Stretch It but Don't Break It: The Hidden Cost of Contract Framing*

Richard R. W. Brooks† Columbia Law School

Alexander Stremitzer‡ UCLA Law School

Stephan Tontrup§ New York University School of Law

July 19, 2016

Abstract

Recent research suggests that loss-framed contracts are an effective instrument for principals to maximize the effort of their agents. Framing effects arise from defining thresholds that vary the salience of losses and gains while preserving payoff equivalence of the underlying contract. While plausible interpretations of Prospect Theory's loss-aversion insight suggest that a loss frame would lead to more effort, we show that contract thresholds also exert a norm-framing effect on performance that can trump the impact of loss aversion. Loss framing therefore carries a risk: poorly selected thresholds may reduce effort. Principals may prefer to avoid this risk by offering contracts that impose no threshold at all.

Keywords: contracts, loss-framing, norm-framing, loss aversion, prospect theory, expressed thresholds, worker productivity.

JEL-Classification: C91, D02, J33, K12.

*The authors are grateful to Johannes Abeler, Thomas Baumann, Joe Doherty, Florian Ederer, Christoph Engel, Florian Englmaier, Robert Gold, Peter Kuhn, Bentley McLeod, Ted Parson, Matteo Rizzoli, Andreas Roider, Patrick Schmitz. We are also grateful to Seminar Audiences at UC Santa Barbara, the University of Mannheim and at the ALEA conference at Vanderbilt. We thank Jaimini Parekh, James Davis, Yijia Lu and Henry Kim for excellent research assistance. Financial Support by the Lowell Milken Center for Business Law and Business Policy is gratefully acknowledged.

†Columbia Law School, 435 West 116th Street, New York, NY 10025, United States, rbrooks@law.columbia.edu.

‡UCLA Law School, 385 Charles E. Young Drive, 1242 Law Building, Los Angeles, CA 90095, United States, stremitzer@law.ucla.edu.

§New York University, School of Law, 40 Washington Square S, New York, NY 10012 s.tontrup@googlemail.com.

1 Introduction

Law sanctions conduct falling short of what is normatively established, but seldom rewards behavior that goes beyond the legal threshold. It is a basic asymmetry that pervades in public and private law. Contract law, for instance, protects buyers from downward deviations in sellers' performances, whether in terms of quantity, quality or timing of delivery among other nonconforming shortfalls of contract stipulations. Aggrieved buyers in these cases are typically afforded monetary compensation or other relief, such as a right to demand that sellers cure their nonconforming performances or a right to rescind the contract. There is, however, no right to additional compensation for a seller that produces a greater quantity or quality than required or who delivers earlier than stipulated to the buyer's advantage.[1] As a consequence, a kink results in performance incentives, where the seller has high-powered incentives below the contractual specifications and weak to no incentives to perform above and beyond what is specified in the contract.

Buyers may remove the kink by offering contractual terms stipulating performance at very high levels. If the buyer offers a high enough price then sellers would accept and, as a consequence, the contract preserves the sellers' performance incentives over a greater range of possible performance. Notwithstanding the simplicity of this solution, it does present some challenges.[2] Breach – with all its associated transaction costs – may occur more frequently.

[1] Under certain strict conditions, an "overperfroming" seller may have a right to recovery under a theory of restitution, quantum meruit, quantum valebat, quasi-contract and such, which all largely rely on the buyer being unjustly enriched by the overperformer. Without *unjust* enrichment, there is generally no right to recover compensation for engaging in conduct not legally required. "If a performance is rendered by one person without any request by another, it is very unlikely that this person will be under a legal duty to pay compensation." Arthur Corbin, *Contracts* §234 (1951).

[2] Take, for example, the case of Oeresund A/S, a consortium overseeing the construction of a $10 billion dollar bridge and tunnel link connecting the cities of Copenhagen, Denmark and Malmo, Sweden. Since the planned link would accommodate rails, as well as cars, Oeresund A/S feared that vibrations caused by trains, particularly on the suspension bridge part of the link, would make drivers feel unsafe and thereby reduce their demand. To address this concern Oeresund A/S contracted with a firm (hereafter "supplier") competent in the design and use of advanced cable and damper technology. Oeresund specified an installed vibration level (i.e., an extremely low vibration level) in the contract, which was impossible to achieve with current technology. Although Oeresund offered a high up-front price for meeting the specified vibration level, all parties anticipated that the supplier would in all likelihood fail to achieve this specification, which would trigger a schedule of stipulated damages paid to Oeresund based on how far short of the quality level the supplier's performance fell.

Some of these costs may be avoided by stipulating damages in the contract, but courts may still refuse to enforce stipulated damages as "penalties." Moreover, a seller's behavioral motivation to perform may be weakened by knowing that, despite its effort, breach and penalties are almost inevitable.[3]

To avoid these potential costs the parties may choose an alternative arrangement. Rather than allowing only for sanctions when performance falls short, whether through stipulated damages or by relying on background legal remedies for downward deviations, they may agree to an intermediate performance level with a bonus for going beyond, i.e., rewarding overperformance, in addition to punishing shortfalls. This was the contractual scheme that Caltrans, the California transportation agency, offered to its contractors following the 1994 Northridge earthquake that resulted in the collapse of two bridges on the Santa Monica Freeway, one of the world's busiest roadways. Caltrans offered a contract with substantial performance incentives and penalties: a $200,000 per day bonus for completing the project ahead of schedule and a $200,000 penalty for each day the project was behind schedule. Under this incentive scheme, reconstruction of the bridges was completed in a little over two months, 74 days ahead of the stipulated deadline (Eggers, 1997).[4] The contract created a symmetric incentive scheme, where the threshold stipulated in the contract no longer defined a cut-off point beyond which additional performance goes uncompensated, but instead became a framing device, allowing the drafting party to vary the salience of losses and gains.

To illustrate the point, consider a contract calling for 50 units of specified goods for 50 dollars. Under the default regime, if the seller delivers 40 units the buyer is entitled to

[3]The initial thought at Oeresund was that a nearly impossible to reach threshold matched with an astronomical contract price would give the supplier high-powered incentives to produce at its best. Yet, when the contract was underway, Oeresund became aware that the contract design may have sapped the motivation of the supplier to exert its greatest effort, knowing that failure was practically inevitable. Following this experience, officials at Oeresund A/S came to believe that it might have been more effective to specify an intermediate quality threshold and offer a bonus for exceeding it, rather than a contract where a penalty was certain no matter how hard the supplier worked. Notice the effect of Oeresund's updated view on the preferred contract offer which accounts for the asymmetry in law discussed above.

[4]To complete the project so early, the contractor used up to 400 workers a day and kept crews on the job 24 hours a day. The $13.8 million the contractor received in performance bonuses was more than offset by the estimated $74 million in savings to the local economy and $12 million in contract administration savings thanks to the shortened schedule.

damages of 10 dollars (assuming a linear apportionment between the shortfall and damages.) However, if the seller delivers 60 units, she cannot demand 60 dollars from the buyer. The incentive scheme is asymmetric providing compensation only until the threshold but not beyond. By increasing the number of delivered units up to the threshold, the seller can increase her payoffs by reducing damage payments, but beyond the threshold any increase in the number of units will not increase her payoff. If, however, the contract stipulates 50 units for 50 dollars and a bonus of 1 dollar per unit delivered beyond the threshold, the seller is entitled to an amount which equals the number of units times \$1. The same would hold true if the threshold were set at 70 or 100 units. The threshold, no longer has any effect on the parties' monetary payoffs as in an asymmetric payoff regime. In a symmetric payoff regime the threshold becomes a pure framing device, expressing a reference point or a norm. The key question is then which frame produces the best performance incentives. Our analysis seeks to answer this question.

We propose a novel theory of contract thresholds. We posit that contract thresholds produce framing effects not only through loss aversion (*loss-framing*) as was previously argued in the literature (Hossain and List, 2012), but also through a separate mechanism, which we label the "norm effect of thresholds" (*norm-framing*). Besides serving as a pivot to frame a loss or gain, contract thresholds also have a suggestive effect which influences behavior. Thresholds may communicate a norm of either how others tend to perform (a positive norm) or what performance is expected (a prescriptive norm) or both.

In order to test our theory we conducted an online experiment involving real effort. Participants were presented with a table with 200 digits between 1 and 9. The task consisted in counting how often a specified digit, e.g., "1" or "4", occurred in the table. After entering the correct number, participants could decide whether to continue or to stop the task. If they chose to continue they were shown a new screen, with a different table of digits.[5] Participants earned €1 for every completed screen. The framing of their earnings, however, was varied

[5]Participants could, in principle, continue with the experiment indefinitely if they so chose, without any arbitrary upper limit imposed by time or anything else, that is, the "effort" consists only of the subject's willingness to keep counting tables.

across four treatment conditions. The first condition offered participants a "plain-vanilla" bonus contract, which promised participants €1 per completed screen. The contract did not prescribe any target quantity. By contrast, the three other treatments, offered contracts with expressed thresholds—5, 15, or 50 screens, each associated with earnings of €5, €15, and €50, respectively. We label the easy-to-meet expressed threshold (i.e., 5 screens) the "low-bar" contract, the more demanding intermediate threshold (of 15) the "stretching" contract and the highest threshold (i.e., 50) the "extreme-effort" contract.[6] Under each of these three contracts, participants earned a bonus of one additional Euro for each screen they completed beyond the stipulated threshold and they faced a penalty of one Euro for each screen they fell short of the stipulated threshold. In other words, under all four contracts that constitute our treatment conditions, participants were offered exactly the same payoffs, namely €1 per completed screen.

To illustrate, assume that a participant completes 12 screens. Then under the plain-vanilla contract he gets €12 (12 x €1). Under the low-bar contract with the threshold of 5 he gets €5 plus a bonus of €7. Under the stretching contract with threshold 15 he gets €15 minus a penalty of €3. Under the extreme-effort contract stipulating threshold 50 he gets €50 minus a penalty of €38. At every level of performance, participants receive exactly the same payoff under each contract. They face the same linear incentive scheme in each condition; only the framing of the contract differs.

The basic results from our experiment indicate that contract framing does indeed matter, as observed in the prior literature. Our findings, however, suggest something largely overlooked in the literature: there is a non-monotonic relationship between the threshold and the performance. Compared to plain-vanilla terms, which state no explicit thresholds, our manipulations produce significantly different mean effort levels, damping or increasing effort conditional on the suggested threshold. Specifically, the low-bar and extreme-effort treatments are associated with lower mean effort and the stretching treatment results in

[6] These labels were not shared with the subjects. We use them here to conveniently distinguish the four treatment conditions: "plain-vanilla", "low-bar", "stretching" and "extreme-effort".

greater mean effort than observed under the plain-vanilla condition.

A trade-off is apparent. While setting a threshold at a demanding yet realistic level seems to increase effort levels of agents, a threshold too low or too high will lead to lower levels of effort. Given the risk associated with selecting a threshold which is too low or too high, it may be better not to specify a threshold at all**.** Hence, framing is not a costless strategy for principals. When the principal lacks good information about agent's production functions, the best strategy may be to play it safe and offer a linear plain-vanilla contract without expressing thresholds. By highlighting this tradeoff, our paper gives caution to the suggestion that loss-framing generally leads to better performance (Hossain and List, 2009, 2012).[7] Moreover, through an additional treatment designed to disentangle loss-framing and norm-framing, we find that our results are driven by both a loss-framing and a norm-framing effect.

Our paper contributes to the larger literature on self-control incentives where firms may impose work targets like production minimums or artificial deadlines in order to help their workers overcome time inconsistencies in preferences (Kaur et al., 2010). However, different from this literature we study pure framing where agents do not suffer disproportionate monetary penalties for failing to meet their targets.

Our paper proceeds as follows: Section 2 presents a simple model with regard to the effect of different contract thresholds on agents' effort from which we derive our hypotheses. Section 3 presents our experimental design. Our results are presented and discussed in Section 4. Section 5 concludes.

2 Theory and Hypotheses

The effectiveness of contract framing is usually explained by appealing to Prospect Theory's concept of loss aversion (Hossain and List, 2012). Prospect Theory tells us that 'losses loom larger than gains,'—that is, individuals prefer avoiding losses to acquiring commensurate

[7] Our observations are nicely summarized in Voltaire's comment that perfection is the enemy of the good ("Le mieux est l'ennemi du bien").

gains—a concept commonly referred to as loss aversion (Kahnemann and Tversky, 1979). Therefore, framing a transaction in terms of a loss, should provoke a stronger response than an economically equivalent transaction framed in terms of a gain. Building on this insight, researchers have recently turned their attention to the framing of contracts. By establishing thresholds in contracts around which earnings are presented as losses or gains, researchers have found that individuals (Brooks, Stremitzer, and Tontrup, 2011; Fryer, Levitt, List, and Sadoff, 2012) and teams (Hossain and List, 2009, 2012) exert greater effort under loss-framed contracts than under payoff-equivalent contracts framed in terms of gains.

We hypothesize an additional influence operating through norm-framing where reasonable thresholds are assumed to set a norm (e.g., by communicating expectations) with which parties tend to comply. Norm-framing assumes that subjects interpret the thresholds as communicating expectations about what the subject should do. Subjects may then experience a disutility from falling short and exceeding the threshold, as both would run counter the expectations communicated through the threshold with which they want to conform.

Both loss-framing and norm-framing can be captured by a utility function of the following form:

$$U(y) = yp + \delta(y-\bar{y})\left[\lambda_0 p(y-\bar{y}) + \lambda_1 p(y-\bar{y})^2\right] - \left[1-\delta(y-\bar{y})\right]\lambda_2 p(y-\bar{y})^2, \qquad (1)$$

where y is the output chosen by the agent, p is the payment per unit of output, $\delta(x)$ is a function indicating whether the output is below or above the threshold $\bar{y}$:

$$\delta(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \end{cases},$$

and λ_0, λ_1, λ_2 are parameters which determine the shape of the utility function below and above the threshold. Let the subject's cost function, $c(\cdot)$, be an increasing function of effort. Then, the optimal effort choice y^* is given by:

$$c'(y^*(\bar{y},\lambda_0,\lambda_1)) = \begin{cases} \left[1+\lambda_0-2\lambda_1(\bar{y}-y^*)\right]p & \text{if } y^* < \bar{y} \\ \left[1-2\lambda_2(y^*-\bar{y})\right]p & \text{otherwise} \end{cases}.$$

Prospect theory's main property that losses loom larger than gains, leading to a kink in the utility function at the reference point, can be captured by assuming $\lambda_0 > 0$ and $\lambda_2 \geq 0$.

If $\lambda_2 > 0$, this would allow for the possibility that an agent's utility function is concave in gains, that is, gains which are more distant from the agent's reference point carry less weight. The common assumption that an agent's utility function is convex in the losses, that is, losses which are more distant from an agent's reference point carry less weight, would follow from $\lambda_1 > 0$.

Norm-framing's main idea that subjects experience disutility from falling below the threshold can be captured by assuming $\lambda_0 > 0$. It might also be plausible to assume under norm-framing that the agent's utility function is convex below the threshold establishing the norm, $\lambda_1 > 0$. This is because a subject that is about to miss the threshold by far may not care much whether or not it is one unit closer to the threshold, whereas a subject being close to meeting the threshold may care a lot. Finally, assuming $\lambda_2 > 0$ captures the possibility that an agent experiences disutility exceeding the threshold.

Assume that the low-bar threshold ($\bar{y}_{LB}$) is set at a fraction of the plain-vanilla effort level (y^*_{PV}), the stretching threshold ($\bar{y}_S$) is set at a level moderately above the plain-vanilla effort level, and that the extreme-effort threshold ($\bar{y}_E$) is set at an unreasonably high level.

Then, under norm-framing, predictions about how the effort level under the plain-vanilla contract compares to effort under the other contracts depend on what we assume about the utility function induced by the plain-vanilla contract. As no express expectations are communicated, it is plausible to assume that the agent's utility function just coincides with the monetary payoff function, $U(y) = yp$. Assuming $\lambda_0 > 0$ and $\lambda_1 > 0$ would then predict that compared to the plain-vanilla contract, the stretching contract increases the effort while the extreme-effort contract decreases the effort. This is because, at effort level $y^*_{PV} < (\bar{y}_S) < (\bar{y}_E)$ the slope of the plain-vanilla utility function is p, while the slope of the utility function induced by the stretching contract and the extreme-effort contract is $[1 + \lambda_0 - 2\lambda_1(\bar{y} - y^*_{PV})]\, p$ which is higher than p for $\bar{y} = \bar{y}_S$ (stretching threshold) and lower than p for $\bar{y} = \bar{y}_E$ (extreme-effort threshold). Finally, assuming $\lambda_2 > 0$ would predict that the low-bar condition could lead to lower effort than the plain-vanilla condition. This is because at $y^*_{PV} > \bar{y}_{LB}$, the slope of the utility function induced by the low-bar contract is

$(1 - 2\lambda_2(y^* - \bar{y}_{LB}))\, p < p$.

Under Prospect Theory, predictions about how mean effort under the plain-vanilla contract compares to the mean effort under the other contracts depend on assumptions about the reference point induced by the plain-vanilla contract. It is plausible to assume that the plain-vanilla contract sets the reference point at the status quo, $\bar{y} = 0$. The slope of the utility function induced by the plain-vanilla contract under Prospect Theory would then be $(1 - 2\lambda_2 y)\, p$ instead of yp under norm-framing. Assuming $\lambda_0 > 0$ and $\lambda_1 > 0$ would then predict that, compared to the plain-vanilla contract, the stretching contract increases the mean effort level, while the extreme-effort contract decreases the mean effort level.[8] Finally, Prospect Theory would predict that the low-bar condition leads to higher effort than the plain-vanilla contract. This is because at $y^*_{PV} > (\bar{y}_{LB})$, the slope of the utility function induced by the low-bar contract is higher than under the plain-vanilla contract, because of concavity of the utility function above the threshold, $(1 - 2\lambda_2(y^*_{PV} - \bar{y}_{LB}))\, p > (1 - 2\lambda_2 y^*_{PV})\, p$.

Under reasonable assumptions, both norm-framing and loss-framing therefore predict that the stretching contract should *increase* and that the extreme-effort contract should *decrease* observed effort compared to the plain-vanilla contract. However, in the low-bar condition the effect predicted by Prospect Theory and the hypothesized norm-framing effect of the threshold should be *countervailing*. Prospect theory predicts that a contract that expresses a threshold and thereby creates a loss frame increases effort levels compared to the plain-vanilla condition. Norm-framing predicts that, compared to the plain vanilla condition, the low-bar threshold will drag down effort levels. Assuming we have set the threshold low enough in the low-bar treatment, we expect the drag-down effect predicted by norm-framing to trump the positive effect predicted by loss-framing.[9] This leads us to the following

[8] Although this is driven by another behavioral channel the mathematical argument is the same as for norm-framing: The argument for the stretching threshold applies *a fortiori* as the slope under the plain-vanilla condition is $(1 - 2\lambda_2 y)\, p$. And also the argument about the extreme-effort contract goes through for a high enough threshold.

[9] As the threshold is set lower, the drag-down effect increases while loss-framing becomes weaker. Therefore, by setting the threshold low enough, the drag-down effect will dominate loss aversion. To see this, note that the difference between $(1 - 2\lambda_2(y^*_{PV} - \bar{y}_{LB}))\, p$ and $(1 - 2\lambda_2 y^*_{PV})\, p$ converges to 0 for $\bar{y}_{LB} \to 0$, whereas the difference beween $(1 - 2\lambda_2(y^* - \bar{y}_{LB}))\, p$ and p converges to $-2\lambda_2(y^*)$.

hypothesis.

Hypothesis 1 *In the low-bar treatment, mean effort is lower than under the plain-vanilla contract (H1.1). In the stretching treatment, mean effort is higher than under the plain-vanilla contract (H1.2). In the extreme-effort condition, the mean effort level decreases relative to the stretching condition and even falls below the level observed in the plain-vanilla contract (H1.3).*

If we were to find that mean effort is lower under the low-bar contract than under the plain-vanilla contract (H1.1) we may conclude that this is driven by norm-framing, given that loss-framing predicted the opposite effect. However, this argument depends crucially on the assumption that the plain-vanilla contract sets the reference point at the status quo, $\bar{y} = 0$.

While this assumption seems plausible, there are other plausible assumptions that would allow to account this pattern (H1.1) within the framework of Prospect Theory. Assume, e.g., that the plain-vanilla contract sets the reference point at an agent's payoff expectations, $\bar{y} = \frac{e}{p}$, as suggested by Koszegi and Rabin (2006). If those payoff expectations happen to set the reference point between the low-bar and the stretching threshold, assuming $\lambda_0 > 0$ and $\lambda_2 \geq 0$ predicts that the low-bar contract leads to *lower* mean effort than the plain-vanilla contract. In other words, it can be argued that Prospect Theory's prediction about how mean effort under the plain-vanilla contract compares to the mean effort under the low-bar contract is ambiguous. This reflects a general concern with Prospect theory which is very sensitive to different assumptions about the location of the reference point.

In order to establish that the drag down effect is indeed driven by norm-framing, we need to run an additional treatment designed to "switch off" the norm-framing effect of the contractual threshold. We label this the "random-threshold" treatment, wherein subjects are offered either a randomly selected low-bar or extreme-effort contract. Since the subjects know the threshold is determined randomly, they should not infer that the threshold communicates a norm of what is expected of them. If only norm-framing causes the drag-down effect under

the low-bar condition then, under a randomly chosen low-bar threshold, mean effort should rise to the same level as under plain-vanilla. This leads us to the following hypothesis:

Hypothesis 2 *The drag-down effect predicted by Hypothesis 1.1 disappears under the low-bar random treatment.*

3 Design

Participants were given the opportunity to enter into a contract to perform a real effort task in exchange for money.[10] They were instructed that if they rejected the contract they would receive no payment based on the real effort task. In that case subjects were immediately directed to the second stage of the experiment where they and all other participants were asked to complete a pair of monetarily incentivized economic and psychological tests (see more details below).

Participants who accepted the contract offer were directed to begin the task. The task consisted of counting how often a specified digit, e.g., "1" or "4", occurred in a table containing 200 digits ranging from 1 to 9. Given that the task did not require any special skills, we assume that performance was a function only of effort (see Abeler, Falk, Goette, and Hoffman, 2011). After scanning the table and counting the occurrences of the specified digit, participants entered the count into a screen dialog box. An answer was considered correct if it fell within a range of +2/–2 of the true value. For example, if the correct number of "3" digits was 42 while the participant counted 40, the result was treated as correct. If participants gave an answer outside of this tolerated margin they could retry counting the table as often as they wished, but had to wait 15 seconds after each failed trial, before they

[10] Subjects were given instructions that emphasized the experiment involved real contracts. As contract law requires, but unlike in most other experiments, we did not impose the contract, but participants were free to accept or reject the offers. After being presented with the contract terms, participants decided whether to accept or reject the offer (see contractual terms described in detail below). Much of the literature does not endogenize the acceptance decision but assumes that participants have entered into a contract under given terms (see, e.g., Falk and Kosfeld, 2006; Fehr, Klein, and Schmidt, 2007). In our design we account for the possibility that behavior changes depending on whether parties explicitly assent to the contract (see Hoppe and Schmitz, 2011, for a similar design.) See also Lazear, Malmendier, and Weber (2012). This design allowed us to elicit participants' willingness to enter into the contract in addition to measuring effort under the contract.

could make a new input.[11] Participants were required to make some screen input at least every three minutes; failure to do so terminated the experiment. After each successfully completed screen, participants were asked whether they wanted to continue on to the next screen. If they decided to go on, a new table was displayed asking participants to count another randomly specified digit; if subjects chose to stop the task, they were directed to the incentivized economic and psychological tests.

As previously described, we implemented four treatments that varied the threshold and one baseline condition—the so-called plain-vanilla condition—which did not express a threshold, but simply offered €1 for each successfully completed screen. We conducted the plain-vanilla treatment in advance and found that participants completed an average of 10.4 screens. We used this data to calibrate the thresholds of the treatments, which have the same linear payment scheme as the plain-vanilla condition. The low-bar threshold (5 screens) was set clearly below the mean effort participants exerted under the plain-vanilla contract. The stretching threshold (with 15 screens) obliged participants to increase their efforts by nearly 50% in comparison to the average performance under the plain vanilla contract. The extreme-effort threshold (50 screens) called for an extremely high level, demanding approximately 5 times the mean performance observed under the plain-vanilla contract.[12] Finally, in the random-threshold treatment, the threshold was randomly determined. Subjects were offered either a low-bar or an extreme-effort contract. Subjects learned that the threshold was determined randomly before they had to decide whether they wanted to accept or reject the offer. However, they were not informed that only thresholds 5 and 50 could be selected in order to avoid the possibility that they would infer some target quantity the principal might expect them to reach (e.g., some quantity between 5 and 50).

[11] We introduced this wait time after each failed trial in order to dissuade participants from guessing repeatedly without actually counting any digits. Counting one table takes between 45 seconds and one minute such that a wait time of 15 seconds seems sufficient to achieve this goal.

[12] Furthermore, it required participants to invest notably more time than is usually expected in online experiments. Online studies are typically completed within approximately 15 to 20 minutes rather than a full hour. Note, though, that participants were informed about the duration of the experiment in the email that invited them to the study. They also learned that the actual time needed would depend on their performance (see Musch and Reips, 2000).

We evaluated the effort participants' exerted under the different treatments along two dimensions: quantity and quality. Quantity is measured by the number of successfully completed screens. Quality of performance is determined by the accuracy with which participants completed the task. We recorded each count that participants entered and calculated by how much the entry deviated from the true number of digits. We considered both failed trials (where participants had to repeat, if they wanted to continue) and successful, but not necessarily accurate, trials (recall success was achieved with an error tolerance of +/-2).

In addition to effort measures, we also elicited participants' willingness to enter into the contract. The efficiency of a particular contract depends not only on how people perform under its terms, but also on how likely they are to agree to incur the contractual obligation in the first place. Unlike in most other experiments, but in line with the formation requirements of a legally enforcable contract, participants were free to accept or reject the contract offers.[13] This real contract setting also allows us to study the impact of different terms on acceptance rates. If rates differ, it follows that participants are more willing to accept some terms than others, which is relevant for a principal who wants to reduce the likelihood that offers are rejected.

The experiment was conducted online using the server of the Max Planck Institute of Economics. We decided to conduct the experiment online because we wanted participants to have real opportunity costs when deciding whether to continue with the task. We measure effort by eliciting the point where participants prefer some other activity over continuing with the experiment. In a laboratory setting, participants have little opportunity costs since they cannot leave until the session is over.[14] By contrast, at home, participants can easily stop and pick up a preferred activity. Even if we had conducted individual sessions in order

[13]Most of the literature does not endogenize the acceptance decision but assumes that participants have entered into a contract under given terms (see, e.g., Falk and Kosfeld, 2006; Fehr, Klein, and Schmidt, 2007). In our design we want to account for the possibility that behavior changes depending on whether parties explicitly assent to the contract (see Hoppe and Schmitz, 2011, for a similar design.) Lazear, Malmendier, and Weber (2012) have recently shown in an experimental study how the possibility to opt out of an economic environment create selection effects (in their case the willingness to share money with other participants.)

[14]If participants were allowed to leave the session early in a laboratory setting, this would be observed by the participants. Observing others to leave would likely influence participants' decisions when to stop the task. The only way, how this could have been avoided is by running 251 individual sessions.

to allow participants to leave without influencing others, coming to the laboratory causes sunk costs that make participants less likely to reject a contract without earning an adequate compensation for their participation.[15]

All 251 participants were students of the University of Münster, Germany, with various majors; 90% of them had not participated in an economic experiment before. Participants were sent an invitation by email via the mail-server of the university. The email did not describe the purpose of the study in order to avoid participants self-selecting in an experiment they were interested in. The invitation contained a link that directed participants to the website of the experiment. It was active only for a single login. Participants had to complete the stages of the experiment within strict time limits and were kept informed of this fact with constant screen messages. If participants logged out or did not finish stages within these time limits, the experiment was automatically aborted and participants were notified that they were excluded from the experiment. We set those time limits to force participants to focus on the task and block internet distractions that can easily distort results of online studies. Participants were informed up-front about the amount of time they would need to complete the whole study thereby reducing the likelihood that participants would have to break off the experiment because they ran out of time. The online instructions given to participants are presented in Appendix C.[16]

The real effort experiment was followed by two psychometric tests. The first test measures loss aversion by giving the participants an opportunity to participate in two lotteries.[17] The first lottery presents participants with a 1/2 probability of winning €8 and a 1/2 probability of losing €5. The second lottery has the same payoffs, but is repeated six times and thus lowers the probability of suffering an overall loss. An unbiased participant should play both lotteries, since participating yields a gain in expectations. We classify participants into two

[15] Show-up fees in Lab experiments can mitigate this problem but can never be tailored to the invividual participant.

[16] We confirmed that participants could understand the instructions by posing control questions in an offline pilot session. Participants had no difficulties in correctly calculating their earnings based on the different contract terms. We programmed the experiment using the open-source survey application LimeSurvey. See http://www.limesurvey.org/.

[17] The test was developed and used by Goette, Hoffman, and Fehr (2004).

different categories: 1) The "loss averse" type who rejects at least one of the lotteries, 2) and the unbiased "rational" type who participates in both lotteries. The second test, a cognitive reflection test, measures participants' impulsiveness (CRT, see Frederick, 2005). The test consists of a set of three questions, which participants have a total of 90 seconds to complete. Questions are designed such that participants' initial impulses lead them to an incorrect answer. The test therefore measures the participants' ability to think beyond their initial impulse and reach the correct but counter-intuitive answer.

All aspects of the experiment, including the follow-up behavioral tests, were incentivized and payoffs depended on the participants' own decisions. When calibrating the payoffs we made sure that participants could expect to earn slightly more than in a regular student job in and outside of the university in order to make incentives comparable to a normal working environment. A student job in Münster would offer approximately €8 for an hour. In our real effort task, participants need less than one minute to count a table and move to the next screen. They could read the instructions in 5 minutes which leaves them with an hourly wage of €55. Since only every 5th participant was randomly selected for payment participants earn €11 in expectation. On average, our participants therefore earn 1/3 more than in a regular student job.[18]

4 Results

4.1 Treatment Effect on Quantity

We first present how participants' mean effort levels differed across treatments (Table 1 and Figure 1). Observe that the plain-vanilla contract, which does not express a quantity threshold, leads to mean effort of 10.4 screens, the low-bar contract (with expressed quantity 5) leads to mean effort of 6.2 screens, the stretching contract (with expressed quantity of 15) leads to a mean effort of 14.3 screens, and the extreme-effort contract (with expressed

[18] Stochastic payouts are routinely used in experimental economics as there is evidence that paying out larger amounts at a lower probability simulates high stakes: that is, incentives for participants are stronger than in the case where they are paid smaller amounts at a higher probability, even though expected payoffs are equivalent (Laury, 2006; Laury and Holt, 2008).

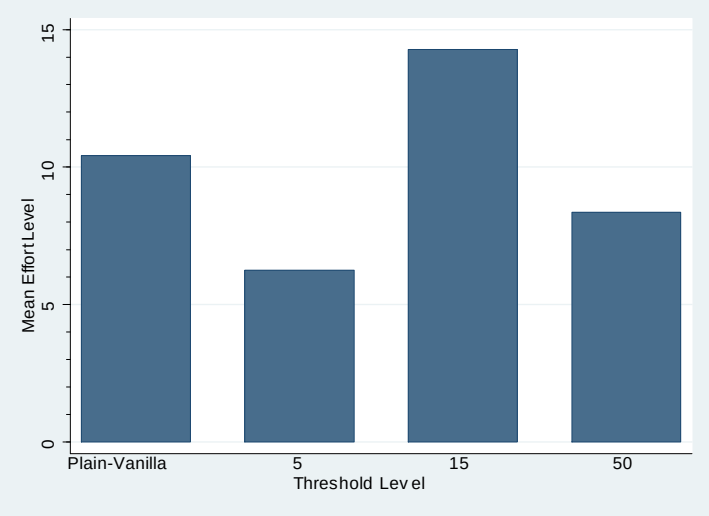

Figure 1: Mean Effort Levels Depending on Specified Quantity Threshold.

quantity 50) leads to mean effort of 8.4 screens.

Compared to the plain-vanilla contract, mean effort decreases under the low-bar contract (H1.1), increases under the stretching contract (H1.2), and decreases under the extreme-effort contract (H1.3). These differences are statistically significant (Mann-Whitney ranksum, Table 1 reports Z and p values).

Table 1: Mean Effort Levels Compared Across Treatments (Mann-Whitney)

			contract type			
contract type	threshold	observed mean	plain-vanilla	low-bar	stretching	extreme-effort
plain-vanilla	—	10.4 (N=50)	—	$Z = 3.3^{***}$ $p < 0.01$	$Z = -2.5^{**}$ $p = 0.01$	$Z = 2.2^{**}$ $p = 0.03$
low-bar	5	6.2 (N=50)		—	$Z = -5.2^{***}$ $p < 0.01$	$Z = 0.1$ $p = 0.96$
stretching	15	14.3 (N=54)			—	$Z = 3.6^{***}$ $p < 0.01$
extreme-effort	50	8.4 (N=39)				—

Considering only those participants who accepted the contract reflects the perspective of a principal who faces an unlimited supply of labor and does not care whether potential workers are turned off by the terms of a particular contract. There are, however, situations

where a principal wants to increase the mean effort level of a fixed group of people. Then, participants rejecting the proposed contract impose a cost on the principal. We incorporate this scenario in our analysis by treating all participants who reject the contract as exerting zero effort and find that our qualitative results are unchanged (See Table 4 in Appendix A).

4.2 Disentangling loss-framing from norm-framing

Beyond our hypotheses about the aggregate pattern of mean effort levels we also had a specific prediction about the behavioral channels driving our effect. We argued that only norm-framing explains the drag-down effect under the low-bar condition, while loss-framing pushes in the opposite direction. Having found support for H1.1 already offered some evidence that this conjecture is correct. However, as we explained, this conclusion rests on the plausible assumption that the plain-vanilla contract sets the reference point at the status quo.[19] In order to find direct evidence, which does not rely on this assumption, we ran the *random* threshold treatment. The treatment is designed to switch off norm-framing by suppressing subjects' beliefs that the principal communicates an expectation of performance. This allows us to cleanly disentangle loss-framing from norm-framing. As loss-framing is switched off, the drag-down effect should disappear under a randomly chosen low-bar threshold. This is what we predicted in H2.

Conditional on opt in, we find that mean effort under the low-bar random treatment is 10.7 screens, while it was 6.2 screens under the low-bar treatment. This effect is statistically highly significant (Mann-Whitney ranksum, $p<0.01$).[20] As subjects on average complete 10.4 screens under the plain-vanilla contract the drag down effect disappeared consistent with H2. Treating all participants who reject the contract as exerting zero effort, we even find evidence for a reversal of the drag-down effect as the mean effort level under the low-bar random treatment (12.2 screens) is higher than under the plain-vanilla treatment (8.4

[19] We report evidence consistent with this assumption in Appendix B.2. Loss averse types indeed exert more effort in the low-bar treatment than rational types.

[20] Treating all participants who reject the contract as exerting zero effort, the mean effort levels are 12.2 and 5.0 screens, respectively. Also this difference is statistically significant (Mann-Withney ranksum, $p<0.01$).

screens). This effect is statistically significant (Mann-Whitney ranksum, p=0.03).[21]

Our argument rests on the assumption that the random threshold manipulation only affected norm-framing, while leaving loss-framing fully operative. If this is true, we can indeed conclude from the fact that the drag-down effect observed under the low-bar treatment disappears under the random low-bar treatment that the drag-down was entirely due to norm-framing. However, it could be argued that the random treatment not only switches off norm-framing but also affects loss-framing, as it might weaken subjects' perception of the threshold as a reference point. This would introduce a confound as our manipulation would have simultaneously weakened loss-framing *and* norm-framing. The vanishing of the drag-down effect under the random low-bar treatment could then be entirely due to the weakened effect of loss-framing.

In order to counter this challenge, we compare the change in mean effort under the random low-bar treatment compared to the plain-vanilla contract for subjects scoring high on the psychometric test of loss aversion (loss averse types) with the change in mean effort for those scoring low (rational types). We find that mean effort of loss averse types is 2.7 screens higher than under the plain-vanilla contract, while mean effort for rational types is 3.1 screens lower. This difference in difference is statistically significant (Mann-Whitney ranksum, p<0.01) suggesting that loss-framing is operative under the random low-bar treatment and pushes mean effort up rather than dragging it down.[22] This suggests that we don't have to worry

[21] Under the *random* extreme-effort treatment mean effort is 11.8 while it was 8.4 under the extreme-effort treatment. Treating all participants who reject the contract as exerting zero effort, the mean effort levels are 10.3 and 5.3 screens, respectively. Also these two differences are statistically significant (Mann-Withney ranksum, p<0.01). In both cases, effort levels are not significantly different from the effort levels induced by the plain-vanilla contract (Mann-Withney ranksum, p=0.35 and p=0.17). This result is less interesting than the results from the random low-bar treatment, as it does not allow us to disentangle norm-framing from loss framing: Both norm-framing and loss-framing should drag dowen effort in the extreme-effort treatment. Thus, eliminating norm-framing should not change the direction of the effect in comparison to the plain-vanilla contract. However, what we observe is consistent with our theory. Participants are likely to be discouraged when they realize that they cannot - with reasonable effort - conform with the norm. Thus switching off norm-framing leads subjects to increase their effort, as they are no longer discouraged by the overly ambitious threshold.

[22] It might seem a little odd that, for rational types, the mean effort level under the low-bar-random treatment seems to be lower than under the plain-vanilla treatment. However, this effect is not statistically significant (Mann-Withney ranksum, p=0.49) while the effect that the mean effort level is higher for loss averse types is statistically significant (Mann-Withney ranksum, p=0.02).

about the possibility of a confound. A weakening of loss-framing could only account for the vanishing of the drag-down effect if loss framing had a negative effect on mean effort.[23]

4.3 Treatment Effect on Quality

In addition to effort measured in terms of *quantity* (that is, screens successfully completed), we also explore *quality* effort (measured by the accuracy with which participants perform the counting task). One concern could be that quantity and quality may be in conflict. In other words, contracts leading to higher quantity might lead to lower quality (lower accuracy of performance). In our experiment, we can distinguish two cases of inaccuracy. The first case of error occurs if a participant's answer deviates from the true value by a margin of less or equal to +/-2. In this case, he can proceed to the next screen. In the second case, the participant's answer falls outside the +/-2 margin of error, which forces the participant to recount the table if he wants to proceed to the next screen. We form an inaccuracy score for each participant by adding up the deviations in both successful and failed attempts and dividing them by the number of successfully completed screens.[24] Table 2 presents how the mean inaccuracy score differs across treatments.

Table 2: Quality Levels Across Treatments

contract type	threshold	inaccuracy score mean	inaccuracy score median
plain-vanilla	– (N=50)	0.75	0.33
low-bar	5 (N=50)	1.23	0.77
stretching	15 (N=54)	0.60	0.21
extreme-effort	50 (N=39)	1.45	1.82

We find that effort quality is the lowest under the extreme-effort contract, closely followed by the low-bar contract, with mean inaccuracy scores of 1.45 and 1.23, respectively. Participants are significantly more accurate under both the plain-vanilla contract (0.75) and

[23] In Appendix B we make further use of psychometric measures to offer additional evidence consistent with our results on the hypothesized interplay of loss-framing and norm-framing.

[24] As we do not record the exact deviation for failed attemps, we multiply every failed attempt by 3, which equals the minimum amount of deviation which counts as a failure.

the stretching contract (0.60). The differences are statistically significant at below the 5% level (Mann-Whitney ranksum).[25] Moreover, comparing Tables 1 and 4, we observe that the plain-vanilla and stretching contracts induce higher effort levels than the low-bar and the extreme-effort contract along both the quantity and the quality dimensions. The positive relationship between quality and quantity is significant (Linear regression, Coef=-1.35, $p<0.051$).[26] This finding seems to run counter the intuition of the multitasking model (Holmstrom and Milgrom, 1991), which would predict that stronger incentives to increase quantity lead to more shirking on the quality dimension.[27] An account consistent with our results might be that contract frames which motivate participants more make them exert more effort on both the quality and the quantity dimension of effort.[28]

4.4 Contract Rejection Rates

Thresholds might also communicate information about the difficulty of the task. Participants might infer from a low threshold that they are only expected to count very few screens suggesting that the task is difficult. In this case, subjects take a high threshold to mean that the task is easy and a low threshold that the task is difficult. This effect should only be present at the time of the opt out decision while later, when subjects have started the task, they will have learned about its difficulty. We would therefore expect that opt out rates increase for lower thresholds as participants avoid the more difficult task.

Table 3 reports the contract rejection rates for the different treatments. Rejection rates under the plain-vanilla and the low-bar contract are 19.4%. They decrease for the stretching

[25] Participants under the plain-vanilla treatment make fewer mistakes than under the extreme-effort treatment (Mann-Whitney test, $p<0.01$) and the low-bar treatment ($p=0.01$). The same holds true when comparing the stretching contract to the extreme-effort ($p<0.01$) and the low-bar condition ($p<0.01$).

[26] Note that our quality measure is an inaccuracy score. Therefore a negative value of the coefficient implies a positive relationship between quanity and quality. The number of observations is 193.

[27] Note that in our context, incentives become stronger because framing affects the mapping of payoffs into agent's utility, while in the original Holmstrom and Milgrom (1991) paper, incentives become stronger because of a direct manipulation of payoffs.

[28] However, our design is not well suited to test the multi-tasking theory. While participants might be able to save time by lowering the accuracy with which they perform the task, it might also slow them down if participants cannot perfectly control their level of accuracy. This is becauses lower accuracy might increase the probability of having to recount screens as answers fall outside the margin of error.

contract (16.9%) and then increase dramatically for the extreme-effort contract (37.1%). While rejection rates under the stretching and the low-bar contract do not differ statistically from one another or from those under the plain-vanilla contract, we find that the extreme-effort contract leads to a significantly higher rejection rate (37.1%) than all other contracts.[29] This behavior cannot be explained by assuming that participants infer from a high threshold that the required task is likely to be easy.

Table 3: Rejection Rates Across Treatments

contract type	threshold	rejection rate
plain-vanilla	– (N=50)	19.4%
low-bar	5 (N=50)	19.4%
stretching	15 (N=54)	16.9%
extreme-effort	50 (N=39)	37.1%

Imas et al (2015) find that subjects might prefer loss contracts over gain contracts, suggesting that subjects select into loss contracts as a commitment device to improve performance. Our (statistically not significant) finding that subjects reject stretching contracts less often than low-bar contracts is consistent with this result. However, the fact that the rejection rate shoots up for the extreme-effort contract suggests that preferences for loss contracts might be non-monotonic.

5 Conclusion

Prospect Theory suggests that if principals frame incentives as losses rather than gains, agents should exert greater effort. We show that this implication is not so simple. When thresholds are used to trigger loss aversion, contract framing entails a risk. In this paper we demonstrate that people tend to meet expressed thresholds around which earnings are framed as losses or gains. This norm-framing effect can trump the impact of loss aversion.

[29] The two-sided Fisher exact test is significant at the 5% level for the difference between the extreme-effort contract and all the other contracts (0<50: p<0.05; 5<50: p<0.05; 15<50: p=0.02). The test is far from being significant for the difference between the other contracts (0<5: p=1.00; 0<15: p=0.82, 5<15: p=0.82).

By stipulating a demanding but reasonable stretching threshold, a principal can maximize the mean effort of her employees. However, demanding too little may dampen the effort level compared to the expected performance under a simple contract that does not impose a threshold. In addition, a threshold pushed too high can similarly depress effort: An extreme loss frame may push the agent in an area of his value function where the slope is lower than under the plain-vanilla contract and undermine intrinsic motivation as the agent is unlikely to meet the threshold. This risk of contract framing is not marginal. The optimal stretching threshold will depend on the production function of the agent. But this information is often private and unobservable. Unless the principal has nuanced information about her agent's abilities and the technology available to the agent, or can invest in obtaining such information, a second-best plain-vanilla contract will often be a better choice.

Apart from informing contract drafters how to set quantity/quality thresholds in contracts, the insights from this study can be applied to other areas. Companies are setting production targets for their manufacturing workers and sales targets for their sales personnel. Law firms are setting billable hours targets for associates.[30] Often, in these cases, there is an informal target below which disciplinary action is taken, a formal target where bonus payments kick in, and an even higher informal target above which an employee is considered for promotion. Similarly, regulators concerned about lowering health care costs often allocate prescription drug budgets to doctors. If doctors prescribe more drugs they face a penalty. If they prescribe fewer drugs, they get a bonus. Our study informs regulators how tightly to set those budgets.

[30] Targets at law firms vary between 1700 and 2300 billable hours, but it is not clear whether this might be law firms trying to screen applicants for different abilities.## http://www.law.yale.edu/studentlife/cdoadvice_truthaboutthebillablehour.htm

References

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

Altmann, S., and A. Falk (2009): "The Impact of Cooperation Defaults on Voluntary Contributions to Public Goods," *mimeo*.

Brooks, R. W.; Stremitzer, A., and S. Tontrup (2011): "Contracts - Why Loss Framing Increases Effort," *Journal of Institutional and Theoretical Economics*, 168, 62–82.

Eggers, W. D. (1997): *How-to Guide No. 17, May 1997*. Reason Foundation (Mar. 7, 2013, 9:45 PM), http://reason.org/files/e0f57a04efae97a83b33c7520dc65dc8.pdf.

Falk, A., and M. Kosfeld (2004): "Distrust: The Hidden Cost of Control," *Zurich, IEW Working Paper No. 193*.

Fehr, E.; Klein, A., and K. M. Schmidt (2007): "Fairness and Contract Design," *Econometrica*, Vol. 75, No. 1, 121–54.

Frederick, S. (2005): "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19(4), 25–42.

Fryer, R. G. J., S. D. Levitt, J. List, and S. Sadoff (2012): "Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment," *NBER Working Paper No. 18237*.

Goette, L; Huffman, D., and E. Fehr (2004): "Loss Aversion and Labor Supply," *Journal of the European Economic Association*, 2(2-3), 216–28.

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

Hoppe, E. I., and P. W. Schmitz (2011): "Can Contracts Solve the Hold-Up Problem? Experimental Evidence," *Games and Economic Behavior*, 73(1), 186–199.

Hossain, T., and J. List (2009): "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations," *NBER Working Paper*, No. 1562.

Hossain T.; List, J. (2012): "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations," *Management Science*, vol. 58 no. 12, 2151–2167.

Imas, A., S. Sadoff, and A. Samek (2015): "Do People Anticipate Loss Aversion?," *CESifo Working Paper No. 5277*.

Kahneman, D., and A. Tversky (1979): "Prospect Theory: An Analysis of Decisions under Risk," *Econometrica*, March 1979, 47(2), 263–91.

KAUR, S.; KREMER, M., AND S. MULLAINATHAN (2010): "Self-Control and the Development of Work Arrangements," *American Economic Review*, 100(2), 624–28.

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

LAURY, S. K. (2006): "Pay One or Pay All: Random Selection of One Choice for Payment," *Andrew Young School of Policy Studies Research Paper Series*, No. 06-13.

LAURY, S. K., AND C. A. HOLT (2008): "Payoff Scale Effects and Risk Preference Under Real and Hypothetical Conditions," in *Handbook of Experimental Economics*, ed. by C. A. Plott, and V. Smith, pp. 1047–53.

LAZEAR, E. P., U. MALMENDIER, AND R. A. WEBER (2012): "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 4(1), 136–63.

MUSCH, J.; REIPS, U.-D. (2000): "A Brief History of Web Experimenting," in *Psychological Experiments on the Internet*, ed. by M. Birnbaum, chap. 3, pp. 61–88. Verlag Academic Press.

VOLTAIRE, F.-M. (1770): *La Begueule (The Prude Woman)*. http://fr.wikisource.org/wiki/La$_B$

A Appendix A: Robustness-Checks

Table 4: Mean Effort Levels Compared Across Treatments (including rejections)

contract type	threshold	observed mean	contract type: plain-vanilla	contract type: low-bar	contract type: stretching	contract type: extreme-effort
plain-vanilla	—	8.4 (N=62)	—	$Z = 2.4^{**}$ $p = 0.02$	$Z = -2.1^{**}$ $p = 0.04$	$Z = 3.1^{***}$ $p < 0.01$
low-bar	5	5.0 (N=62)		—	$Z = -4.1^{***}$ $p < 0.01$	$Z = 1.8^{*}$ $p = 0.08$
stretching	15	11.9 (N=65)			—	$Z = 4.2^{***}$ $p < 0.01$
extreme-effort	50	5.3 (N=62)				—

B Appendix B: Additional Evidence Based on Psychometric Measures

B.1 Hypotheses

As loss aversion is correlated with a subject's sensitivity to reference points, the effects predicted by Prospect Theory should be larger for subjects scoring higher on a psychometric test of loss aversion. We can derive the following hypotheses:

Hypothesis 3 *The higher a participant's loss aversion,*
i) the smaller is the drag-down effect under the low-bar contract predicted in Hypothesis 1.1 (H3.1),
ii) the larger is the pull-up effect under the stretching contract predicted in Hypothesis 1.2 (H3.2.),
iii) the larger is the drag-down effect under the extreme-effort contract predicted in Hypothesis 2 (H3.3),

We assume that subjects who are less able to resist the initial impulse to fulfill expectations are more likely to be influenced by the suggestive force of the threshold. The CRT test is a psychometric test measuring the ability to resist an initial impulse. We therefore predict that the suggestive effect of the threshold should be higher for subjects scoring low on the CRT test:

Hypothesis 4 *The lower the CRT score, the higher the fraction of subjects that will exactly match the threshold (H4).*

If (H4) holds, setting a threshold is a double-edged sword. Those participants, who under a plain-vanilla contract would complete fewer screens than the threshold demands should increase their effort and do more, while those subjects who would have exceeded the threshold, should reduce their effort on average. As we set the low-bar quantity below the mean effort level under plain-vanilla and the stretching quantity above the mean effort level under plain-vanilla we expect a drag-down effect for the low-bar contract and a pull-up effect for the stretching contract. Given that we assume convexity below the threshold $\lambda_1 > 0$, we also expect a drag-down effect for the extreme-effort contract. Again assuming that the effect is larger for subjects with low CRT scores, we can derive the following hypotheses:

Hypothesis 5 *The lower the CRT score*
i) the larger the drag-down effect in the low-bar contract predicted in Hypothesis 1.1 (H5.1),
ii) the larger the pull-up effect in the stretching contract predicted in Hypothesis 1.2 (H5.2),
iii) the larger the drag-down effect in the extreme-effort contract predicted in Hypothesis 2 (H5.3).

If the results based on our psychometric measures suggest that norm-framing drags down effort as we move from a plain-vanilla to a low-bar contract (H5.1), while Prospect Theory pushes in the opposite direction (H3.1), we could conclude that the reduced effort we observe under the low-bar contract is exclusively driven by norm-framing.

B.2 Results

We first test Hypothesis 3 that Prospect Theory explains at least part of our result. In Table 5, we compare participants with low loss aversion scores ("rational types") with those with high loss averse scores ("loss averse types") and observe that the mean drag-down effect under the low-bar treatment relative to the plain-vanilla treatment is smaller for loss averse types (2.8 screens) than for rational types (6.0 screens). We also find that the average pull-up under the stretching contract relative to the plain-vanilla treatment is 6.5 screens for loss averse types and only 0.7 screens for rational types. Both differences in difference are statistically significant at the 1% level (Mann-Whitney ranksum, $p<0.01$) supporting Hypotheses H3.1 and H3.2. We find no evidence for loss aversion to affect the drag-down effect under the extreme-effort treatment (Mann-Whitney ranksum, $p=0.5$) so that we cannot reject the null hypothesis for H3.3.

These results suggest that Prospect Theory is a behavioral channel affecting participants' behavior in the low-bar and in the stretching treatment, but not in the extreme-effort condition. This latter finding might suggest that extreme thresholds fail to establish a loss frame as agents may not take them seriously. Rather the extreme-effort treatment may reduce effort because it undermines the intrinsic motivation to comply with the norm when participants realize that they will not meet the threshold. The result that loss aversion has a positive effect on mean effort in the low-bar treatment is consistent with assuming that the reference point under the plain vanilla contract is set at the status quo level 0.[31] It also follows, that loss-framing cannot cause the drag-down effect under the low-bar treatment, suggesting that norm-framing drives this result.

[31] See discussion in Subsection 2.

Table 5: Loss Aversion and Effort Levels

		threshold contract type			
		plain-vanilla	low-bar	stretching	extreme
Rational (N=99)		11.5 (N=26)	5.5 (N=27)	12.2 (N=24)	9.5 (N=22)
	Diff		-6.0	+0.7	-2.0
Loss Averse (N=151)		9.7 (N=36)	6.9 (N=35)	16.2 (N=40)	7.9 (N=40)
	Diff		-2.8	+6.5	-1.8
Mann-Whitney			$Z = -4.1^{***}$ p<0.01	$Z = -2.7^{***}$ p<0.01	$Z = -0.7$ p=0.50

In the next step we test Hypothesis 4 which assumes that participants experience a disutility from both, falling short and exceeding the threshold, as both run counter the norm communicated through the threshold. We assumed that the norm-framing effect of the expressed threshold is due to the fact that participants take the threshold as a norm they want to conform to. It has been shown before that cognitive reflection correlates with peoples' tendencies to conform to norms (see, Altmann and Falk, 2009). If people indeed conform to the norm the threshold expresses we would expect that participants with low scores on the cognitive reflection test would be more likely to follow their initial impulse and match the threshold than participants with high scores. Table 6 compares the frequency with which low (0 correct answers) and high (1-3) CRT types exactly match the threshold in our three treatments. For the stretching contract, we observe that 31% of the low CRT types exactly meet the threshold, while 13% of the high types match it. The difference is statistically significant (one-sided Fisher exact test, p=0.02). For the low-bar treatment we observe that 33% of the low CRT types match the threshold, but only 20% of the high CRT types. However, the effect is only significant at the 10% level for the one-sided Fisher exact test (p=0.09). We thus find some support for Hypothesis 4. By contrast, the results for the extreme-effort treatment suggest that the unreasonable threshold is easier to ignore, as neither high nor low CRT types matched the threshold. The reported results are robust to different coding where we define high CRT types as those participants with 2-3 correct answers (see Table 7).

If Hypothesis 4 holds, setting a threshold becomes a double-edged sword. Participants, who under a plain-vanilla contract would do less than the threshold will do more, while those who would do more than the threshold, will do less. We therefore predicted in Hypothesis 5 that there would be a drag-down effect for the low-bar contract and a pull-up effect for the stretching contract. In Table 8, we compare high CRT types with low CRT types and observe that the mean drag-down effect under the low-bar treatment relative to the plain-vanilla treatment is higher for low CRT types (5.6 screens) than for high CRT types (2.3 screens). This difference in difference is highly statistically significant (Mann-Whitney ranksum, p<0.01) supporting Hypothesis 5.1. We also find that the average pull-up under

Table 6: Cognitive Reflection and Matching Threshold

		threshold contract type		
		low-bar	stretching	extreme
High CRT				
	match	4 (16%)	0	0
	exceed/undercut	21 (84%)	17	15
Low CRT				
	match	12 (29%)	13 (27%)	0
	exceed/undercut	29 (71%)	36 (73%)	24
Fischer exact		p=0.129	p<0.01***	-

Table 7: Robustness Check: Cognitive Reflection and Matching Threshold (recoded)

		threshold contract type		
		low-bar	stretching	extreme
High CRT				
	match	9 (20%)	5 (13%)	0
	exceed/undercut	36 (80%)	35 (87%)	24
Low CRT				
	match	7 (33%)	8 (31%)	0
	exceed/undercut	14 (77%)	18 (69%)	15
Fisher exact		p=0.09	p=0.02**	

the stretching contract relative to the plain-vanilla treatment is 6.8 screens for low CRT types but only 3.6 screens for high CRT types. However, while the trend points in the predicted direction, this difference is not statistically significant ((Mann-Whitney ranksum, p=0.21, H5.2).[32]

It follows that only norm-framing can explain the lower level of effort we observe in the low-bar treatment compared to the plain-vanilla treatment (H5.1), since loss aversion pushes in the opposite direction, increasing rather than decreasing effort (H3.1).

[32]The reported results are robust to different coding where we define high CRT types as those participants with 2-3 correct answers (see Table 6 in Appendix A).

Table 8: CRT and Effort Levels

		threshold contract type			
		plain-vanilla	low-bar	stretching	extreme
High CRT (N=99)		8.8 (N=26)	6.0 (N=21)	12.2 (N=18)	7.3 (N=15)
	Diff		-2.8	+2.8	-1.5
Low CRT (N=151)		11.2 (N=36)	6.4 (N=29)	15.6 (N=36)	9.0 (N=24)
	Diff		-4.8	+4.4	-2.2
Mann-Whitney			$Z = -4.3^{***}$ p<0.01	$Z = -1.1$ p=0.27	$Z = -1.2$ p=0.22