sum, based on a meta-analysis and a new experiment, we conclude that the productivity effects of loss-framed incentive contracts may not be as large as they appear in the behavioral science literature. Moreover, in our experiment, the productivity-enhancing effect of loss-framed contracts is concentrated among a minority of workers (∼20%) who prefer loss-framed contracts over gain-framed contracts. These workers, on average, are less productive than workers who either prefer gain-framed contracts or are indifferent between the two contracts. Organizations that only offer loss-framed incentive contracts would experience higher costs of recruitment and replacement than similar organizations offering gain-framed contracts, with no countervailing increase in average worker productivity. These organizations could offer both types of contracts, and let workers choose their preferred contract, but the costs of administering two parallel payment systems might outweigh the benefits.

In addition to attempting to replicate our results, future studies should explore whether these patterns are observed outside of high-income nations - most of the studies in our meta-analysis come from the U.S. Moreover, our experiment and all the studies in our meta-analysis have very short time horizons in which worker output was observed. Most of the studies observed behavior for fewer than two hours, and the longest period was four weeks (Hossain and List, 2012). Longer time horizons are needed to determine if the modest effects of loss-framed contracts decline (or grow) with repetition over longer horizons. For example, in one study, loss aversion could not be detected in contexts in which workers experienced repeated losses and gains (Erev et al., 2008). Should the loss-framed contract productivity impacts be ephemeral, the potential gains from field applications of loss-framed contracts in naturally occurring environments would be even smaller.

References

Armantier O, Boly A (2015) Framing of Incentives and Effort Provision. International Economic Review 56(3):917, DOI 10.1111/iere.12126

Baker GP, Jensen MC, Murphy KJ (1988) Compensation and Incentives: Practice vs. Theory. Journal of Finance 43(3):593–616

Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature News 533(7604):452

Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. Biometrics pp 1088–1101

Brooks RR, Stremitzer A, Tontrup S (2012) Framing contracts: Why loss framing increases effort. Journal of Institutional and Theoretical Economics JITE 168(1):62–82, URL http://www.ingentaconnect.com/content/mohr/jite/2012/00000168/00000001/art00008

Brooks RRW, Stremitzer A, Tontrup S (2017) Stretch It but Don't Break It: The Hidden Cost of Contract Framing. The Journal of Legal Studies 46(2):399–426, DOI 10.1086/694234, URL http://www.journals.uchicago.edu/doi/10.1086/694234

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14(5):365

Chrisman JJ, Patel PC (2012) Variations in R&D investments of family and nonfamily firms: Behavioral agency and myopic loss aversion perspectives. Academy of management Journal 55(4):976–997

Church BK, Libby T, Zhang P (2008) Contracting frame and individual behavior: Experimental evidence. Journal of Management Accounting Research 20(1):153–168, URL http://aaajournals.org/doi/abs/10.2308/jmar.2008.20.1.153

Convery F, McDonnell S, Ferreira S (2007) The most popular tax in Europe? Lessons from the Irish plastic bags levy. Environmental and Resource Economics 38(1):1–11, URL http://link.springer.com/article/10.1007/s10640-006-9059-2

DellaVigna S, Pope D (2018) What Motivates Effort? Evidence and Expert Forecasts. The Review of Economic Studies 85(2):1029–1069, DOI 10.1093/restud/rdx033, URL https://academic.oup.com/restud/article/85/2/1029/3861288

Dolan P, Metcalfe R, Navarro-Martinez D (2012) Financial incentives and working in the education sector. Department for Education Research Report DFE–RR251 URL https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR251.pdf

Duval S, Tweedie R (2000a) A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association 95(449):89–98

Duval S, Tweedie R (2000b) Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56(2):455–463

Eckartz K (2014) Task enjoyment and opportunity costs in the lab: The effect of financial incentives on performance in real effort tasks. Tech. rep., Jena Economic Research Papers, URL http://www.econstor.eu/handle/10419/98451

Erev I, Ert E, Yechiam E (2008) Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. Journal of Behavioral Decision Making 21(5):575–597

Evers ERK, Inbar Y, Blanken I, Oosterwijk LD (2017) When Do People Prefer Carrots to Sticks? A Robust "Matching Effect" in Policy Evaluation. Management Science 63(12):4261–4276, DOI 10.1287/mnsc.2016.2539, URL https://pubsonline-informs-org.libproxy.chapman.edu/doi/abs/10.1287/mnsc.2016.2539

Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. Experimental economics 10(2):171–178, URL http://link.springer.com/article/10.1007/s10683-006-9159-4

Galizzi MM, Navarro-Martinez D (2018) On the External Validity of Social Preference Games: A Systematic Lab-Field Study. Management Science 65(3):976–1002, DOI 10.1287/mnsc.2017.2908, URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2017.2908

Gächter S, Huang L, Sefton M (2016) Combining "real effort" with induced effort costs: the ball-catching task. Experimental Economics 19(4):687–712, DOI 10.1007/s10683-015-9465-9, URL https://doi.org/10.1007/s10683-015-9465-9

Gelman A, Carlin J (2014) Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science 9(6):641–651

Goldsmith K, Dhar R (2011) Incentives Framing and Task Motivation: The Intuitive Appeal of Gains and the Actual Efficacy of Losses. SSRN Working Paper Series URL https://www.researchgate.net/profile/Ravi_Dhar2/publication/228225386_Incentive_Framing_and_Task_Motivation_The_Intuitive_Appeal_of_Gains_and_the_Actual_Efficacy_of_Losses/links/55226c6a0cf2f9c13052bc2b.pdf

Grolleau G, Kocher MG, Sutan A (2016) Cheating and loss aversion: Do people cheat more to avoid a loss? Management Science 62(12):3428–3438

Haigh MS, List JA (2005) Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis. The Journal of Finance 60(1):523–534, DOI 10.1111/j.1540-6261.2005.00737.x, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2005.00737.x/abstract

Hannan RL, Hoffman VB, Moser DV (2005) Bonus versus penalty: does contract frame affect employee effort? In: Experimental business research, Springer, pp 151–169, URL http://link.springer.com/chapter/10.1007/0-387-24243-0_8

Hardie BGS, Johnson EJ, Fader PS (1993) Modeling Loss Aversion and Reference Dependence Effects on Brand Choice. Marketing Science 12(4):378–394, DOI 10.1287/mksc.12.4.378, URL https://pubsonline.informs.org/doi/abs/10.1287/mksc.12.4.378

Hedges LV (1981) Distribution theory for Glass's estimator of effect size and related estimators. journal of Educational Statistics 6(2):107–128

Hedges LV (1982) Estimation of effect size from a series of independent experiments. Psychological bulletin 92(2):490

Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21(11):1539–1558, DOI 10.1002/sim.1186, URL http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1186, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1186

Homonoff TA (2018) Can Small Incentives Have Large Effects? The Impact of Taxes versus Bonuses on Disposable Bag Use. American Economic Journal: Economic Policy 10(4):177–210, DOI 10.1257/pol.20150261, URL https://www.aeaweb.org/articles?id=10.1257/pol.20150261

Hossain T, List JA (2012) The behavioralist visits the factory: Increasing productivity using simple framing manipulations. Management Science 58(12):2151–2167, URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1544

Imas A (2016) The realization effect: Risk-taking after realized versus paper losses. The American Economic Review 106(8):2086–2109, URL http://www.ingentaconnect.com/contentone/aea/aer/2016/00000106/00000008/art00005

Imas A, Sadoff S, Samek A (2016) Do people anticipate loss aversion? Management Science 63(5):1271–1284

Ioannidis JP (2005) Why most published research findings are false. PLoS medicine 2(8):e124

Ioannidis JP, Stanley TD, Doucouliagos H (2017) The power of bias in economics research. Oxford University Press Oxford, UK

Jakovcevic A, Steg L, Mazzeo N, Caballero R, Franco P, Putrino N, Favara J (2014) Charges for plastic bags: Motivational and behavioral effects. Journal of Environmental Psychology 40:372–380, DOI 10.1016/j.jenvp.2014.09.004, URL http://linkinghub.elsevier.com/retrieve/pii/S0272494414000863

Jarrow R, Zhao F (2006) Downside loss aversion and portfolio management. Management Science 52(4):558–566

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society pp 263–291, URL http://www.jstor.org/stable/1914185

Lazear EP (1991) Labor Economics and the Psychology of Organizations. The Journal of Economic Perspectives 5(2):89

Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? The Journal of Economic Perspectives 21(2):153–174, URL http://www.jstor.org/stable/10.2307/30033722

Levitt SD, List JA, Neckermann S, Sadoff S (2016) The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. American Economic Journal: Economic Policy 8(4):183–219, DOI 10.1257/pol.20130358, URL https://www-aeaweb-org.libproxy.chapman.edu/articles?id=10.1257/pol.20130358

van der Linden S, Goldberg MH (2020) Alternative meta-analysis of behavioral interventions to promote action on climate change yields different conclusions. Nature Communications 11(1):3915, DOI 10.1038/s41467-020-17613-7, URL https://www.nature.com/articles/s41467-020-17613-7, number: 1 Publisher: Nature Publishing Group

List JA, Samek AS (2015) The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption. Journal of Health Economics 39:135–146, DOI 10.1016/j.jhealeco.2014.11.002, URL http://www.sciencedirect.com/science/article/pii/S0167629614001398

Looney CA, Hardin AM (2009) Decision support for retirement portfolio management: Overcoming myopic loss aversion via technology design. Management Science 55(10):1688–1703

Luft J (1994) Bonus and penalty incentives contract choice by employees. Journal of Accounting and Economics 18(2):181–206, DOI 10.1016/0165-4101(94)00361-0, URL http://www.sciencedirect.com/science/article/pii/0165410194003610

Maniadis Z, Tufano F, List JA (2014) One swallow doesn't make a summer: New evidence on anchoring effects. The American Economic Review 104(1):277–290, URL http://www.ingentaconnect.com/content/aea/aer/2014/00000104/00000001/art00010

Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G (2014) Promoting transparency in social science research. Science 343(6166):30–31

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716

de Quidt J (2018) Your Loss Is My Gain: A Recruitment Experiment with Framed Incentives. Journal of the European Economic Association 16(2):522–559, DOI 10.1093/jeea/jvx016, URL http://academic.oup.com/jeea/article/16/2/522/3860644

de Quidt J, Fallucchi F, Kölle F, Nosenzo D, Quercia S (2017) Bonus versus penalty: How robust are the effects of contract framing? Journal of the Economic Science Association 3(2):174–182, DOI 10.1007/s40881-017-0039-9, URL https://doi.org/10.1007/s40881-017-0039-9

Raudenbush SW, Bryk AS (1985) Empirical Bayes Meta-Analysis. Journal of Educational Statistics 10(2):75–98, DOI 10.2307/1164836, URL http://www.jstor.org/stable/1164836, publisher: [Sage Publications, Inc., American Educational Research Association, American Statistical Association]

Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychological bulletin 86(3):638

Schwarzer G, Carpenter JR, Rücker G (2015) Meta-analysis with R, vol 4724. Springer

Simmons JP, Nelson LD, Simonsohn U (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22(11):1359–1366, DOI 10.1177/0956797611417632, URL https://doi.org/10.1177/0956797611417632

Simonsohn U, Nelson LD, Simmons JP (2014) p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspectives on Psychological Science 9(6):666–681, DOI 10.1177/1745691614553988, URL http://journals.sagepub.com/doi/10.1177/1745691614553988

Tannenbaum D, Valasek CJ, Knowles ED, Ditto PH (2013) Incentivizing wellness in the workplace: Sticks (not carrots) send stigmatizing signals. Psychological science 24(8):1512–1522

Thaler RH, Johnson EJ (1990) Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. Management science 36(6):643–660, URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.36.6.643

Thaler RH, Sunstein CR (2008) Nudge : improving decisions about health, wealth, and happiness. New Haven : Yale University Press

Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in medicine 18(20):2693–2708

Supplementary Materials

S.1 Meta-analysis

In Figure S.1, we present the forest plot from Figure 1 alongside the data that underlie the information displayed. Both figures were produced using the R package *meta* (Schwarzer et al., 2015).

Study	Loss Frame Obs	Loss Frame Mean	Loss Frame SD	Gain Frame Obs	Gain Frame Mean	Gain Frame SD	SMD	95%−CI	Weight
Field									
de Quidt et al. "Announced"	287	38.00	12.40	292	39.00	12.00	−0.082	[−0.245; 0.081]	6.1%
de Quidt et al. "Unannounced"	137	41.00	11.50	137	40.00	9.78	0.093	[−0.144; 0.330]	5.6%
DellaVigna and Pope	532	2155.00	532.57	545	2136.00	575.69	0.034	[−0.085; 0.154]	6.3%
Hossain and List "Indiv"	53	4.37	0.37	45	4.36	0.36	0.030	[−0.367; 0.428]	4.4%
Hossain and List "Team"	47	6.47	0.34	47	6.46	0.34	0.030	[−0.374; 0.435]	4.3%
Levitt et al. "Financial"	948	0.10	0.60	1130	0.08	0.59	0.031	[−0.056; 0.117]	6.4%
Levitt et al. "Non−Financial"	780	0.06	0.61	667	0.03	0.54	0.050	[−0.054; 0.153]	6.4%
Pierce et al.	140	0.91	6.60	154	3.23	8.01	−0.313	[−0.543; −0.082]	5.6%
Random effects model	**2924**			**3017**			**−0.004**	**[−0.081; 0.074]**	**45.1%**
Heterogeneity: I^2 = 31%, τ^2 = 0.0046, p = 0.18									
Lab									
Armantier and Boly "Conventional"	56	82.31	7.80	58	78.68	7.88	0.460	[0.088; 0.832]	4.5%
Armantier and Boly "Unconventional"	34	83.73	6.28	29	80.59	4.82	0.547	[0.042; 1.052]	3.6%
Brooks et al.	54	14.30	8.10	50	10.40	7.30	0.501	[0.110; 0.892]	4.4%
Brooks et al. (Stated)	73	1.86	1.16	72	1.58	0.97	0.260	[−0.067; 0.587]	4.9%
Bulte et al. "Task1 , envelope folding"	800	27.63	8.78	200	20.57	6.71	0.840	[0.680; 0.999]	6.1%
Church et al.	31	1.14	0.17	36	1.11	0.34	0.108	[−0.373; 0.588]	3.8%
Dolan et al.	789	508.62	228.24	784	530.74	219.38	−0.099	[−0.198; 0.000]	6.4%
Goldsmith and Dhar	46	8.02	3.49	46	5.65	3.24	0.698	[0.276; 1.119]	4.2%
Grolleau et al.	150	3.45	2.27	150	3.61	2.48	−0.067	[−0.294; 0.159]	5.7%
Hannan et al. (Stated)	33	9.58	3.31	35	7.40	4.58	0.537	[0.052; 1.021]	3.7%
Imas et al.	43	17.28	3.33	40	15.35	4.50	0.486	[0.048; 0.923]	4.1%
Imas et al. "Pilot"	30	15.27	4.44	32	11.88	5.55	0.664	[0.151; 1.176]	3.6%
Random effects model	**2139**			**1532**			**0.386**	**[0.202; 0.570]**	**54.9%**
Heterogeneity: I^2 = 91%, τ^2 = 0.0712, p < 0.01									
Random effects model	**5063**			**4549**			**0.206**	**[0.064; 0.348]**	**100.0%**
Heterogeneity: I^2 = 87%, τ^2 = 0.0796, p < 0.01									
Test for subgroup differences: χ^2_1 = 14.61, df = 1 (p < 0.01)									

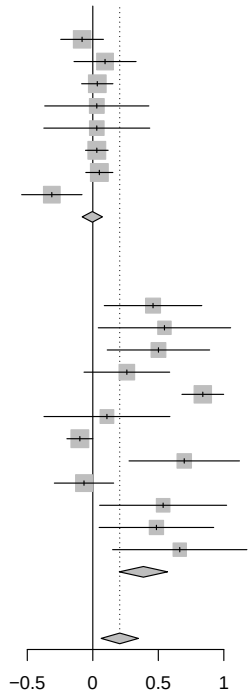


Figure S.1: **Meta-analysis of experimental studies estimating the effect of loss-framed contracts on productivity (effort).** For each study, the figure reports the number of observations (Obs), the mean value of the outcome measure in units reported in the experiment (Mean), and the standard deviation of the outcome measure (SD) for the loss-framed-contract and gain-framed-contract groups. It also reports the standardized mean difference (SMD) between the outcomes in the loss-framed-contract and gain-framed-contract groups (standardized by the pooled standard deviation) and its 95% CI. The final column reports the weight that each study contributes to the summary estimated treatment effect.

S.1.1 Tests of asymmetry in funnel plots

The Begg-Mazumdar method tests if the rank according to effect size is correlated with the rank according to standard error size (Begg and Mazumdar, 1994). The test statistic, z, is the difference in pairs with positive correlation and the pairs of with negative correlation, adjusted for number of studies. In expectation, it is mean zero. The probability of the calculated value in a standard normal distribution is the test's $p - value$. For the studies in the meta-analysis $z = 2.47, p - value = 0.01$.

The Thompson-Sharp test regresses each study's z-scored SMD on the inverse of its standard error (Thompson and Sharp, 1999). It assumes that studies with fewer observations and smaller inverse standard errors should not systematically have larger standardized effects, and thus the regression estimate of the constant term should be zero. The test statistic of the constant term is calculated as it normally is in a regression–via a t-distribution. The estimated constant for our studies is $t = 4.33, p - value = 0.0007$.

The trim-and-fill technique uses an iterative method that first trims the studies that lie outside the expected error range. Then the "true" center of the funnel is then estimated from the remaining subset. Then, for each trimmed study, a "fill" study is added to the full set of studies. The filled study is the same distance from the true center as the trimmed study, but on the opposite side of the funnel. The final step is to asses if symmetry has been achieved; if it has not, there is another iteration. By default, meta uses the "L" method (Duval and Tweedie, 2000a,b), which employs the fixed-effects model estimator of the true effect, and uses a rank statistic to trim the asymmetric studies. It then uses estimate random-effects model to assess if symmetry has been achieved.

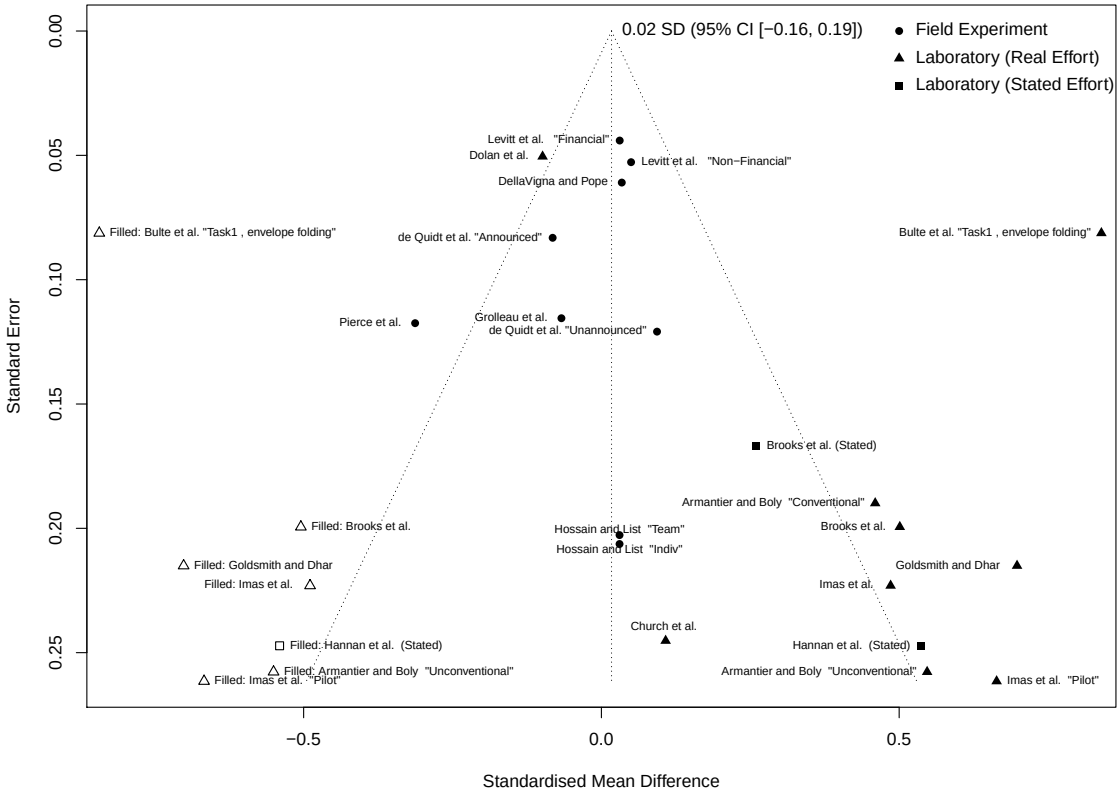


Figure S.2: **Funnel plot after studies are trimmed and filled.** Estimated standardized effect sizes and their standard errors (black shapes) plus counterfactual studies (white shapes) that are added by a "trim-and-fill" approach to generate a more symmetric funnel. The dotted vertical line is the revised summary estimated effect from loss-framed contracts.

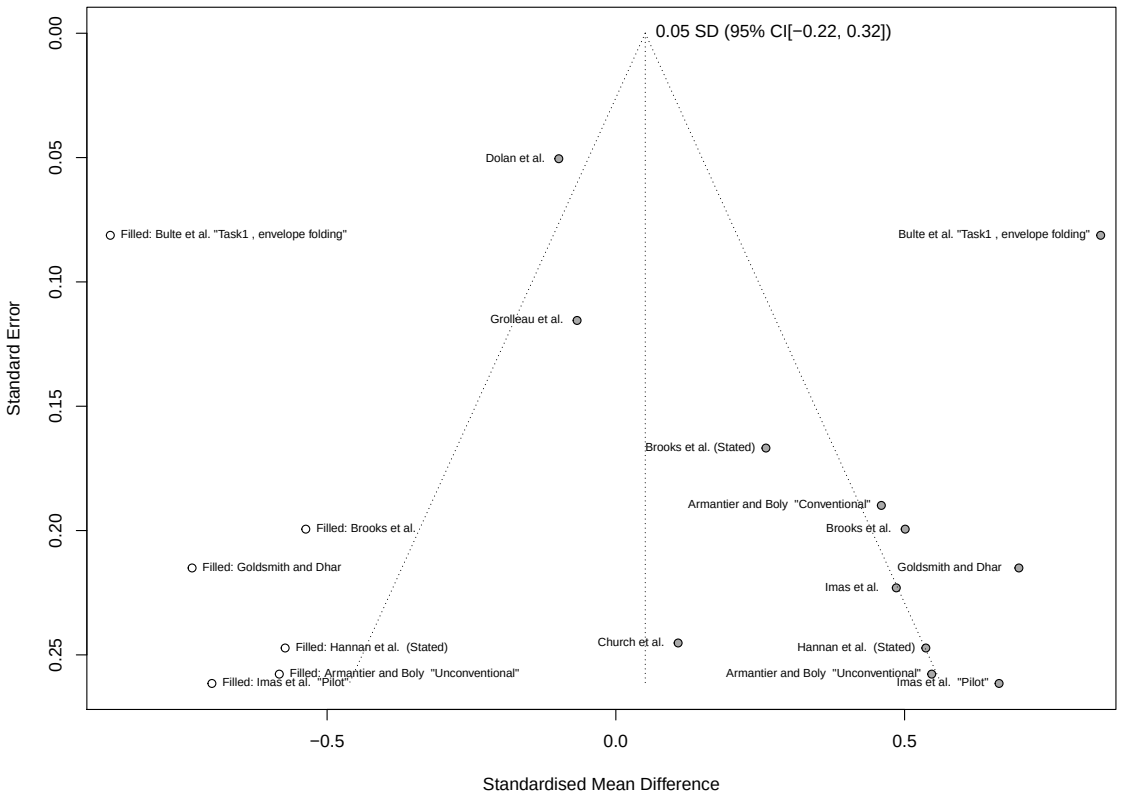


Figure S.3: **Funnel plot after laboratory studies are trimmed and filled.** Estimated standardized effect sizes and their standard errors (black shapes) plus counterfactual studies (white shapes) that are added by a "trim-and-fill" approach to generate a more symmetric funnel. The dotted vertical line is the revised summary estimated effect from loss-framed contracts.

S.1.2 Piece-rate versus threshold

The effect of loss-framed contracting can be impacted by the details of the contract. Specifically, contracts can state either a single "threshold" or a "piece-rate". In the former, unless a worker achieves a threshold, she incurs a penalty. In the latter, the worker is penalized the piece-rate for every unit her output is under a target quota. For instance, a threshold contract might state that unless a worker produces 20 units she will be penalized \$10; whereas, a piece-rate contract might state that the worker is penalized \$1 for every unit she falls short of producing 20. A worker who produced 15 units would be penalized \$10 under the threshold contract, but only \$5 under the piece-rate contract.

Figure S.4 presents a forest plot of the same studies as Figure S.1, except subgroups are divided by contract type (piece-rate or threshold), rather than being divided by setting as in Figure S.1. All of the piece-rate estimates, are laboratory experiments (one is laboratory in field). For piece-rate contracts, the summary estimated effect size is 0.29 SD (95% CI [0.01, 0.57]). For threshold contracts, the summary estimated effect size is only half the size: 0.17 SD (95% CI [0.01, 0.34]).

Designing the right threshold contract to motivate workers is difficult. To illustrate this difficulty, we use the example from the first paragraph above, where the worker was producing 15 units, and suppose she can increase her output a maximum of 3 more units. She would not be motivated by the threshold contract, but she would be motivated by the piece-rate contract; even producing 18 units, she would receive the same \$10 penalty under the threshold contract, but her penalty would be reduced from \$5 to \$2, under piece-rate. Thus, she would choose not to produce the additional 3 units under threshold, but would under piece-rate.

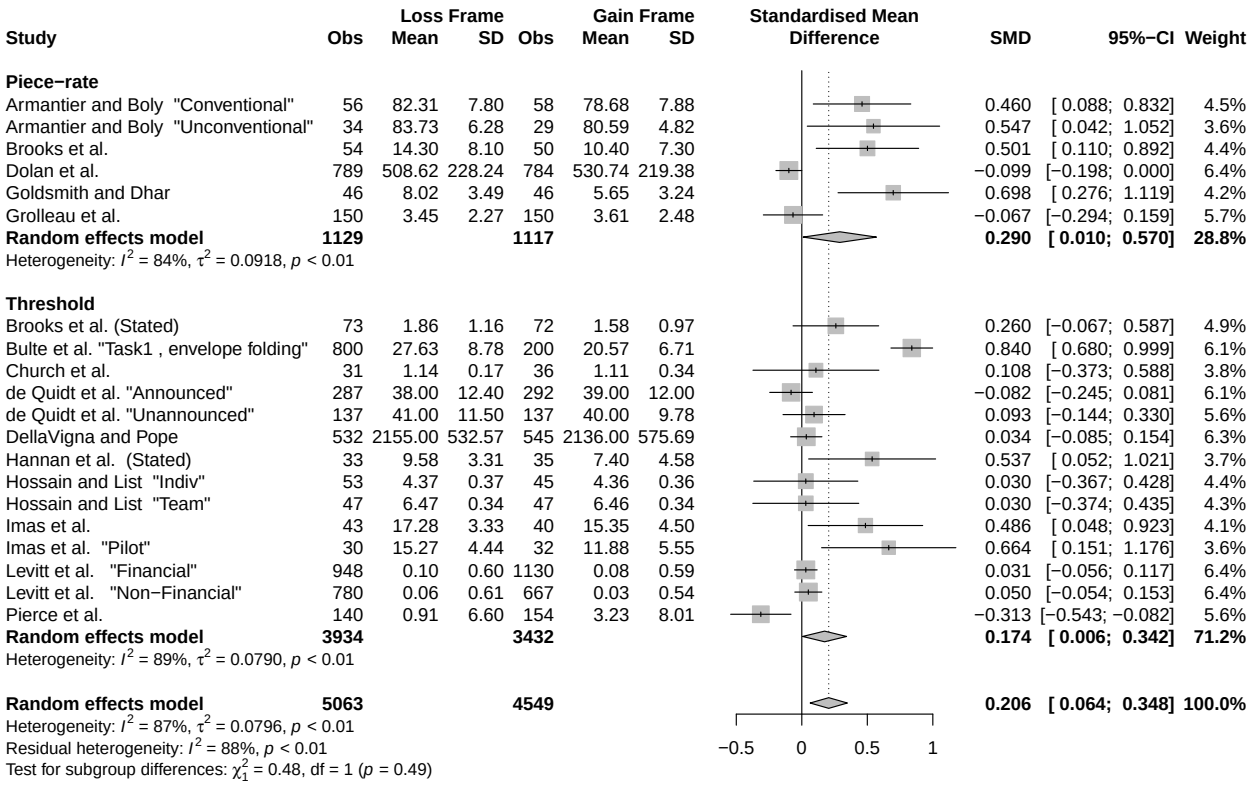


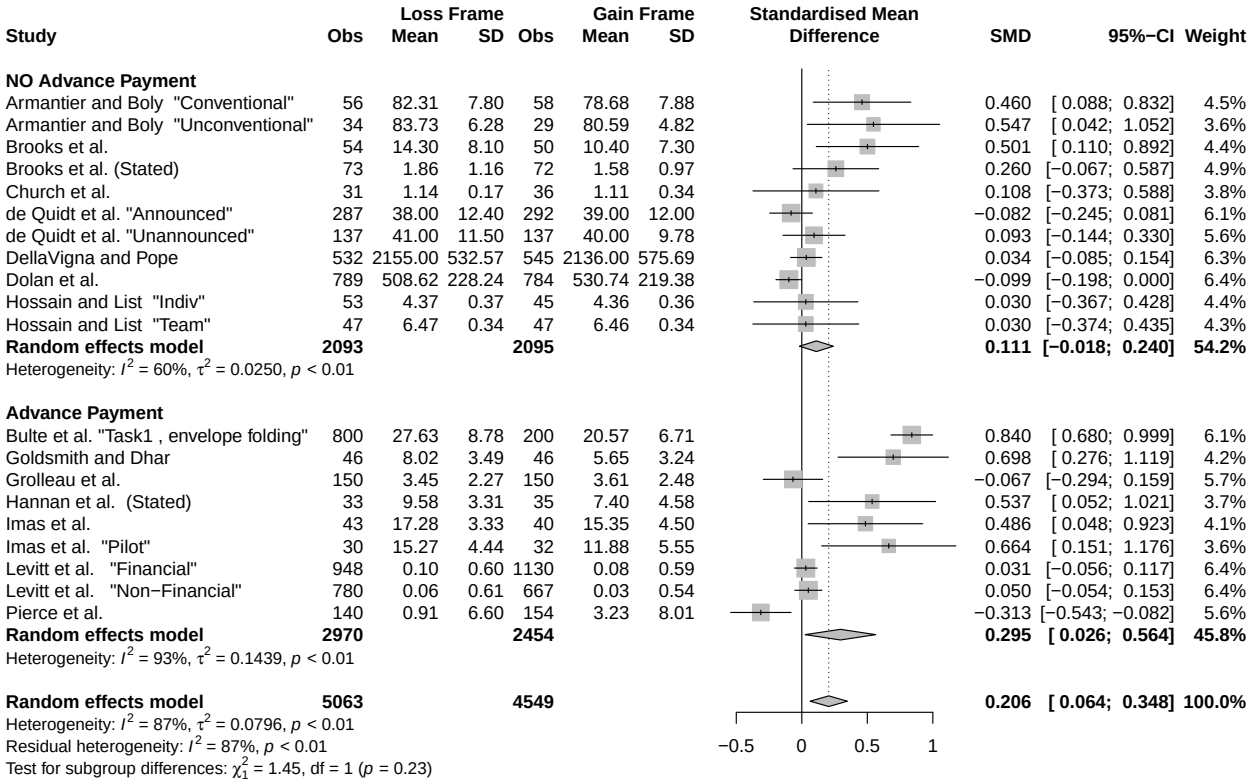
Figure S.4: Meta-analysis of experimental studies of loss-framed contracts grouped by piece-rate versus threshold designs

To motivate this worker, the threshold could be set to 18, rather than 20. Then the threshold contract would more strongly motivate her than the piece-rate contract ($10 is more than $3). However, if worker output is heterogeneous, a threshold that motivated this worker might be too high or low for her peers. A coworker whose natural output is only 12 might only be able to increase his output to 14, so will not be motivated by that threshold. Another coworker, who naturally produces 18, might have increased output to 20 under piece-rate or had the threshold been 20, might only produce 18, if the threshold were reduced. Brooks et al.Brooks et al. (2017) showed that even under piece-rate the quota may be too high or low to be effective, but piece-rate seems more robust to heterogeneous abilities and misestimation of baseline productivity. Given these design challenges, and the lack of evidence from our meta-analysis that the threshold design induces greater effort (in fact, the estimated overall effect is smaller in the threshold designs), we use the piece-rate design.

S.1.3 Advance payment

Table S.5 tests for difference between experiments in which the workers received the reward (payment) in advance versus studies in which they were merely told they would get the reward. While the estimated effect size is larger when the workers get the reward in advance 0.29 SD (95% CI [0.03, 0.56]) than when the do not 0.11 SD (95% CI [-0.02, 0.24]), the differnce is not statistically significant ($\chi^2 = 1.45$, $df = 1$, $p = 0.23$).

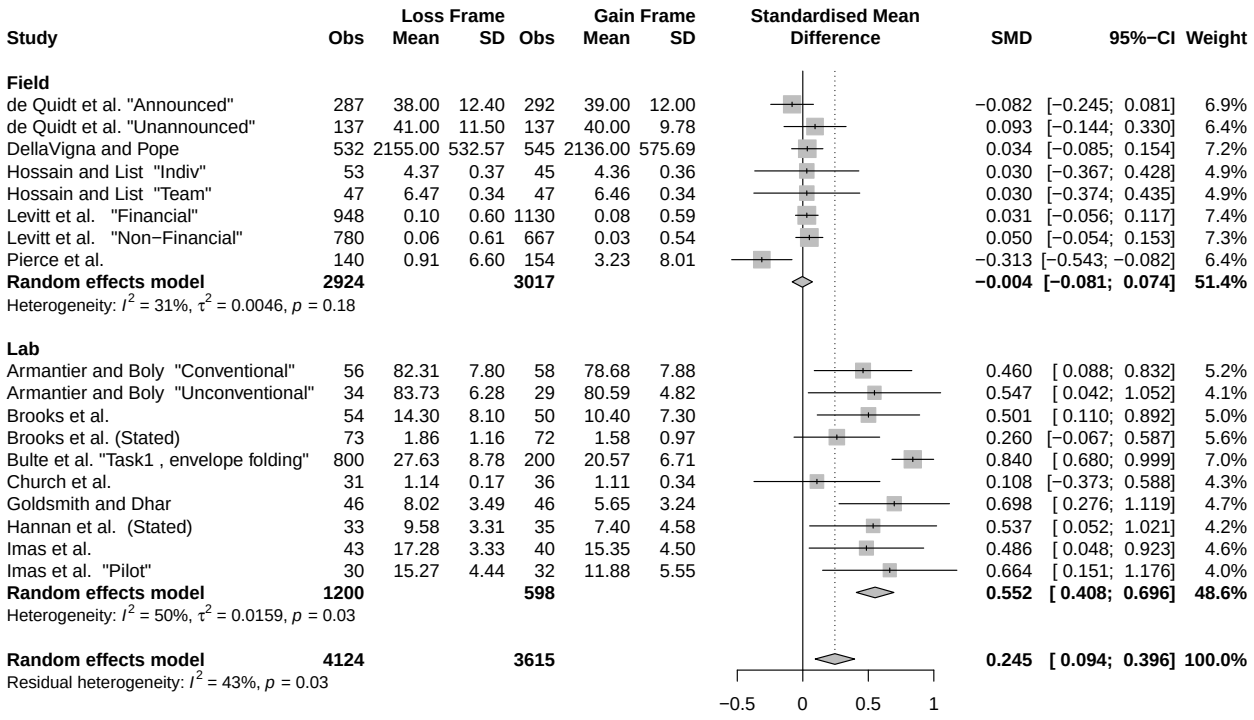
Figure S.5: Meta-analysis of experimental studies of loss-framed contracts grouped by whether subject received payment in advance



S.1.4 Limiting to studies focusing on loss framing effect on effort

To estimate the effect size of the literature scholars typical associate with loss-framed contracts, we remove two studies from our dataset and present the results of that meta-analysis in Figure S.6. We remove Dolan et al. (2012), because their experiment was in a government report but loss-framing was not the focus of the report. We remove Grolleau et al. (2016) because the central claim of the paper is that loss-framed contracts make cheating more likely, not that the authors fail to detect an effect of loss-framed contracts on effort.

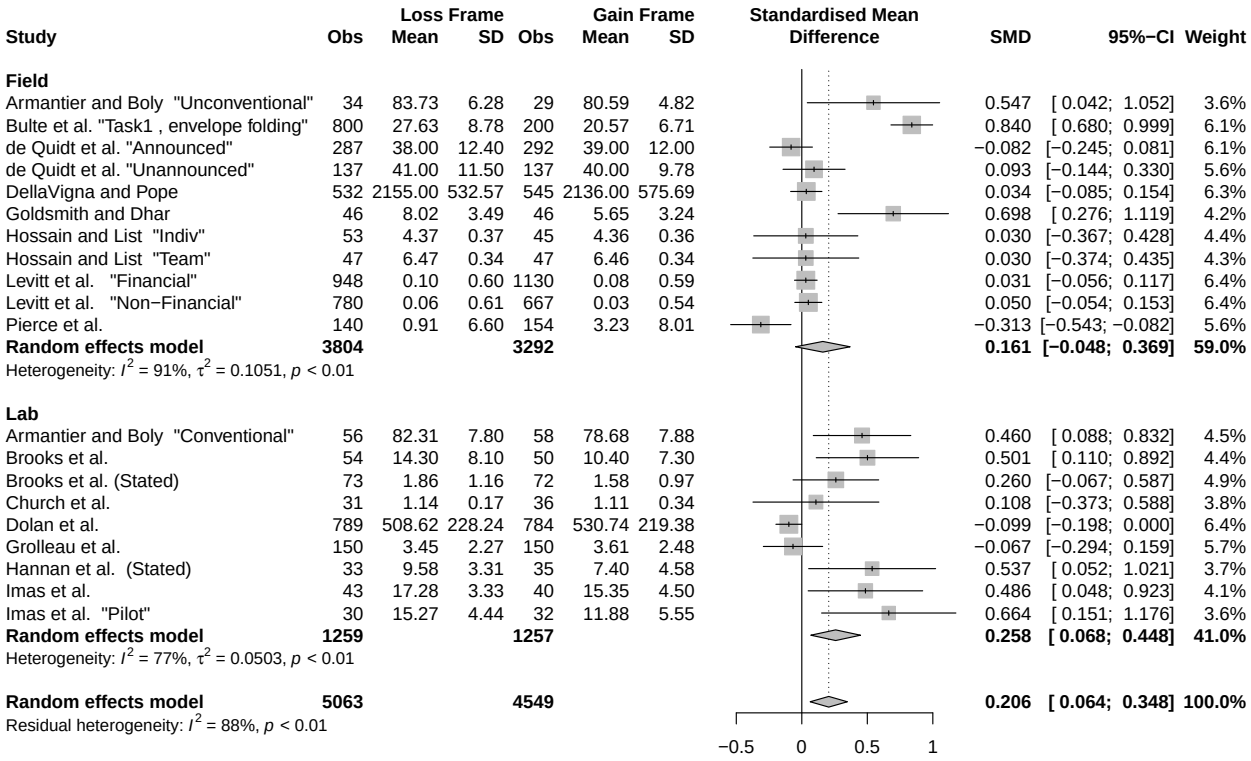
Figure S.6: Meta-analysis of studies focusing on loss framing effect on effort



S.1.5 Re-classifying "lab-in-the-field" experiments as "field experiments"

Because some scholars classify laboratory experiments conducted with non-standard subjects as "artefactual" field experiments, we re-do the meta-analysis after re-classifying these lab-in-the-field experiments as "field experiments" and present the results in Figure S.7. Because these experiments have some of the largest effect sizes, the gap between the summary effect sizes for laboratory and field experiment decreases substantially, but field experiments still have a 95% CI that includes zero (and now the study estimates from field and laboratory experiments are equally heterogeneous).

Figure S.7: Meta-analysis re-classifying "lab in field" as "field"



S.2 Experiment

S.2.1 Preference versus WTP

Here, we justify with a numerical example the claim that "If the margin by which loss-frame-preferring people are willing to pay more for their preferred contract is larger than the margin by which gain-frame-preferring people are willing to pay more for their preferred contract, then it is possible for $Mean(WTP_{LF}) > Mean(WTP_{GF})$ even if most people prefer the gain-framed contract." For example, say that: (1) 60% prefer gain-framed contracts to loss-framed contracts and, for each person, $\$0.85 = WTP_{LF} < WTP_{GF} = \$0.90$; (2) 25% prefer loss-framed contract to gain-framed contracts and, for each person, $\$2 = WTP_{LF} > WTP_{GF} = \$0.50$; and (3) 15% of the people are indifferent and, for each person, $\$1 = WTP_{LF} = WTP_{GF}$. Then, despite only a minority preferring the loss frame, $Mean(WTP_{LF}) > Mean(WTP_{GF})$ i.e., $Mean(WTP_{LF}) = .15 * 1 + .6 * 0.85 + .25 * 2 = \$1.16 > Mean(WTP_{GF}) = .15 * 1 + .6 * .90 + .25 * .5 = \$0.815$.

S.2.2 Alternative regression specifications

Tables S.1, S.2 and S.3 report regression estimates of the models in Table 6 with alternate clustering of errors. Each table contains only a single column from Table 6. All regressions were run in Stata 16. Column 1, like Table 6 uses the `xtreg` command; however, rather than clustering errors on subjects, it reports SE estimates for clustering on session (xtset index is set session rather than subject ID). Column 2 uses the `reghdfe` command, which allows subject level effect to be "absorbed" into session effect. Column 3 uses the `cgmreg` command, which allows two-way clustering of errors. Two-way clustering accounts for covariance of error both by individual and session, but does not "nest" the former in the latter. Column 4 uses the `svy` preface, designed to organize data by primary sampling units. Across the various methods, the standard errors of

the estimates for *Loss Framed* change only at the second decimal digit.

Table S.1: Estimated effect of loss-framed contracts on grids completed, using alternative variance estimators

	(1) xtreg	(2) reghdfe	(3) cgmreg	(4) svy:
Loss Framed	0.89	0.89	0.89	0.89
	[0.36,1.42]	[0.32,1.45]	[0.36,1.42]	[0.33,1.45]
Observations	536	536	536	536
Clustering	Session	Nested	Two-Way	Nested

95% CI in brackets, based on heteroskedastic-robust standard errors. All regressions also included dummy variables for order effects (=1 if started in loss frame), and for round effects (=1 if second round), whose estimated coefficients are suppressed for clarity.

Table S.2: Estimated effect of loss-framed contracts by contract preference on grids completed, using alternative variance estimators

	(1) xtreg	(2) reghdfe	(3) cgmreg	(4) svy:
Loss Framed	0.18	0.18	0.18	0.18
	[-0.48,0.84]	[-0.53,0.89]	[-0.48,0.84]	[-0.52,0.88]
Prefer Loss Frame	-2.60		-2.60	-2.60
	[-5.03,-0.18]		[-5.03,-0.18]	[-5.19,-0.01]
Prefer LF	3.28	3.28	3.28	3.28
& Loss Framed	[1.41,5.15]	[1.28,5.28]	[1.41,5.15]	[1.29,5.28]
Observations	536	536	536	536
Clustering	Session	Nested	Two-Way	Nested

95% CI in brackets, based on heteroskedastic-robust standard errors. All regressions also included dummy variables for order effects (=1 if started in loss frame), and for round effects (=1 if second round), whose estimated coefficients are suppressed for clarity.

Table S.3: Estimated effect of loss-framed contracts by contract preference (with indifference) on grids completed, using alternative variance estimators

	(1) xtreg	(2) reghdfe	(3) cgmreg	(4) svy:
Loss Framed	0.03	0.03	0.03	0.03
	[-0.82,0.87]	[-0.88,0.93]	[-0.82,0.87]	[-0.87,0.93]
Prefer Loss Frame	-2.71		-2.71	-2.71
	[-5.26,-0.15]		[-5.26,-0.15]	[-5.43,0.02]
Indifferent	-0.55		-0.55	-0.55
	[-2.65,1.56]		[-2.65,1.56]	[-2.79,1.70]
Prefer LF	3.46	3.46	3.46	3.46
& Loss Framed	[1.53,5.38]	[1.39,5.52]	[1.53,5.38]	[1.40,5.51]
Indifferent	0.91	0.91	0.91	0.91
& Loss Framed	[-0.62,2.45]	[-0.73,2.55]	[-0.62,2.45]	[-0.72,2.55]
Observations	536	536	536	536
Clustering	Session	Nested	Two-Way	Nested

95% CI in brackets, based on heteroskedastic-robust standard errors. All regressions also included dummy variables for order effects (=1 if started in loss frame), and for round effects (=1 if second round), whose estimated coefficients are suppressed for clarity.

S.2.3 Breaks

Table S.4 reports summary statistics for the number of breaks taken in the experiment.

Table S.4: Breaks by Frame

Frame	Round	Obs	Mean	SD	Min	Max
Gain Frame	Both	268	0.32	0.97	0	11
Gain Frame	Both	268	0.14	0.76	0	11
Gain Frame	1	135	0.34	1.20	0	11
Loss Frame	1	133	0.12	0.37	0	2
Gain Frame	2	133	0.29	0.66	0	3
Loss Frame	2	135	0.16	1.01	0	11