EUI DEPARTMENT OF ECONOMICS

Three Essays in Experimental Economics

Essi Kujansuu

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Florence, 21 June 2021

European University Institute
Department of Economics

Three Essays in Experimental Economics

Essi Kujansuu

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Examining Board

Michele Belot, Cornell University, Supervisor
Arthur Schram, EUI and University of Amsterdam, Co-Supervisor
Heike Hennig-Schmidt, University of Bonn
Christina Gravert, University of Copenhagen

Researcher declaration to accompany the submission of written work Department Economics - Doctoral Programme

I Essi Susanna Kujansuu certify that I am the author of the work Three Essays in Experimental Economics I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party.

I declare that this work consists of 61250 words.

Statement of inclusion of previous work:

I confirm that chapter 1 was jointly co-authored with Dr Arthur Schram and I contributed 60% of the work.

I confirm that chapter 1 is being published in the Journal of Economic and Behavioral Organization.

Signature and date:

In Pöytyä, Finland, 14 July 2021

Abstract

The first chapter studies how a gift exchange labor market reacts to the occurrence of negative shocks. One-round shocks may hit either workers' wages or employers' earnings. In our model, other-regarding preferences suffice to predict gift exchange and wages above the competitive level. The model predicts wage rigidity if we add wage illusion and loss aversion. Using a real-effort laboratory experiment, we find support for this model. When there are no shocks, there is gift exchange. After a wage shock we see strong nominal wage rigidity and no impact on workers' effort, as predicted. Rigidity is also observed after a productivity shock, but here we observe increases in effort, especially at low wages. The latter is contrary to the model predictions and suggests that productivity shocks alter gift-exchange patterns.

The second chapter studies how workers' effort responds to wage cuts and whether employers anticipate these reactions correctly. Hiring happens by initial offers before shocks are announced. The non-binding offers can then be adjusted. To model responses to wage cuts, I add negative reciprocity to the previous model. Without shocks, employers should never cut wages. Adjustment might be, however, justifiable after a shock if sharing its burden is considered to be fair. With a laboratory experiment, I find that although wage cuts are counterproductive, their effects are insignificant in the absence of shocks. After shocks, wage cuts are punished, regardless of whether the shock hits employers or workers.

The third chapter studies if nudges influence behavior as effectively when people are aware of nudging as when they are unaware. I use a limited attention model that distinguishes between two kinds of nudges. System 1 nudges (e.g., defaults) provide quick decision-making shortcuts, while System 2 nudges encourage reflective thinking, e.g., cost-benefit analysis. Transparency is predicted to reduce the effectiveness of System 1 nudges but not that of System 2 nudges. Moreover, conditional on Choice Architects having image concerns, transparency is predicted to reduce the use of System 1 nudges while increasing the use of System 2 nudges. With an online framed field experiment, I find that transparency does not change how Choice Architects use nudges. The effects of System 1 nudges are somewhat weakened by transparency, but System 2 nudges are unaffected.

Acknowledgements

The first acknowledgements rightfully belong to my advisors. Thank you, Arthur Schram, for teaching experimental economics at EUI and introducing me to the methodology. Your presence there from the beginning to the end has been much appreciated. Thank you, Michèle Belot, for all your advice and enthusiasm. Thanks to Andrea Galeotti for the early guidance and for inspiring me to not cut my hair until the thesis is defended. Christina Gravert and Heike Hennig-Schmidt, thank you for all the valuable comments and for agreeing to be in the committee.

Special thanks to all those of you, too many to mention by name even if it was not against all ethical rules, for participating in my many pilot experiments. Christine Alamaa, Martina Vecchi, Egon Tripodi, Christian Meyer and Junze Sun, you were an inspiration and source of wisdom. Phillipp Chapkovski, I learned a lot from you. Chiara and Chiara, many thanks for the work you put into the translations. I owe you.

Thank you Rossella, Lucia, and Sarah, for keeping things flowing, and Antonella and Lori for the sympathy, coffee, and cake.

There are many people that I met at the EUI through either music or football. You gave a great counterbalance to the PhD. Thank you, Coppa Pavone, my teammates in Mad Cows, Hoops band, Fiasco, Villa YOLO, and all the people there. I have great memories with you all!

I made many great friends during these years, I wish I had the space to mention you all. Thank you Agnès and Chiara for inventing the tea table in the greenhouse, Oliko and Simon for helping me with my theories, Chiara for all the moments of music, Motyo and Palma for the wine and cheese, Rafa for all the random discussions, Yorgos for always staying up late. We had a great cohort!

The thesis was completed during the covid-19 pandemic. Agnès, thank you for keeping me sane through the first Italian lockdown. Ale and Dalila, thank you for the pandemic friendship. Thanks to the Academy of Finland for continuing to fund my PhD also through the pandemic.

Many thanks for my friends and family back home, for the many New Year's Eves and Christmases, for the great times and constance. Mom and dad, thank you for the support throughout the years. Last, a special mention to all of those who call me an aunt — you

have been a great source of brightness and fun! Hugs.

Contents

Introduction

In the first chapter *Shocking Gift Exchange*, written together with Arthur Schram, we study how a gift exchange labor market reacts to the occurrence of negative shocks. One-round shocks may hit either workers' wages or employers' earnings (via worker productivity). In our model, other-regarding preferences suffice to predict gift exchange and wages above the competitive level (no assumption of reciprocity is required). The model predicts wage rigidity if we add wage illusion and loss aversion. Using a real-effort laboratory experiment, we find support for the model. When there are no shocks, there is gift exchange. We see strong nominal wage rigidity after wage shocks and no impact on workers' effort, as predicted. Rigidity is also observed after a productivity shock, but here we observe increases in effort, especially at low wages. The latter is contrary to the model's predictions and suggests that productivity shocks alter gift-exchange patterns. We conclude that the wage rigidity often observed in the field can be explained by boundedly rational workers with social preferences.

In the second chapter *Fairness of Wage Cuts*, I study how workers' effort responds to wage cuts and whether employers anticipate these reactions correctly. As in the previous chapter, workers and employers engage in a real-effort gift exchange market with negative shocks. This time, however, hiring happens by initial offers before potential shocks are announced. The non-binding offers can then be adjusted. Effort is non-contractible and determined after workers learn their final wages. To model responses to wage cuts, I add negative reciprocity to the gift exchange model from the previous chapter. As a result, employers should never cut wages without the shocks. Adjustment might, however, be justifiable after a shock if sharing the burden with the other party is considered to be fair. I also consider maintaining the status quo as an alternative standard for fairness and, interestingly, this standard gives predictions about real wage cuts that are similar to nominal illusions. I test the predictions with a laboratory experiment. I find that wage cuts are counterproductive as long as the initial wage offer is not unreasonably high, however, the effect is small and statistically insignificant. Wage cuts after a shock, on the other hand, are punished. Surprisingly, it does not matter whether the shock hits predominately the employer or the worker: in both cases the cuts lead to considerably lower effort. The shocks on their own, in the absence of wage cuts, do not have a significant effect on effort, meaning that workers do not reduce effort in response to real wage shocks. Finally, employers cut wages more often than what is optimal, suggesting that they do not fully anticipate the effort responses by workers. In conclude that the results indicate that keeping the status quo is a stronger reference point for fairness than equity when adjusting to shocks.

The two first chapters are are closely related, for which reason, I add a few observations here comparing the two settings and their results. First, the possibility to adjust wages appears to make gift exchange stronger in the second setting. My interpretation is that the mere possibility of adjustment invokes positive reciprocity towards those employers that keep their promise of the initial wage offer. Perhaps related to this, the asymmetric gift exchange pattern is maintained in the second setting even when shocks and wage cuts occur, while in the first chapter we find that this pattern is weakened by the shocks. Furthermore, in the second setting it is generally speaking profitable for employers to pay wages up to the level we call 'the objectively fair wage level'. In the first setting, employers do not make more money paying the fair wage, although neither do they significantly lose money. So although one might intuitive expect the possibility to adjust wages to dampen gift exchange, we observe the opposite, that is, (unused) opportunities to adapt wages strengthens gift exchange.

This observation fits the literature that compares different wage-setting institutions. Charness (2004) finds that when the wage is not set by the (self-interested) employer but by a random mechanism or a dis-interested person, wage cuts do not trigger negative reciprocity. Similarly, Brandts and Charness (2003) look at look the role of intentions communicated via cheap talk and they find strong negative reciprocity (frequent punishment after deceptive communication) and weak positive reciprocity (less frequent rewarding behavior after truthful communication). Bartling and Schmidt (2015) demonstrate that renegotiation is not a neutral procedure. They find that having a previous contract leads to lower markups and higher rejection rates than when there is no history, regardless of the level of competition. Also Fehr et al. (2011) find a tradeoff between contract rigidity and the ability to adjust terms of trade in adverse conditions. Rigidity makes adverse conditions more harmful, but on the other hand, when contracts are more flexible, workers also expect firms to be more generous, predisposing the firms for negative reciprocity. Similarly, imposing minimum requirements may be considered to be a sign of mistrust and therefore punishable (Falk and Kosfeld, 2006).

In the light of this literature, it is not surprising that wage-setting procedures change the nature of the worker-employer interaction, which can be seen particularly in how the workers' respond to the shocks. While a shock on the employers' earnings leads to higher effort in the first chapter, no such effect is observed in the second chapter. Instead, workers cut effort to punish any cuts that happen after a shock, including shocks on the employers' earnings. What the procedures change is how certain actions and outcomes are interpreted. In general, wage cuts have huge potential to be interpreted as *unkind* rather than as being *justified* or *fair.* The procedures can make thus one interpretation more prevalent than the other.

In the last chapter *Choice Architecture and Transparency*, I investigate if nudges influence behavior as effectively when people are aware of nudging, that is, nudged transparently, as when they are unaware of nudging. I also investigate if Choice Architects decisions over nudges change with transparency. I model behavior with a limited attention model that distinguishes between two kinds of nudges, in line with Kahneman's framework for fast and slow thinking (System 1 and System 2). System 1 nudges provide quick decision-making shortcuts, for instance, default options, while System 2 nudges encourage reflective thinking, for example, by encouraging cost benefit analysis. I show how the framework of slow and fast nudges can explain some contradicting results reported in the previous literature. The model predicts that transparency increases attentiveness and thus reduces the effectiveness of System 1 nudges. Transparency is not expected to make System 2 nudges weaker. Moreover, conditional on Choice Architects having image concerns, transparency is predicted to reduce the use of System 1 nudges but increases the use of System 2 nudges. I test these predictions in an online framed field experiment. I find that transparency does not change how Choice Architects use nudges. The effects of System 1 nudges are weakened by transparency, but System 2 nudges remain unaffected. I also show how the framework of System 1 and System 2 nudges can explain some contradicting results reported in the previous literature.

Bibliography

Bartling, Björn, and Klaus M Schmidt. 2015. "Reference points, social norms, and fairness in contract renegotiations." *Journal of the European Economic Association*, 13(1): 98–129.

Brandts, Jordi, and Gary Charness. 2003. "Truth or consequences: An experiment." *Management Science*, 49(1): 116–130.

Charness, Gary. 2004. "Attribution and reciprocity in an experimental labor market." *Journal of labor Economics*, 22(3): 665–688.

Falk, Armin, and Michael Kosfeld. 2006. "The hidden costs of control." *American Economic Review*, 96(5): 1611–1630.

Fehr, Ernst, Oliver Hart, and Christian Zehnder. 2011. "Contracts as reference points—experimental evidence." *American Economic Review*, 101(2): 493–525.

Chapter 1

Shocking Gift Exchange

joint with Arthur Schram[1]

1.1 Introduction

Labor markets are often considered *rigid*; wages do not adjust quickly to changes in market conditions, particularly if there is downward pressure to cut them (Bewley 1999, Dickens et al. 2007). Rigidity can be harmful when it stops markets from clearing, which, for example, can bring about involuntary unemployment. Rigidity would not occur if market forces determined wages in the labor market. Labor relations, however, are often characterized by incomplete contracts on the one hand (Milgrom and Roberts, 1992), and trust and reciprocity on the other. There are many examples of how the effects of moral hazard are mitigated by trust and reciprocity between employers and workers and mutual regard for each other's well being, rather than by attempts to reach more complete contracts (Fehr et al. 1997; Gächter and Fehr 2002). This may be partly attributed to psychological reactions to market forces. For example, workers often perceive wage cuts as unfair and demotivating (Bewley 1999; Kahneman et al. 1986). Fairness and motivation are concepts that are deemed to play a central role in labor relations. Indeed, experiments have shown that cuts to nominal wages are considered unfair and lead to lower effort exerted by the worker (e.g., Hannan 2005, Kube et al. 2013, Cohn et al. 2015, and Koch 2021).

If labor relations are not purely market interactions, they might not follow the conventional rules of supply and demand when adjusting to shocks. A seminal and simple theory to explain why this might occur was presented as the 'fair wage-effort hypothesis' by Akerlof and Yellen (1990). The basic idea is that workers have a notion of a wage level

[1] We are grateful for the comments by Andrea Galeotti, Simon Skipka, and Aljaž Ule and thank Philipp Chapkovski for his help with programming. We gratefully acknowledge funding from the Academy of Finland and the European University Institute.

that is deemed 'fair'. They will respond negatively (by reducing effort) to wages below this level. The effect is, however, asymmetric; there is no effort response to wages above the fair level. The fair wage is thus defined as the point at which a positive relationship between wage and effort levels out. Empirical support for such a kink in the effort-wage relationship is provided by, among others, Mas (2006), Gächter and Thöni (2010), Kube et al. (2013), Cohn et al. (2015), and Sliwka and Werner (2017). The existence of such a fair wage level may induce employers to offer wages that are higher than the market-clearing level. They then trust that workers will reciprocate with their effort levels. Note, however, that it is not a priori obvious at what wage level the kink will occur (that is, what constitutes a fair wage). Moreover, it remains unknown if and how such fair-wage reference points adjust to shocks in a gift exchange market. We explore these issues in this paper.

This paper applies the accumulated knowledge about labor relations with incomplete contracts in an attempt to better understand wage rigidity; it studies how one-time negative shocks in earnings are absorbed in a gift exchange labor market. Gift exchange describes a two-player interaction where the first mover offers a benefit ('gift') to a second mover without any certainty that the second mover will honor the expectation of a counter-gift. In the labor market context with moral hazard, this means that a wage above the market-clearing wage is offered while expecting this to be responded to with higher than minimal effort. Under the fair wage-effort hypothesis, this gift exchange is observed only for wages below the fair-wage level (Gächter and Thöni, 2010). Gift exchange is thus based on social relations, such as the above-mentioned trust and reciprocity or the other-regarding preferences that we use later in our model. It can improve moral hazard situations in which standard rational behavior would cause the market to fail (Akerlof 1982, Mauss 2002). We refer to the observed relationship between wages and effort in a setting of moral hazard as the 'gift-exchange pattern'.

Gift-exchange patterns point to the possible advantages of wage rigidity. In preventing wage cuts, rigidity could simultaneously prevent subsequent drops in labor productivity that would occur in response to wages that are lower than those deemed fair. Indeed, depending on the specific pattern, rigidity may even be an optimal strategy for employers (Fehr et al. 1993). Thus, we are interested here in the gift exchange patterns that occur, and specifically in the extent to which these can explain the observed wage rigidity after negative shocks. Importantly, we do not consider the gift-exchange pattern as given; we recognize that shocks may affect the pattern itself. Note that in studying this, we abstract away from institutional factors that prevent wage adjustments, such as unions, collective bargaining or binding contracts. This allows us to isolate gift exchange patterns, and more specifically the role they play in wage rigidity.

We derive predictions from a simple model of gift exchange with a fair-wage reference point. These predictions are subsequently tested in a laboratory experiment. The structure of the model and the experimental design build on the seminal experiment by Fehr et al. (1993). It adds to the original by using a real-effort task to measure workers' productivity and, in particular, by adding one-time negative shocks. A novelty of this paper is also to vary the side of the market that receives the negative shock in the gift exchange labor market. The random shocks come in two types, a cut in (all) workers' wages and a reduction in (all) workers' productivity. The wage shock causes a real wage cut that keeps the nominal (gross) wages intact but reduces the net wages for all workers. The productivity shock reduces all employers' earnings for any given effort level. These two shocks allow us to alternate who benefits most from maintaining the status quo. Note that our interest lies in temporary shocks that affect either net wages or productivity for one round only. This is because we see the labor market that we create as matching workers to employers for a length of time (e.g., a year) in which shocks (like a pandemic) may happen that will have faded away by the time the next round starts.[2]

The earlier non-experimental literature shows that real wage cuts through inflation are not perceived to be as unfair and demotivating as nominal wage cuts are (e.g., Kahneman et al. 1986, Kaur 2019). This is observed even when the economic consequences are equal, that is, when the achievable bundle of consumption goods is equally reduced. Related experimental literature has studied the effects of nominal wage changes while holding real wages constant, the opposite case to ours. For instance, a real-effort experiment by Fochmann et al. (2013) finds that subjects work harder and longer, the higher the nominal wage is, even when this higher wage is accompanied by a change in the tax rate that keeps the real wage constant. This is referred to as 'net wage illusion'. One extension of our model will allow for net wage illusion.

Our paper is not the first to study wage rigidity in the laboratory. In a gift-exchange context, strong wage rigidity in response to shocks is not a common experimental result. For example, Koch (2021) finds that the average wage is lower after a shock has occurred than when there is no shock, although some rigidity remains as wages do not adjust fully. Gerhards and Heinz (2017) use a two-round laboratory market where the employer might be hit by an external shock in the second round. In their experiment, employers pay on average lower second-round wages if a shock is realized and workers do not subsequently reduce effort in response to the lower wages. They also observe that the mere possibility of a second-round shock makes both first-round wage and effort adjust upwards. We will see, however, that our results over time show strong learning effects in the first two rounds. Reference points (and rigidity) require time to develop, but once so, they remain stable. This casts some doubt on the external validity of previous studies that rely on

[2] As an anonymous reviewer pointed out, one could also consider permanent shocks, which may be seen as a regime change. We leave this for future research.

only one or two rounds. Last, Buchanan and Houser (forthcoming) find that about half of the employers cut wages, and they are punished for it by reduced effort. With hindsight, they estimate that rigidity is the optimal policy for employers.

In two related experiments, by Rubin and Sheremeta (2015) and Davis et al. (2017), a gift exchange market is shocked with on-average neutral events that vary how well effort translates to output. Both papers find that such shocks reduce wages. Rubin and Sheremeta (2015) conclude that welfare is reduced by these shocks despite the fact that they have zero impact on average productivity. Davis et al. (2017) speculate that the reason underlying the lower welfare is not the shocks themselves but the history of shocks that in some cases triggers hysteresis. Our data do not allow us to study the role of hysteresis.

Finally, the experimental literature on shocks to employers' earnings has established that workers' effort is sensitive to the surplus of the employer (e.g. Hannan 2005, Hennig-Schmidt et al. 2010, Koch 2021). This is what we also observe. Interesting here is the asymmetry: our results show that wages are not 'required' to adjust after the workers have been hit by a shock, yet the workers do adjust to the shocks experienced by the employer. To our knowledge, we are the first to observe this. Moreover, from the welfare comparisons between treatments, we find a clear indication that 'shock-fairness' matters; welfare is highest in the setup where either party can experience a shock.

Our paper aims to contribute to the literature on how labor markets adjust to shocks in various ways. What we have in common with the above-mentioned studies is that we study this in a gift-exchange context, building on the work of Fehr et al. (1993) and Fehr et al. (1997). While Rubin and Sheremeta (2015) and Davis et al. (2017) add (on-average neutral) shocks to labor productivity and conclude that these reduce gift exchange, Koch (2021), Gerhards and Heinz (2017), and Buchanan and Houser (forthcoming) introduce purely negative shocks (in the latter two studies these shocks are permanent). Our work differs from these other studies in various important ways. First, we believe to be the first to consider equivalent (temporary) shocks on both sides of the labor market. Second, to the best of our knowledge, our paper is the first to experimentally study the effects of real wage cuts while keeping nominal wages and employer profits constant.[3] Third, in order to make the effects of shocks more salient we use a real effort task instead of stated effort. Real effort allows one to also capture subconscious effort responses, such as reductions in motivation that might negatively affect prolonged concentration; it also allows for an intrinsic motivation to work. In the world outside the laboratory, effort is real and workers typically desire this to be recognized by their employer. It is unclear whether such elements can be captured in a stated-effort design. Finally, we stay close

[3] Buchanan and Houser (forthcoming) do consider the case of real wage cuts when there are permanent shocks.

to the original Fehr et al. (1993) design by matching participants anonymously through a market with excess supply of labor. These other studies are based on pre-determined pairs and often on repeated within-pair interactions. While this makes those studies relevant for principal-agent relationships within firms, ours aims at studying the effects of shocks on gift exchange patterns in the labor market more generally, where periods of unemployment and relative inactivity are also possible.

Our theoretical model starts with a simplified version of the Charness and Rabin (2002) framework. It allows individuals to derive disutility from inequality in payoffs, similar to the approach of Benjamin (2015). We then introduce loss aversion and net wage illusion (as explained below). This is in contrast to Dickson and Fongoni (2019), who model gift exchange based on *work morale* and reference points. Their model does not provide much rationale for how and why effort would react to shocks. We are interested in such predictions, which our approach provides. When there are no shocks on either side of the market, the model predicts wages above the competitive level together with gift exchange. A wage shock is predicted to have no effect on wages or effort. In this way, the model predicts wage rigidity. Wage rigidity is also predicted if a productivity shock occurs, but this also leads to a reduction in effort for the simple economic reason that effort is less productive.

Our experimental results in the absence of shocks confirm previous findings on gift exchange. The fact that we do so in a real-effort experiment is evidence of the robustness of the traditional results. Our experimental treatments with shocks show three main findings. First, we confirm the model's predictions on wages as we observe strong wage rigidity. Wages do not react systematically to realized shocks. Second, although we do not find that wages are significantly higher when shocks *might* occur, neither do we find that the shocks significantly reduce welfare in ex-ante terms. The market seems to adjust to the risk of shocks in a way that largely stabilizes welfare. Our third main finding is that gift exchange (the workers' effort responses to wages) is not affected by real wage cuts. Productivity shocks, however, lead to increases in effort (where decreases were predicted), especially at lower wage levels. This suggests that productivity shocks cause a shift in workers' fairness standards.

The remainder of this paper is organized in the traditional way. Our model is presented and analyzed in Section 1.2. Section 1.3 presents the experimental design and procedures. The results are presented and discussed in Section 1.4 and a concluding discussion is in Section 1.5.

1.2 Theory

In this section, we present a model of gift exchange to analyze the interaction between a worker and an employer when both (may) have social preferences. We will subsequently use this model to predict the effects of shocks. The basic setup is a simple one-shot, two-player gift exchange game between an employer and a worker.[4] A minimum wage level applies, which we normalize to zero.

The game consists of two-stages:

- In the *first stage* the employer sets a wage $w \geq 0$ for the worker.
- In the *second stage* the worker observes w and chooses effort $e \geq 0$; that is, effort is non-contractible.

We will start with a model of gift exchange in which actors exhibit other-regarding preferences and then discuss the effects of the two shocks. Then we introduce well-established elements of bounded rationality into the model and study how they change the way the shocks are absorbed. We conclude with a set of theoretical predictions derived from the models.

1.2.1 A Model of Gift Exchange

Following the logic of backward induction, we first consider how workers in the second stage respond with effort to a given wage, which is independent of the effort. We then model how employers set the wage in the first stage, given the workers' best response function. At this point, we are not yet considering shocks.

Worker's Effort Choice

Utility. The worker's utility, denoted by u^W, is captured by the expression:

$$u^W = (1 - \beta(e))w + \beta(e)(f(e) - w) - c(e). \tag{1.1}$$

Utility thus depends on the worker's monetary payoff (wage, w); the (utility) costs of exerting effort, $c(e)$, and a social preference term reflecting the difference between the employer's monetary earnings and the wage. Employers' earnings consist of the (monetary) benefits that the worker's effort generates, depicted by $f(e)$, minus the wage. We interpret that $f(e)$ captures worker productivity, which depends on the effort that she exerts.

[4] In the experiment, employers are linked to workers via an anonymous hiring market. For simplicity, we assume here that the two are already linked. We think of the equilibrium wage in our model as the wage offered (and accepted) on the market.

The function β is derived from a simplified version of the Charness and Rabin (2002) model. This allows one to capture various types of social preferences in a single framework.[5] For example, it allows individuals to derive a disutility from an inequality in payoffs. The reaction to the inequality may differ, depending on whether they are earning more or less than the employer. Inequality here is simply defined by the monetary earnings.[6] An often-made assumption introduced by Fehr and Schmidt (1999) is that individuals dislike disadvantageous inequality more than they dislike advantageous inequality. When the worker earns less than the employer, $w < f(e) - w$, the preference in the Fehr and Schmidt (1999) model is captured by a parameter $\sigma < 0$; when the worker earns more than the employer, $w > f(e) - w$, the preference is captured by parameter $\rho > 0$. We follow Charness and Rabin (2002), however, and allow for a more general class of other-regarding preferences by not restricting σ to be negative and only assume $\rho > \sigma$ and $\rho > 0$.[7] In summary,

$$\beta(e) = \begin{cases} \sigma, & \text{if } w < \frac{f(e)}{2} \\ 0, & \text{if } w = \frac{f(e)}{2} \\ \rho, & \text{if } w > \frac{f(e)}{2}. \end{cases} \tag{1.2}$$

Before we derive a best response function for a worker, we make some functional assumptions. The costs of effort are assumed to be a strictly convex function of the effort exerted, $c'(e) > 0$ and $c''(e) > 0$. In addition, we assume $c(0) = 0$. The benefit that effort generates is assumed in turn to be a concave function of the effort, $f'(e) > 0$ and $f''(e) \leq 0$, while no effort means no benefits, $f(0) = 0$. To ensure that a positive level of effort is efficient, we assume $\lim_{x \downarrow 0} f'(0) > \lim_{x \downarrow 0} c'(0)$.

Best Response. A worker maximizes u^W in eq. (1.1), that is, for any given w she chooses e such that

$$\frac{c'(e)}{f'(e)} = \beta(e). \tag{1.3}$$

[5] We do not include reciprocal preferences, which are also part of the Charness and Rabin (2002) model.

[6] We assume that workers do not take into account social preferences that the employer may have, nor do they account for their own social preferences or effort costs when comparing themselves to the employers. This is grounded in the so-called availability heuristic (Kahneman et al. 1982), as payoffs are the only comparative metric readily available in the experiment. While this assumption simplifies the analysis, extending the model by, for example, including effort costs to the inequality comparison does not qualitatively change the predictions.

[7] When the payoffs are equal ($w = \frac{f(e)}{2}$), the weight β is assumed equal to zero. This does not mean that the employer's income plays no role; as long as the earnings remain equal, changes in one's own payoff are perfectly aligned with changes in the employer's. Of course, as soon as a change causes differences in the earnings, the worker will attribute a non-zero weight to the employer's earnings.

The best response of a worker, $\hat{e}$, thus depends on her social preferences. Note that $\hat{e}$ varies with w because $\beta(e)$ depends on w (eq. (1.2)). Denote by $\hat{e}_\sigma$ the solution to eq. (1.3) for $\beta(e) = \sigma$, and by $\hat{e}_\rho$ the solution for $\beta(e) = \rho$. For $\sigma < 0$ we have a corner solution $\hat{e}_\sigma = 0$. Beyond this corner solution, the solution is increasing in β because $\partial(\frac{c'(e)}{f'(e)})/\partial e > 0$. Thus, $\sigma < \rho$ together with $\rho > 0$ implies that $\hat{e}_\sigma < \hat{e}_\rho$; that is, optimal effort is lower with disadvantageous inequality than with advantageous inequality. Finally, denote by $\hat{e}_0(w)$ the effort level that equalizes earnings between worker and employer; this is implicitly defined by $w = \frac{f(\hat{e}_0)}{2}$.[8]

Result 1. The worker's best response function is given by

$$\hat{e}(w) = \begin{cases} \hat{e}_\sigma, & \text{if } w < \frac{f(\hat{e}_\sigma)}{2} \\ \hat{e}_0(w), & \text{if } \frac{f(\hat{e}_\sigma)}{2} \leq w \leq \frac{f(\hat{e}_\rho)}{2} \\ \hat{e}_\rho, & \text{if } w > \frac{f(\hat{e}_\rho)}{2}. \end{cases} \tag{1.4}$$

Eq. (1.4) implies that effort is non-decreasing in wage.[9] Moreover, the second line on the r.h.s. shows that (because $\hat{e}_\sigma < \hat{e}_\rho$) there is a range of wages for which workers choose an effort level that equalizes earnings. Figure 1.1 illustrates this best response function.[10]

The effort function is non-decreasing in wage and is reminiscent of the fair wage-effort hypothesis mentioned in the introduction (Akerlof and Yellen, 1990) that argues that effort responds positively to wages up to a wage level that is deemed 'fair'. Above the fair wage, workers are assumed to provide a constant effort level that Akerlof and Yellen call "normal". The kink in the response at a wage of $\frac{f(\hat{e}_\rho)}{2}$ defines the *objectively fair wage* in our model. The characteristics of this fair wage depend on the worker's disutility parameter ρ and the assumptions we make for the unobservable functions $c(e)$ and $f(e)$.[11]

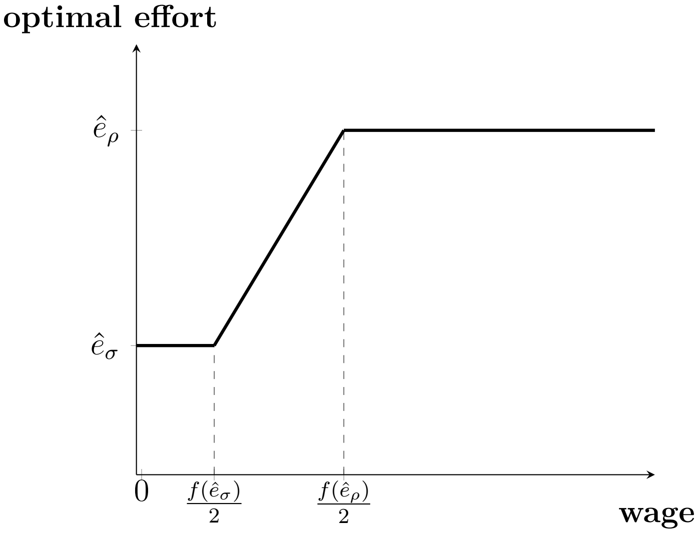
[8] To avoid further corner solutions, we assume that there exists an $\hat{e}_0$ for which this equality holds. For ease of notation, we further assume that $\sigma < \frac{c'(\hat{e}_0(w))}{f'(\hat{e}_0(w))} < \rho, \forall w$. This assures that $\hat{e}_\sigma < \hat{e}_0(w) < \hat{e}_\rho, \forall w$, thus avoiding cumbersome notations.

[9] We note that although the discontinuity of the beta function (1.2) shapes the gift exchange function $e(w)$, it does not drive the predictions of this paper. Our predictions only require that for some positive levels of wages, optimal effort increases in wage with diminishing returns $f(e)$. The latter ensures that there is a local maximum in employer's utility. We choose the discontinuous Charness and Rabin (2002) function because of its prominent place in the literature.

[10] For presentational purposes, $f(e)$ is assumed to be linear. A non-linear $f(e)$ would add curvature to the intermediate segment of the best response function.

[11] For similar patterns, see Benjamin (2015) (using a model based on other-regarding preferences) and Dickson and Fongoni (2019) (a model of 'worker morale').

Figure 1.1: Worker's response curve e(w) as a function of wage



Notes: The optimal effort (vertical axis) is shown as a function of the wage (horizontal axis). $\hat{e}_\rho$ ($\hat{e}_\sigma$) depicts the solution to the first order condition (1.4) in case the worker faces (dis)advantageous inequality. In this example, $\sigma > 0$.

Employer's Wage Setting

Utility. Employers choose a wage at the first stage of the interaction. Their utility, denoted by u^F (where 'F' stands for 'firm'), is assumed to be given by

$$u^F = (1 - \alpha)(E[f(e(w))] - w) + \alpha w. \tag{1.5}$$

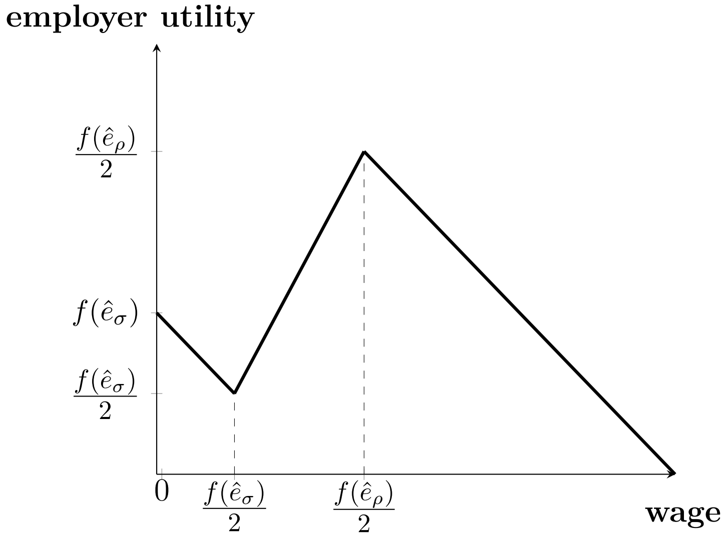
The utility thus consists of the expected monetary earnings (expected revenue $E[f(e(w))]$ minus the wage) plus a social preference term reflecting concern for the worker, and in particular, the worker's wage (weighted by α).[12] In a Subgame Perfect Equilibrium (SPE), employers expect the workers to best respond to the wage offered, that is, $E[f(e(w))]$ is determined by eq. (1.4). In other words, $E[f(e(w))] = f(\hat{e}(w))$.

We first consider the role of α. Recall from worker's best response, eq. (1.4), that for the low and high wage ranges, effort does not respond to changes in the wage. A wage increase within either of these ranges then raises the worker's earnings without affecting her productivity, $f(e)$. The utility-maximizing wage for the employer in each of these wage ranges is then a corner solution of either the lowest wage (in case the employer cares more for her own payoff, $\alpha < 0.5$) or the highest wage (when the employer cares more for the worker's payoff ($\alpha > 0.5$). From here onward, we will assume the former scenario, that is, the employer cares more for her own payoff than that of the worker.

[12] As with the worker, we assume that the employer's other-regarding preferences are fully based on monetary earnings. The employer does not take into account the worker's other-regarding preferences or her effort costs.

In the intermediate wage range, the worker responds with effort in a way that equalizes the net monetary benefits. Substituting $w = E[f(e(w))] - w$ in (1.5) gives $u^F = E[f(e(w))] - w$. This means that the other-regarding preferences drop out. For this intermediate range, we thus set $\alpha = 0$ without loss of generality. The utility maximizing wage is then the wage that maximizes the employer's monetary earnings.

Figure 1.2: Employer's utility as a function of wage



Notes: The employer's utility is shown as a function of the wage (horizontal axis), assuming $\alpha < 0.5$. $\hat{e}_\sigma$ ($\hat{e}_\rho$) depicts the solution to the first order condition (1.4) in case the worker faces (dis)advantageous inequality.

Optimal Wage Setting. Figure 1.2 summarizes the discussion above and shows how the employer earnings, given by $f(\hat{e}(w)) - w$, vary with the wage offered in the SPE. Increasing low wages (below $\frac{f(\hat{e}_\sigma)}{2}$) does not affect the worker's chosen effort level (which stays at the low $\hat{e}_\sigma$), so the employer's earnings drop linearly in w.[13] A wage equal to zero then yields a local maximum in the employer's utility. Similarly, the linear negative relation between this utility and wages above $\frac{f(\hat{e}_\rho)}{2}$ follows from workers not responding to increased wages with higher effort. Only the intermediate range provides an opportunity for further gift exchange, that is, a marginal increase in effort in response to a wage increase. In this range, a wage increase leads to higher effort that benefits the employer. Revenue can rise up to a level of $\frac{f(\hat{e}_\rho)}{2}$ for a wage of $\frac{f(\hat{e}_\rho)}{2}$. This provides a second local maximum of the employer's utility. A comparison of the two local maxima yields our next result.

[13] If $\sigma < 0$ then $f(\hat{e}_\sigma) = 0$.

Result 2. The utility maximizing wage for an employer is

$$\hat{w} = \begin{cases} 0, & \text{if } \frac{f(\hat{e}_\rho)}{2} < f(\hat{e}_\sigma) \\ \frac{f(\hat{e}_\rho)}{2}, & \text{if } f(\hat{e}_\sigma) \leq \frac{f(\hat{e}_\rho)}{2}, \end{cases} \tag{1.6}$$

where we assume that an employer chooses the higher wage whenever indifferent.

Recall that we call $w = \frac{f(\hat{e}_\rho)}{2}$ the objectively fair wage. Result 2 shows that whether the employer prefers the minimum wage of zero or the objectively fair wage depends on σ and ρ, which are the worker's social preference parameters. This is because the employer's optimal action depends on the extent to which she can stimulate sufficient gift exchange from the worker's side. We conclude that whenever ρ is large enough relative to σ, the SPE involves gift exchange: employers set wages above the minimum and workers respond with an effort level that equalizes earnings. Note that this gift exchange model does not require workers to have reciprocal preferences, which would yield even higher wages and effort levels. Moreover, gift exchange is observed in equilibrium even if employers have selfish preferences. All that is needed for gift exchange is that the worker cares about the employer's earnings.

Incomplete Information

Thus far, we have assumed that this is a game of complete information. In particular, this assumes that employers know the workers' preference parameters σ and ρ. In practice, workers' preferences will be heterogeneous with respect to these parameters and employers will update their beliefs about workers' (social) preferences based on experienced effort choices. Our goal, however, is not to provide a full-fledged analysis of this game. Instead, our aim is to derive directional predictions with respect to the effects of shocks on gift exchange. The complete-information SPE derived here suffices to do so.

1.2.2 The Impact of Shocks

We now consider shocks in monetary earnings. These may occur randomly with known probability. When a shock occurs, it reduces the monetary income of either all workers or all employers, thus affecting one side of the market. Think for example of an externally enforced tax. We consider two potential common shocks:

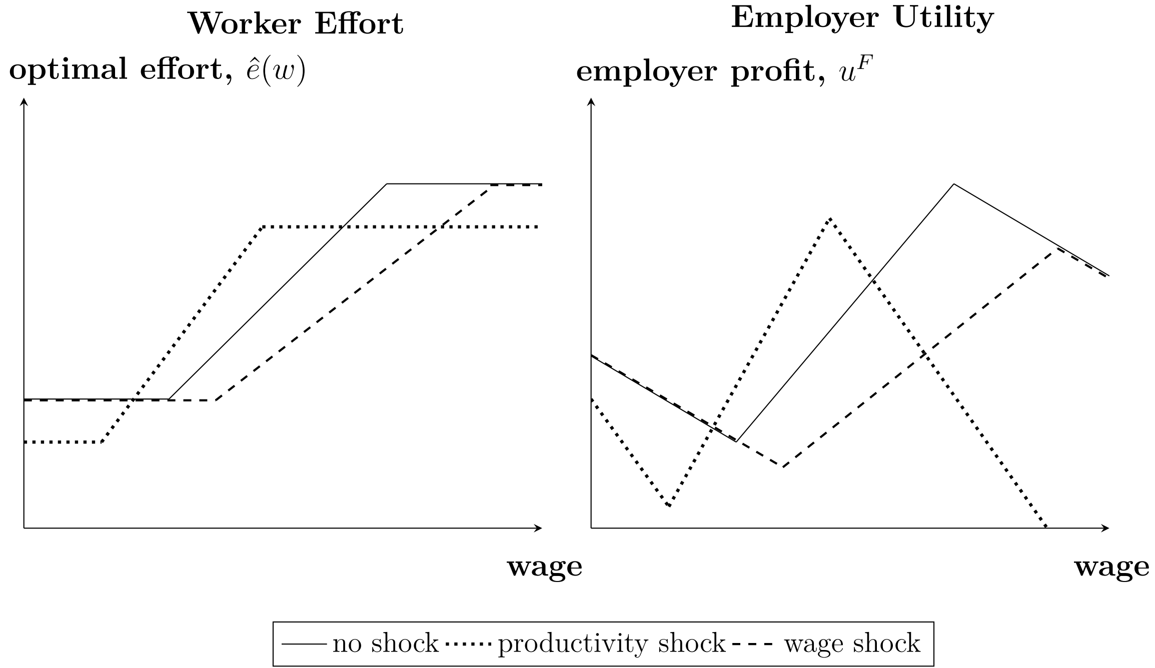
Wage shock: reduces the wage (w) received by the workers, leaving employer earnings unaffected.

Productivity shock: reduces the employers' revenues ($f(e)$), leaving worker earnings unaffected.

A detailed description of the model with shocks is presented in Appendix 1.A. Here we provide an overview of the model's implications.

Figure 1.3 illustrates the effects of shocks on the worker's best response function (left panel) and the employer's utility (right panel). For presentational purposes, we again assume a linear $f(e)$ (cf. fn. 10).

Figure 1.3: The Effects of Shocks



Notes: The left panel shows optimal effort (vertical axis) as a function of the nominal wage (horizontal axis). The right panel shows employer's utility as a function of the nominal wage (horizontal axis).

Observe that a wage shock (dashed line) shifts the worker's best response to the right because a higher wage is needed to equalize earnings (left panel). Moreover, the upper bound shifts further to the right than the lower bound (cf. Appendix 1.A). As a consequence, the intermediate wage area with gift exchange is larger than without the shock. There is no vertical shift of the response function, because this is determined by the f.o.c. (1.3), which is not affected by a wage shock. Because the wage shock does not affect effort levels at low wages and because it does not reduce employers' revenues for given effort, employer utility (right panel) at the minimum wage is the same with and without wage shock. As wages increase, u^F develops in the same way in both cases. However, it takes a higher wage for the worker to start equalizing earnings as effort does not increase until the net wage is equal to the (minimum) employer profit. This occurs at a higher wage than when there is no shock. The employer's utility subsequently reaches its maximum at a higher objectively fair wage and lower level of utility due to the increased wage expenses.

A productivity shock (dotted line) shifts the area of wages where the worker wants to equalize earnings to the left. Moreover, it shifts the upper bound further to the left than the lower bound (cf. Appendix 1.A), yielding a smaller range of wages where gift exchange is observed. The productivity shock also shifts the worker's best response curve downward. This is because the worker recognizes that each unit of effort gives less return to the employer and internalizes this by lowering the provided effort such that the marginal cost of effort matches the lowered marginal benefit to the employer. As a consequence, a productivity shock reduces employer's utility (right panel) at the minimal wage (here normalized to $w = 0$). Utility then declines linearly until the worker starts to respond to wage increases by equalizing earnings. This gift exchange takes place up to the objectively fair wage, but this is lower than the objectively fair wage in the case without shocks. As wages increase beyond this level, employer's payoff decreases linearly because effort no longer increases in response to higher wages.

One will observe gift exchange in the SPE if the utility achieved at the objectively fair wage is higher than the utility achieved at the minimum wage. Appendix 1.A derives precise conditions for this to occur.[14] The theoretical predictions in the following subsection are based on the assumptions that these conditions for the occurrence of gift exchange are met.

1.2.3 Theoretical Predictions

We start with the employer-worker interaction when there are no shocks. The possibility of gift exchange in the SPE gives the following theoretical predictions. As discussed above, these have found support in numerous laboratory and field experiments.

Theoretical Prediction 1: (*Wages*) Employers offer wages above the minimum level.

Theoretical Prediction 2: (*Gift Exchange*) The relationship between wages and effort is positive up to a fair wage level. No relation is expected at wages above the fair wage level.

Based on the subgame-perfect equilibria depicted in Figure 1.3 and the analysis of Appendix 1.A, we derive the following comparative static predictions for the effects of shocks.

Theoretical Prediction 3: (*Wage shock*) Compared to the case without shocks, a negative wage shock yields higher wages and does not affect (equilibrium) effort.

Theoretical Prediction 4: (*Productivity shock*) Compared to the case without shocks, a negative productivity shock yields lower wages and lower (equilibrium) effort.

Note that these hypotheses do not predict wage rigidity. This is because the objectively fair wage, based on equity and cost-benefit calculations, varies with the shocks. In the

[14] We also show in the appendix, that if worker preferences yield an SPE with gift exchange when there is a wage shock, then there is also gift exchange in the equilibrium for the case without a shock.

next subsection we discuss alternative behavioral models that do predict wage rigidity.[15]

1.2.4 Alternative behavioral models

Net wage illusion

Various experimental studies on labor market responses to taxes observe that workers respond more to gross wages than to net (after-tax) wages (Fochmann et al. 2013; Weber and Schram 2017). In an environment of shocks, this would mean that a worker neglects the effects of a shock on her real wage (if it leaves the nominal wage unchanged) and therefore does not change her effort. As a consequence, the effort response function and the employer's utility in Figure 1.3 do not shift after a wage shock compared to the no-shock case.

Loss aversion

Our static model assumes that the worker responds to wages independently of any prior expectations she might have had about a 'reasonable' wage level. Instead, a worker might consider a wage that is lower than what she expected to be a 'loss', irrespective of whether this lower wage might be justified by a shock. We rationalize this by applying the Kőszegi and Rabin (2006) notion of reference-dependent preferences. Utility is measured against some reference point. If the outcome falls short of the expected, the individual experiences a loss even if the outcome is positive in absolute terms. It is worth noting that this formulation of loss aversion is closely related to the formulation of a negative reciprocity term in Charness and Rabin (2002). Here, 'misbehaving' is essentially understood as setting a wage below the relevant reference point.

We assume that for a worker the objectively fair wage in the no-shock case serves as a reference.[16] We now denote this by $\tilde{w}$. Recall that $\tilde{w} = \frac{f(\hat{e}_\rho)}{2}$. The worker then experiences a loss if the current wage falls short of this reference point. In our model, we capture this by adding a loss term to the social preference function β(e) in the worker's utility function (1.1). Once again, we set $\beta = 0$ for the range of wages where the worker equalizes earnings (cf. fn. 7).

[15] The asymmetry in the Predictions 3 and 4 with respect to the effects on effort stems from the fact that optimal effort is given by an equilibrium condition on which a productivity shock has an impact, but a wage shock does not. This asymmetry will also be observed in the model extensions discussed below.

[16] We make this assumption to stay within the realm of our model. All that is needed for the effects described in what follows is that people have some idea of what is a 'fair' wage in the absence of shocks.

$$\beta(e) = \begin{cases} \sigma - \lambda, & \text{if } w < \frac{f(e)}{2} \wedge w < \tilde{w} \\ 0, & \text{if } w = \frac{f(e)}{2} \wedge w < \tilde{w} \\ \rho - \lambda, & \text{if } w > \frac{f(e)}{2} \wedge w < \tilde{w} \\ \rho, & \text{if } w > \frac{f(e)}{2} \wedge w \geq \tilde{w}, \end{cases} \tag{2'}$$

where parameter λ measures the degree of loss aversion. In the first line of (2'), the worker faces disadvantageous inequality and a wage that is lower than the reference point. In the second, earnings between the worker and employer are equal, but the wage is still below the reference. The latter also holds in the third line, but here the worker is earning more than the employer. Finally, the fourth line covers the situation where the worker faces advantageous inequality and at the same time a wage that is larger than or equal to the reference point.

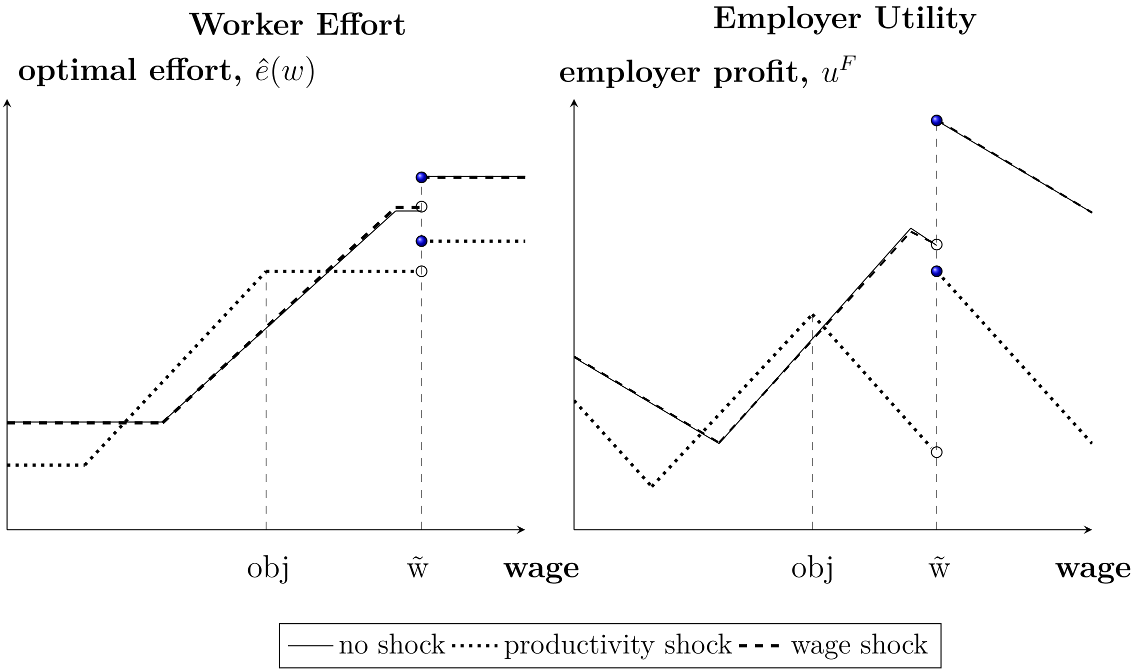
Without shock, the parameter λ shifts the worker's best response function downward (because $\hat{e}_{\sigma-\lambda} < \hat{e}_{\sigma}$ and $\hat{e}_{\rho-\lambda} < \hat{e}_{\rho}$) and to the left (because $f(\hat{e}_{\sigma-\lambda}) < f(\hat{e}_{\sigma})$ and $f(\hat{e}_{\rho-\lambda}) < f(\hat{e}_{\rho})$). Otherwise, the predictions of the static model remain unaltered. When there is a productivity shock the *objectively fair wage* diminishes (cf. Figure 1.3). With loss aversion we assume that the worker does not adjust her reference point accordingly. We provide more details in appendix 1.B. Here, we summarize the combined effects of net wage illusion and loss aversion.

Combined effects

Figure 1.4 shows the best response and employer utility functions when there is both net wage illusion and loss aversion. Note the discontinuity in both graphs at $\tilde{w}$. The 'jump' at this reference point is caused by loss aversion (measured by λ) no longer playing a role in the worker's effort decision (left panel). This has direct consequences for the employer's utility (right panel). We call the point at which this occurs the 'subjectively fair wage'. Note that when there is a productivity shock, this subjectively fair wage $\tilde{w}$ is larger than the objectively fair wage, which is determined by the upper kink in the worker's effort function. When there is a wage shock, the two are equal, due to the net wage illusion.

The right panel of Figure 1.4 shows that when there is a productivity shock there are three local maxima in the employer's utility. They are at the minimum wage (0), the objectively fair wage (the peak in utility for $w = obj < \tilde{w}$) and the subjectively fair wage ($\tilde{w}$). Assuming that the objectively fair wage yields higher utility than the minimum wage, it is straightforward to formulate conditions under which the employer will prefer to keep wages at the subjectively fair level (cf. Appendix 1.B). In the right panel of Figure 1.4, utility is higher for the subjectively fair wage than for the objectively fair wage. If this

Figure 1.4: The Effects of Net Wage Illusion and Loss Aversion



Notes: The left panel shows optimal effort (vertical axis) as a function of the nominal wage (horizontal axis). The right panel shows employer's utility as a function of the wage (horizontal axis). $\tilde{w}$ depicts the subjectively fair wage, which is defined as the objectively fair wage in the no-shock case and which serves as a reference point for the worker. *obj* is the objectively fair wage when there is a productivity shock.

holds, the model predicts wage rigidity, that is, employers prefer to hold wages constant even if they face a shock on their income. With a productivity shock, wage rigidity arises from loss aversion; if employers were to cut wages, workers would retaliate by cutting effort, making the wage adjustment unprofitable. The model also predicts wage rigidity for wage shocks as the objective fair wage is the same as the subjective one when there is both nominal illusions and loss aversion.

1.2.5 Alternative theoretical predictions

Based on the relationships illustrated in Figure 1.4 and the elaboration in Appendix 1.B, we can formulate alternatives to Hypotheses 3 and 4, for the case where workers exhibit net wage illusion and loss aversion that is strong enough to cause wage rigidity.

Theoretical Prediction 3A: (*Wage shock under net wage illusion*) A negative wage shock has no effect on wages or effort.

Theoretical Prediction 4A: (*Productivity shock under loss aversion*) A negative productivity shock has no effect on wages and yields lower effort.

Off the equilibrium path

Note that the behavioral model predicts no effects of a wage shock, on or off the equilibrium path (cf. Figure 1.4). The case is different with a productivity shock, where equilibrium effort is lower with than without shock (Theoretical Prediction 4A). Out of equilibrium, however, one might observe the opposite. Consider the upward sloping part of the gift exchange curve for the no-shock and productivity shock cases. The worker's best response to a wage in this range (out of equilibrium) yields higher effort after a shock than when there is none. This is because the worker equalizes payoffs on this part of the curve. As the return on effort is lower, a higher level is needed to achieve balance.

1.3 Experimental design and procedures

1.3.1 Design

The design builds on Fehr et al. (1993). In contrast to their seminal paper, we use a computerized experiment and implement a real-effort task to measure productivity. The experiment is framed as a labor market and consists of eight rounds. Shocks are framed as one-round taxes. Each round consists of the following stages, which are elaborated below.

1. If tax shocks are possible, the (common) tax scheme (or the lack thereof) is announced
2. Employers hire workers in an auction
3. Workers conduct a real effort task
4. Payoffs are determined and reported

We start with a description of the hiring stage. Hiring happens in real time, via a one-sided auction. Employers post wage offers between 30 and 100 points, in intervals of 5, on a public platform observable by all employers and workers in the market. Offers can be updated while not yet accepted. Once a worker accepts an offer, the offer is removed and the worker is hired by the employer in question. The market consists of five employers and seven workers and each participant can have only one hiring contract per round.[17] As a consequence, at least two workers are unemployed in each round. The hiring stage lasts at most two minutes and finishes as soon as all five employers have hired a worker.

[17] Following the original design of Fehr et al. (1993), the market consists of 7 workers and 5 employers. Brandts and Charness (2004) show that the market conditions (whether labor is in excess supply or demand) do not matter for the occurrence of gift exchange.

After the auction, anonymized information is provided to all market participants about the number of hired workers and the realized wages (wages are given in random order).

At the start of the second stage, each hired worker thus knows her wage and whether or not a shock has occurred. She then works for five minutes on a real effort task. For the task (introduced by Weber and Schram 2017), two 10x10 matrices appear on the computer monitor. Each matrix cell contains a two-digit number. The worker needs to find the highest number in each matrix and add these two up. A correct answer yields a reward of 20 points to the employer (part of which may be taxed, as explained below). Whether the answer is correct or incorrect, a new pair of matrices appears. The maximum number of tasks that can be attempted is limited to ten.[18]

In some rounds, one-round shocks might be implemented. These are framed as 'taxes', which are announced before the hiring auction and are known to hold for all workers or employers in that round. Note that this means that all participants are fully informed before they make any decisions in a round. The taxes impact participants' earnings. We distinguish between (1) a wage tax; this reduces the wage that the worker receives from the employer in that round by 20%; and (2) a productivity tax; this reduces the revenue that the employer receives from the hired worker's correctly solved tasks in that round by 20% (from 20 to 16 points). Tax revenues are not returned to participants in any way; proceeds are returned to the experimenter.

The experiment consist of four treatments that are varied between subjects. These differ in the type of tax that *might* occur. The four treatment options are 1) no tax (denoted by NT), 2) productivity tax (ET, for 'Employer Tax'), 3) wage tax (WT) and 4) employer or wage tax (AT, for 'All Taxes'). In treatments where taxes are possible, they happen in any round with an probability equal to $\frac{1}{3}$. When both taxes are possible, each tax is equally likely but they cannot occur simultaneously. All of this is common knowledge. The sequence of taxes was drawn randomly beforehand and was fixed in order for all sessions to have a directly comparable history.[19]

It is important to distinguish between tax *treatments* (tax environments) and the tax *outcomes.* Throughout this paper, we indicate treatments with capital letters; they define which tax shocks (outcomes) are possible. Tax outcomes are realized per round; we indicate these with lower case letters. Table 1.1 summarizes all possible cases.

Each round ends with a payoff report for that round. Participants learn their own payoffs and if hired or hiring, the payoff of the partner to which they had been linked, as

[18] This limit is set to discourage a strategy of guessing one answer and repeatedly entering this number at a very high pace. The limit is not binding; from previous projects, we know that even when incentivized with piece-rate rewards, fewer than 1% of the subject pool is able to reach this limit.

[19] The shocks occur in rounds 2, 4, and 5. In AT, half of the sessions had a one-round productivity tax in round 2 and a one-round wage tax in rounds 4 and 5; the remaining sessions had the reverse. Note that the productivity tax is an example of the productivity shock that we modeled above, while the wage tax is a wage shock.

Table 1.1: Treatments and outcomes

Treatment	NT	ET	WT	AT
possible tax outcomes	nt	nt, et	nt, wt	nt, et, wt

Notes: NT/nt = 'no tax'; ET/et = 'productivity tax'; WT/wt = 'wage tax'; AT = 'all taxes'.

well as the number of tasks attempted and the number of tasks correctly solved. Payoffs depend on the hiring status and the tax outcome and are summarized in Table 1.2. If an employer hires a worker, the employer receives 40 points and all of the revenue from the task but must pay the worker's wage from this income. A worker's payoff consists entirely of the wage. If unmatched, employers earn nothing and unemployed workers receive an unemployment benefit of 20 points, regardless of the tax outcome. When taxes apply, they directly affect only one side, either the employer or the worker. The productivity tax is collected from the revenue that the employer receives, which means that when taxed, instead of the usual 20 points, the employer receives only 16 points for each task correctly completed by the worker. When the wage tax applies, the workers receive only 80% of the wages paid by their employer.[20]

Table 1.2: Payoffs

	employer payoff	worker payoff
no tax (nt)	$40 - w + 20 * e$	w
productivity tax (et)	$40 - w + 16 * e$	w
wage tax (wt)	$40 - w + 20 * e$	$0.8 * w$
outside option (no contract)	0	20

Notes: Cells show payoffs in points for employers and workers, depending on the outcome of the tax shock.

At the end of the experiment, two rounds are randomly selected for payment.[21] The exchange rate used is one euro for every ten points earned in those two rounds. Note that for employers negative earnings in a round are possible. Because two rounds are paid, this can be compensated. In the end, only very few participants had negative earnings, and everyone who did was able to cover these with the show-up fee.

[20] It follows from the payoffs in Table 1.2 that (if one does not consider effort costs) equal payoffs are not possible for odd wages (35, 45, ...). We nevertheless chose to restrict the set of possible wages to the set with intervals of five to avoid employers signaling their identity by repeatedly making the same 'unusual' offer (like 41).

[21] In the first three sessions, due to computational errors the incentive scheme rewarded three rounds instead of two (which was only known to the participants ex post) and a shock occurred in fewer rounds than intended (which is not expected to affect choices because the occurrence of a shock is common knowledge before any decision is made).

1.3.2 Procedures

The experiment was run at the BLESS laboratory of the University of Bologna, in 2017 - 2018. Participants were primarily students and recruited using ORSEE (Greiner, 2004). The experimental software was programmed in oTree (Chen et al., 2016). We had 312 participants in 13 sessions. Each session had 2 groups (each consisting of 5 employers and 7 workers).[22] Average earnings (including a five euro show up fee) were 14.5 euros.

Reading the instructions and getting familiar with the software took approximately 20 minutes and the main experiment lasted about one hour. A translation of the instructions is presented in Appendix 1.C. During the software tutorial, the participants did the real effort task for five minutes to get acquainted with it. At the end of the instructions, the participants had a comprehension test (cf. Appendix 1.C).

1.3.3 Testable hypotheses for the experimental design

We apply our theoretical predictions to this experimental environment. Note that – as is common when using laboratory data to test hypotheses – our predictions are concerned with the comparative statics that follow from the theoretical discussion in the previous section. We keep the same order and start with the baseline in which no shock is realized (note that the occurrence of a shock is common knowledge at the start of a round). Recall that our first theoretical prediction is that employers will offer wages above the minimum level. We test this against a null hypothesis based on the rational choice equilibrium of no gift exchange. This involves employers offering a minimum wage and workers exerting no effort.

Hypothesis 1: **No Tax: Wages**

- H_0^1: In no shock (nt) rounds, employers offer the minimum wage of 30 points.
- H_1^1: In no shock (nt) rounds, employers offer wages above the minimum level of 30 points.

Closely related to this is the second theoretical prediction that the relationship between wages and effort is positive up to a fair wage level. For our environment, this gives

Hypothesis 2: **No Tax: Effort**

- H_0^2: In nt rounds, there is no relationship between wages and effort.
- H_1^2: In nt rounds, there is a positive relationship between wage and effort up to the objectively fair wage and no relationship beyond that.

[22] For three groups we have 11 participants instead of 12, due to recruitment failures. In these cases, the experiment proceeded with six workers and five employers in the group. Our conclusions do not change if we drop these groups from the analyses.

For the reactions to shocks we have two sets of hypotheses, depending on whether or not the model includes net wage illusion and loss aversion. For wages, a model without net wage illusion predicts that a wage shock will yield an increase while net wage illusion predicts wages that do not respond to such shocks.[23] The latter is also predicted by the rational model with selfish preferences.

Hypothesis 3: **Wage Tax: Wages**

- H_0^3: *Rational-selfish model and social preferences with net wage illusion.* Wages are the same in *wt* rounds as in *nt* rounds.
- H_1^3: *Social preferences without net wage illusion.* Wages are higher in *wt* rounds than in *nt* rounds.

For effort, we focus on the equilibrium case where wages are as predicted. When analyzing the data, we will also consider the wage-effort relationship more generally (that is, including out-of-equilibrium wages), but our hypotheses are derived from the equilibrium predictions. Recall that none of our models predict that equilibrium effort will be affected by a wage shock. For the model with net wage illusion, this is trivial (workers do not 'recognize' the change in net wage).

Hypothesis 4: **Wage Tax: Effort**

- H_0^4: Effort is the same in *wt* rounds as in *nt* rounds.

The predictions for a productivity shock again depend on the model. As with a wage shock, the rational-selfish model predicts no effects on wages or effort. The same holds for the model with loss aversion. The model with social preferences (but without loss aversion), however, predicts that the productivity tax will yield lower wages. Thus,

Hypothesis 5: **Productivity Tax: Wages**

- H_0^5: *Rational-selfish model and social preferences with loss aversion.* Wages are the same in *et* rounds as in *nt* rounds.
- H_1^5: *Social preferences without loss aversion.* Wages are lower in *et* rounds than in *nt* rounds.

[23] It might seem counterintuitive that net wage illusion takes away the effect of wage shock. The underlying mechanism is that the burden of the shock is shared equally when the shock is noticed. When there is net wage illusion, no effect is expected as the illusion 'hides' the changed market situation.

Finally the productivity shock is predicted to reduce equilibrium effort by the social preference models with and without loss aversion.

Hypothesis 6: **Productivity Tax: Effort**

- H_0^6: *Rational-selfish model.* Effort is the same in *et* rounds as in *nt* rounds.
- H_1^6: *Social preferences (with and without loss aversion).* Effort is lower in *et* rounds than in *nt* rounds.

1.4 Results

We have data for a total of 130 employers, 179 workers, and 934 employer-worker matchings. These matchings include, however, eight rounds of observations for each worker and employer (though an observation may consist of nothing more than not having a contract in a round). To correct for such multiple observations, we treat – unless specified otherwise – the average observation for an employer over the rounds as the unit of observation. We choose to aggregate over the employers because they cannot be selected out of a round to the same extent that workers can. This gives us 30 observations each for NT, ET, and WT, and 40 for AT, though not every employer has an observation in every round.[24]

Unless indicated otherwise, test results are based on non-parametric permutation t-tests (cf. Schram et al. 2018), here referred to as PtT. In order to obtain an impression of the power of our statistical tests, we use information from a different experiment we ran where wages could be changed after the initial contract (more information about this experiment is available upon request). The mean wage observed there in NT was 41.7, with a standard deviation of approximately 10. An underlying treatment effect of 15% (observed in the other experiment) would then give us a power of 66% for a standard t-test with 30 observations per treatment. We nevertheless expect our tests to be sufficiently powered, because (i) the PtT is a higher-powered test than the standard t-test (Moir 1998, Schram et al. 2018)[25]; and (ii) we expect the standard deviation to be lower in sessions where the wage cannot be altered within a round.

We organize the discussion around two key elements in our data, the realized wages and the exerted effort. For the latter, much of our focus will be on the occurrence of

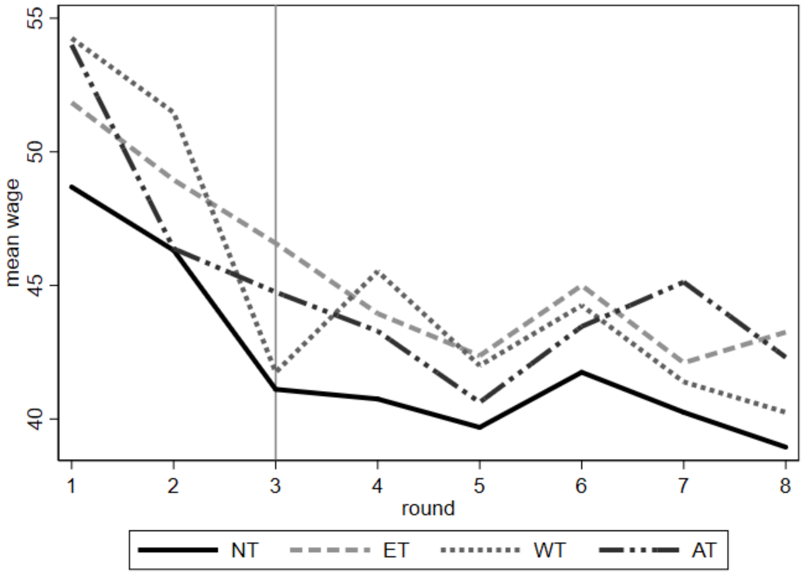
[24] In rare occasions, an employer did not succeed in hiring a worker before the two-minute auction deadline. In early sessions, we also lost some of the late-round data and the post-experiment survey results due to technical problems.

[25] We know of no method to directly calculate the power of a PtT.

gift exchange (that is, the relationship between realized wage and exerted effort). We distinguish between treatments and shocks. As before, treatments (indicated by capital letters) are environments in which shocks (lower-case letters) may occur.

1.4.1 Realized Wages

Figure 1.5: Average wage



Notes: Lines show average realized wage over the eight rounds of the experiment. The minimum wage is 30. *NT*: no taxes possible; *WT*: wage tax possible; *ET*: productivity tax possible; *AT*: both taxes possible. Tax shocks occurred in rounds 2, 4, and 5.

In all treatments, the average wage starts relatively high and drops over the first two rounds, stabilizing around a level of 40-45 points from round 3 onward.[26] Our interpretation of the wage drop in the first two rounds is learning; employers adjust their wage offers quickly once they experience the workers' responses and the behavior of the other employers. Interestingly, this learning period casts some doubt on the results in previous papers that draw conclusions about wage rigidity based on only one or two rounds (e.g., Gerhards and Heinz 2017).

Because our predictions are based on equilibria, we lay aside the learning effects in the first rounds and focus our analysis on rounds 3-8. As a consequence there are two

[26] In all treatments, the wage of round 1 is significantly higher than that of round 8. The $p-$values for the null of no difference are for *NT*: PtT, $p = 0.001$ ($N = 16$); *ET*: PtT, $p = 0.025$ ($N = 18$); *WT*: $p < 0.001$ ($N = 30$) and *AT*: PtT, $p =< 0.001$ ($N = 39$). The wage is not significantly different in round 3 from that in round 8 in any treatment. The $p-$values are for *NT*: PtT, p = 0.850 ($N = 16$); *ET*: PtT, $p = 0.094$ ($N = 18$); *WT*: PtT, $p = 0.104$ ($N = 30$); and *AT*: PtT, $p = 0.340$ ($N = 39$). For these comparisons, note that rounds 1, 3, and 8 are all without shock. Also, recall that we have some missing values for round 8, due to technical problems in early sessions.

rounds (4 and 5) with realized shocks in our analysis of treatments ET, WT, and AT. For completeness, Appendix 1.D presents the analysis using data from all rounds; the results are very similar. Throughout the experiment, almost all wage offers were accepted. The acceptance rate of the first offer made by an employer in rounds 3-8 varies across treatments between 85% and 93%. This means that variations that we observe in realized wages can by-and-large be attributed to variations in wage offers. To start, Table 1.3 shows average wages per treatment and tax shock. In this table, we use the fact that AT consists of two sub treatments that are mirror images of each other. This was done to balance the number of observations under each shock. AT_{et} has one wt shock in round 2 followed by two et shocks in rounds 4 and 5, while AT_{wt} has one et shock in round 2 followed by two wt shocks in rounds 4 and 5.[27]

Table 1.3: **Wages, treatments, and shocks**

tax outcome	NT	ET	WT	AT_{et}	AT_{wt}	**pooled**
nt	**40.8**	**43.6**	**42.3**	**47.4**	**40.5**	**42.8**
obs.	30	30	30	20	20	130
et		**42.4**		**44.6**		**43.4**
obs.		25		20		45
wt			**43.5**		**38.8**	**41.1**
obs.			20		20	40
PtT (p-values)						
nt vs et	-	*0.434*	-	*0.034*	-	
nt vs wt	-	-	*0.325*	-	*0.117*	

Notes: Results are for rounds 3-8. Tax shocks occurred in rounds 4 and 5. The unit of observation is the mean wage paid by an employer across rounds. Paired tests between shock and no-shock rounds are reported. We do not conduct tests for the pooled data because these combine paired with unpaired comparisons. Mean wages across employers are in bold. 'obs.' shows the number of employers. NT: no taxes possible; nt: no tax shock realized; WT: wage tax possible; wt: wage tax shock realized; ET: productivity tax possible; et: productivity tax shock realized; AT_{et}: both taxes possible, only et realized; AT_{wt}: both taxes possible, only wt realized. 'pooled' combines treatments. PtT: permutation t-test.

The results show that average wages within a treatment vary little with realized tax shocks. Results of the PtT (shown in the lower panel of Table 1.3) indicate that shocks have no significant effect on the wages in ET, WT or AT_{wt}. Though the effect on wage in AT_{et} is relatively small (6%), it is statistically significant. In this treatment, employers that face a productivity shock manage to pay lower wages. Note, however that in the pooled data average wages are even higher after a productivity shock than without shock.

[27] As we are only considering rounds 3-8, this means we have observations of et shocks only under AT_{et} and observations of wt shocks only under AT_{wt}. Because we are using the mean wage per employer as the unit of observation, we use paired-sample permutation tests in Table 1.3 (the mean wage paid in rounds without shock is paired with the mean wage in rounds with a shock). This requires doing the tests for AT_{et} and AT_{wt} separately.

A comparison between ET and AT_{et} shows that in the latter case the apparent negative effect of a shock on wages is not caused by low wages after et, but that, instead, average wages in nt are relatively high.[28] All in all, we find little evidence that the wage systematically adjusts to tax shocks. Note also that in all treatments the mean wages are far from the minimum level of 30 points. The 95% confidence intervals for outcome nt are (36.9, 44.6), (39.4., 47.8), (38.7, 46.0), (40.5, 54.3), and (36.7, 44.2) for NT, ET, WT, AT_{et}, and AT_{wt}, respectively.

These results can be directly applied to our hypotheses regarding wages. The confidence intervals for nt indicate that wage offers are not at the minimum, which rejects H_0^1 in favor of H_1^1. This leads us to reject the standard rational model with selfish preferences. The result that wages are not significantly different after a wage shock (wt) than in nt means that we cannot reject H_0^3 in favor of H_1^3. Given our support (from the first hypothesis) for social preferences over the standard model, the difference between H_0^3 and H_1^3 is that the former assumes net wage illusion while the latter does not. This suggests that net wage illusion affects decisions in this environment. Finally, we conclude that loss aversion also plays a role, because we cannot systematically reject H_0^5 in favor of H_1^5 (wages are not different in et than in nt). We will summarize the results for all hypotheses below.

Our results provide evidence of nominal wage rigidity. We therefore pool the wage results across the tax shock outcomes. Table 1.4 shows the mean wages per treatment that this gives.

Table 1.4: Wages and treatments

	NT	**ET**	**WT**	**AT**
all	**40.8**	**43.0**	**42.6**	**43.2**
obs.	30	30	30	40
PtT for differences against NT				
p-value	na	*0.427*	*0.475*	*0.364*

Notes: Results are for rounds 3-8. The unit of observation is the mean wage of an employer across rounds (presented in bold). NT: no taxes possible; WT: wage tax possible; ET: productivity tax possible and AT: both taxes possible. PtT: (unpaired) permutation t-test.

We observe higher wages in the treatments where tax shocks are possible (AT, ET, and WT) than in NT, but none of the differences are statistically significant. If we pool the three treatments with possible shocks, the difference with NT is still insignificant

[28] As explained in the table footnote, no pairwise test can be performed for the data pooled across all treatments. We can, however, pool only ET and AT_{et}. This gives mean wages of 45.3 for nt and 43.4 for et, a marginally significant difference (PtT, $p = 0.062$, $N = 45$). In a similar vein, pooling WT and AT_{wt} gives mean wages of 41.1 (nt) and 41.3 (wt). The difference is insignificant (PtT, $p = 0.818$, $N = 40$).

(PtT, $p = 0.322$). Whereas the results in Table 1.3 show wage rigidity in response to shocks, the results here indicate that the *possibility* of tax shocks also does not lead to an increase in wages. Before turning to possible effort responses to shocks, we summarize our results on wages.

Result 1: Realized wages are systematically higher than the minimum wage (30 points) in all treatments.

Result 2: The occurrence of a tax shock does not systematically affect wages.

Result 3: The possibility of a tax shock does not systematically affect wages.

1.4.2 Effort and Gift Exchange

We measure effort by the number of correct summations in the real-effort task.[29] To start, Table 1.5 summarizes the mean realized effort across treatments and shocks (again using the employer as the unit of observation). Note that this averages effort across distinct wage levels. Below, we investigate the relationship between wage and effort.

Table 1.5: **Effort, treatments, and shocks**

tax outcome	NT	ET	WT	AT_{et}	AT_{wt}	**pooled**
nt	**2.8**	**2.8**	**2.7**	**3.3**	**3.0**	**2.9**
obs.	30	30	30	20	20	130
et		**3.1**		**3.8**		**3.4**
obs.		25		20		45
wt			**2.6**		**3.2**	**2.9**
obs.			20		20	40
PtT (p-values)						
nt vs et	-	*0.170*	-	*0.021*	-	
nt vs wt	-	-	*0.502*	-	*0.588*	

Notes: Results are for rounds 3-8. Tax shocks occurred in rounds 4 and 5. The unit of observation is the mean effort received by an employer across rounds. We do not conduct tests for the pooled data because these combine paired with unpaired observations. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; AT_{et}: both taxes possible, only *et* realized; AT_{wt}: both taxes possible, only *wt* realized. 'pooled' combines treatments. PtT: permutation t-test.

[29] Of course, this 'performance' is determined by a combination of effort and ability. Because of our randomization of participants (and therefore their ability) across treatments, we attribute any treatment differences to effort. Note that we do not provide a graph depicting performance over time. Performance may differ across rounds because wages vary or because the response to given wages changes. To correct for the former, we checked the effort-wage ratio, measured as the number of correct sums, divided by the wage. Given that employer's earnings increase by 20 for each additional unit of effort, any ratio higher than 0.05 reflects a profitable mean earnings increase to the employer. The observed effort-wage ratio over time reveals that for each treatment, the margin within which the ratio moves is small (roughly between 0.055 and 0.085; that is, all values are above the break-even point). Importantly, there is no discernible trend for any of the treatments.

In neither of the treatments with wage shocks (wt) is the effort significantly different in rounds with a shock than in rounds without. This means that we do not reject the null hypothesis H_0^4 (for which none of our models predicted an alternative). Formally, H_0^4 predicts a null effect. The PtT in Table 1.5, however, only show that we cannot reject a null effect. This in itself does not provide evidence in favor of the hypothesis. To test H_0^4, we therefore resort to a Bayesian analysis. We base our analysis on linear regressions of effort (the number of correct summations) on a constant term and a dummy indicating that a shock took place (with robust standard errors clustered at the group level). We do this separately for cases where wt and et were possible. The former gives no significant effect of wt, while the coefficient for et is 0.607, which is significant with $p = 0.002$.

The Bayesian analysis for H_0^4 requires an assumption about the prior distribution of the effect of wt on effort (as measured by the regression coefficient) in the cases where wt is possible. To formulate a null hypothesis for wt, we use the results for et and assume a normal distribution for the coefficient with mean and standard deviation determined by the corresponding et regression. This basically assumes that wt's effect on effort has the same distribution as et's effect on effort. We use an alternative hypothesis that the effect of wt centers around 0 (no effect), assuming a normal prior distribution for the coefficient with standard deviation 1 (our conclusions are robust to choosing standard deviation 0.1 instead). This setup allows us to calculate the posterior odds ratio of the alternative hypothesis (no effect of wt) being correct to the null hypothesis (same effect as of et) being correct. Assuming that both models are equally likely a priori, this posterior ratio is more than 2:1. We therefore conclude that a model where a shock wt has the same effect as a shock et is rejected in favor of one where the shock has no effect.

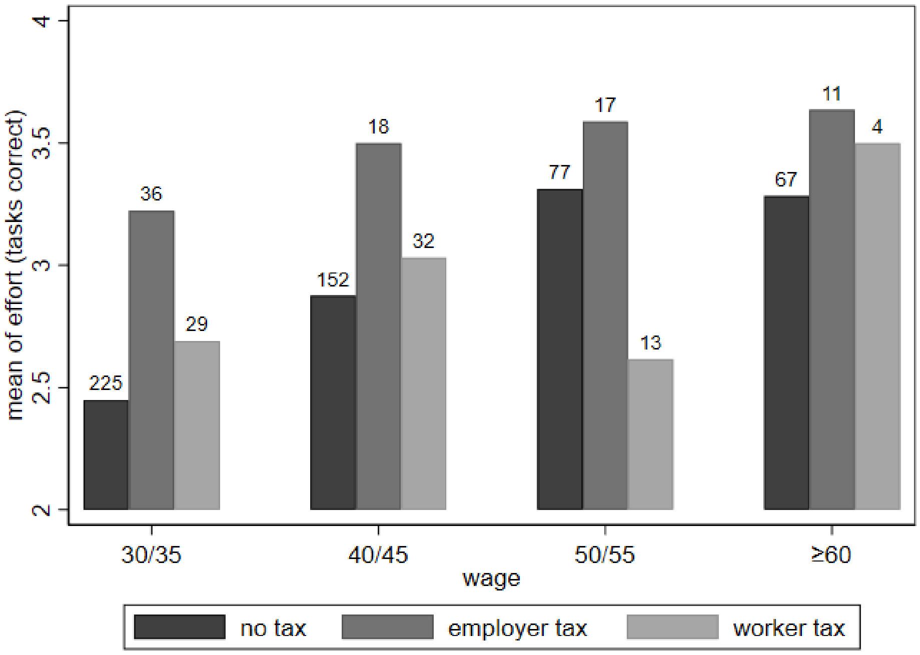
The results in Table 1.5 for productivity shocks (et) are far from the predictions. With ET, we cannot reject the null of no effect (H_0^6). In fact, effort is higher in et than in nt, which is opposite to H_1^6. The difference is, however, insignificant. In AT_{et}, effort is also higher in et; here the difference of 0.6 units is significant. Although this result is contrary to the prediction, it does not reject the social preference model per se, particularly if one allows for behavior off the equilibrium path. As argued at the end of Section 1.2.5, one may expect to see higher effort in et if the wage is below the equilibrium level. More generally, a positive reaction of effort to a productivity shock seems to indicate that fairness considerations play a role in the effort decision.

The only hypothesis that we have not yet formally tested is Hypothesis 2, where H_1^2 predicts a positive relationship between wage and effort up to a fair wage level. To get a first impression, Figure 1.6 relates effort to nominal wages.[30]

The baseline nt is represented by the black bars. It has the shape predicted by the

[30] For this analysis, we do not use the employer as the unit of observation but the labor contract. This is because effort is assumed to respond non-linearly to realized wage (and therefore not to average wage). Moreover, we pool wages over 60 because we have few high wage observations.

Figure 1.6: Gift Exchange



Notes: The number of observations in each bin is reported above each bar.

fair wage-effort hypothesis; at lower wage levels we observe clear evidence of gift exchange (effort increasing in wage) but no further increase is observed beyond the 50/55 wage bin. We interpret 50 as the fair wage level. Indeed, in the 50/55 bin mean earnings of workers (52.1) and employers (53.9) are more or less equal; employers earn more than workers at lower wages and vice versa for higher wages. Note that the gift exchange up to this wage level is substantial. At a wage of 30 or 35, the mean effort is 2.45, while it is 3.31 for wages of 50 or 55. This is an increase of 35%. The effort increases from the 30/35 bin to the 40/45 bin and from the 40/45 bin to the 50/55 bin are both (marginally) statistically significant (PtT, $p = 0.017$, $p = 0.054$, respectively). The slight decrease from 50/55 to 60-100 is insignificant (PtT, $p = 0.914$).[31] Together, this allows us to reject H_0^2 in favor of H_1^2. Without shocks, effort increases with wages (only) up to a fair wage level, which provides support for a model with other-regarding preferences. This result adds to the empirical support that has been found for fair wage-effort hypothesis (e.g., Mas 2006, Gächter and Thöni 2010, Kube et al. 2013, Sliwka and Werner 2017, Cohn et al. 2015).

Recall that we observed in Figure 1.5 that it took two periods for wages to 'settle in'. To see whether a similar learning period is observed for gift exchange, we consider the equivalent of Figure 1.6 – that is, the effort per wage bin – in *nt*, in rounds 1 and 2. In

[31] Considering all wages (as opposed to wage bins), we observe that the correlation between wages and effort between wages 30 and 55 is 0.21. This is statistically significant (Pearson correlation test, $p < 0.001$). For wages 55 and above, the correlation of 0.09 is statistically insignificant (Pearson correlation test, $p = 0.461$).

these rounds, the average effort for wages in the 30/35 bin is 2.95. This increases to 3.40 for the 40/45 bin and 3.58 for wages of 50/55. For wages of 60 or more, the average effort is 3.54. The increase from 30/35 to 50/55 is 21%. Thus, the gift exchange is weaker in early rounds than thereafter. None of the differences between adjacent bins is statistically significant (PtT, all $p > 0.216$). Moreover, the difference between the 30/35 and 50/55 bins is also statistically insignificant (PtT, $p = 0.108$). We conclude that it indeed takes time for gift exchange patterns to develop.

Observations for *et* are represented by the dark gray bars. The productivity tax shock has a positive effect on the effort provided at low wages (30/35), where effort under the productivity tax is 31% higher than when no shock has occurred. The difference is statistically significant (PtT, $p = 0.011$). This difference is +22% (PtT, $p = 0.109$), +8% (PtT, $p = 0.543$) and +11% (PtT, $p = 0.549$) for wages 40/45, 50/55, and 60-100, respectively (all are statistically insignificant). The graph suggests that, as in *nt*, there might be gift exchange up to a fair wage level. None of the steps between adjacent bins, however, is statistically significant (PtT, $p = 0.505$, $p = 0.902$, $p > 0.999$, respectively).[32] We conclude that in rounds with a productivity tax, increased worker effort compensates the loss for employers. There is no evidence, however, of further gift exchange.

Finally, the wage tax does not seem to have any systematic effect on the effort compared to *nt*, though this might be related to the low number of relatively high wages observed. At wages 30/35 and 40/45, effort is, respectively, 10% and 5% higher in *wt*, but the differences are insignificant (PtT, $p = 0.499$ for 30/35; $p = 0.632$ for 40/45). At wages 50/55 average effort is about 21% lower in *wt* (PtT, $p = 0.135$), while the low number of very high wages in *wt* (4) makes a comparison with *nt* meaningless. None of the three pairwise comparisons between adjacent bins is statistically significant (PtT, $p = 0.458$, $p = 0.449$, $p = 0.135$, respectively).[33] We conclude that gift exchange in not observed when a wage tax occurs.

In summary, there is clear evidence of gift exchange in *nt*, which confirms many results in the previous literature. When there is a shock on employers' earnings, workers compensate by exerting more effort (especially for low wages), but this diminishes the pattern of gift exchange. A tax on the worker's wage, on the other hand, does not effect mean effort, but it does seem to eliminate gift exchange. This gives the following results.[34]

[32] The correlation is positive (0.12) for wages up to 55, but this is statistically insignificant (Pearson correlation test, $p = 0.336$). For wages of 55 and above, there is a negative (–0.06), but statistically insignificant (Pearson correlation test, $p = 0.857$) correlation with effort.

[33] Though there is a positive correlation between wages and effort up to a wage of 55, and also for wages above 55 (0.01 and 0.58, respectively), neither is statistically significant (Pearson correlation test, $p = 0.904$, $p = 0.423$, respectively).

[34] It is noteworthy that a productivity shock has a stronger impact on effort than a wage shock. Both shocks are exogenous, that is, neither party can be 'blamed' for them. A possible explanation is that the wage-effort relationship is more complicated than assumed here. In separate analyses we regress effort on wages and find that the effects of the tax shocks are robust to various non-linear relationships between

Result 4: Without shocks, there is gift exchange.

Result 5: A productivity shock yields an increase in worker effort for low wages and crowds out gift exchange.

Result 6: There is no gift exchange when there is a wage shock.

1.4.3 Overview of Results

The big picture is that we reject the nulls of the Hypotheses 1 and 2 concerning the rounds without shocks, *nt*. This confirms the results in the existing literature that gift exchange occurs when there are no shocks. We add to this previous literature by showing that gift exchange also occurs when workers conduct a real-effort task.

We cannot reject the null of Hypothesis 3 (*wt*), but our Bayesian analysis does provide support for the null prediction of Hypothesis 4 (*wt*). We find no support for Hypotheses 5 or 6 (*et*). Considering the underlying theories used to develop the hypotheses in Section 1.3.3, these non-rejections suggest that the behavioral elements of our model in Section 1.2 play an important role in the interaction between employer and worker. Indeed, net wage illusion (H_0^3), loss aversion (H_0^5) and social preferences (H_0^6) underlie the null hypotheses that we fail to reject.

1.4.4 Welfare Consequences

Our results suggest that effort responds more strongly to shocks than wages do. The strength of gift exchange depends, however, on which shocks occur. Realized productivity shocks lead to increased effort, while realized wage shocks have no effect on effort provision. To investigate the net effects of this complex employer-worker interaction, Table 1.6 summarizes the earnings of hiring employers (left panel) and hired workers (right panel) in each treatment and tax outcome. As before, we take for each tax outcome the average earnings across rounds 3-8 as the unit of observation for the employer. Similarly, for worker earnings we use the average (across rounds) earning per worker (and per tax outcome) as the unit of observation.

In all cases, employers earn more on average than workers. This might be partially explained by the fact that the employers are on the short side of the market. Furthermore, employers bare more risks. Indeed, their payoffs vary more[35] and – unlike workers' payoffs – employers' earnings in a round may be negative.

We calculate theoretical ex-ante payoffs per treatment as the average payoffs in rounds

the two. More information is available upon request.

[35] The standard deviation of average (across rounds) employer payoffs is 18.6 points while it is only 9.1 points for workers.

Table 1.6: After tax earnings by treatment and tax outcome

	Panel A: Employer earnings					Panel B: Worker earnings			
	NT	**ET**	**WT**	**AT**		**NT**	**ET**	**WT**	**AT**
nt	55.1	52.1	51.8	58.6	**nt**	40.7	43.7	41.3	42.6
obs.	30	30	30	40	obs.	41	38	42	54
et		47.2		55.4	**et**		42.8		44.0
obs.		25		20	obs.		30		26
wt			47.5	65.8	**wt**			34.8	31.5
obs.			20	20	obs.			25	28
	Ex-ante payoffs					Ex-ante payoffs			
	55.1	48.9	51.4	59.2		40.7	43.6	38.9	40.9
se	(3.93)	(3.41)	(5.07)	(2.20)	**se**	(1.45)	(1.73)	(1.22)	(1.33)

Notes: Unit of observation is the employer (averaged across rounds 3-8) in the left panel and the worker (averaged across rounds 3-8) in the right panel. Cells show mean earnings. Ex-ante payoffs are determined by weighting realized earnings with the probability of a shock. Standard errors are in parenthesis.

with and without shocks weighted by the probability of each shock occurring. These theoretical before-tax-announcement payoffs are reported in the lower panel of Table 1.6. None of the differences for employers are significantly different when shocks are possible than when they are not (PtT, $p = 0.250$, $p = 0.560$, $p = 0.327$, for ET, WT, AT, respectively). This result is surprising given that the productivity shock directly reduces employers' payoffs. We know from our results on gift exchange, however, that workers respond to the productivity shocks by increasing effort. Ex-ante worker earnings show that they are not significantly worse off in tax treatments than in NT. In fact, workers earn slightly more when employers can be taxed (ET) but the difference is insignificant (PtT, $p = 0.183$); the other two comparisons to NT yield $p = 0.381$ for WT and $p = 0.916$ for AT.

By combining the numbers in the two panels of Table 1.6, we obtain a measure of aggregate surplus. This varies between 93.3 in WT and 100.6 in AT.[36] This difference is marginally significant (PtT, $p = 0.062$); all other pairwise differences in aggregate surplus are statistically insignificant (PtT, all $p > 0.21$). Tax revenues also differ across treatments. They are higher with a productivity tax (12.3 in ET and 15.2 in AT) than for a wage tax (8.7 in WT and 7.8 in AT). In AT, this gives an average tax revenue of 11.5. Together with the measured aggregate surplus, this suggests that due to gift exchange, a tax system with only wage taxes is less efficient than one with taxes on both sides of the labor market.

[36] This aggregate is slightly different than the sums of averages for employers and workers in Table 1.6. This is because we need to change the unit of observation to enable testing. Specifically, we determine here per employer for each contract the sum of her and the worker's earnings. We then use the mean per employer across rounds 3-8 as the unit of observation.

1.5 Concluding Discussion

We study gift exchange in a market where one-round negative shocks may occur. The predictions of our gift exchange model depend on whether we allow for other-regarding preferences, net wage illusion, or loss aversion. We test these predictions in a laboratory experiment. Our data for the case without shocks allow us to conclude that wages are set above the minimal level and that gift exchange takes place. This replicates the traditional gift exchange results in a real-effort environment. Our model shows that such gift exchange can take place even in the absence of reciprocal motives (cf. Charness and Rabin 2002). This result is reminiscent of models by Benjamin (2015) and Dickson and Fongoni (2019) who also predict gift exchange without reciprocity. The former, however, relies on previous transactions to determine the fairness of current choices. The latter introduces the notion of 'worker morale', which forms a ground for gift exchange. In contrast to both, gift exchange in our model is the result of other regarding preference even when these affect only current decisions and without the need to introduce novel concepts. Instead, our model applies well-established behavioral regularities. When we introduce wage or productivity shocks, the pattern of behavior we observe allows us to conclude that social preferences, net wage illusion and loss aversion all play a role in workers' decision making.

Though somewhat speculative, we can attempt to compare the three behavioral elements that we distinguish between. To start, given the broad literature on gift exchange, it should not come as a surprise that gift exchange is observed in the no-shock treatment. This shows that other-regarding preferences play an important role here, like they have been shown to play in many environments. Moreover, the occurrence of a wage shock has little effect on effort for low wage bins. This suggests that net wage illusion is also a strong force (which is also in line with much of the literature referred to above). The precise role of loss aversion is less clear. Though the results of our hypothesis testing show support for a model that includes loss aversion, it is not directly clear (or measurable) how strong the effect is when wage rigidity occurs. One interesting pattern in our data is that workers increase effort at low wages when their employers are hit by a shock. This might mean that workers have an aversion to their employer's losses. Whether such 'other-regarding loss aversion' exists and plays a role seems an interesting topic for future research. Finally, we can compare our approach to Dickson and Fongoni (2019)'s worker morale function. Our view is that the social preferences and the worker morale function play largely similar roles in the models as both bring about the fair wage-effort hypothesis. While in the worker morale case, loss aversion is a key assumption needed for creating the kink at the reference point, in our setting this kink arises already from the other-regarding preferences. Loss aversion's role is then to explain why tension arises in response to shocks, as is captured by the difference between the objective and the subjective fair wage. Adding worker morale to our model would, therefore, not change the results concerning wage rigidity.

Our results highlight how involved the interaction between shocks, wages, and effort responses can be. In rounds where no shock is realized we observe strong gift exchange, that is, a strong response of effort to wage levels. If a shock is actually realized, its effect on this effort response depends on which side of the market it hits. A negative wage shock has very little effect, while a negative productivity shock – which affects employers' earnings – makes workers exert much more effort (especially at low wages), compared to when no shock is realized. Employers do not appear to take these effort responses into account when setting a wage. They do not adjust their wage offers to the realization of a shock. This causes wage rigidity when shocks appear. For the wage shock, this is rationalizable because workers do not adjust their efforts. With a productivity shock, the workers compensate the employers by increased effort, and the latter have no reason to adjust the wages downward to compensate the shock. In fact, if they did reduce wages to cushion the shock, workers might not be as generous.

All in all, our results show that an understanding of the complexities of the labor market goes beyond the simple rational choice model with selfish preferences and requires more than simply allowing for gift exchange. Wage rigidity has been observed in the field (Kaur 2019) and we observe it in the laboratory. Additional insights from behavioral economics are needed to reconcile such data patterns even if one allows for other-regarding preferences. Nevertheless, the effects seem to evolve around a pattern of gift exchange and employers' expectation of this pattern. Our study hopes to contribute to a better understanding of the interactions involved.

Bibliography

Akerlof, George A. 1982. “Labor contracts as partial gift exchange.” *The quarterly journal of economics*, 97(4): 543–569.

Akerlof, George A, and Janet L Yellen. 1990. “The fair wage-effort hypothesis and unemployment.” *The Quarterly Journal of Economics*, 105(2): 255–283.

Benjamin, Daniel J. 2015. “A theory of fairness in labour markets.” *The Japanese Economic Review*, 66(2): 182–225.

Bewley, Truman. 1999. *Why don’t wages fall in a recession.* Harvard University Press Cambridge.

Brandts, Jordi, and Gary Charness. 2004. “Do labour market conditions affect gift exchange? Some experimental evidence.” *The Economic Journal*, 114(497): 684–708.

Buchanan, Joy, and Daniel Houser. forthcoming. “If wages fell during a recession.” *Journal of Economic Behavior & Organization.*

Charness, Gary, and Matthew Rabin. 2002. “Understanding social preferences with simple tests.” *The Quarterly Journal of Economics*, 117(3): 817–869.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. “oTree—An open-source platform for laboratory, online, and field experiments.” *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2015. “Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment.” *Management Science*, 61(8): 1777–1794.

Davis, Brent J, Rudolf Kerschbamer, and Regine Oexl. 2017. “Is reciprocity really outcome-based? A second look at gift-exchange with random shocks.” *Journal of the Economic Science Association*, 3(2): 149–160.

Dickens, William T, Lorenz Goette, Erica L Groshen, Steinar Holden, Julian Messina, Mark E Schweitzer, Jarkko Turunen, and Melanie E Ward. 2007. “How wages change: micro evidence from the International Wage Flexibility Project.” *Journal of Economic Perspectives*, 21(2): 195–214.

Dickson, Alex, and Marco Fongoni. 2019. “Asymmetric reference-dependent reciprocity, downward wage rigidity, and the employment contract.” *Journal of Economic Behavior & Organization*, 163: 409–429.

Fehr, Ernst, and Klaus M Schmidt. 1999. "A theory of fairness, competition, and cooperation." *The quarterly journal of economics*, 114(3): 817–868.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1993. "Does fairness prevent market clearing? An experimental investigation." *The quarterly journal of economics*, 108(2): 437–459.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a contract enforcement device: Experimental evidence." *Econometrica*, 65: 833–860.

Fochmann, Martin, Joachim Weimann, Kay Blaufus, Jochen Hundsdoerfer, and Dirk Kiesewetter. 2013. "Net wage illusion in a real-effort experiment." *The Scandinavian Journal of Economics*, 115(2): 476–484.

Gächter, Simon, and Christian Thöni. 2010. "Social comparison and performance: Experimental evidence on the fair wage–effort hypothesis." *Journal of Economic Behavior and Organization*, 76(3): 531–543.

Gächter, Simon, and Ernst Fehr. 2002. "Fairness in the labour market." In *Surveys in Experimental Economics*. 95–132. Springer.

Gerhards, Leonie, and Matthias Heinz. 2017. "In good times and bad–Reciprocal behavior at the workplace in times of economic crises." *Journal of Economic Behavior & Organization*, 134: 228–239.

Greiner, Ben. 2004. "The online recruitment system ORSEE 2.0." *A Guide for the Organization of Experiments in Economics*.

Hannan, R Lynn. 2005. "The combined effect of wages and firm profit on employee effort." *The Accounting Review*, 80(1): 167–188.

Hennig-Schmidt, Heike, Abdolkarim Sadrieh, and Bettina Rockenbach. 2010. "In search of workers' real effort reciprocity—a field and a laboratory experiment." *Journal of the European Economic Association*, 8(4): 817–837.

Kahneman, Daniel, Jack L Knetsch, and Richard Thaler. 1986. "Fairness as a constraint on profit seeking: Entitlements in the market." *The American economic review*, 728–741.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge university press.

Kaur, Supreet. 2019. "Nominal wage rigidity in village labor markets." *American Economic Review*, 109(10): 3585–3616.

Koch, Christian. 2021. "Can reference points explain wage rigidity? Experimental evidence." *Journal for Labour Market Research*, 55(1): 1–17.

Kőszegi, Botond, and Matthew Rabin. 2006. "A model of reference-dependent preferences." *The Quarterly Journal of Economics*, 121(4): 1133–1165.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2013. "Do wage cuts damage work morale? Evidence from a natural field experiment." *Journal of the European Economic Association*, 11(4): 853–870.

Mas, Alexandre. 2006. "Pay, reference points, and police performance." *The Quarterly Journal of Economics*, 121(3): 783–821.

Mauss, Marcel. 2002. *2.* Routledge.

Milgrom, Paul R, and John Donald Roberts. 1992. *Economics, organization and management.* Prentice-Hall.

Moir, Robert. 1998. "A Monte Carlo analysis of the Fisher randomization technique: reviving randomization for experimental economists." *Experimental Economics*, 1(1): 87–100.

Rubin, Jared, and Roman Sheremeta. 2015. "Principal–agent settings with random shocks." *Management Science*, 62(4): 985–999.

Schram, Arthur, Jordi Brandts, and Klarita Gërxhani. 2018. "Social-status ranking: a hidden channel to gender inequality under competition." *Experimental Economics.*

Sliwka, Dirk, and Peter Werner. 2017. "Wage increases and the dynamics of reciprocity." *Journal of Labor Economics*, 35(2): 299–344.

Weber, Matthias, and Arthur Schram. 2017. "The Non-equivalence of Labour Market Taxes: A Real-effort Experiment." *The Economic Journal*, 127(604): 2187–2215.

Appendix 1.A Shocks

In this appendix we discuss the effects of shocks in the model. The size of a shock is captured by parameter δ^j, $j \in \{W, F\}$ such that $0 < \delta^j < 1$. In our experimental design, shocks are realized before wages are set, so all effects are known before the employer and worker interact. δ^W then reduces the worker's payoff to $(1 - \delta^W)w$ while leaving the employer's earnings unchanged at $f(e) - w$. δ^F reduces the employer's payoff to $(1 - \delta^F)f(e) - w$ and leaves the worker's earnings at w. We call the latter a productivity shock.

1.A.1 Worker Effort Choice

First consider a productivity shock on the employer side, which changes the second term in the worker's utility eq. (1.1) to $\beta((1 - \delta^F)f(e) - w)$. This affects both the f.o.c. (1.3), where the r.h.s. is replaced by $\beta(1 - \delta^F)$ and the inequalities in (1.2), where $f(e)$ is replaced by $(1 - \delta^F)f(e)$. Denote by $\hat{e}^\delta_\sigma$ ($\hat{e}^\delta_\rho$) the solution to the f.o.c. for $\beta = \sigma$ ($\beta = \rho$).[37] Because $\frac{c'(e)}{f'(e)}$ is increasing in e, it holds that $\hat{e}^\delta_\sigma < \hat{e}_\sigma$ and $\hat{e}^\delta_\rho < \hat{e}_\rho$. For equal earnings ($\beta = 0$), we have optimal effort $\hat{e}^\delta_0$ implicitly determined by $w = \frac{(1-\delta^F)f(\hat{e}^\delta_0)}{2}$, with $\hat{e}^\delta_0 < \hat{e}_0$. For the worker's best response to wage w when a productivity shock δ^F occurs, this gives

$$\hat{e}^\delta(w) = \begin{cases} \hat{e}^\delta_\sigma, & \text{if } w < \frac{(1-\delta^F)f(\hat{e}^\delta_\sigma)}{2} \\ \hat{e}^\delta_0(w), & \text{if } \frac{(1-\delta^F)f(\hat{e}^\delta_\sigma)}{2} \leq w \leq \frac{(1-\delta^F)f(\hat{e}^\delta_\rho)}{2} \\ \hat{e}^\delta_\rho, & \text{if } w > \frac{(1-\delta^F)f(\hat{e}^\delta_\rho)}{2}. \end{cases} \tag{1.A.1}$$

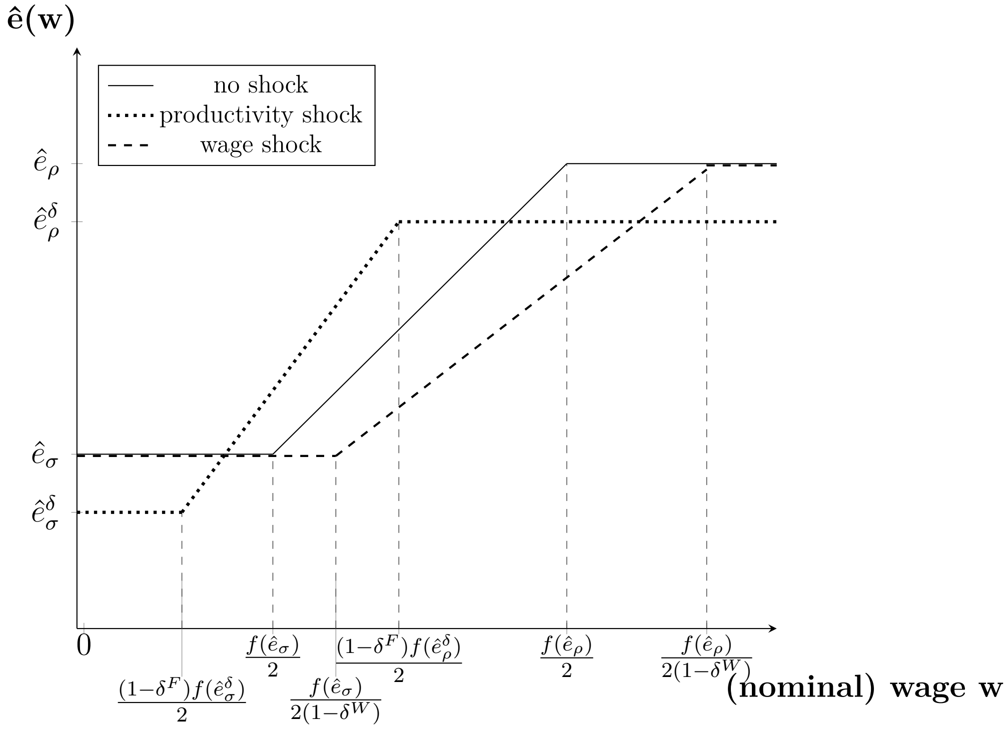
With a wage shock δ^W, on the other hand, the first term on the r.h.s. of utility eq. (1.1) is replaced by $(1 - \beta)(1 - \delta^W)w$. Because the wage the worker receives is sunk when she makes the effort decision, this shock does not affect f.o.c. (1.3). It does, however, affect the inequality conditions in eq. (1.2), where w is replaced by $(1 - \delta^W)w$.

Figure (1.A.1) illustrates the effects of shocks on either side of the market on the worker's best response function. For presentational purposes, we again assume a linear $f(e)$ (cf. fn. 10 in the main text). Observe that a shock at the employer side (dotted line) shifts the area of wages where the worker wants to equalize earnings to the left. Moreover, it shifts the upper bound ($\frac{(1-\delta^F)f(\hat{e}^\delta_\rho)}{2}$) further to the left than the lower bound ($\frac{(1-\delta^F)f(\hat{e}^\delta_\sigma)}{2}$), because $\hat{e}^\delta_\rho > \hat{e}^\delta_\sigma$ and f is monotonically increasing. As a consequence, the intermediate wage area where earnings are equalized is smaller with the shock than when $\delta^F = 0$. Moreover, the productivity shock shifts the worker's best response curve downward. This is because the effect of effort on the employer's income is diminished, which the worker

[37] The optimal effort level $\hat{e}$ is only affected by a shock on the employer side, not by a wage shock (as explained below); a superscript δ for the optimal effort therefore always refers to δ^F.

internalizes through the social preferences that enter worker's utility.

Figure 1.A.1: Worker's best response with shocks



A shock to worker's wages (dashed line), on the other hand, shifts the best response to the right because a higher wage is needed to equalize earnings. Here, the upper bound ($\frac{f(\hat{e}_\rho)}{2(1-\delta^W)}$) shifts further to the right than the lower bound ($\frac{f(\hat{e}_\sigma)}{2(1-\delta^W)}$) because $\hat{e}_\sigma < \hat{e}_\rho$ and f is monotonically increasing. With a wage shock, there is no vertical shift of the response function, because this is determined by the f.o.c. (1.3), which is not affected by δ^W.

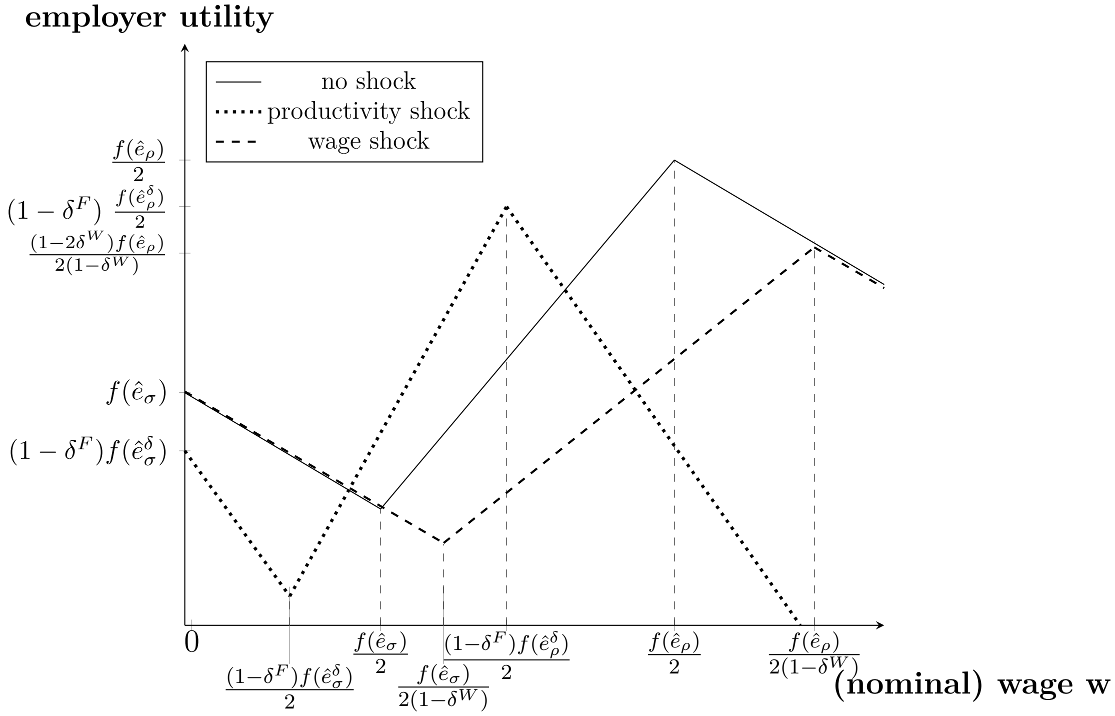
1.A.2 Employer Wage Setting

The effects of shocks on wage setting at stage 1 are illustrated in Figure 1.A.2.

Following a productivity shock at the employer side and the expected best response by the worker, employer's utility is given by $u^F = (1-\delta^F)f(\hat{e}^\delta(w)) - w$. At the minimal wage (here normalized to $w = 0$), this gives $u^F = (1-\delta^F)f(\hat{e}_\sigma^\delta)$. Utility then declines linearly until $w = \frac{(1-\delta^F)f(\hat{e}_\sigma^\delta)}{2}$, after which the worker responds by equalizing earnings. This gift exchange takes place up to the objectively fair wage $w = \frac{(1-\delta^F)f(\hat{e}_\rho^\delta)}{2}$. At this point, the employer obtains $u^F = (1-\delta^F)f(\hat{e}_\rho^\delta) - \frac{(1-\delta^F)f(\hat{e}_\rho^\delta)}{2} = (1-\delta^F)\frac{f(\hat{e}_\rho^\delta)}{2}$. As wages increase beyond this level, employer's payoff decreases linearly because no further gift exchange takes place.

A wage shock yields employer utility $u^F = f(\hat{e}((1-\delta^W)w)) - w$. At the minimum wage $w = 0$, optimal effort is $\hat{e}_\sigma$ and as wages increase, the u^F develops as with $\delta^W = 0$.

Figure 1.A.2: Employer's utility with shocks



It takes a higher wage for the worker to start equalizing earnings, however, as effort does not increase until the net wage reaches the minimum employer profit, which is a higher wage than when $\delta^W = 0$ (cf. Figure 1.A.1). The employer's utility subsequently reaches its maximum at a higher (objectively fair) wage ($\frac{f(\hat{e}_\rho)}{2(1-\delta^W)}$), at a lower level of utility at $\frac{f(\hat{e}_\rho)}{2} - \frac{f(\hat{e}_\rho)}{2(1-\delta^W)} = \frac{f(\hat{e}_\rho)(1-2\delta^W)}{2(1-\delta^W)}$ due to the increased wage expenses.

Note that one will observe gift exchange in the SPE if the utility achieved at the objectively fair wage is higher than the utility achieved at the minimum wage. With a productivity shock this requires $(1-\delta^F)\frac{f(\hat{e}^\delta_\rho)}{2} > (1-\delta^F)f(\hat{e}^\delta_\sigma)$, which occurs iff $\frac{f(\hat{e}^\delta_\rho)}{2} > f(\hat{e}^\delta_\sigma)$. In case of a wage shock, the objectively fair wage yields higher employer utility than the minimum wage if $\frac{1-2\delta^W}{1-\delta^W}\frac{f(\hat{e}_\rho)}{2} > f(\hat{e}_\sigma)$. Because $\frac{1-2\delta^W}{1-\delta^W} < 1$, this condition also implies $\frac{f(\hat{e}_\rho)}{2} > f(\hat{e}_\sigma)$. Thus, if worker preferences yield an SPE with gift exchange when there is a wage shock, then there is also gift exchange in the equilibrium for the case without a shock.

Appendix 1.B Loss Aversion

In this appendix, we adapt the model to allow for loss aversion. Recall from the main text that the subjectively fair wage is the objectively fair wage in the absence of shocks, that is, $\tilde{w} = \frac{f(\hat{e}_\rho)}{2}$.[38] The best response function $\hat{e}^\delta(w)$ now becomes:

$$\hat{e}^\delta(w) = \begin{cases} \hat{e}^\delta_{\sigma-\lambda}, & \text{if } w < \frac{(1-\delta^F)f(\hat{e}^\delta_{\sigma-\lambda})}{2} (< \tilde{w}) \\ \hat{e}^\delta_0(w), & \text{if } \frac{(1-\delta^F)f(\hat{e}^\delta_{\sigma-\lambda})}{2} \leq w \leq \frac{(1-\delta^F)f(\hat{e}^\delta_{\rho-\lambda})}{2} (< \tilde{w}) \\ \hat{e}^\delta_{\rho-\lambda}, & \text{if } \frac{(1-\delta^F)f(\hat{e}^\delta_{\rho-\lambda})}{2} < w < \tilde{w} \\ \hat{e}^\delta_\rho, & \text{if } w \geq \tilde{w} (> \frac{(1-\delta^F)f(\hat{e}^\delta_\rho)}{2}). \end{cases} \tag{9'}$$

The first line in the r.h.s. of eq. (9') describes the case where the current wage is lower than the subjectively fair wage and lower than the employer payoff; this is responded to in a way that gives minimal effort while accounting for loss aversion. In the second line, the worker equalizes earnings for the current wage, which is lower than the subjectively fair wage. In the third line, the current wage is lower than subjectively fair wage, but the optimal response creates advantageous inequality for the worker, as wage is above the *objectively fair wage*. In the final line, the subjectively fair wage is such that the optimal effort response creates higher earnings for the worker than for the employer, while the actual wage is even higher.

Figure 1.B.1 demonstrates the best response functions $\hat{e}(w)$ when there is both net wage illusion and loss aversion. Note the discontinuity at the subjectively fair wage $w = \frac{f(\hat{e}_\rho)}{2}$. The 'jump' at this wage level equals λ in all three cases. Because of the jump in the effort response at the subjectively fair wage, a similar discontinuity occurs for the employer's utility. This is illustrated in Figure 1.B.2.

Now there are potentially three local maxima in the employer's utility. They are at the minimum wage (0), the objectively fair wage ($\frac{(1-\delta^F)f(\hat{e}^\delta_{\rho-\lambda})}{2}$) and the subjectively fair wage ($w_{t-1} = \frac{f(\hat{e}_\rho)}{2}$). Assuming that the objectively fair wage yields higher utility than the minimum wage (which holds if $f(\hat{e}_{\rho-\lambda}) > 2f(\hat{e}_{\sigma-\lambda})$), the employer will prefer to keep wages at the subjectively fair level if and only if

$$(1-\delta^F)\frac{f(\hat{e}^\delta_\rho)}{2} - \frac{f(\hat{e}_\rho)}{2} \geq (1-\delta^F)\frac{f(\hat{e}^\delta_{\rho-\lambda})}{2}, \tag{1.B.1}$$

where we assume that the wages will be unchanged if the employer is indifferent. Eq. (1.B.1) is a condition for wage rigidity. If it holds, then employers will prefer to hold wages constant, even if they face a shock on their income.

[38] Note that $\tilde{w} = \frac{f(\hat{e}_\rho)}{2} > \frac{f((1-\delta^F)\hat{e}^\delta_\rho)}{2} > \frac{f((1-\delta^F)\hat{e}^\delta_{\rho-\lambda})}{2}$. The first inequality is illustrated in Figure 1.A.1, the second follows because the worker puts less weight on the employer's earnings and therefore exerts less effort. As a consequence, $w < \frac{f((1-\delta^F)\hat{e}^\delta_{\rho-\lambda}}{2})$ implies $w < \tilde{w}$.

Figure 1.B.1: Optimal response with net wage illusion and loss aversion

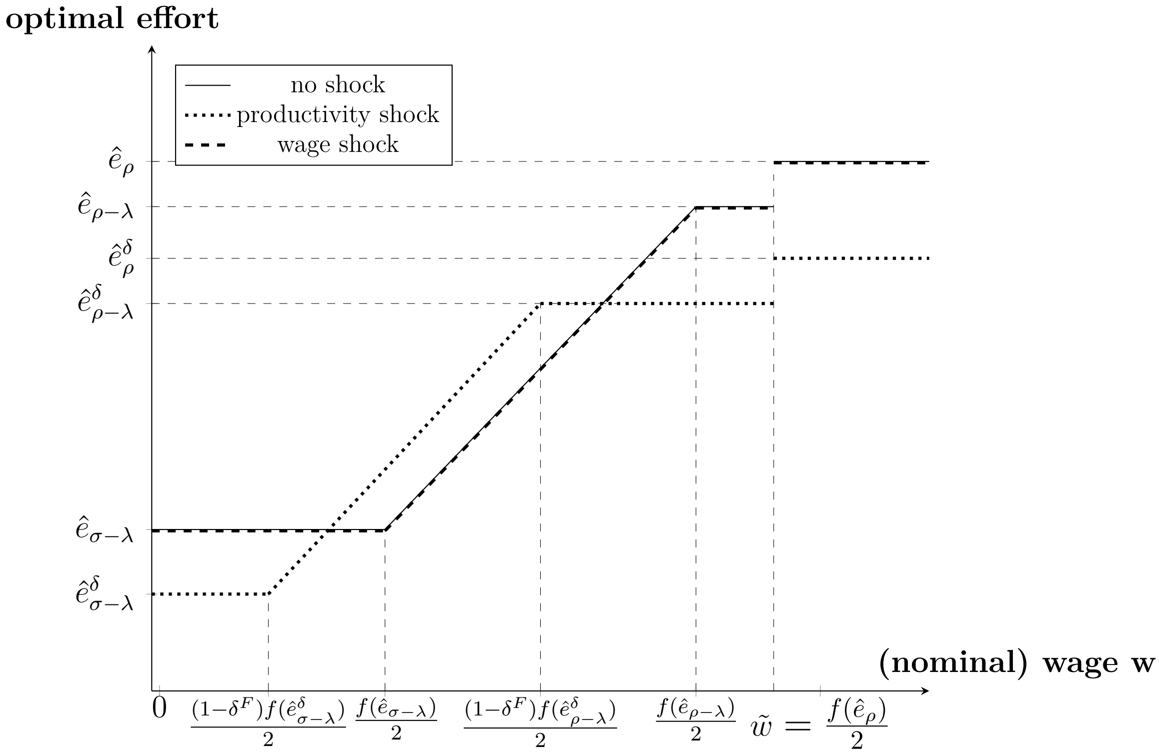
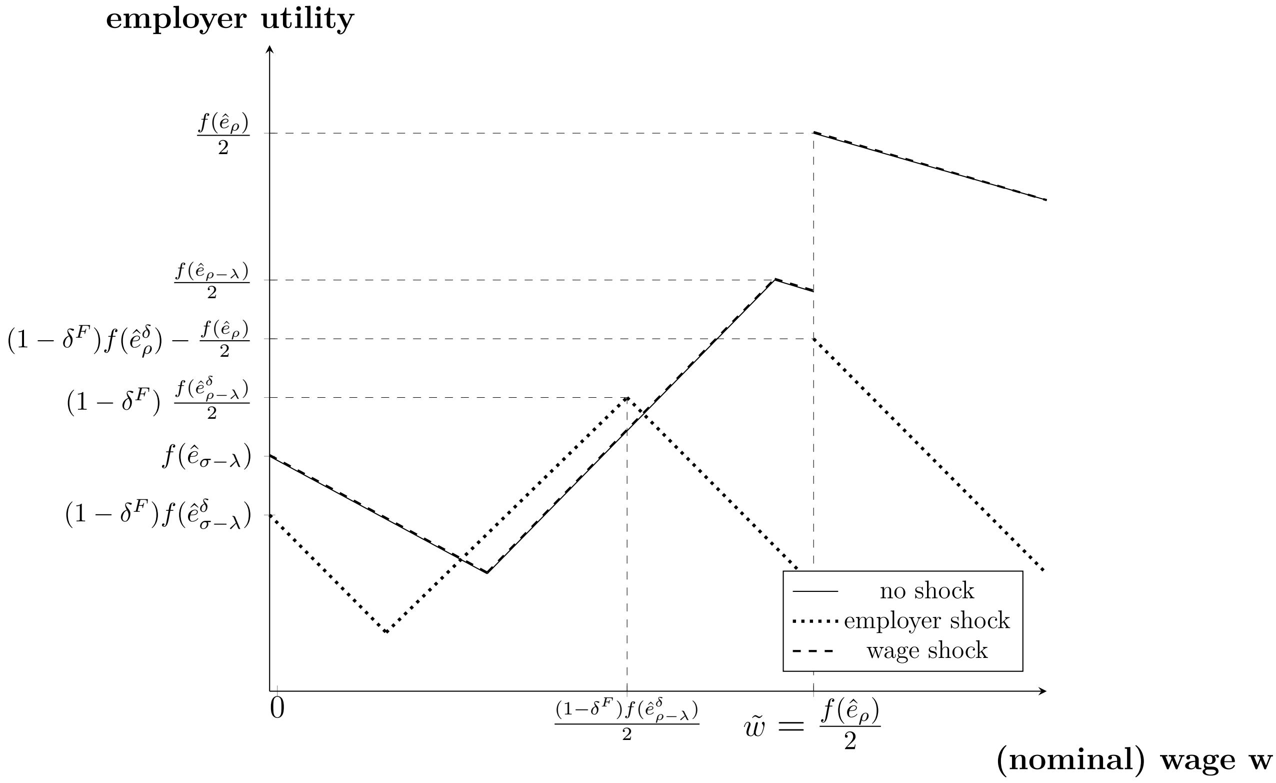


Figure 1.B.2: Employer's utility with net wage illusion and loss aversion



Appendix 1.C Experimental Instructions [Original in Italian]

The instructions differ for each treatments. When appropriate, we indicate additional texts by the following system. "When taxes" refers to all treatments that allow taxes: AT, ET, and WT. "In AT" refers to the tax treatment with all taxes, "ET" refers to the tax treatment with only employer taxes and "WT" refers to the tax treatment with only wage taxes.

Welcome to the experiment!

From now on, please, do not talk with the other participants. If you have any questions, please, raise your hand. Place your phone in your bag: you are not allowed to use it during the experiment. In case you want to revisit the instructions after the software tutorial, you can use the paper version on your desk where you also find a pen and a paper.

Your payoff from the experiment will consist of two parts: the 5 euro show-up fee and the earnings (or losses) from 2 rounds out of the 8 rounds in total. These 2 rounds will be chosen at random.

Role

You participate in a labor market that has 5 employers and 7 employees. After the tutorial and a questionnaire on the instructions, you will be randomly assigned to either the role of an employer or the role of a worker, and you will keep the same role for the entire duration of the experiment.

Overall structure

The experiment consists of 8 rounds.

[**When taxes:** *In the beginning of each round, the taxation scheme of the round will be announced. After the announcement,*] each round will have the following stages:

1st Stage: Hiring

Each employer can make a wage offer on a public platform, and each worker can accept one of these offers. Once an offer becomes accepted, the hired worker will work that round for the employer that made the offer. All the hiring results of the round will be made public.

2nd Stage: Work

Each hired worker has 5 minutes to work on the tasks. After the 5 minutes, the work results will be communicated to the respective worker and employer, and the earnings are calculated.

Detailed instructions

Hiring Stage

[**When taxes:** *Before the hiring stage begins, there will be a 10 second announcement that reveals the taxation scheme that is effective during the round (more information on the possible taxation schemes in the next page of instructions).*]

The hiring stage lasts at most for 2 minutes. There are 5 employers and 7 workers in the market. Each employer can announce a wage offer on a public platform. The offer must be between 30 and 100 points, in steps of 5 points, and it can be modified while not yet accepted, but cannot be withdrawn entirely once made.

A worker can accept one of the available offers. Once accepted, the worker is immediately hired by the employer for the reminder of the round and the offer is removed from the platform. If more than one worker attempts to accept the same offer, it is granted to the fastest. All of the offers and subsequent modifications are updated to the platform in real time and published in a random order.

If an offer is not accepted within the 2 minutes, the employer is not able to hire anyone. In the same way, if a worker does not accept an offer within the 2 minutes or if all of the 5 offers made have been accepted by other workers, the market closes and these workers will be unemployed for the round. Out of the 7 workers, at least 2 will be unemployed every round.

Without a contract, the workers and employers will not participate in the remaining stages of the round: an employer earns 0 points and a worker earns 20 points as an unemployment benefit. Both will resume the experiment again in the beginning of the next round.

If an employer hires a worker, the employer receives 40 points and any earnings from the work of the hired worker. The worker's wage will then be subtracted from these earnings. The worker's earnings consist of the wage. [**When taxes:** *AT: Both payoffs/ET: employer's payoff/ WT: worker's payoff may be subject to taxes, as explained in the next part.*]

The experiment is anonymous: the worker will not know the identity of the employer, and likewise, the employer will not know the identity of the worker.

After the hiring stage, all of the participants see the overall results of the hiring stage: how many workers were hired and at what wages.

[When taxes:] ***Taxes***

[The options and probabilities depend on which taxes are possible. The following section is written for AT unless otherwise specified]

The taxation scheme is announced before the hiring stage, it is randomly chosen by a computer, and it can be one of 3 ***[In ET or WT: 2]*** *possibilities:*

- ***No taxes*** *(probability 66.7%)*
- ***Tax*** *of 20% on the revenues of the employer (probability 1/6 = 16,7%)* ***[In ET 1/3 = 66.7%, not mentioned in WT]***
- ***Tax*** *of 20% on the wage of the worker (probability 1/6 = 16,7%)* ***[In WT 1/3 = 66.7%, not mentioned in ET]***

In total, there is a 33% probability that a tax is applied, and a 67% probability that there are no taxes; on average, 1 in 3 rounds has taxes. ***[In AT only:*** *The type of the tax is randomly chosen by computer, each type being equally likely.*]

[**In AT and ET only:** *The tax on the revenues of the employer reduces the earnings from the worker tasks: each correctly completed task is worth 16 points, instead of the 20 points when there is no tax. The tax does not impact the 40 points received from hiring.*]

[**In AT and WT only:** *The tax on the earnings of the worker reduces the amount of wages received by 20%. Each employer however pays the full salary.*]

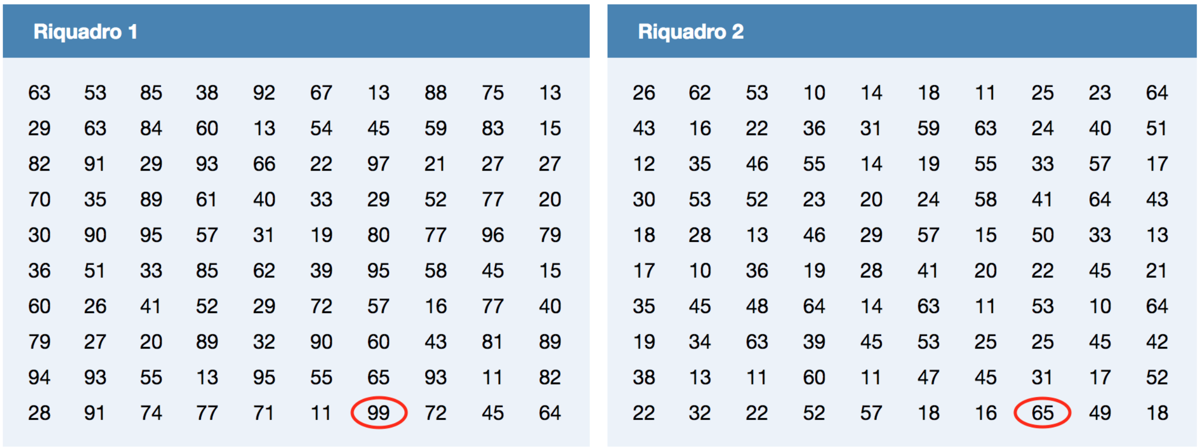
The collected taxes will be returned to the experimenter.

Work Stage

The hired workers have 5 minutes to work, during which they can attempt at most 10 tasks in total. Each task consists of two boxes, each containing 100 numbers: the task is to find the largest number in each box and then sum them together.

Each correctly completed task will give the employer 20 points [**In AT and ET:** *if there are no taxes on the employer's taxes, in which case, each correctly complete task is worth 16 points*]. Wrong answers do not affect payoffs but count as 'attempted tasks'. The workers can submit only one answer per task.

Example: The largest number in the left box is 99 and the largest number in the right box is 65, both are circled with red. Summed together they give 99 + 65 = 164: **164** is the correct answer to be submitted!



Riquadro 1

63	53	85	38	92	67	13	88	75	13
29	63	84	60	13	54	45	59	83	15
82	91	29	93	66	22	97	21	27	27
70	35	89	61	40	33	29	52	77	20
30	90	95	57	31	19	80	77	96	79
36	51	33	85	62	39	95	58	45	15
60	26	41	52	29	72	57	16	77	40
79	27	20	89	32	90	60	43	81	89
94	93	55	13	95	55	65	93	11	82
28	91	74	77	71	11	[✓ 99]	72	45	64

Riquadro 2

26	62	53	10	14	18	11	25	23	64
43	16	22	36	31	59	63	24	40	51
12	35	46	55	14	19	55	33	57	17
30	53	52	23	20	24	58	41	64	43
18	28	13	46	29	57	15	50	33	13
17	10	36	19	28	41	20	22	45	21
35	45	48	64	14	63	11	53	10	64
19	34	63	39	45	53	25	25	45	42
38	13	11	60	11	47	45	31	17	52
22	32	22	52	57	18	16	[✓ 65]	49	18

The Payoffs

After 5 minutes or after having tried all 10 tasks, all of the participants are directed to a results page. The worker and the employer who has hired the worker get to know the number of correct and attempted tasks, and the resulting payoffs of both, but will not get to know the results of the other participants.

Scenario A:

If the participant does not have a contract:

- Employer's payoff = **0 points**
- Worker's payoff = **20 points**

Scenario B:

If the participant has a contract [When taxes: *and there are no taxes]:*

- Employer's payoff = **40 − wage + 20 * number of tasks correct**
- Worker's payoff = **wage**

In other words, the employer receives 40 points when hiring a worker, pays the wage and receives the revenues from each correctly completed task. What remains is the earnings of the employer, and note that this can also be negative. Conversely, the earnings of the worker consists of the wage.

[Only in AT and WT: ***Scenario C:***

If the participant has a contract and there is a 20% tax on the earnings of the employer*, the payoff from each correctly completed task is reduced to 16 (from 20) and thus the payoffs are given as:*

- Employer's payoff = **40 − wage + 16 * number of tasks correct**

The worker's payoff is the same as under Scenario B.]

[**Only in AT**] ***Scenario D:*** [**OR Only in ET**] ***Scenario C;***
***If the participant has a contract and there is a 20% tax on the earnings of the worker*, the payoff of the worker is given by the salary less the taxes:**

- Worker's payoff = **wage − 20% of the wage**

The employer's payoff is the same as under scenario B.]

[**Only in AT**] *The two taxation systems are alternatives, they can never apply simultaneously.*

The points earned in the laboratory will be converted into Euros with the following exchange rate: **10 points = 1 euro**. On top of the 5 euro show-up fee, the participants are remunerated for only two rounds (out of the 8 in total) that are randomly selected in the end of the experiment.

Comprehension test

The comprehension test consisted of 12 true or false statements. The first 10 questions were the same for all tax treatments. The correct answer is reported in the parenthesis.

1. If a worker is unemployed for a round, she or he does earns nothing. (FALSE)
2. If an employer does not manage to hire a worker for a round, the employer earns nothing. (TRUE)
3. Accepting an offer, the worker commits to work for that employer for that round. (TRUE)
4. An employer who has hired someone earns 40 points. (TRUE)
5. In general, the salary is deducted from the earnings of the employer and given to the worker. (TRUE)
6. The number of tasks that a worker can try is unlimited. (FALSE)
7. The workers obtain a higher salary if they complete more tasks. (FALSE)
8. Other than the worker himself/herself, only the employer will get to know how many tasks were completed. (TRUE)
9. You will be compensated for all of the 8 rounds. (FALSE)
10. There are always unemployed workers. (TRUE)

The last two questions depend on what taxes are possible.

When no taxes are possible (NT):

11. Your earnings will depend on your decisions and those of the other participants. (TRUE)

12. The earnings of an employer cannot be negative for a round. (FALSE)

If only productivity taxes are possible (ET):

11. 20% of 20 points is 4 points. Thus, when we have taxes on the employers, the earnings per each correct task is 16 instead of 20 points. (TRUE)

12. The earnings of an employer cannot be negative for a round. (FALSE)

If only worker taxes are possible (WT):

11. The earnings of an employer can be negative for a round. (TRUE)

12. The taxes on the worker's earnings are always 20 points. (FALSE)

If both taxes are positive (AT):

11. 20% of 20 points is 4 points. Thus, when we have taxes on the employers, the earnings per each correct task is 16 instead of 20 points. (TRUE)

12. The taxes on the worker's earnings are always 20 points. (FALSE)

Appendix 1.D All Rounds

In this appendix, we provide the most important results of the main text when using data from all eight rounds.[39] We start by investigating how wages respond to shocks. Table 1.D.1 shows average wages per treatment and tax shock.

Table 1.D.1: Wages, treatments and shocks, all rounds

tax outcome	NT	ET	WT	AT_{et}	AT_{wt}	**pooled**
nt	**42.6**	**46.6**	**44.6**	**48.7**	**43.2**	**45.0**
obs.	30	30	30	20	20	130
et		**43.6**		**44.6**	**42.0**	**43.4**
obs.		25		20	20	65
wt			**48.2**	**50.8**	**38.8**	**46.2**
obs.			30	20	20	70
PtT (p-values)						
nt vs et	-	*0.093*	-	*0.003*	*0.208*	
nt vs wt	-	-	*0.073*	*0.153*	*0.001*	

Notes: Results are for rounds 1-8. Tax shocks occurred in rounds 2, 4, and 5. The unit of observation is the mean wage paid by an employer across rounds. Paired tests between shock- and no-shock rounds are reported. We do not conduct tests for the pooled data because these combine paired with unpaired comparisons. Mean wages across employers are in bold. 'obs.' shows the number of employers. NT: no taxes possible; nt: no tax shock realized; WT: wage tax possible; wt: wage tax shock realized; ET: productivity tax possible; et: productivity tax shock realized; AT_{et}: wt realized in round 2, et realized in rounds 4 and 5; AT_{wt}: et realized in round 2, wt realized in rounds 4 and 5. 'pooled' combines treatments. PtT: permutation t-test.

The results are similar to those observed for rounds 3-8 in Table 1.3, but somewhat statistically stronger.[40] An exception is that now a wage shock wt yields higher wages in WT (and also in AT_{wt}). It appears that wage shocks in the second round (before wages in general have settled) are compensated by higher wages. In all treatments the mean wages are again far from the minimum level of 30 points (which is not surprising, because wages in the first two rounds are higher than in subsequent rounds). For comparison to Table 1.4, Table 1.D.2 shows the mean wages per treatment. As we found for rounds 3-8, we observe no treatment differences.

1.D.1 Effort and Gift Exchange

To start, Table 1.D.3 summarizes the mean realized effort across treatments and shocks.

[39] Unless indicated otherwise, we use the same methods as in the main text.

[40] Combining ET and AT, the mean wages are 46.1 (nt) and 43.4 (et); the difference is significant (PtT, $p = 0.001$, $N - 65$). Pooling WT and AT, mean wages are 45.4 (nt) and 46.2 (wt) and differ insignificantly (PtT, $p = 0.488$, $N = 70$.

Table 1.D.2: Wages and treatments

	NT	**ET**	**WT**	**AT**
all tax outcomes	**42.6**	**45.9**	**45.2**	**45.0**
obs.	30	30	30	40
PtT for differences against NT				
p-value (p=c/n)		0.247	0.350	0.397

The unit of observation is the mean wage of an employer across rounds (presented in bold). *NT*: no taxes possible; *WT*: wage tax possible; *ET*: productivity tax possible and *AT*: both taxes possible. PtT: (unpaired) permutation t-test.

Table 1.D.3: **Effort, treatments, and shocks**

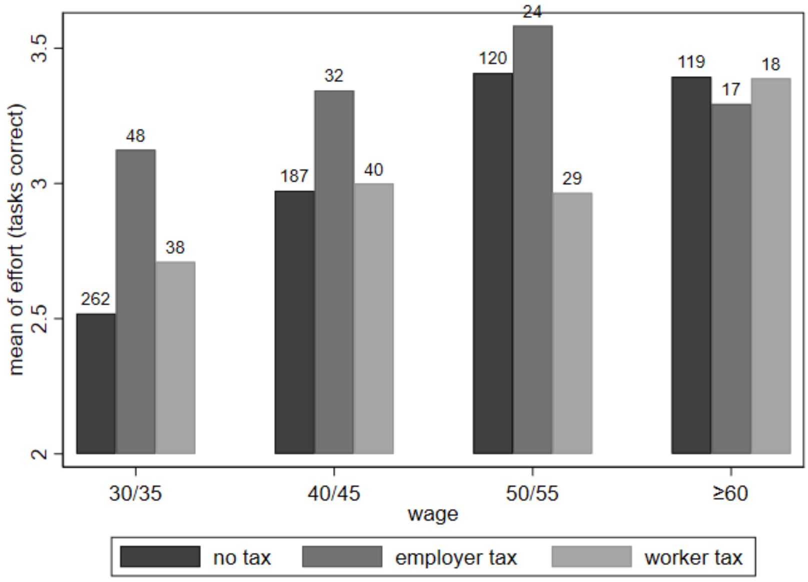
tax outcome	*NT*	*ET*	*WT*	AT_{et}	AT_{wt}	**pooled**
nt	**2.9**	**3.1**	**2.9**	**3.3**	**3.1**	**3.0**
obs.	30	30	30	20	20	130
et		**3.1**		**3.8**	**3.0**	**3.3**
obs.		25		20	20	65
wt			**2.8**	**3.4**	**3.2**	**3.1**
obs.			30	20	20	70
PtT (p-values)						
nt vs et	-	*0.334*	-	*0.017*	*0.733*	
nt vs wt	-	-	*0.854*	*0.668*	*0.700*	

Notes: Results are for rounds 1-8. Tax shocks occurred in rounds 4 and 5. The unit of observation is the mean effort received by an employer across rounds. We do not conduct tests for the pooled data because these combine paired with unpaired observations. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; AT_{et}: *wt* realized in round 2, *et* realized in rounds 4 and 5; AT_{wt}: *et* realized in round 2, *wt* realized in rounds 4 and 5. 'pooled' combines treatments. PtT: permutation t-test.

Again, the results are similar to those reported in the main text. At this level of aggregation (across wages) there is little variation of effort across shocks.

Finally, Figure 1.D.1 relates effort to wages. This shows the same pattern of gift exchange as observed in Figure 1.6 of the main text. In *nt*, the effort increases from bin 30/35 to 40/45 and 40/45 to 50/55 are both statistically significant (PtT, $p = 0.005$ and $p = 0.018$, respectively) while the step from 50/55 to higher wages is not (PtT, $p = 0.967$). In *et* and *wt* we observe no gift exchange; none of the differences between adjacent wage bins is statistically significant (PtT, all $p > 0.37$). Comparing the effects of shocks on effort within wage bins shows for *et* that effort is significantly higher than in *nt* for the lowest wages (PtT, $p = 0.024$) while the differences in the other three bins are all insignificant (PtT, all $p > 0.22$). For *wt*, none of the differences with *nt* is statistically significant (PtT, all $p > 0.18$). All of these results are qualitatively the same as those reported in the main text for rounds 3-8.

Figure 1.D.1: Average effort across wages



Note: The number of observations in each bin is reported above the bar.

Chapter 2

Fairness of Wage Cuts[1]

2.1 Introduction

Labor markets are considered to be rigid; wages rarely adjust downwards (Dickens et al., 2007). Do employers avoid cutting wages because they *believe* that wage cuts are detrimental to work motivation? Bewley (1999) interviews managers and labor market experts and finds evidence for the fear that wage cuts hurt worker morale. Is this fear based on facts? It is difficult to address this question with observational data from the field. Wage cuts are rare, and when they do happen, they might be particularly justified and accepted, for which reason changes in work morale might not be observed. Furthermore, when wages are cut, great measures are taken at times to frame them as removals of recently negotiated extra benefits or additions to work time rather than as direct cuts to the pay.[2] In a similar vein, nominal wages often lose value over time through phenomena like inflation or currency devaluation unless wages are actively updated. Surveys by Kahneman et al. (1986) and by Kaur (2019) suggest that such "hidden" cuts to real wages are largely perceived as acceptable. It thus follows that these cuts due to inaction are not expected to impact work morale. Similarly in a lab experiment, Charness (2004) shows that wages set by self-interested firms have different effects on workers' effort than same wages set by a random or disinterested mechanism.

This chapter sets to investigate how different types of wage cuts affect work morale, and if such effects are correctly anticipated by those in the role of the employers. I am in particular interested in finding the extend to which wage cuts reduce effort, whether neg-

[1] I am grateful for the comments by Arthur Schram and thank Philipp Chapkovski for his help with programming. I gratefully acknowledge funding from the Academy of Finland and the European University Institute.

[2] For example, a Finnish collective bargaining agreement *kilpailukykysopimus* in 2016 aimed to increase the competitiveness and employment rate by freezing pay levels, by additions to work time, and by cutting holiday pay. In 2011-2014 in Portugal, aside from nominal wage cuts, wage reductions were made also by reducing and suspending holiday pay and by increasing weekly working hours from 35 to 40.

ative shocks may justify wage cuts, and how the effects of real wage cuts compare to those of nominal wage cuts. I do so by combining theoretical and experimental analysis. The theoretical model builds on that of the previous chapter but adds a negative reciprocity parameter, a component of Charness and Rabin (2002)'s social preference function, that was not included in the previous chapter. Such negative reciprocity may be triggered by nominal wage cuts. I use this model to theorize how workers react to different kinds of wage adjustments and how employers should behave, anticipating the workers' reactions. The predictions are tested in a laboratory experiment that also builds on the framework set in the earlier chapter. The major difference between the two settings is that nominal cuts are possible in the current one. The workers continue to be hired through an auction, but now, the initial wage offer with which the worker is hired is non-binding and the employer can change this at will afterwards. Shock announcements are made after the hiring stage but before the final wage is decided, thus potentially offering justifications for wage adjustments. Changes to wages are not, however, conditional on a shock occurring. Shocks that hit the workers capture the essential features of *real wage cuts* that happen through inflation – if the employer takes no action, the workers experience a real wage cut even if nominal wage is kept the same – allowing one to compare the effects of real wage cuts to those of nominal wage cuts. Last, as in the previous chapter, effort is measured as performance in a real effort task. Real effort allows on to capture 'subconscious' effort responses, such as a loss of motivation to concentrate on the effort task, that stated effort does not necessarily capture. On the other hand, real effort also incorporates other features, including intrinsic motivation to work and seeking recognition and approval from the employer, that are arguably important in the work that people do outside the laboratory, adding realism to the experimental design.

The theoretical model provides an interesting framework to study the effects of wage cuts as it produces an asymmetric wage-effort relation in line with the fair wage hypothesis of Akerlof (1982) and Akerlof and Yellen (1990). Effort responds to wages up to a level called the objectively fair wage. Beyond this level, further wage increases no longer increase effort. The model suggests that wage cuts are mostly unprofitable; because cuts can trigger negative reciprocity, employers should immediately offer workers the final wage. The only exception occurs when workers find it justifiable to share the burden of the shock. Then, it is justifiable to cut wages in response to a shock that makes employer worse off. By the same logic, employers should actively increase wages after the negative shock hits the workers. In contrast, if maintaining the *status quo* is a stronger fairness norm than equally spreading the impact of shocks, no wage adjustment is considered fair after a shock, making all wage cuts unacceptable to the workers. The results from the experiment suggest the latter more than the former. In the absence of shocks, wage cuts have a relatively small or no impact on effort, while when wages are cut after the announcement of a shock, regardless of its type, effort is significantly reduced.

This study contributes, on one hand, to the literature on wage cuts and their effects on worker morale and productivity and, on the other hand, the literature on gift exchange labor markets. Studying wage cuts is difficult in the field; effort can consists of many factors and is therefore difficult to measure. Moreover, there are many confounding factors to control for; for example, individuals may be concerned about reputation and future job opportunities, making them less sensitive to wage changes. Field settings do not easily allow one to study what people would do in the role of employers. Last, there are more legal limitations to what can be done in the field than what can be done in the lab. So while Kube et al. (2013) find that wage cuts have negative and persistent effects on productivity in a field experiment and that wage increases do not have similar positive effects, others have found weaker or no effect. To mention a few, Gneezy and List (2006) find temporary increases in effort after an increase in hourly wages that disappear quickly. Hennig-Schmidt et al. (2010) find wages to affect effort in the lab but not in the field setting. Cohn et al. (2015) show that wage increases have modest effects on average. There is, however, a lot of heterogeneity; Cohn and coauthors demonstrate how personal fairness perceptions explain the reactions to the wage adjustments. In the lab, I expect a cleaner measurement of the wage effort relationship.

This chapter contributes more specifically to the literature on wage cuts in the laboratory and online settings. In particular, Hannan (2005) and Chen and Horton (2016) vary the justifications given for wage cuts and find that when a wage cut is “better justified”, its effect on effort is weaker. Chen and Horton (2016) study the spot labor market workers in MTurk also using a real effort task (transcribing). The excuse of maximizing profits is not seen as a proper justification for cutting the payment, while other reasons seem to work better. Reactions are measured by accepting or rejecting a further task at a lower price. Hannan (2005) show that “more justified” wage cuts due to reduced employers’ earnings lead to smaller reactions in effort than when wage cuts follow an increase in the employers’ earnings. Koch (2021) also finds that wage cuts lead to drops in effort even after employer shocks. This study differs from the last two, however, by also allowing cuts when there is no shock. Another key difference is that I look at the effects of a wage shock that makes workers worse off, in contrast to the employer profit shocks that are featured in most studies.

Last, I contribute to the literature and shocks and gift exchange markets. Aside from Koch (2021) and Hannan (2005), as discussed above, Gerhards and Heinz (2017); Rubin and Sheremeta (2015); Davis et al. (2017) and Buchanan and Houser (forthcoming) belong to this literature. Buchanan and Houser (forthcoming) is the closest to this study. It does not, however, allow wages to adjust every round but only after the first round and after the shock. Like this paper, Buchanan and Houser (forthcoming) considers both types of shock, one on employers (recession) and another on both employers and workers (inflation).

In contrast to these studies (with the exception of Hannan 2005), this experiment follows the design of the original Fehr et al. (1993) in that workers are hired in a common market and not simply paired at the beginning of the round. Although the original wage offers are to an extent only cheap talk, it is reasonable to expect that being hired at this initial offer in an active market makes the offer a stronger reference point than if there had not been a market.[3] The fact that the initial offer is not guaranteed also creates a situation where keeping the initial offer can trigger some positive reciprocity on top of the basic gift exchange.

I observe a strong pattern of gift exchange. Up to a level, wage increases are profitable to the employer as a higher wage incites increased effort. Effort plateaus after this level, leading to decreased employer earnings. I call this profit-maximizing wage level the objectively fair wage. This point of asymmetry is important in understanding the results of this study. I also observe that the gift exchange pattern is stronger here than in the setting of the previous chapter. Although the initial offer has no direct impact on the payoffs, it has effects on effort are observable. I interpret is this result as positive reciprocity from keeping the promise of the wage offer.

Regarding the wage cuts, I find that cuts do not have strong effects on their own (in the absence of shocks). The workers are relatively rational in the sense that they are strongly affected by the final wage rather than the wage adjustment. In other words, the shape of the effort-wage curve holds largely even after wage cuts. When wage cuts follow shocks, I find that they lead to significant reductions in effort. This is the case for both employer and worker shocks, which suggests that a desire to sustain the status quo is a stronger reference point for wages than splitting the surplus (and the shock) evenly. Hannan (2005) finds similar results when considering (negative and positive) employer shocks. Hannan (2005) also finds that the reactions to a positive profit shock are sharper in magnitude than the reactions to a negative profit shock, suggesting that cuts are more acceptable when the firm is hit by a negative shock. Note, however, that the reference point in the study is a positive profit shock; in my study, the reference point is a no shock outcome.

I also observe that employers cut wages more frequently than expected, and on average offer wages below the fair wage level. Employers cut wages more frequently after they have experienced a shock than when no shock occurs. Wage cuts are least frequent when

[3] The market makes the opportunity costs (alternative offers) more salient, which in turn may increase sensitivity towards wages also early on in the experiment. For example, Greiner et al. (2011) find that peer comparisons are important in invoking the gift exchange pattern in their setting, where individuals experience two different wage conditions. If individuals are not aware of the wage differences with their peers, they do not respond to wage changes in effort.

workers have experienced a negative shock. Similarly, wage increases are most frequent following a worker shock and least frequent after an employer shock. The changes in the frequencies of adjustment due to shocks are not, however, large enough to significantly impact the average realized wages per shock outcome. For wages below and up until the objectively fair wage level, wage cuts are unprofitable.

Finally, the shocks on their own, absent a wage cut, do not seem to have a significant effect on effort. This means that wage cuts through indirect mechanisms, so called real wage cuts, do not lead to significant changes in effort (or work morale).

The chapter continues by explaining the theory in Section 2.2. This is followed by Section 3 on the experimental design, and Section 4 on the results. Section 5 concludes.

2.2 Theory

In this section, I set up a model for gift exchange based on other-regarding preferences and negative reciprocity, building on Charness and Rabin (2002). The basic setup is a one-shot, two-player gift exchange game between an employer and a worker. The game consists of the following stages. Workers are first hired in a market with non-binding wage offers. After hiring, potential shocks are realized and employers may adjust the wage by confirming the final wage for the round. Then, knowing the realization of shocks and how the wage was adjusted, the workers chooses a (costly) effort level that determines the employer's payoff. In the experiment that will be used to test my theoretical predictions, the worker and the employer are linked in the first stage and stay together until the end of the process. The equilibrium relevant wage is the wage after potential adjustments. A minimum wage level applies; it has been normalized here to zero.

To summarize the game:

1. The employer hires a worker with a non-binding wage offer, $w_0 \geq 0$

2. Potential shocks are announced. These shocks affect payoffs.

3. The employer decides the binding 'final' wage, $w_1 \geq 0$.

4. Observing the potential shock, the initial wage offer w_0, and the final wage w_1, the worker chooses effort $e \geq 0$, which is non-contractible.

I start by analyzing a model without shocks; this highlights how gift exchange arises with other regarding-preferences and reciprocity. I then explain how shocks affect the gift exchange. This is followed by a discussion on the effects of nominal (net) wage illusion and of the ways in which negative reciprocity maybe triggered.

2.2.1 A Model of Gift Exchange

Following the logic of backward induction, I first consider how workers respond with effort to the final wage and the potential wage adjustment. The wage offer and the final wage are both independent of the effort. I then model how employers set the final wage and make their initial wage offers, given the workers' best response function. At this point, I do not yet consider shocks.

Worker's Effort Choice

Utility. The worker's utility, denoted by u^W, is captured by the expression:

$$u^W = (1 - \beta(e))w_1 + \beta(e)(f(e) - w_1) - c(e). \tag{2.2.1}$$

Utility thus depends on the worker's monetary payoff (final wage, w_1); the (utility) costs of exerting effort, $c(e)$, and a social preference term reflecting the employer's net earnings. Employers' net earnings consist of the (monetary) benefits that the worker's effort generates, depicted by $f(e)$, minus the final wage. The function $f(e)$ captures worker productivity, which depends on the effort that she exerts.

As in the previous chapter, the other-regarding preferences, $\beta(e)$, are represented by a function derived from the Charness and Rabin (2002) model. This captures several types of social preferences, including inequity aversion, competitive preferences, and reciprocity. The basic idea is that individuals can react to relative payoff differences in different ways. For simplicity, I consider differences in monetary earnings only, and not in overall utility, as the former is readily available and comparable, while the latter is to a large extent unobservable.[4] When the worker is earning more than the employer, she assigns weight ρ to the earnings of the employer in their utility function. When the worker is earning less, she assigns weight σ. It is common to assume that individuals dislike disadvantageous inequality more than they dislike advantageous inequality (Fehr and Schmidt, 1999). This can be captured by a simple parametric assumption: $\rho > \sigma$. I furthermore assume that $\rho > 0$, to ensure that there are positive feelings towards the employer at least when the worker is better off, that is, when $w_1 > \frac{f(e)}{2}$. When the payoffs are equal ($w_1 = \frac{f(e)}{2}$), the weight β is assumed equal to zero. This does not mean that the employer's income plays no role; as long as the earnings remain equal, changes in one's own payoff are perfectly aligned with changes in the employer's and β becomes irrelevant. Of course, as soon as a change causes differences in the earnings, the worker will attribute a non-zero weight to the employer's earnings. The previous chapter shows that other-regarding preferences without negative reciprocity (that is, a model with only the other-regarding preferences

[4] See fn 5 in the previous chapter for more discussion.

σ and ρ) are enough to facilitate gift exchange.

Finally, workers are allowed to react to *perceived misbehavior*: if the employer acts unkindly, the worker can respond by 'punishing' the employer. In the context of this model, I define an unkind action as setting an actual wage w_1 lower than the original agreement w_0. Note that w_0 has no actual effect on payoffs; hence, the effect of w_0 on behavior is purely psychological. Workers punish the employers by reducing the weight that they give to the employers' earnings by a parameter $\theta \geq 0$. As mentioned before, when the worker is equalizing payoffs, β loses significance and hence it may still be set to 0. Otherwise, effort is lowered per wage level. If no wage cut happens, $w_1 \geq w_0$, then $\theta = 0$. β is, therefore, given by:

$$\beta(e) = \begin{cases} \sigma, & \text{if } w_1 < \frac{f(e)}{2} \wedge w_1 \geq w_0 \\ 0, & \text{if } w_1 = \frac{f(e)}{2} \wedge w_1 \geq w_0 \\ \rho, & \text{if } w_1 > \frac{f(e)}{2} \wedge w_1 \geq w_0 \\ \sigma - \theta, & \text{if } w_1 < \frac{f(e)}{2} \wedge w_1 < w_0 \\ 0, & \text{if } w_1 = \frac{f(e)}{2} \wedge w_1 < w_0 \\ \rho - \theta, & \text{if } w_1 > \frac{f(e)}{2} \wedge w_1 < w_0. \end{cases} \tag{2.2.2}$$

I make the following functional assumptions. The costs from effort are assumed to be strictly convex in effort, $c'(e) > 0$ and $c''(e) > 0$. In addition, I assume $c(0) = 0$. The benefit that effort generates is assumed in turn to be a concave function of effort, $f'(e) > 0$ and $f''(e) \leq 0$, while no effort means no benefits, $f(0) = 0$. To ensure that a positive level of effort is efficient, I assume $\lim_{x\downarrow 0} f'(0) > \lim_{x\downarrow 0} c'(0)$.

Best Response. As in the previous chapter, a worker maximizes u^W in eq. (2.2.1), that is, for any given w_1 she chooses e such that

$$\frac{c'(e)}{f'(e)} = \beta(e). \tag{2.2.3}$$

The best response of a worker, $\hat{e}$, thus depends on her social preferences. Note that $\hat{e}$ varies with w_1 and w_0 because $\beta(e)$ depends on w_1 and w_0 (eq. (2.2.2)). Suppose first that $w_1 \geq w_0$. Denote by $\hat{e}_\sigma$ the solution to eq. (2.2.3) for $\beta(e) = \sigma$, and by $\hat{e}_\rho$ the solution for $\beta(e) = \rho$. For $\sigma < 0$, the solutions is a corner, $\hat{e}_\sigma = 0$. Beyond this corner, the solution is increasing in β because $\partial(\frac{c'(e)}{f'(e)})/\partial e > 0$. Thus, $\sigma < \rho$ together with $\rho > 0$ implies that $\hat{e}_\sigma < \hat{e}_\rho$; that is, optimal effort is lower with disadvantageous inequality than with advantageous inequality. Denote by $\hat{e}_0(w_1)$ the effort level that equalizes earnings between worker and employer; this is implicitly defined by $w_1 = \frac{f(\hat{e}_0)}{2}$.[5] Last, suppose that the wage has been cut, $w_1 < w_0$. Then, there are reductions in effort compared to the no cut case such that $\hat{e}_{\sigma-\theta} < \hat{e}_\sigma$ and $\hat{e}_{\rho-\theta} < \hat{e}_\rho$. Note that if $\rho - \theta > 0$, there exists a wage range for w_1, for which the worker will equalize earnings, even after a wage cut.[6] Otherwise, no gift exchange will take place after a wage cut and the optimal wage and effort are set at their minimum levels. From here onward, I assume that $\rho - \theta > 0$. For expositional purposes, I also assume that $\rho - \theta > \sigma$, but allowing otherwise does not affect my conclusions. As the maximum and minimum effort levels move down with the triggering of the negative reciprocity, so does the range of wages at which equalizing happens.

Result 1. The worker's best response function is given by

$$\hat{e}(w_1, w_0) = \begin{cases} \hat{e}_\sigma, & \text{if } w_1 < \frac{f(\hat{e}_\sigma)}{2} \quad \wedge\, w_1 \geq w_0 \\ \hat{e}_0(w_1), & \text{if } \frac{f(\hat{e}_\sigma)}{2} \leq w_1 \leq \frac{f(\hat{e}_\rho)}{2} \wedge\, w_1 \geq w_0 \\ \hat{e}_\rho, & \text{if } w_1 > \frac{f(\hat{e}_\rho)}{2} \quad \wedge\, w_1 \geq w_0 \\ \hat{e}_{\sigma-\theta}, & \text{if } w_1 < \frac{f(\hat{e}_\sigma)}{2} \quad \wedge\, w_1 < w_0 \\ \hat{e}_0(w_1), & \text{if } \frac{f(\hat{e}_{\sigma-\theta})}{2} \leq w_1 \leq \frac{f(\hat{e}_{\rho-\theta})}{2} \wedge\, w_1 < w_0 \\ \hat{e}_{\rho-\theta}, & \text{if } w_1 < \frac{f(\hat{e}_\rho)}{2} \quad \wedge\, w_1 < w_0. \end{cases} \tag{2.2.4}$$

As in the previous chapter, eq. (2.2.4) implies that effort is non-decreasing in wage, that there is a range of wages for which workers choose to equalizes earnings, and that the model captures the fair wage hypothesis of Akerlof and Yellen (1990). Figure 2.2.1 illustrates this best response function and how it reacts to the wage cuts.[7] The solid line represents the worker's response when no wage cut has happened and the dashed line represents the response after a wage cut. The two lines overlap in the middle of the part where payoffs are equalized in both cases. Effort per final wage is always weakly higher without a wage cut than with one.

Employer's Wage Setting

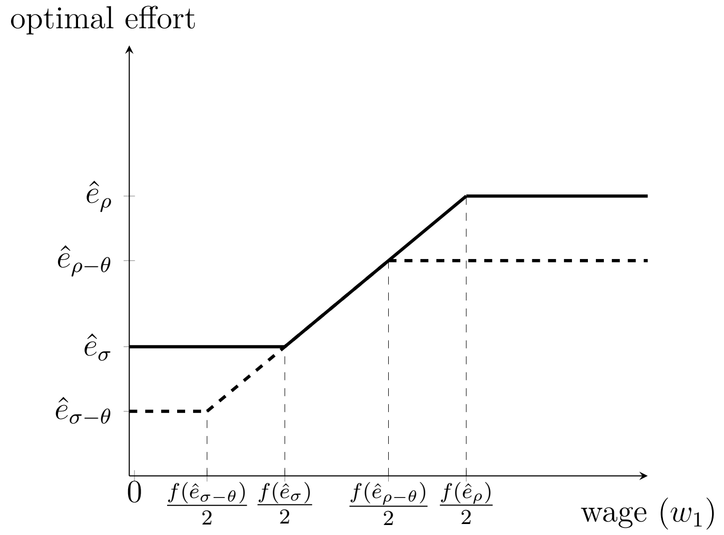
Utility. Employers first make a non-binding wage offer, w_0, with which the worker is hired and only later decide on the final binding wage w_1. Their utility, denoted by u^F, is

[5] To avoid further corner solutions, I assume that there exists an $\hat{e}_0$ for which this equality holds. For ease of notation, I further assume that $\sigma < \frac{c'(\hat{e}_0(w_1))}{f'(\hat{e}_0(w_1))} < \rho, \forall w_1$. This assures that $\hat{e}_\sigma < \hat{e}_0(w_1) < \hat{e}_\rho, \forall w_1$, thus avoiding cumbersome notations.

[6] This follows from a comparison to the situation in the previous chapter. Assume that in the previous chapter the worker's preferences are characterized by $\rho' = \rho - \theta$ and that w_1 is the original wage (which cannot be altered). If $\rho - \theta > 0$, the analysis of the previous chapter can be directly applied.

[7] For presentational purposes, $f(e)$ is again assumed to be linear (cf. previous chapter). A non-linear $f(e)$ would add curvature to the intermediate segment of the best response function.

Figure 2.2.1: Worker's response curve e(w) as a function of wage



Notes: The optimal effort (vertical axis) is shown as a function of the final wage (horizontal axis). $\hat{e}_\rho$ ($\hat{e}_\sigma$) depicts the solution to the first order condition (2.2.4) in case the worker faces (dis)advantageous inequality. The dashed line presents the optimal effort levels after a wage cut. $\hat{e}_{\rho-\theta}$ ($\hat{e}_{\sigma-\theta}$) depicts the solution to the first order condition (2.2.4) in case the worker faces (dis)advantageous inequality and perceives misbehavior. In this example, $\sigma > 0$.

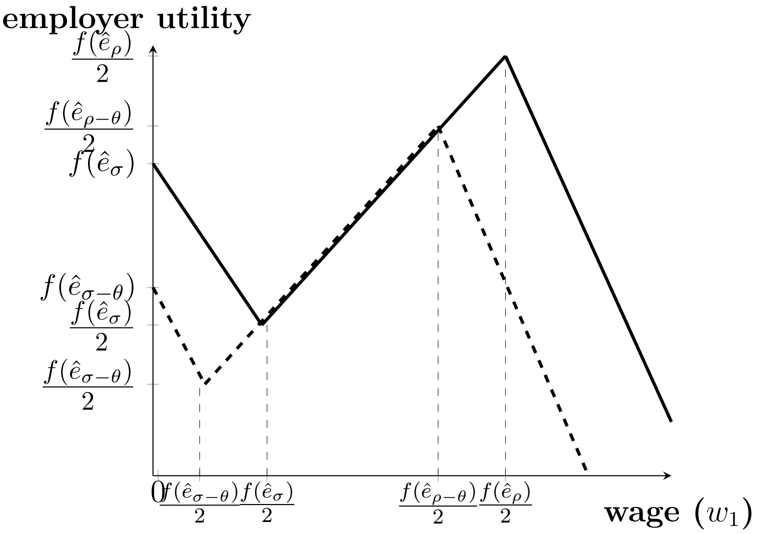
assumed to be given by

$$u^F = (1-\alpha)(E[f(e(w_1, w_0))] - w_1) + \alpha w_1. \tag{2.2.5}$$

The utility thus consists of the expected monetary earnings (expected revenue $E[f(e(w_1, w_0))]$ minus the final wage) plus a social preference term reflecting concern for the worker and his or her final wage (weighted by α).[8] If I assume that the employer cares more for the own monetary earnings than those of the worker, $\alpha < 0.5$, the other-regarding preferences can be set to $\alpha = 0$ without loss of generality. As explained in the previous chapter, this is due to the fact that other-regarding preferences do not affect profit maximizing behavior in this model as long as $\alpha < 0.5$. In a Subgame Perfect Equilibrium (SPE), employers expect the workers to best respond to the wage offer and the final wage, that is, $E[f(e(w_1, w_0))]$ is determined by eq. (2.2.4). In other words, $E[f(e(w_1, w_0))] = f(\hat{e}(w_1, w_0))$. Note that the wage offer w_0 enters into the employer's utility function only through the worker's effort response.

[8] As with the worker, I assume that the employer's other-regarding preferences are fully based on monetary earnings. The employer does not take into account the worker's other-regarding preferences or her effort costs.

Figure 2.2.2: Employer's utility as a function of wage



Notes: The employer's utility is shown as a function of the wage (horizontal axis), assuming $\alpha < 0.5$. $\hat{e}_\sigma$ ($\hat{e}_\rho$) depicts the solution to the first order condition (2.2.4) in case the worker faces (dis)advantageous inequality. $\hat{e}_{\rho-\theta}$ ($\hat{e}_{\sigma-\theta}$) depicts the solution to the first order condition (2.2.4) in case the worker faces (dis)advantageous inequality and the employer cuts the initial wage offer. The dashed line represents employer's utility after a wage cut.

Optimal Wage Setting. Figure 2.2.2 depicts how the employer earnings, given by $f(\hat{e}(w_1, w_0)) - w_1$, vary with the final wage and whether wages were cut in the SPE. The solid lines represent the case where the wage is not cut and the dashed lines show the case after the wage cut. At low wages (until $\frac{f(\hat{e}_\sigma)}{2}$ or $\frac{f(\hat{e}_{\sigma-\theta})}{2}$), increases to wages do not affect the worker's chosen effort level (which stays at the low $\hat{e}_\sigma$ or $\hat{e}_{\sigma-\theta}$), so the employer's earnings drop linearly in w_1.[9] In this section of the wage curve, a final wage of zero gives the local maximum of the employer's utility. When there is gift exchange, that is, when wages are not cut or when gift exchange can be sustained after a cut, then, in the intermediate range between $\frac{f(\hat{e}_\sigma)}{2}$ and $\frac{f(\hat{e}_\rho)}{2}$ (or $\frac{f(\hat{e}_{\sigma-\theta})}{2}$ and $\frac{f(\hat{e}_{\rho-\theta})}{2}$ after a wage cut), wage increases lead to higher effort by workers and rising profits for the employers. The revenue is maximized locally at $\frac{f(\hat{e}_\rho)}{2}$ at the final wage of $\frac{f(\hat{e}_\rho)}{2}$ in the no wage cut case, and at $\frac{f(\hat{e}_{\rho-\theta})}{2}$ and the final wage of $\frac{f(\hat{e}_{\rho-\theta})}{2}$ in the case wages are cut, conditional on sustained gift exchange. Beyond these points, employers' earnings start to drop because the workers are providing effort at their maximum levels and any further wage increases no longer increase effort and only reduce the employer's earnings.

Whether the global maximum for the employer is at the zero wages or at $\frac{f(\hat{e}_\rho)}{2}$ depends on the parameters σ and ρ, or $\frac{f(\hat{e}_\rho)}{2}$ and $f(\hat{e}_\sigma)$. If gift exchange is profitable, $\frac{f(\hat{e}_\rho)}{2} > f(\hat{e}_\sigma)$, then an employer cannot reach the maximum utility with a wage cut. In general, the utility

[9] If $\sigma < 0$ then $f(\hat{e}_\sigma) = 0$; if $\sigma - \theta < 0$, $f(\hat{e}_{\sigma-\theta}) = 0$.

without a wage cut is always at least as good as the utility after a wage cut, making wage cuts always suboptimal. If the employer cuts wages, employer utility is maximized at $\frac{f(\hat{e}_{\rho-\theta})}{2}$ with $w_1 = \frac{f(\hat{e}_{\rho-\theta})}{2}$, conditional on $\frac{f(\hat{e}_{\rho-\theta})}{2} > f(\hat{e}_{\sigma-\theta})$ being true. Otherwise, utility is maximized at zero wage.

Result 2. The utility maximizing wage for an employer is

$$\hat{w}_1 = \hat{w}_0 = \begin{cases} 0, & \text{if } \frac{f(\hat{e}_\rho)}{2} < f(\hat{e}_\sigma) \\ \frac{f(\hat{e}_\rho)}{2}, & \text{if } f(\hat{e}_\sigma) \leq \frac{f(\hat{e}_\rho)}{2}, \end{cases} \tag{2.2.6}$$

where I assume that an employer chooses the higher wage whenever indifferent. Recall from the previous chapter that the wage $w_1 = \frac{f(\hat{e}_\rho)}{2}$ is called the objectively fair wage.

Full information

The purpose of the model is to illustrate the interaction between employers and workers. The model assumes complete information, meaning employers and workers know the parameter values before they make their decisions. I further assume, for simplicity, that these parameters do not vary between employers and workers.

2.2.2 The Impact of Shocks

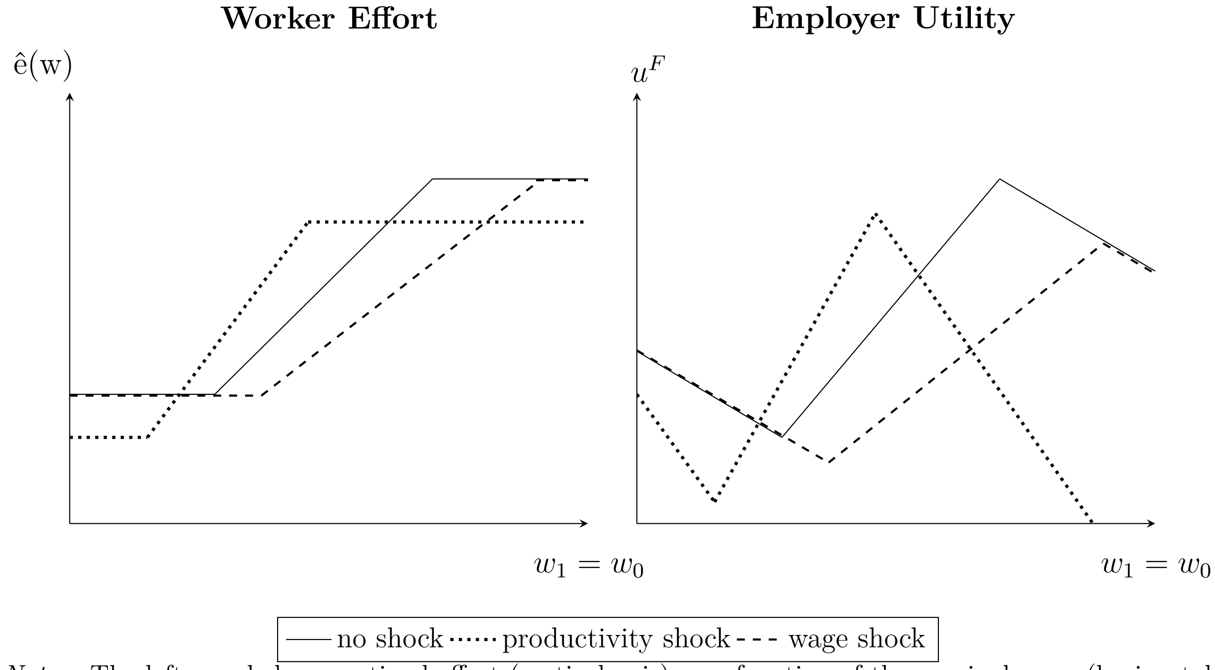
The previous section shows that wage cuts come with a substantial risk to the employer, making it suboptimal for employers to cut wages. To provide more rationale for wage cuts, negative shocks are added to the market. The effects of these shocks are discussed in detail in the previous chapter (Appendix A). Here, I summarize the findings derived there. In the next section, I discuss how shocks may interact with wage cuts.

The shocks occur randomly with a known probability. They reduce monetary earnings, affecting at a time either workers' wage earnings or employers revenue derived from workers' effort. There are three potential shock outcomes:

- *No shock occurs.* The market works as explained above.
- *Wage shock*: the shock reduces workers' wage earnings w_1 by a proportion $0 < \delta < 1$. Employers' earnings are unaffected.
- *Productivity shock*: the shock reduces employers' earnings $f(e)$ by $0 < \delta < 1$. Workers' earnings are unaffected.

Figure 2.2.3 illustrates how the shocks affect worker's best response function (left panel) and employer's utility (right panel). For presentational purposes, we again assume a linear $f(e)$ (cf. fn. 7), and that no wage cut has occurred.

Figure 2.2.3: The Effects of Shocks



Notes: The left panel shows optimal effort (vertical axis) as a function of the nominal wage (horizontal axis). The right panel shows employer's utility as a function of the nominal wage (horizontal axis).

In essence, the shocks change the wage that maximizes the employer's utility and can, therefore, provide a rationale for the employer to adjust wages. The productivity shock reduces the employers' earnings and moves the optimal wage downwards. This is driven by two factors. First, as employers become relatively worse off, workers need to provide more effort per wage to equalize payoffs whenever that is the appropriate strategy. Second, as the marginal benefit from each unit of effort drops, the equilibrium effort level drops correspondingly. The new utility maximum for employers is at a lower level and achieved with a lower wage. Similarly, a wage tax makes workers relatively worse off, meaning that in order to reach the same outcome in terms of effort, employers need to pay a higher wage. The SPE gives a lower maximum utility level that is achieved with a higher final wage.

Given how the shocks move the optimal wage of the employers, one can venture that wage cuts become more acceptable with shocks, in particular, if it is considered fair that the burden of the shocks is shared. One way to operationalize this idea is to assume that negative reciprocity is not triggered by wage adjustments that aim to equalize payoffs at the new subgame perfect equilibrium, that is $\theta = 0$ is the equilibrium response if a wage cut occurs after a shock. This leads to the following hypotheses:

Hypothesis 2: Wage cuts become more acceptable after a productivity shock has occurred.

H2a: Employers cut wages more often.

H2b: Workers do not punish wage cuts.

In a similar vein, the wage shock moves the optimal wage upwards for the employers. Hence the expectation is that wage cuts become even less acceptable after a wage shock. This hypothesis could also be formulated differently: that wage increases become the expectation and hence wage increases become more common with the wage shocks.

Hypothesis 3: Wage increase becomes the expectation after a wage shock has occurred.

H3a: Employers increase wages more often.

H3b: Workers punish unadjusted wages.

Note that Hypothesis H3b considers the case of a real wage cut when the nominal wage is kept intact. The basic model predicts that this leads to a drop in effort: as workers pursue equal payoffs, they respond to a shock that reduces their own payoffs by equally reducing the employers' payoffs.

To study Hypotheses 2 and 3, I observe the behavior of both employers and workers. It is particularly interesting to see if their conceptions of fairness revealed through behavioral responses coincide.

2.2.3 Shocks and the Norm of the *Status Quo*

The hypotheses above are derived with inequity aversion in mind: the point of departure is that the burden of a shock is shared evenly between the parties. Equity is not, however, the only conceivable way to "fairly" allocate the burden of the shocks. An alternative benchmark is to maintain the status quo. That would mean that is it deemed acceptable to not increase wages after a wage shock, while it is considered unacceptable to decrease wages after a productivity shock.

Alternative Hypothesis 2: Wage cuts do not become more acceptable after a productivity shock has occurred.

AH2a: Employers do not cut wages more often.

AH2b: Workers punish wage cuts after a productivity shock.

Alternative Hypothesis 3: Wage increases do not become the expectation after a wage shock has occurred.

AH3a: Employers do not increase wages more often.

AH3b: Workers do not punish unadjusted wages.

Keeping the status quo might be a norm that extends beyond just wage cuts. Status quo may offer guidance on how to behave also after a shock: continue as before, in both wages and effort. In this case, the burden of the shock is primarily expected to be carried by the party that nominally receives the shock. The shocks are exogenous and there is nothing that the workers or employers can do to change the shocks' frequency or impact. This fact may suggest no party is particularly responsible – suggesting sharing the burden equally – but one could also argue that the 'other party' is not responsible for the shock, for which reason, it might be a reasonable expectation that the shock is primarily carried by the party that receives it.

Note that Alternative Hypothesis AH3b is in line with the *nominal illusion* hypothesis. A real wage shock that makes workers worse off but keeps the nominal wage intact is not expected to lead to an effort change under the status-quo norm.

2.3 Experimental Design and Procedures

2.3.1 Design

The experimental design builds on the experiment of Fehr et al. (1993) and is closely related to the design used in the previous chapter. Workers are hired in a one-sided auction where employers make competing wage offers. However, in contrast to the design used in the previous chapter, this wage at which the worker gets hired is 'cheap talk' – once hiring has taken place, the employers may change the wage *at will* to any other level. I maintain the real effort task used in the previous chapter.

The experiment is framed as a labor market, and participants keep their roles as a worker or a employer throughout the experiment. Shocks are framed as temporary, one-period taxes. The experiment consists of eight rounds. Each round consists of the following stages, each of which is elaborated below.

1. Employers hire workers in an auction
2. If shocks are possible, the common tax scheme (or the lack thereof) is announced
3. Employers confirm the wage offer or adjust the wage – the final wage is communicated to the worker
4. Workers conduct a real effort task
5. Payoffs are determined and reported

Hiring happens in real time, via a one-sided auction. Employers post wage offers between 30 and 100 points, in steps of 5, on a public platform observable to all employers and workers in the market. Offers can be updated while not yet accepted. Once a worker accepts an offer, the offer is removed from the platform and the worker is hired by the employer in question. The market consists of seven workers and 5 employers and each participant can have only one hiring contract per round. As a consequence, at least two workers are unemployed each round. The hiring stage lasts at most 2 minutes and finishes as soon as all five employers have hired a worker. After the auction, anonymized information is provided to all market participants about the wage offers at which workers have been hired and the number of hired workers.

In some rounds, shocks might be implemented. If this is the case, these shocks are framed as ‘taxes’, and they are announced right after the hiring stage. The taxes impact participants’ earnings. I implement two kinds of taxes: 1) a wage tax; this reduces the final wage that the worker receives from the employer by 20% and 2) a productivity tax; this reduces the revenue that the employer receives from the hired worker’s effort by 20% (each correctly solved task rewards the employer 16 points instead of the usual 20 points). The taxes are discussed in more detail below, when I discuss the payoffs and the treatments. The tax revenues are not returned to the participants in any way; proceedings are returned to the experimenter.

Once the (potential) shocks have been announced, the employers must either confirm the earlier wage offer or adjust it. Adjusting the wage is a possibility regardless of whether or not a shock occurred. The wage can thus be changed at will to any wage level within the initial limits, that is, between 30 and 100 points, in intervals of 5 points. The final wage is then communicated to the hired worker before they start to work on the real effort task.

To summarize, each worker knows the initial wage offer, the final wage, and what kind of shock, if any, has occurred before they start on the real effort task. Each worker has five minutes to work. The task (introduced by Weber and Schram, 2017) has two 10x10 matrices appearing on the computer monitor. Each matrix cell contains a two-digit number. The worker needs to find the highest number in each matrix and add them up. A correct answer yields a reward of 20 points for the employer (which may be subject to a tax). Regardless of whether the answer is correct or incorrect, a new set of matrices appear as a new task. To discourage guessing, the number of tasks that can be attempted is limited to ten.

In the end of each round, each participant learns their potential earnings for that round (at the end of the experiment, two out of the eight rounds are paid at random), as well as the potential earnings of the partner that they have been linked with. Furthermore, each worker-employer pair will learn the worker’s results in the real effort task, namely, the number of correct tasks and the number of tasks attempted.

Before discussing how the payoffs are constructed, it is important to distinguish between tax **treatments** (*tax environments*) and tax **outcomes**. Throughout this chapter, I indicate tax treatments by capital letters; they define which tax shocks (outcomes) are possible. Tax outcomes are realized per round; I indicate these with lower case letters. Table 2.3.1 summarizes all possible cases.

Table 2.3.1: Treatments and outcomes

Treatment	NT	ET	WT	AT
possible tax outcomes	nt	nt, et	nt, wt	nt, et, wt

Notes: NT/nt = 'no tax'; ET/et = 'productivity tax'; WT/wt = 'wage tax'; AT = 'all taxes'.

The experiment consist of four treatments that are varied between subjects. These differ in the type of tax that *might* occur. The four treatment options are 1) no tax (denoted by NT), 2) productivity tax (ET, for 'Employer Tax'), 3) wage tax (WT) and 4) employer or wage tax (AT, for 'All Taxes'). In treatments where taxes are possible, they happen with a probability equal to $\frac{1}{3}$. When both taxes are possible, each tax is equally likely but they cannot occur simultaneously. All of this is common knowledge. The sequence of taxes was drawn randomly beforehand and was fixed in order for all sessions to have a directly comparable history.[10]

Payoffs depend on the hiring status and the tax outcome and are summarized in Table 2.3.2. If an employer hires a worker, the employer receives 40 points and all of the revenue from the task but must pay the worker's wage from this income. An employer's payoff can thus be negative in a round. A worker's payoff consists entirely of the wage. Keep in mind that the payoff-relevant wage is not the wage at which the worker is hired, the so called 'wage offer', but the *final wage* that the employer determines after the hiring stage and the shock announcement. If unmatched, employers earn nothing and unemployed workers receive an unemployment benefit of 20 points, regardless of the tax outcome. When taxes apply, they directly affect only one side, either the employer or the worker. The productivity tax is collected from the revenue that the employer receives, which means that when taxed, instead of the usual 20 points, the employer receives only 16 points for each task correctly completed by the worker. When wage tax applies, the workers receive only 80% of the wages paid by their employer.

Table 2.3.2: Payoffs

	employer payoff	worker payoff
no tax (nt)	$40 - w + 20 * e$	w
productivity tax (et)	$40 - w + 16 * e$	w
wage tax (wt)	$40 - w + 20 * e$	$0.8 * w$
outside option (no contract)	0	20

Notes: Cells show payoffs in points for employers and workers, depending on the outcome of the tax shock. w refers to final wage.

At the end of the experiment, two rounds are randomly selected for payment.[11] The

[10] The shocks occur in rounds 2, 4, and 5. In AT, half of the sessions had a productivity tax in round 2 and a wage tax in rounds 4 and 5; the remaining sessions had the reverse.

[11] In some of the early sessions, due to computational errors the incentive scheme rewarded three rounds

exchange rate used is one euro for every ten points earned in those two rounds.

2.3.2 Procedures

The experiment was run at the BLESS laboratory of the University of Bologna, in 2017 - 2018. Participants were primarily students and recruited using ORSEE (Greiner, 2004). The experimental software was programmed in oTree (Chen et al., 2016). I had 312 participants in 13 sessions. Each session had two groups (each consisting of five employers and seven workers). Average earnings (including a five euro show up fee) were 14 euros.

Reading the instructions and getting familiar with the software took approximately 20 minutes and the main experiment lasted about one hour. A translation of the instructions is presented in Appendix 2.B. During the software tutorial, the participants did the real effort task for five minutes to get acquainted with it. At the end of the instructions, the participants had to take a comprehension test (cf. Appendix 2.B).

2.4 Results

I have data for 130 employers, 181 workers, and 967 employer-worker matchings. The data structure is such that these matchings may include up to eight rounds of observation for each worker and employer, although an observation may consists also of not having a contract at all for a round. To correct for multiple observations, I will use either random effects estimations to correct for individual effects (primarily with the workers), cluster standard errors at group level, or use the average observation over the rounds as the unit of observation, in particular, with the employers. In particular, I will mainly aggregate over the employer observations because they cannot select out from a round to the same extent as the workers can. This gives us 30 observations each for NT, ET, and WT, and 40 for AT, though not every employer has an observation in every round. Unless indicated otherwise, test results are based on non-parametric permutation t-tests (cf. Schram et al., 2018), here referred to as PtT.

2.4.1 Wage Offers, Final Wages, Adjustments, and Effort over Time

Figure 2.4.1 depicts the development of wage offers, final wages, wage adjustments and effort over the eight rounds for each of the treatments NT, ET, WT, and AT. In all treatments, the average initial wage offer is higher than the average final wage, meaning

instead of two (which was only known to the participants ex post) and a shock occurred in fewer rounds than intended (which is not expected to affect choices because the occurrence of a shock is common knowledge before any decision is made). Two sessions ended at round 6.

that employers frequently adjust wages downwards. In spite of these adjustments, the final wage stays clearly above the minimal wage of 30 in all treatments, averaging between 40 and 55 points. Most wage offers are accepted immediately, for which reason one can interpret that the wage setting is mostly driven by employers' behavior.[12] Note that in *NT*, final wages and effort both drop steadily across the rounds. Moreover, wage adjustments in *NT* are small and stable. It appears that in an environment where no shocks are possible, market wages are rarely altered, but do decrease over time, with a corresponding decrease in effort. When shocks are possible, it is more difficult to discern at this aggregate level the patterns in the development of wages and effort, though it is noteworthy that market wages in WT are adjusted less as the rounds proceed.[13]

Table 2.4.1 explores if final wages respond to the occurrence of negative shocks. To account for the fact that each employer has multiple observations, I use the average observation of the employer per each shock outcome. I find that final wages are rigid: they do not systematically react to the occurrence of the tax shocks in any of the treatments, nor when the data are pooled across treatments.

Pooling across the tax outcomes, the average final wages per treatment are the following: *NT*: 42.3 (sd 10.3), *ET*: 46.5 (sd 14.0), *WT*: 49.0 (sd 11.2), and *AT*: 44.9 (sd 12.7). The difference between *NT* and *ET* is insignificant ($p = 0.191$), as is the difference between *NT* and *AT* ($p = 0.368$), but the difference between *NT* and *WT* is significant ($p = 0.019$). This means that when workers may be hurt by a shock, they are compensated by higher wages than when no shock is possible.[14]

Of course, the final wages are a combination of the initial market wage and the wage adjustment. We therefore consider whether employers adjust wages in reaction to shocks. To start, Table 2.4.2 reports the average wage adjustments by treatment and shock outcome. I find that the average wage cut is largest when the shock hits productivity, that is, with *et* shocks. The average cut in *ET* is 1.9 points and this is marginally significant ($p = 0.069$). The cuts are larger, 3.3 points, and highly significant in *AT* ($p < 0.001$). The difference is significant also in the pooled data.[15] Wage adjustments are slightly smaller

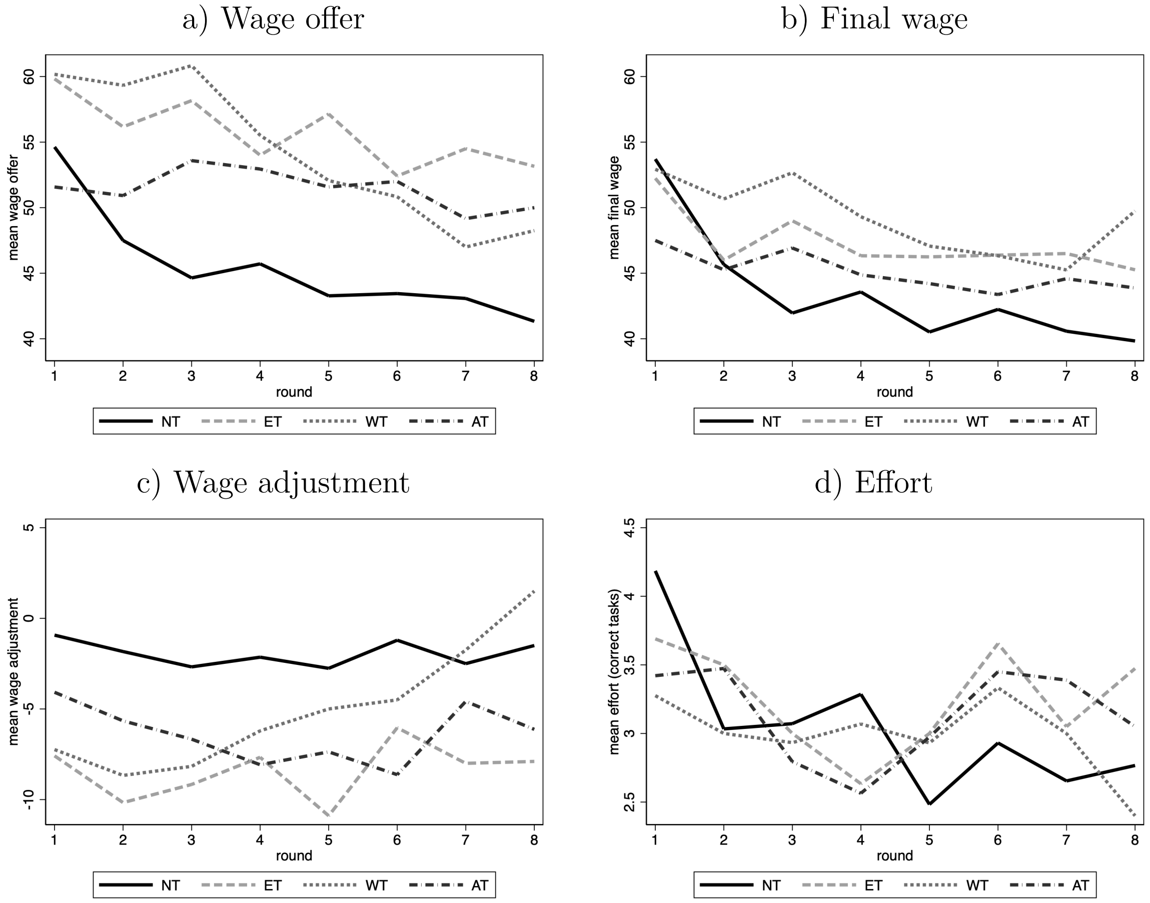
[12] The average number of offers made per round is 1.07 for *NT*, $n = 228$, 1.22 for *ET*, $n = 215$, 1.12 for *WT*, $n = 218$, and 1.17 for *AT*, $n = 309$. In *NT*, 94% of the first offers are accepted before they can be adjusted in the hiring market. For the other treatments, the numbers are 90% in *ET*, 91% in *WT*, 90% in *AT*.

[13] In the previous chapter, we observed across all treatments that employers lower the wage offers in the first few rounds, after which they stabilize. We interpret this as learning and therefore exclude these rounds from the statistical analysis. I do not observe similar patterns here and, therefore, I include all rounds in the subsequent analysis.

[14] Note, however, that when the final wage is set, employers and workers both know whether or not a shock has occurred. The compensation is therefore aimed at something that could have happened. Table 3 shows that final wages after a shock has occurred (*wt* in *WT*) is only slightly higher than when it could have, but did not occur (*nt* in *WT*).

[15] Note that the PtT tests are pairwise, which means that only a subgroup of the *nt* observations are relevant for the comparison, that is, only those who also experience the *et* outcome. For this group of

Figure 2.4.1: Average wage offers, final wages, wage adjustment, and effort over the rounds 1-8


a) Wage offer

b) Final wage

c) Wage adjustment

d) Effort

Notes: Lines show average realized wage offer, final wages, wage adjustments, and effort over the eight rounds of the experiment. The minimum wage is 30. *NT*: no taxes possible; *WT*: wage tax possible; *ET*: productivity tax possible; *AT*: both taxes possible. Tax shocks occurred in rounds 2, 4, and 5.

in most cases, but still negative, after a *wt* shock, however, none of these differences are significant.

To summarize, we do not observe significant variation in the final wages but we find some significant variation in the wage adjustment data. One explanation is that the wage offers are different by tax outcome (though the average wage offers do not differ significantly, see Table 2.A.1 in Appendix 2.A). Another explanation could be that there might be a difference in the average size of the adjustments. Conditional on there being a wage cut, its size is 18 points in *nt* (n=257), 20 points in *et* ($n = 85$), and 24 points in *wt* ($n = 38$). Conditional on there being a wage increase, its size is 9 points in *nt* ($n = 150$), 7 points in *et* ($n = 17$), and 10 points in *wt* ($n = 38$). These average cuts and increases are largely similar across shocks, and do not explain the pattern.[16] Rather, it would seem

people the mean adjustments are -6.9 in *nt* and -9.6 in *et*, $n = 70$.

[16] None of the differences is significant, except the difference in cuts between *wt* and *nt*: cuts are marginally larger in *wt* than in *nt*, $p = 0.066$, two-sided PtT-test. However, this does not explain the

Table 2.4.1: **Wages, treatments, and shocks**

tax outcome	*NT*	*ET*	*WT*	*AT*	**pooled**
nt	**43.7**	**48.1**	**49.5**	**45.3**	**46.6**
obs.	30	30	30	40	130
et		**45.9**		**43.9**	**44.8**
obs.		30		40	70
wt			**50.4**	**46.4**	**48.2**
obs.			30	38	68
PtT (p-values)					
6 **nt vs et**	-	*0.352*	-	*0.318*	*(0.140)*
nt vs wt	-	-	*0.181*	*0.506*	*(0.194)*

Notes: Tax shocks occurred in rounds 2, 4, and 5. The unit of observation is the mean wage paid by an employer across rounds. Paired tests between shock- and no-shock rounds are reported. Tests for the pooled data are conducted on the paired data, while unpaired averages shown on the table. Mean wages across employers are in bold. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *AT*: both taxes possible, 'pooled' combines treatments. PtT: permutation t-test.

Table 2.4.2: **Adjustments, treatments, and shocks**

tax outcome	*NT*	*ET*	*WT*	*AT*	**pooled**
nt	**-2.0**	**-7.7**	**-4.9**	**-6.3**	**-5.3**
	30	30	30	40	130
et		**-9.6**		**-9.6**	**-9.6**
obs.		30		40	70
wt			**-5.7**	**-3.4**	**-4.4**
obs.			30	38	68
PtT (p-values)					
nt vs et	-	*0.069*	-	*<0.001*	*(<0.001)*
nt vs wt	-	-	*0.528*	*0.481*	*(0.696)*

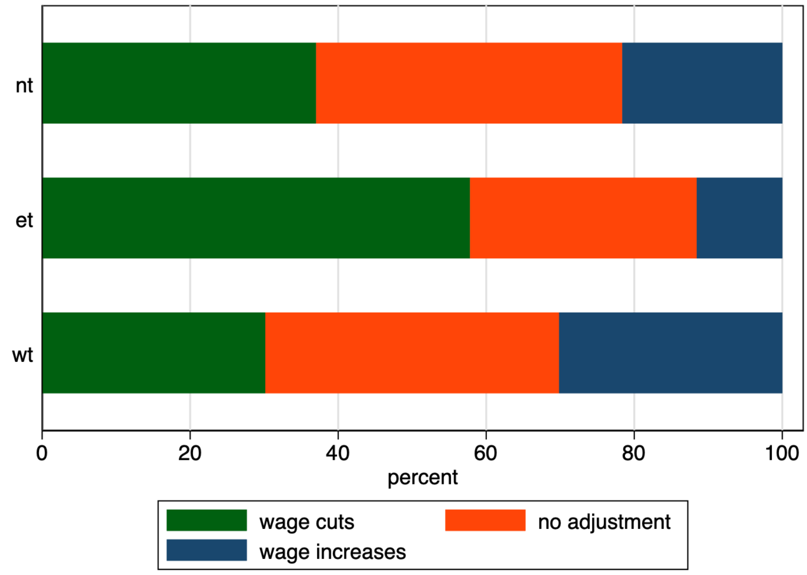
Notes: Tax shocks occurred in rounds 2, 4, and 5. The unit of observation is the mean adjustment by an employer across rounds. Paired tests between shock- and no-shock rounds are reported. Tests for the pooled data are conducted on the paired data, while unpaired averages shown on the table. Mean wages across employers are in bold. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *AT*: both taxes possible, 'pooled' combines treatments. PtT: permutation t-test.

that employers make cuts more frequently after *et* than *nt*.

To increase my understanding of when and how wages are adjusted, I categorize them as being either a wage cut, no adjustment, or a wage increase. Figure 2.4.2 illustrates the results. In *nt*, 37% of the wage offers are cut, 42% are not adjusted, and 22% are

observation about *et* shocks.

Figure 2.4.2: Percentage of wage cuts, no adjustments and wage increases by shock outcome



Notes: All data pooled. *nt*: no tax shock realized; *et*: productivity tax shock realized; *wt*: wage tax shock realized.

increased. In *et* (when the employer is hurt by the shock), I observe many more wage cuts; a majority of 58% of the wages are cut; 31% are not adjusted and only 12% are increased. In *wt*, the workers are hit by a shock and about 30% of the wages are cut, 40% are not adjusted, and 30% are increased.

To further study these patterns, I construct a variable $ADJ_{t,i}$ for the three wage adjustment types (cut, no adjustment, increase), where t is an indicator for the time and i is an indicator for the employer. I then estimate a multinominal logit regression of $ADJ_{t,i}$ on the shock outcomes (taxes). Table 2.4.3 reports the results of this regression as relative risk ratios.[17] The *no adjustment* outcome is set as the baseline, and the analysis are done separately for each treatment and jointly for the pooled data. In the pooled data, I find that the productivity shock on employers, *et*, increases the risk of a wage cut by doubling it, while the shock on workers, *wt*, has no significant effect on the likelihood of a wage cut – the risk ratio is very close to 1. On the other hand, wage increases are significantly more likely after a *wt* shock: the risk ratio is 1.78 in the pooled data, while *et* shocks have no significant effect on wage increases. Last, note that the magnitude of the effect in *et* is larger than that in *wt*: employers seem to more frequently do adjustments that are favorable to them than those that are costly.

To conclude, I firmly reject H1 (Employers do not cut wages). I find that wage cuts

[17] For multinominal logits, relative risk ratios are an intuitive way to report the results similar but not identical to odds ratios. The relative risk ratios are constructed with respect to the base outcome and account for the fact that other outcomes are possible.

Table 2.4.3: Risk ratios of wage cuts and wage increases by tax shocks and treatments

Relative risk ratios	*NT*	*ET*	*WT*	*AT*	**pooled**
wage cut					
nt (constant)	**1.04**	**1.23**	**0.75**	**0.72**	**1.04**
p-value	0.900	0.444	0.456	0.218	0.892
et	na	**1.80**	na	**2.16**	**2.02**
p-value	na	0.052	na	0.002	0.000
wt	na	na	**1.33**	**0.78**	**1.00**
p-value	na	na	0.424	0.089	0.990
no wage adjustment					
	(the base outcome)				
wage increase					
nt (constant)	**0.64**	**0.72**	**0.55**	**0.32**	**0.64**
p-value	0.255	0.111	0.175	0.005	0.222
et	na	**0.84**	na	**0.43**	**0.72**
p-value	na	0.371	na	0.321	0.172
wt	na	na	**1.91**	**1.64**	**1.78**
p-value	na	na	0.028	0.065	0.001
treatment FE	na	na	na	na	yes
Pseudo R2	0.000	0.013	0.007	0.028	0.024
obs.	227	215	217	308	967

Notes: The standard errors are robust, clustered at group level. *NT*: no taxes possible; *nt*: no tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *AT*: both taxes possible. 'pooled' combines treatments.

appear more frequently after a productivity shock hitting the employers and that wage increases become more frequent when there is a wage shock on the workers. Hence, I fail to reject hypotheses H2a and H3a in favor of the alternatives. This pattern, however, is not strong enough to significantly affect the average wages; this can be attributed to the high variance in initial wage offers and final wages.

Result 1: Employers cut wages frequently, even when no shocks have occurred.

Result 2: Wage cuts become more frequent when employers experience productivity shocks *et*.

Result 3: Wage increases become more frequent when workers experience wage shocks *wt*.

2.4.2 Effort and Gift Exchange

Next, consider how workers react in effort to the different tax shocks. I measure effort as the number of correct summations in the real effort task.[18] Table 2.4.4 summarizes the

[18] Measured this way, performance captures not only effort but also ability. However, as participants are randomly assigned to the treatments, so should ability be randomly assigned, and differences in

effort levels across the different treatments and shock outcomes. I find no significant differences in average effort levels across the different tax outcomes. Note that this averages effort across different wage levels, and although the wage differences across treatments are not significant, reporting effort this way might hide some interesting patterns in the data.

Table 2.4.4: **Effort, treatments, and shocks**

tax outcome	*NT*	*ET*	*WT*	*AT*	**pooled**
nt	**3.0**	**3.4**	**3.1**	**3.2**	**3.2**
obs.	30	30	30	40	130
et		**3.1**		**3.1**	**3.1**
obs.		30		40	70
wt			**3.0**	**3.1**	**3.1**
obs.			30	38	68
PtT (p-values)					
nt vs et	-	*0.154*	-	*0.482*	*(0.128)*
nt vs wt	-	-	*0.698*	*0.654*	*(0.556)*

Notes: Tax shocks occurred in rounds 2, 4, and 5. The unit of observation is the mean effort received by an employer across all rounds. The tests for the pooled data are conducted with the paired data , while unpaired means are shown in the table. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; *AT*: both taxes possible. 'pooled' combines treatments. PtT: permutation t-test.

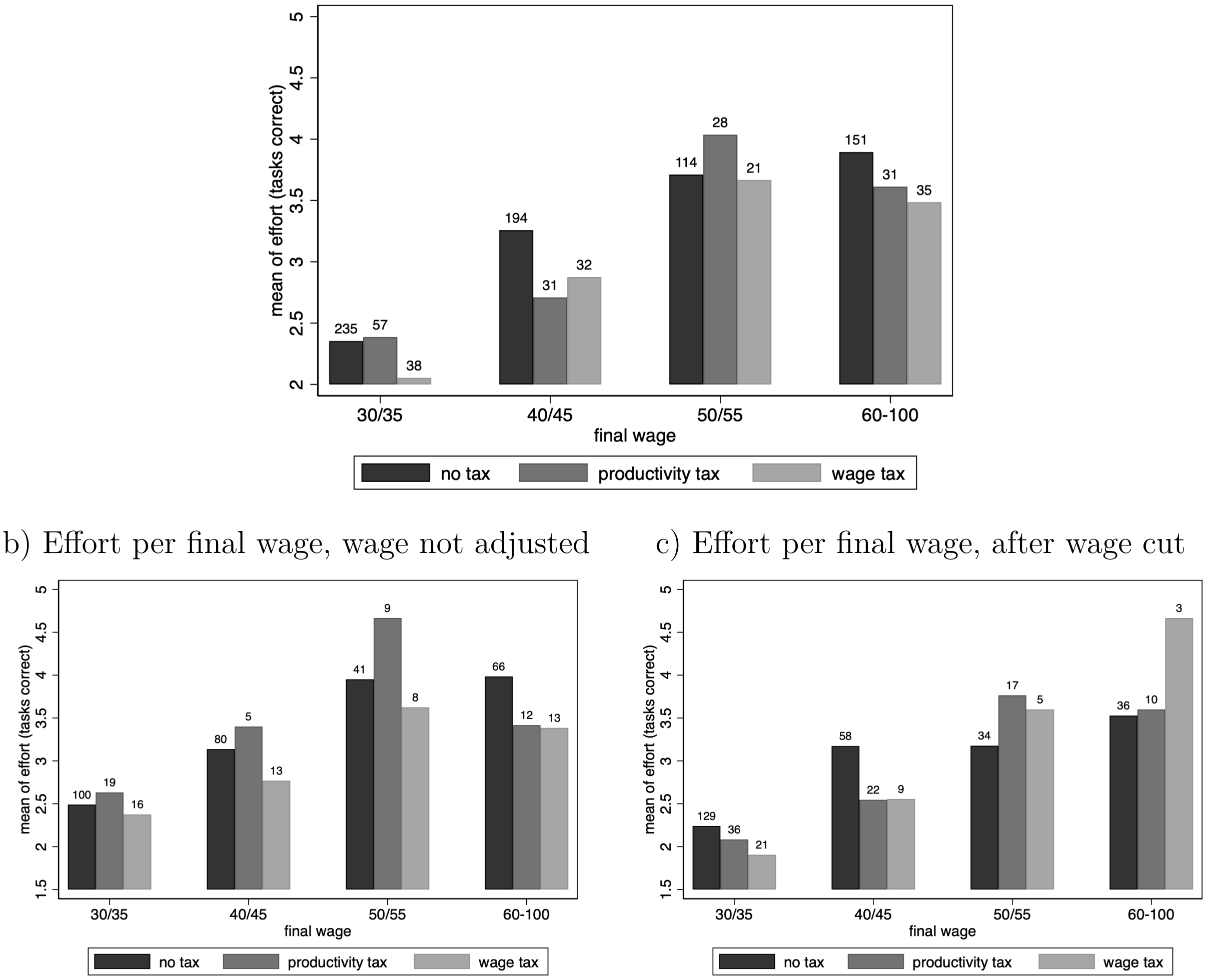
A more interesting question is if effort reacts to wages. I find significant gift exchange in the data, meaning that workers respond to higher wages with higher than minimal effort.[19] Figure 2.4.3 Panel a) demonstrates this by depicting the average effort for sets of final wage levels. The graph pools all of the treatments, and separates the effort by the tax outcomes (*nt*, *et*, and *wt*). Similar graphs by treatment can be found in Figure 2.A.1 in Appendix 2.A.

I discuss first the results concerning the *nt* baseline that is depicted by the black bars in the Figure 2.4.3. Effort increases in response to wages in particular at the lower wage levels. In line with the *fair wage* hypothesis, increases in effort beyond wages of 50/55 points do not happen. Hence this wage level can be interpreted as the objectively fair wage. At the wage of 50, the mean earnings of workers is 50 points, and the mean earnings of employers is 62 points. At the wage of 55, the means are 55 and 63 points, respectively. Increasing the wage from 30/35 to 40/45 in *nt* leads to a significant increase of 0.9 units

performance can be interpreted as differences in effort.

[19] For this analysis, I do not use the employer as the unit of observation but the labor contract. I do this because effort is assumed to respond non-linearly to the realized final wage (and not to the average wage). Moreover, I pool wages over 60 here as there are only a few high wage observations. I also combine the 0's and 5's to make the graphs easier to read.

Figure 2.4.3: Gift exchange



a) Effort per final wage, all data

b) Effort per final wage, wage not adjusted

c) Effort per final wage, after wage cut

Notes: The number of observations in each bin is reported above the bar.

in effort, $p < 0.001$. A further increase from 40/45 to 50/55 increases effort on average by 0.45 units, which is also a significant increase, $p = 0.028$. Any further wage increase from 50/55 to 60-100 points increases effort by 0.18 units, but this effect is no longer significant, $p = 0.350$. Hence also the interpretation that 50/55 is the fair wage level.

Next, consider the results concerning the tax shock outcomes. In general, the gift exchange patterns after *et* and *wt* shocks are similar to what they are in *nt*. In *et*, effort increases 0.32 units when wage is increased from 30/35 points to 40/45 points, however, this increase is insignificant, $p = 0.418$. A further wage increase from 40/45 to 50/55 increases effort on average by 1.33 units, which is a significant increase in effort, $p < 0.001$. Similar to the *nt* case, further wage increases from 50/55 to 60-100 points does not bring significant improvements to effort. In fact, the average effort drops by 0.42 units, $p = 0.333$. For *wt*, effort increases significantly by 0.82 units when wage increases from 30/35 to 40/45 points, $p = 0.048$. Effort increases further by 0.79 units when wage is increased from 40/45 to 50/55, which is marginally significant, $p = 0.066$. Again, further increases have no significant effect on effort and a wage increase from 50/55 to 60-100 points decreases average effort by 0.18 units, $p = 0.761$.

To summarize, the gift exchange pattern is strong and present also when shocks occur. Interestingly, this is not what we observe in the previous chapter, where no gift exchange was found after shocks. In general, the results suggest that gift exchange is stronger with the 'cheap talk' market stage preceding shocks and the subsequent opportunity for wage adjustments after. It would seem that the possibility for workers to react to wage adjustments might add not only negative reciprocity but also positive reciprocity. The latter occurs, for example, when a worker notices that an employer could have reduced the wage (even to the minimum) but did not.

I consider in more detail how workers respond to wage adjustments. First, consider the effect of wage cuts on gift exchange. Panel b) of Figure 2.4.3 shows the gift exchange relations for contracts where the initial offer was kept as the final wage. The gift exchange pattern seems slightly stronger than in Panel a) where all adjustments are pooled. Panel c) of Figure 2.4.3 illustrates the gift exchange pattern after the wage has been cut. The pattern seems weaker after the wage cuts, even if the graph controls for the final wages. For example, after a shock, effort is higher if wages are not cut than if they are, even when comparing identical final wages in the 40/45 bin. I will further investigate this in the next section.

Result 4: There is significant gift exchange: effort increases with the final wage up to a fair wage level.

Result 5: Gift exchange is also observed after shocks have taken place.

2.4.3 Workers' Response to Wage Cuts

In this section, I explore how workers respond to wage cuts. In particular, I am interested in the interactions between wage adjustment and the occurrence of tax shocks to see if tax shocks can be used as a justification for wage cuts. The second interest is to compare the effects of nominal wage cuts and real wage cuts, where the latter happen through the imposed wage tax, to see if they have a similar impact on work effort.

I use as a baseline a random effects model where effort is explained by the final wage.[20] I use a log wage model to parsimoniously capture the non-linear shape of the wage-effort relationship. The parameters of interest are the interaction terms between the

[20] Random effects model is appropriate over a fixed effects model, in particular, more efficient, if the unobserved worker characteristics are uncorrelated with the treatment effects. As individuals are assigned to the treatments randomly, it is expected that this assumption holds. Moreover, the hiring market is anonymous and non-binding, for which reason wage should not be correlated with the worker's unobserved characteristics. A Hausman test further confirms that the fixed effects model estimates and the random effects model estimates of the coefficients are not systematically different from each other, $p = 0.615$.

tax shock outcomes and the wage adjustment categories that indicate whether a wage cut, no adjustment or a wage increase occurred. Note that by controlling for the final wage, we are not measuring the whole impact of a wage cut, but the punishment for making a higher wage offer and then cutting it at a lower level versus the case where that lower wage is offered immediately in the first stage and kept constant.

The equation to be estimated is given by:

$$e_{it} = \beta_0 + \beta_1 lnw_{it} + \Sigma_{adj}\Sigma_s\beta_{adj,s}\mathbb{1}_{adj,s} + u_i + \epsilon_{it}. \tag{2.4.1}$$

Variable e stands for effort, lnw for log wages, and $\mathbb{1}_{adj,s}$ is an indicator variable for each combination of wage adjustment adj and shock s capturing $adj \in \{cut, noadj, inc\}$ for wage cut, no adjustment and wage increase, and $s \in \{nt, et, wt\}$. Subscript i is for the individual worker and t is for the round. The term u_i captures the individual effects, while ϵ_{it} represents the noise.

Table 2.4.5 reports the model estimates. Observe that wage cuts have a negative but insignificant effect on effort when there is no shock. When the wage cut occurs in conjunction with a tax shock, regardless of type, it has a (marginally) significant negative impact on effort. When the wage is cut after an employer shock, et, this reduces effort by 0.42 units ($p = 0.034$). A wage cut after a worker shock, wt, reduces effort by 0.45 units ($p = 0.080$). This corresponds to a 15% drop in average effort just due to the employer's wage setting method (recall that I am correcting for the final wage). The shocks on their own (that is, when employers do not adjust the wage), have no significant effect: et shock's impact is estimated at 0.053 units ($p = 0.801$), and wt shock's impact at -0.136 units ($p = 0.563$). This means that workers do not significantly reduce effort in response to a real wage cut if the employer does not alter the nominal wage. Finally, a wage that is increased from the first stage offer does not increase effort beyond the level that would have been achieved had the higher wage been offered immediately.

To summarize, I reject hypothesis H2b in favor of the alternative AH2b – workers punish wage cuts after et shocks. I also reject the hypothesis H3b in favor if its alternative AH3b – workers do not cut effort in response to the work tax shock when the wage is not adjusted – meaning real wage cuts without nominal adjustments do not lead to significant changes in worker morale.

Result 6: Wage cuts after a shock are punished by reduced effort, even after correcting for the final wage itself.

Result 7: A real wage cut without a nominal wage cut does not lead to reductions in effort (work morale).

Table 2.4.5: Workers' response to wage cuts

	coefficient	standard error	p-value
wage cut × *nt*	-0.219	0.145	0.131
wage cut × *et*	-0.417	0.197	0.034
wage cut × *wt*	-0.452	0.259	0.080
wage increase × *nt*	0.072	0.160	0.653
wage increase × *et*	0.230	0.442	0.603
wage increase × *wt*	-0.142	0.269	0.598
no adjustment × *et*	0.053	0.210	0.801
no adjustment × *wt*	-0.136	0.235	0.563
log wage	1.662	0.235	0.000
constant	-3.071	0.916	0.001
σ_u	0.965		
σ_e	1.398		
ρ (fraction of var due to u_i)	0.323		
R-squared (overall)	0.113		
obs.	967		
number of workers	181		

Notes: *nt*: no tax shock realized; *et*: productivity tax shock realized; *wt*: wage tax shock realized. Standard errors are bootstrapped. The case of no shocks and no adjustment is absorbed in the constant term.

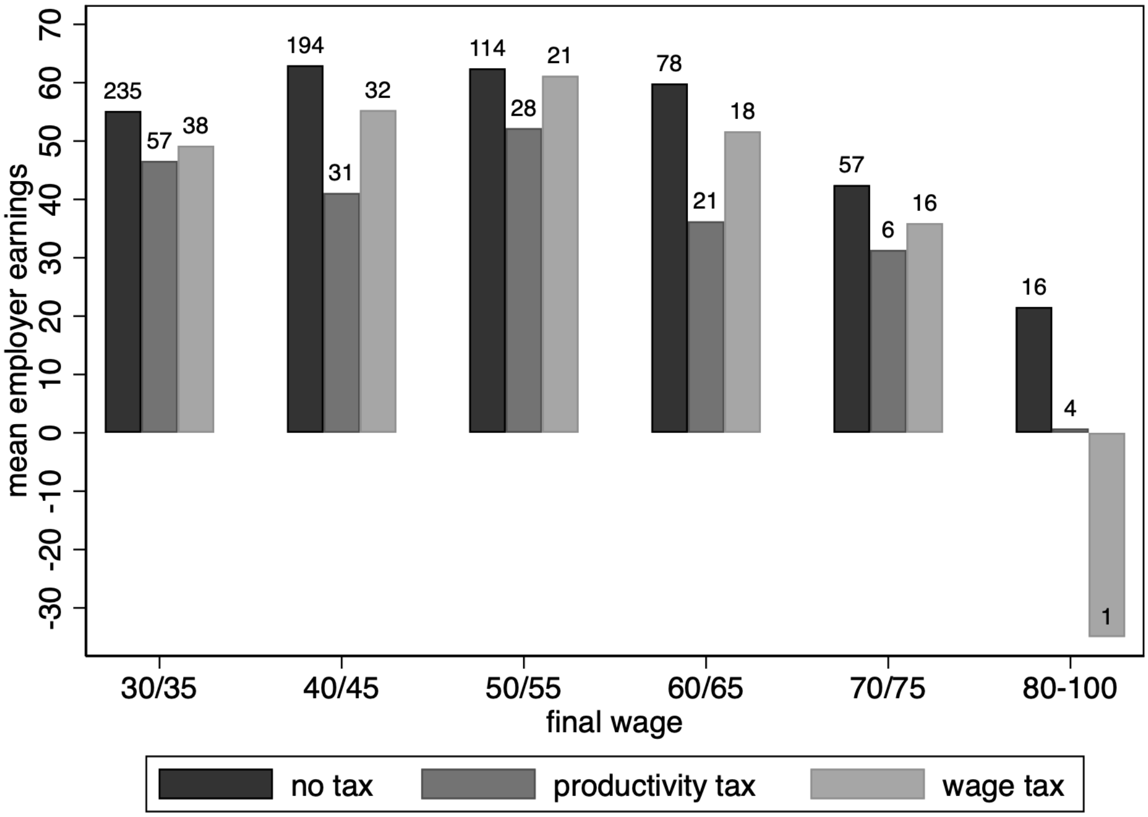
2.4.4 Employer's Earnings

With this overview of how workers react to shocks and wage adjustments, I can reconsider employers' behavior and investigate whether they are best responding in this experimental labor market. Based on the model, in the absence of shocks, the best policy is to immediately offer the workers their final wage. The model does not predict positive reciprocity beyond the *objectively fair wage* point and, hence, additional wage increases are not expected to bring extra effort. The experimental results support this view. In the absence of shocks wages beyond the fair-wage level yield no additional worker effort.

The picture is more involved when there are shocks. If workers consider wage cuts under shocks to be justified, then employers' best response might be to cut wages. If workers punish wage cuts by reducing effort, employers' best policy is to not cut wages, even after a shock.

Consider first if paying a high final wage is a profitable policy. Figure 2.4.4 depicts the average employer earnings in different tax shock conditions over the final wages. This shows that is profitable for the employers to engage in gift exchange. Average profits are higher at wage levels above the minimum of 30 points for the no tax shocks outcome, depicted by the black bars. This is not necessarily true after productivity tax shocks, *et*, as is depicted by the medium gray bars. There, one see that the pattern of employers' earnings is relatively erratic across wages, partly because some wages are observed only

Figure 2.4.4: Employer earnings by final wage



Notes: The number of observations in each bin is reported above the bar.

rarely. The positive gift exchange result does seem to carry over to the cases with wage tax shocks, *wt*.

Result 8: Employers profit from offering higher than minimal wages.

Note that Figure 8 aggregates across the possible wage adjustment scenarios. I am also interested in determining whether an employer is better off in sticking to a particular wage offer or by cutting the wage. For this reason, I depict employer earnings per *wage offer* and across different wage adjustment policies.

Figure 2.4.5 shows the average employer earnings per initial wage level and by how the wage is adjusted. Panel A) shows this for the no tax outcome *nt*. Wage cuts do not seem to have an effect in the lowest wage bin (note that wage cuts are restricted here, as only wages of 35 can be cut here down to 30). Wage cuts are harmful for employer earnings in the wage range of 40-55 points, but the effects are relatively small. When the initial wage offer increases above 50/55, the negative effect of cuts first disappears in the 60-75 range and then reverses for very high wage offers. This effect reversal is intuitive – it is very difficult for workers to complete enough correct tasks to produce a positive profit for the employer at these wages and this may also make it easy for the workers to accept a wage cut without reducing effort. Finally, wage increases have a modest positive effect at low wages and a modest negative effect at high wages.

Next, consider the cases with shocks. First, the productivity tax shock *et* hits the employers directly, which means that for the shocked rounds, the employers income is reduced by about 20%. This is noticeable in the Figure 2.4.5, as the bars in panel B) are much lower across the board in comparison to the other panels. Wage cuts together with *et* seem to have mostly harmful effects at wages 30-55, but positive effects for wage offers above 60. Note, however, that wage cuts are relatively common even at the lower wages. It appears that employers misjudge workers effort reactions. Employers might feel that a wage cut is justified because the productivity shock hits their earnings. At low wages, nevertheless, such a wage cut backfires because it makes the workers reduce their effort.

Last, consider the worker shocks, *wt*, as depicted in panel C) of Figure 2.4.5. Wage cuts following *wt* have negative effects at the lower end of the wage distribution, for wages from 30 to 45 points, while they have a positive effect on profits already at wages 50/55. At the lower levels, it seems that the money saved by reduced wages is not enough to cover the income lost due to of reduced worker effort.

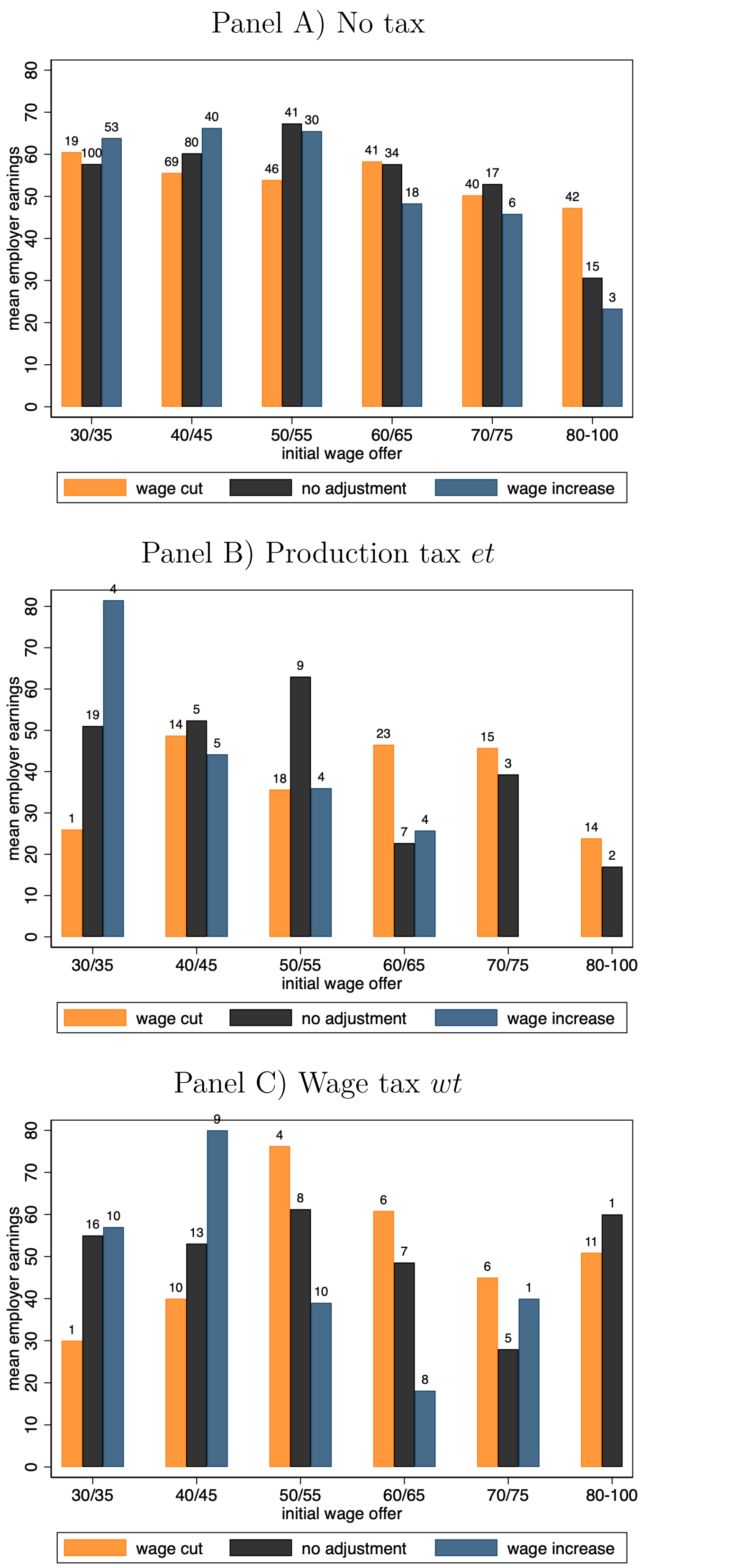
Note that the wages around 50 to 55 points play an important role in these discussions about employer earnings. In both the earlier results on effort per wage and the analysis of the previous chapter, this wage level constitutes the fair wage level in the experimental environment. Therefore, it might be useful to test the effects for wages below and above 55.

Using observations only up to the initial wage of 55 points, the average impact of a wage cut on employer earnings in *nt* is -6 points, controlling for wage offers and clustering the standard errors by market (five employers and seven workers). However, this effect is not significantly different from zero ($n = 479$, $p = 0.183$). In *et*, cutting a wage after an initial wage offer up to 55 points leads to a -12 points average change in earnings, which is also insignificant ($n = 79$, $p = 0.164$). In *wt*, the impact is -11 points and similarly insignificant ($n = 82$, $p = 0.139$). Pooling the shocks *et* and *wt* give -12 points ($n = 161$, $p = 0.005$), indicating that wage cuts in shocks are usually not good for employer earnings. For wages 60-100, cuts have generally positive effects on earnings. In *nt*, a wage cut increases employer earnings on average by 6 points ($n = 217$, $p = 0.301$), in *et*, by 20 points ($n = 68$, $p = 0.003$), and in *wt*, by 10 points ($p = 0.235$). Pooling the shocks together, the effect is 16 points ($n = 113$, $p = 0.002$).

Result 9: Wage cuts are profitably only when the wage offer is “irrationally high”, that is, above the fair wage level.

Result 10: Wage cuts are not profitable at the fair wage or below, nor do they become profitable after shocks.

Figure 2.4.5: Employer earnings by final wage, by tax outcome



Panel A) No tax

Panel B) Production tax *et*

Panel C) Wage tax *wt*

2.5 Discussion and Conclusions

I study a gift exchange labor market where wage cuts can occur. Based on the theoretical model, it is expected that wage cuts have detrimental effects on effort in general. The only exception is that workers are predicted to accept wage cuts in response to employer shocks, with the intention to share the burden. Indirect cuts (that occur due to real-wage changes), on the other hand, are not expected to invoke an effort response.

These predictions are tested with a laboratory experiment. The first observation is that the possibility of adjustment seems to evoke positive reciprocity to the extent that I observe profitable gift exchange in this setting, while it is not observed to the same extent in the setting of the previous chapter. The wage-effort relationship is, however, nonlinear, suggesting again that the objectively fair wage is a useful concept in these types of labor markets.

Somewhat surprisingly, wage cuts in the absence of shocks do not lead to reductions in effort, while wage cuts after any shock do lead to extra punishments by workers after controlling for the final wage. This suggests that the status quo is a strong reference point for fairness. The effect of cuts is similar in size regardless of whether the shock is primarily felt by the employer or the worker, counter to the burden-sharing benchmark (where wage cuts are expected after an employer shock while wage increases are expected after a worker shock). Wage cuts through the worker shock alone, that is, reductions to the real wages while keeping nominal wages intact, do not lead to significant reductions in effort.

Shocks do not justify wage cuts. This might be because such a ‘mechanical’ justification is not in the spirit of gift exchange, which builds on trust and social preferences. It would be interesting to see if rapport between the worker and the employer, for example through personalized messages explaining the need to adjust after an employer shock, might justify cuts without the negative effort consequences.

Last, employers do not seem to fully anticipate the workers’ reactions. Rather, some employers seem to use an interpretation of fairness that is more in their interests than others, as is suggested by the observation that employers adjust wages more frequently after they experience a shock than when it is experienced by the workers. It is worth pointing out, however, that in one treatment (WT), employers average wage adjustments become positive over the 8 rounds of the experiment. Perhaps a greater convergence of fairness norms could be achieved with more experience.

Bibliography

Akerlof, George A. 1982. "Labor contracts as partial gift exchange." *The Quarterly Journal of Economics*, 97(4): 543–569.

Akerlof, George A, and Janet L Yellen. 1990. "The fair wage-effort hypothesis and unemployment." *The Quarterly Journal of Economics*, 105(2): 255–283.

Bewley, Truman. 1999. *Why don't wages fall in a recession.* Harvard University Press Cambridge.

Buchanan, Joy, and Daniel Houser. forthcoming. "If wages fell during a recession." *Journal of Economic Behavior & Organization.*

Charness, Gary. 2004. "Attribution and reciprocity in an experimental labor market." *Journal of labor Economics*, 22(3): 665–688.

Charness, Gary, and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *The Quarterly Journal of Economics*, 117(3): 817–869.

Chen, Daniel L, and John J Horton. 2016. "Research note—Are online labor markets spot markets for tasks? A field experiment on the behavioral response to wage cuts." *Information Systems Research*, 27(2): 403–423.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2015. "Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment." *Management Science*, 61(8): 1777–1794.

Davis, Brent J, Rudolf Kerschbamer, and Regine Oexl. 2017. "Is reciprocity really outcome-based? A second look at gift-exchange with random shocks." *Journal of the Economic Science Association*, 3(2): 149–160.

Dickens, William T, Lorenz Goette, Erica L Groshen, Steinar Holden, Julian Messina, Mark E Schweitzer, Jarkko Turunen, and Melanie E Ward. 2007. "How wages change: micro evidence from the International Wage Flexibility Project." *Journal of Economic Perspectives*, 21(2): 195–214.

Fehr, Ernst, and Klaus M Schmidt. 1999. "A theory of fairness, competition, and cooperation." *The Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1993. “Does fairness prevent market clearing? An experimental investigation.” *The Quarterly Journal of Economics*, 108(2): 437–459.

Gerhards, Leonie, and Matthias Heinz. 2017. “In good times and bad–Reciprocal behavior at the workplace in times of economic crises.” *Journal of Economic Behavior & Organization*, 134: 228–239.

Gneezy, Uri, and John A List. 2006. “Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments.” *Econometrica*, 74(5): 1365–1384.

Greiner, Ben. 2004. “The online recruitment system ORSEE 2.0.” *A Guide for the Organization of Experiments in Economics.*

Greiner, Ben, Axel Ockenfels, and Peter Werner. 2011. “Wage transparency and performance: A real-effort experiment.” *Economics Letters*, 111(3): 236–238.

Hannan, R Lynn. 2005. “The combined effect of wages and firm profit on employee effort.” *The Accounting Review*, 80(1): 167–188.

Hennig-Schmidt, Heike, Abdolkarim Sadrieh, and Bettina Rockenbach. 2010. “In search of workers’ real effort reciprocity—a field and a laboratory experiment.” *Journal of the European Economic Association*, 8(4): 817–837.

Kahneman, Daniel, Jack L Knetsch, and Richard Thaler. 1986. “Fairness as a constraint on profit seeking: Entitlements in the market.” *The American economic review*, 728–741.

Kaur, Supreet. 2019. “Nominal wage rigidity in village labor markets.” *American Economic Review*, 109(10): 3585–3616.

Koch, Christian. 2021. “Can reference points explain wage rigidity? Experimental evidence.” *Journal for Labour Market Research*, 55(1): 1–17.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2013. “Do wage cuts damage work morale? Evidence from a natural field experiment.” *Journal of the European Economic Association*, 11(4): 853–870.

Rubin, Jared, and Roman Sheremeta. 2015. “Principal–agent settings with random shocks.” *Management Science*, 62(4): 985–999.

Schram, Arthur, Jordi Brandts, and Klarita Gërxhani. 2018. “Social-status ranking: a hidden channel to gender inequality under competition.” *Experimental Economics.*

Weber, Matthias, and Arthur Schram. 2017. “The Non-equivalence of Labour Market Taxes: A Real-effort Experiment.” *The Economic Journal*, 127(604): 2187–2215.

Appendix 2.A Additional Data Analysis

Table 2.A.1 reports the initial non-binding wage offers per treatment and per tax setting. The unit of observation is the average initial wage offer per employer and tax outcome (*nt*, *et*, *wt*). The wage offers do not vary significantly per tax outcome. This simply implicates a successful randomization; the participants did not anticipate the random shocks before they were announced (recall that the wage offers are made before shocks are realized).

Table 2.A.1: **Wage offers, treatments, and shocks**

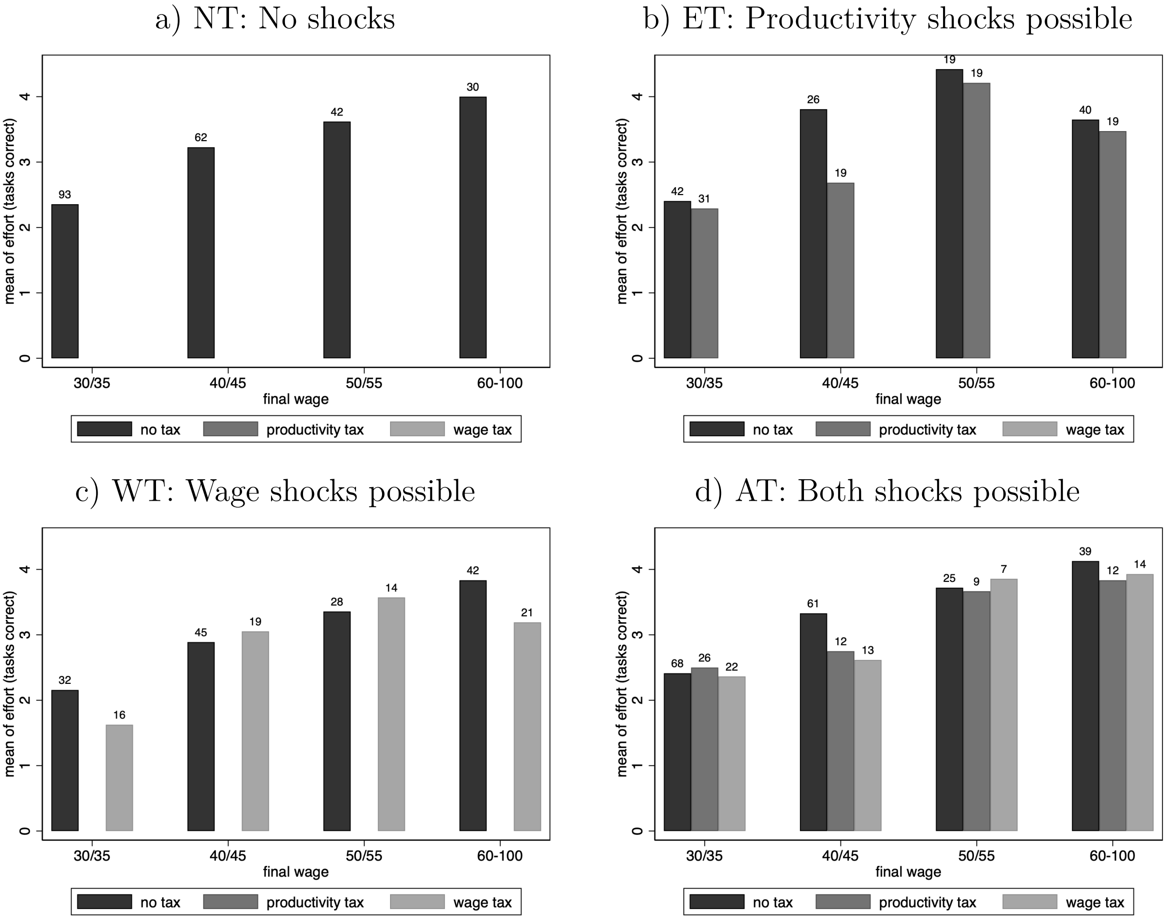
tax outcome	*NT*	*ET*	*WT*	*AT*	**pooled**
nt	**45.7**	**55.8**	**54.4**	**51.6**	**51.8**
	30	30	30	40	130
et		**55.6**		**53.5**	**54.4**
obs.		30		40	70
wt			**56.1**	**49.9**	**52.6**
obs.			30	38	68
PtT (p-values)					
nt vs et	-	*0.899*	-	*0.172*	*(0.325)*
nt vs wt	-	-	*0.170*	*0.675*	*(0.752)*

Notes: Tax shocks occurred in rounds 2, 4, and 5. The unit of observation is the mean adjustment by an employer across rounds. Paired tests between shock- and no-shock rounds are reported. Tests for the pooled data are conducted on the paired data (unpaired averages shown on the table). Mean wages across employers are in bold. 'obs.' shows the number of employers. *NT*: no taxes possible; *nt*: no tax shock realized; *WT*: wage tax possible; *wt*: wage tax shock realized; *ET*: productivity tax possible; *et*: productivity tax shock realized; *AT*: both taxes possible, 'pooled' combines treatments. PtT: permutation t-test.

Figure 2.A.1 shows the basic gift exchange results separately for each treatment.

To continue the comparisons of Figure 2.4.3, Figure 2.A.2 depicts gift exchange after there has been a wage increase. However, as this is the least frequent wage adjustment outcome, the number of observations per bar can be very small. Note also that the scale of the y-axis is different than in the main text due to a few outliers in the data.

Figure 2.A.1: Gift exchange by treatment, all wage adjustments


a) NT: No shocks

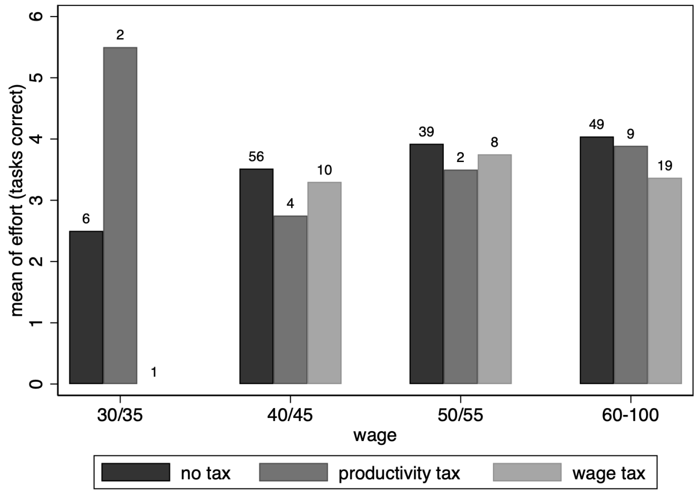
b) ET: Productivity shocks possible

c) WT: Wage shocks possible

d) AT: Both shocks possible

Notes: Effort by final wage levels. The number of observations in each bin is reported above the bar.

Figure 2.A.2: Gift exchange after a wage increase



Appendix 2.B Experimental Instructions [Original in Italian]

The instructions differ for each treatments. When appropriate, we indicate additional texts by the following system. "When taxes" refers to all treatments that allow taxes: AT, ET, and WT. "In AT" refers to the tax treatment with all taxes, "ET" refers to the tax treatment with only employer taxes and "WT" refers to the tax treatment with only wage taxes.

Welcome to the experiment!

From now on, please, do not talk with the other participants. If you have any questions, please, raise your hand. Place your phone in your bag: you are not allowed to use it during the experiment. In case you want to revisit the instructions after the software tutorial, you can use the paper version on your desk where you also find a pen and a paper.

Your payoff from the experiment will consist of two parts: the 5 euro show-up fee and the earnings (or losses) from 2 rounds out of the 8 rounds in total. These 2 rounds will be chosen at random.

Role

You participate in a labor market that has 5 employers and 7 employees. After the tutorial and a questionnaire on the instructions, you will be randomly assigned to either the role of an employer or the role of a worker, and you will keep the same role for the entire duration of the experiment.

Overall structure

The experiment consists of 8 rounds.

1st Stage: Hiring

Each employer can make an initial wage offer on a public platform, and each worker can accept one of these offers. Note that the wage can be adjusted in the next stage. Once an offer becomes accepted, the hired worker will work that round for the employer that made the offer.

All the hiring results of the round will be made public.

2nd Stage: Taxes and Final Wages

All the hiring results of the round will be made public and possible taxes are announced. The employers can now decide what the final wage they will pay.

3rd Stage: Work

Each hired worker has 5 minutes to work on the tasks. After the 5 minutes, the work results will be communicated to the respective worker and employer, and the earnings are calculated.

Detailed instructions

Hiring Stage

The hiring stage lasts at most for 2 minutes. There are 5 employers and 7 workers in the market. Each employer can announce an initial wage offer on a public platform. The offer must be between 30 and 100 points, in steps of 5 points, and it can be modified while not yet accepted, but cannot be withdrawn entirely once made.

A worker can accept one of the available offers. Once accepted, the worker is immediately hired by the employer for the reminder of the round and the offer is removed from the platform. If more than one worker attempts to accept the same offer, it is granted to the fastest. All of the offers and subsequent modifications are updated to the platform in real time and published in a random order.

If an offer is not accepted within the 2 minutes, the employer is not able to hire anyone. In the same way, if a worker does not accept an offer within the 2 minutes or if all of the 5 offers made have been accepted by other workers, the market closes and these workers will be unemployed for the round. Out of the 7 workers, at least 2 will be unemployed every round.

Without a contract, the workers and employers will not participate in the remaining stages of the round: an employer earns 0 points and a worker earns 20 points as an unemployment benefit. Both will resume the experiment again in the beginning of the next round.

If an employer hires a worker, the employer receives 40 points and any earnings from the work of the hired worker. The worker's wage will then be subtracted from these earnings. The worker's earnings consist of the wage. [**When taxes:** *AT: Both payoffs/ET: employer's payoff/ WT: worker's payoff may be subject to taxes, as explained in the next part.*]

The experiment is anonymous: the worker will not know the identity of the employer, and likewise, the employer will not know the identity of the worker.

Taxes and Final Wages

After the hiring stage, all of the participants see the overall results of the hiring stage: how many workers were hired and at what wage offers.

After these results, the taxation scheme for the round is announced. It is chosen randomly by the computer.

[**The options and probabilities depend on which taxes are possible. The following section is written for AT unless otherwise specified**]

There are 3 ***[In ET or WT: 2]*** *possibilities:*

- ***No taxes*** *(probability 66.7%)*
- ***Tax*** *of 20% on the revenues of the employer (probability 1/6 = 16,7%)* ***[In ET 1/3 = 66.7%, not mentioned in WT]***
- ***Tax*** *of 20% on the wage of the worker (probability 1/6 = 16,7%)* ***[In WT 1/3 = 66.7%, not mentioned in ET]***

In total, there is a 33% probability that a tax is applied, and a 67% probability that there are no taxes; on average, 1 in 3 rounds has taxes. ***[In AT only:*** *The type of the tax is randomly chosen by computer, each type being equally likely.*]

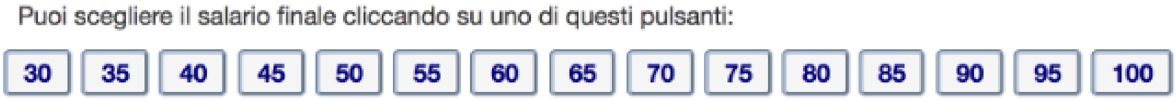
[**In AT and ET only:** *The tax on the revenues of the employer reduces the earnings from the worker tasks: each correctly completed task is worth 16 points, instead of the 20 points when there is no tax. The tax does not impact the 40 points received from hiring.*]

[**In AT and WT only:** *The tax on the earnings of the worker reduces the amount of wages received by 20%. Each employer however pays the full salary.*]

The collected taxes will be returned to the experimenter.

Before the next stage, the employers need to decide the final wages. The final wage must be between 30 and 100 points, in steps of 5. Only the hired worker will learn the final wage – no other participant will know it.

Screenshot from the program



Puoi scegliere il salario finale cliccando su uno di questi pulsanti:

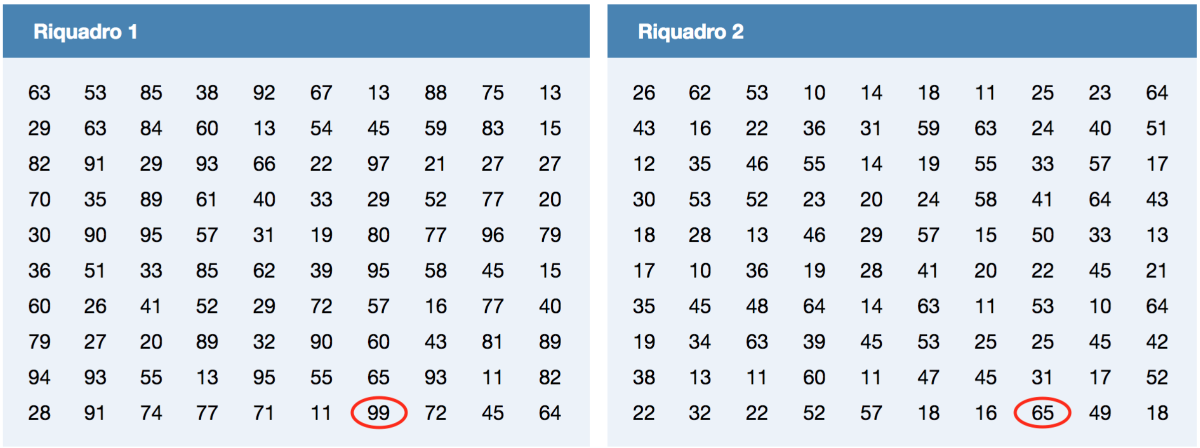
"You can choose the final wage by clicking one of these buttons."

Work Stage

The hired workers have 5 minutes to work, during which they can attempt at most 10 tasks in total. Each task consists of two boxes, each containing 100 numbers: the task is to find the largest number in each box and then sum them together.

Each correctly completed task will give the employer 20 points [**In AT and ET:** *if there are no taxes on the employer's taxes, in which case, each correctly complete task is worth 16 points*]. Wrong answers do not affect payoffs but count as ’attempted tasks’. The workers can submit only one answer per task.

Example: The largest number in the left box is 99 and the largest number in the right box is 65, both are circled with red. Summed together they give 99 + 65 = 164: **164** is the correct answer to be submitted!



Riquadro 1									
63	53	85	38	92	67	13	88	75	13
29	63	84	60	13	54	45	59	83	15
82	91	29	93	66	22	97	21	27	27
70	35	89	61	40	33	29	52	77	20
30	90	95	57	31	19	80	77	96	79
36	51	33	85	62	39	95	58	45	15
60	26	41	52	29	72	57	16	77	40
79	27	20	89	32	90	60	43	81	89
94	93	55	13	95	55	65	93	11	82
28	91	74	77	71	11	99	72	45	64

Riquadro 2									
26	62	53	10	14	18	11	25	23	64
43	16	22	36	31	59	63	24	40	51
12	35	46	55	14	19	55	33	57	17
30	53	52	23	20	24	58	41	64	43
18	28	13	46	29	57	15	50	33	13
17	10	36	19	28	41	20	22	45	21
35	45	48	64	14	63	11	53	10	64
19	34	63	39	45	53	25	25	45	42
38	13	11	60	11	47	45	31	17	52
22	32	22	52	57	18	16	65	49	18

The Payoffs

After 5 minutes or after having tried all 10 tasks, all of the participants are directed to a results page. The worker and the employer who has hired the worker get to know the number of correct and attempted tasks, and the resulting payoffs of both, but will not get to know the results of the other participants.

Scenario A:
If the participant does not have a contract:

- Employer's payoff = **0 points**
- Worker's payoff = **20 points**

Scenario B:
If the participant has a contract [**When taxes:** *and there are no taxes]:*

- Employer's payoff = **40 − final wage + 20 * number of tasks correct**
- Worker's payoff = **final wage**

In other words, the employer receives 40 points when hiring a worker, pays the final wage and receives the revenues from each correctly completed task. What remains is the earnings of the employer, and note that this can also be negative. Conversely, the earnings of the worker consists of the final wage.

[**Only in AT and WT:** ***Scenario C:***
If the participant has a contract and there is a 20% tax on the earnings of the employer*, the payoff from each correctly completed task is reduced to 16 (from 20) and thus the payoffs are given as:*

- Employer's payoff = **40 − final wage + 16 * number of tasks correct**

The worker's payoff is the same as under Scenario B.]

[**Only in AT**] ***Scenario D:*** [**OR Only in ET**] ***Scenario C;***
If the participant has a contract and there is a 20% tax on the earnings of the worker**, the payoff of the worker is given by the final wage less the taxes:**

- Worker's payoff = **final wage − 20% of the final wage**

The employer's payoff is the same as under scenario B.]

[**Only in AT**] *The two taxation systems are alternatives, they can never apply simultaneously.*

The points earned in the laboratory will be converted into Euros with the following exchange rate: **10 points = 1 euro**. On top of the 5 euro show-up fee, the participants are remunerated for only two rounds (out of the 8 in total) that are randomly selected in the end of the experiment.

Comprehension test

The comprehension test consisted of 12 true or false statements. The first 10 questions were the same for all tax treatments. The correct answer is reported in the parenthesis.

1. If a worker is unemployed for a round, she or he does earns nothing. (FALSE)
2. If an employer does not manage to hire a worker for a round, the employer earns nothing. (TRUE)
3. Accepting an offer, the worker commits to work for that employer for that round. (TRUE)
4. An employer who has hired someone earns 40 points. (TRUE)
5. In general, the salary is deducted from the earnings of the employer and given to the worker. (TRUE)
6. The number of tasks that a worker can try is unlimited. (FALSE)

7. The workers obtain a higher salary if they complete more tasks. (FALSE)
8. Other than the worker himself/herself, only the employer will get to know how many tasks were completed. (TRUE)
9. You will be compensated for all of the 8 rounds. (FALSE)
10. There employer can always change the wage offered after hiring stage. (TRUE)

The last two questions depend on what taxes are possible. When no taxes are possible (NT):

11. Your earnings will depend on your decisions and those of the other participants. (TRUE)
12. The earnings of an employer cannot be negative for a round. (FALSE)

If only productivity taxes are possible (ET)

11. 20% of 20 points is 4 points. Thus, when we have taxes on the employers, the earnings per each correct task is 16 instead of 20 points. (TRUE)
12. The earnings of an employer cannot be negative for a round. (FALSE)

If only worker taxes are possible (WT)

11. The earnings of an employer can be negative for a round. (TRUE)
12. The taxes on the worker's earnings are always 20 points. (FALSE)

If both taxes are positive (AT)

11. 20% of 20 points is 4 points. Thus, when we have taxes on the employers, the earnings per each correct task is 16 instead of 20 points. (TRUE)
12. The taxes on the worker's earnings are always 20 points. (FALSE)

Chapter 3

Choice Architecture and Transparency[1]

3.1 Introduction

Choice architecture and nudging have gained prevalence as Behavioral Insights Units have been founded across the world – OECD (2019) counts over 200 government units, initiatives, and partnerships that use this behavioral intervention approach and nudges in particular. Nudges work through *seemingly irrelevant factors* that by basic rational theory should not have any effect on decision-making (Thaler, 2015). They do not meaningfully change the monetary incentives or the options available, yet they can have a significant impact on the subsequent decisions (Thaler and Sunstein, 2008). Because nudges often impact decisions indirectly, it can be difficult to judge the effectiveness and legitimacy of nudging policies (Luc Bovens, 2009; Pelle Guldborg Hansen and Andreas Maaløe Jespersen, 2013), at least, without further experimentation. Indeed, nudges may be considered *manipulative* in the sense that they may lead people to act against their own interests if they misrepresent information to those who lack it or opt people in by default when the people suffer from limited attention.

One way to avoid manipulation is to be transparent about the use of nudges. Little is known, however, of how transparency affects the behavior of those involved. It is common that subjects underestimate the impact that nudges have on others and especially on their own behavior (H. Min Bang, Suzanne B. Shu and Elke U. Weber, 2018; Emily Pronin, Daniel Y. Lin and Lee Ross, 2002), but this result might not hold as people become more used to behavioral intentions or as nudging becomes more transparent. People might

[1] I am grateful for all the advice by Michele Belot and Arthur Schram. I also want to thank Christine Alamaa, Ayala Arad, Agnès Charpin, Junze Sun, Egon Tripodi and Oliko Vardishvili for their useful comments and the seminar audiences at EUI and the ESA Global Online Around-the-Clock Meeting. I thankfully acknowledge funding from the Academy of Finland and the European University Institute.

become immune to nudging, or even resist nudges when they are transparent by choosing the opposite. On the other side, those who design the choice environments, the so-called Choice Architects, might also react in response to transparency, for example, by lessening the use of nudges, in fear of negative judgments. Surprisingly, only very few studies (exceptions are discussed below) have considered the possibility that transparency about the use of nudges might affect the way in which they are applied and the effects they have. I aim to fill this gap. I use two methods to address these question. First, I model decision making in a limited attention model, to make predictions about nudges and transparency. Second, I test these predictions with an online framed field experiment.

My limited attention model distinguishes between two types nudges, following the fast and slow thinking framework of System 1 and System 2 (Daniel Kahneman, 2011; Pelle Guldborg Hansen and Andreas Maaløe Jespersen, 2013). System 1 nudges offer quick shortcuts that make decision-making easy. These nudges affect decision making in particular when Decision Makers are not paying full attention to the problem at hand. For example, default choices opt people in without active decision-making. System 2 nudges, on the other hand, make people pay attention to the problem and the options available. To give a few examples, a reminder that asks people to double-check the information that they are about to submit is a System 2 nudge, and so is a nudge that encourages people to consider the costs and benefits associated with each option before making their choice. In the language of McCrudden and King (2016), a System 1 nudge re-biases while a System 2 nudge de-biases decision-making. Increasing transparency in this model, I hypothesize, is likely to engage the reflective thinking of System 2, making Decision Makers more attentive and thus interrupting the automatic processes of System 1. This leads to my hypothesis that fast System 1 nudges become less effective when nudging is made transparently, while the opposite is expected with the slow System 2 nudges.

This theoretical approach has similarities with Löfgren and Nordblom (2020), who look at pure and preference nudges also in a limited attention model. In their model, nudges work only when people are inattentive, as they do not explicitly consider attention increasing System 2 nudges. Although Löfgren and Nordblom mention opportunities for possible extensions, for example to accommodate *boosts*[2], they do not explore these possibilities. My theoretical approach does this by looking into System 2 nudges that simply increase attentive decision-making.

[2] *Boost* are behavioral interventions similar to nudges. Boosts are, however, conceptually more demanding than for example the System 2 nudges discussed here, as boosts *educate and inform Decision Makers in a behaviorally smart way* (Till Grüne-Yanoff and Ralph Hertwig, 2016; Till Grüne-Yanoff, 2018).Grüne-Yanoff and Hertwig (2016) write that in their view “boosting goes beyond education and the provision of information. For example, in order to boost Decision Makers’ skills, policy designers need to identify information representations that match the cognitive algorithms of the human mind, thus using the environment (e.g., external representations) as an ally to foster insight and decision-making skills.”

The theoretical analysis shows that System 1 nudges are more potent interventions than System 2 nudges, making System 1 nudges the preferred tool of influence for Choice Architects. A Choice Architect who cares for (self-)image concerns and does not want to look manipulative, however, might refrain from using the more manipulative System 1 nudges to sustain a better self-image. This is backed by the empirical findings that the de-biasing System 2 nudges are viewed more favorably than the re-biasing System 1 nudges by those subjected to them (Cass R. Sunstein, Lucia A. Reisch and Micha Kaiser, 2019; Cass R. Sunstein, 2016; Gidon Felsen, Noah Castelo and Peter B. Reiner, 2013; Ayala Arad and Ariel Rubinstein, 2018). In this framework, transparency is assumed to increase image concerns as the Architect's actions become more visible to others and it is predicted to reduce the impact of System 1 nudges, as they reply on opposite modes of thinking. Together these two factors lead to the hypothesis is that image-concerned Choice Architects switch the nudge that they use when transparency is imposed, decreasing the use of System 1 nudges in favor of System 2 nudges.

I test these hypotheses with an online frame field experiment (Harrison and List, 2004). In the experiment, the Decision Makers do a series of real effort tasks to earn money. I am particularly interested in the second round, for which the Decision Makers choose a performance target. They have two options to choose from: a high target is riskier but offers a better reward if obtained while a lower target is safer but pays less if reached. The stakes are of medium size (4-6 USD). The Choice Architects choose how this question over performance targets is framed. In particular, the Architect receives a five-dollar bonus if the Decision Maker chooses the high performance target. The Architects have 3 options to choose from: A) a simple question presentation ('the benchmark'), B) a question with a strong default option (targeting System 1), and C) a presentation that encourages risk-taking yet asks people to do a risk-benefit analysis and then to follow it (targeting System 2). The setup is designed such that the simple question promotes the low target by listing it first, while the default nudge (System 1) and the risk-benefit analysis nudge (System 2) are set to promote the high target. In the transparent setting, it is common information that Choice Architects may nudge the Decision Makers through the question presentation. In the non-transparent setting, Decision Makers know that another player exists, but they do not know about the nudging.

I find that transparency has no effect on the Choice Architects' behavior. A majority of the Choice Architects choose the default nudge (System 1 nudge), while the risk-benefit analysis nudge (System 2) is the second most popular option. This holds for both non-transparent and transparent settings. In general, the Choice Architects choose the nudge that they expect to the most effective in making people choose the high target. This also coincides with what they believe to be the most manipulative intervention. Hence, I do not find that image concerns play a large role in this setting, or that people would instinctively differentiate between System 1 and System 2 interventions and how manipulative they are.

With the Decision Makers, I find that the default nudge (targeting System 1 thinking) has a large effect on subsequent choices and that a considerable proportion of this effect is eroded by transparency. The risk-benefit analysis nudge (slower System 2 thinking) does not have a consistent effect and this is unchanged by transparency. However, exploratory analysis of the effects by subgroups of confidence and ability shows that the risk-benefit analysis nudge may impact decision-making. For example, among correctly confident, high ability people, the System 2 nudge leads to more people choosing the high target. More importantly, I find that the System 1 nudge loses its effectiveness mainly among those participants who report low confidence and ability levels. These individuals are much more likely to choose the high target with the non-transparent default than with the transparent default. I do not observe similar patterns among the other confidence-ability groups: they are equally like to choose the high target with non-transparent and transparent System 1 nudges.

This paper contributes to the limited literature on the transparent use of nudges. In particular, this project's take on transparency, making Choice Architects' role in the nudging common knowledge, has not been studied before in this literature. Defining transparency in this way is natural in the sense it does not require extensive explanations of what is meant by a nudge (which can be difficult to grasp sometimes) nor does it give any additional nudges, such as positive arguments, for one option over the other. It is assumed to engage critical System 2 thinking, for which reason it is also expected to have an effect on the subsequent decision-making.

There are a few experimental studies in Economics and Psychology that investigate how transparency impacts the effectiveness of nudges, however, in this strand of literature, transparency has many different meanings, including explanations of 1) what a default is, 2) how defaults can influence decisions, and 3) what purpose the nudge serves. The most common finding is that transparency has no impact on the effectiveness of nudges (Hendrik Bruns, Elena Kantorowicz-Reznichenko, Katharina Klement, Marijane Luistro Jonsson and Bilel Rahali, 2018; George Loewenstein, Cindy Bryce, David Hagmann and Sachin Rajpal, 2015; Floor M. Kroese, David R. Marchiori and Denise T.D. De Ridder, 2016; Mary Steffel, Elanor F. Williams and Ruth Pogacar, 2016; Patrik Michaelsen, Lars-olof Johansson and Martin Hedesström, 2020).[3] A few studies find that transparency strengthens the default nudge, however, in these setting, transparency is mainly concerned with the merits of the default option (Sandro Casal, Francesco Guala and Luigi Mittone, 2019; Yavor Paunov, Michela Wänke and Tobias Vogel, 2018). The opposite result is

[3] A further difference is that many of the studies are hypothetical. Michaelsen et al. (2020) find that adding small stakes (0.40 USD) changed how participants felt about nudging, although it did not affect behavior.

found, for example, by a meta-study that shows that awareness of experiment participation lessens the impact of the plate-size nudge (Holden et al., 2016).[4] Most of these studies focus on the default nudge, which can be a particularly difficult behavioral intervention to counteract as defaults work through several mechanisms, making it difficult to cancel all of them at the same time. For example, defaults enjoy the status of a recommendation (McKenzie et al., 2006), act as reference points for loss aversion (Johnson and Goldstein, 2003), and benefit from inertia, procrastination, and present bias (Keith Marzilli Ericson, 2017; Gabriel D Carroll, James J Choi, David Laibson, Brigitte C Madrian and Andrew Metrick, 2009). For example, Casal et al. (2019) try several transparency messages and find that while some have an impact on the default nudge, others do not. Steffel et al. (2016) find transparency interventions to be ineffective but discover effective de-biasing tools that partially counteracted the effects of a default nudge. Therefore, while some ways to be *transparent* have no impact on decision-making, others can fundamentally change how people interpret the situation.

The second contribution of this paper relates to the Choice Architect literature, where transparency has not yet been explicitly studied. A common result in this literature is that cognitive biases appear also in choice architecture decisions. Daniels and Zlatev (2019) find that Choice Architects prefer certain and positive frames; Ambuehl et al. (2021) find that architects often impose on others what they would have chosen for themselves. Altmann et al. (2013) find that Choice Architects are more willing to misguide others through default setting than by direct advice and Blount and Larrick (2000) show that although Choice Architects often choose frames suboptimally, they behave more selfishly in an ultimatum game when they have the power to choose the frame than when they do not. These studies suggests that actions in the Choice Architecture game might be guided by a different set of norms than, for example, actions in general communications games, making it an important area of study.

This chapter is organized in the following way. Section 3.2 defines the Choice Architecture Game and analyses it in a limited attention model. Section 3.3 explains the experimental design, its main treatment, and controls for potential confounders. Section 3.4 presents and discusses the results and Section 3.5 concludes.

3.2 Theory

The Choice Architecture Game has two players: a Choice Architect who designs the choice environment, and a Decision Maker who later operates in this environment. The Decision

[4] When people have a smaller plate or a bowl, a common result is that they self-serve or consume less food. However, when participants are aware that their food habits are being studied, the plate-size effect is much weaker than otherwise Holden et al. (2016).

Maker chooses between two options, A and B. The optimal choice for the Decision Maker is not known as there is some uncertainty. The Architect decides how to present this choice to the Decision Maker by selecting a question formulation out of 3 options: no nudge, a System 1 nudge and a System 2 nudge. Each question formulation provides the Decision Maker with the same two options and the related payoffs, yet, the formulations have an impact on which option the Decision Maker is likely to choose. The architect prefers that the Decision Maker chooses A and choice architecture is the only tool through which they can impact the outcome.

Using backward induction, I first determine the response of the Decision Maker in the case without transparency. I deploy a limited attention model; when the Decision Maker is inattentive she uses mental shortcuts and otherwise she maximizes utility. I use as a benchmark a situation where people's inattentive choice is the option least preferred by the Choice Architect so that the nudges are likely to have more of an impact. System 1 nudges redirect the mental shortcut from this option to the other, while System 2 nudges reduce inattention altogether. Using Decision Makers' responses as an input, I derive the best response of the Choice Architect. The Architect decides which type of nudge, if any, to use. Image-concerned Choice Architect might refrain from using nudges in the fear of looking 'manipulative'. Last, I add transparency and derive hypotheses about how it impacts the game.

3.2.1 The Problem of the Decision Maker

The Decision Maker maximizes her utility by choosing an action, $a \in \mathcal{A}$: $max_a\ U(a)$. Suppose for simplicity that there are only two mutually exclusive options to choose from: $\mathcal{A} = \{A, B\}$. Neither option is optimal for everyone and the choice is made before experiencing the full consequences, that is, it is made over expectations. A decision with uncertainty is interesting because individuals are potentially less certain about their choice and thus more susceptible to influencing and nudging. Introducing only uncertainty keeps the setting otherwise simple; for example, social norms, coordination and information play little role. The lack of universal optimal choice gives nudging a bigger role in the decision-making process, which allows me to study the effects of transparency on nudging itself.

To model boundedly rational behavior, I use a simple limited-attention model.[5] With probability δ, an individual is distracted (inattentive) and follows some simple mental shortcut present in the environment. The shortcut could be an unintended feature in the environment or an intentional nudge planted there by a Choice Architect. It is assumed

[5] People have limited capacity to do rational decisions before getting fatigued or tired. The use of this limited resource could be endogenously determined. For our proposes, however, it is sufficient to have attentiveness endogenously given but affected by two factors, transparency and System 2 nudges, both of which increase attentiveness.

that in the initial environment, denoted by η_0, people tend to choose B inattentively. In general, an attentive decider (noted by $d = 0$) will choose according to his or her preferences and chooses A if $E[U(A)] \geq E[U(B)]$ and B otherwise.[6] An inattentive ($d = 1$) Decision Maker, however, chooses A if there is a cue to choose A, and B if there is a cue to choose B.[7]

3.2.2 Aggregate Perspective

Define ρ as the probability that a Decision Maker holds the expectation $E[U(A)] \geq E[U(B)]$. This is also the proportion of people for which A is the rational choice. The likelihood that a random Decision Maker chooses A in a choice environment η can thus be expressed as:

$$P(A) = \delta P(A|d = 1, \eta) + (1 - \delta)\rho, \tag{3.2.1}$$

where $P(A|d = 1, \eta)$ is the probability that an individual chooses A when she is inattentive ($d = 1$) in the environment η, and δ is the probability of being inattentive (distracted). This approach has similarities with that of Bernheim and Rangel (2009), who divide the choice process into two components, a set of alternatives ($\mathcal{A}$ in our setup) and the ancillary conditions, which I capture with the choice environment η and the rate of inattentiveness δ. Note that the initial environment, η_0 is not necessarily neutral in the sense that people would on average choose according to their true preferences on average. That is, $P(A|d = 1, \eta_0) \neq \rho$. I assume here that $P(A|d = 1, \eta_0) < \rho$, meaning that in the initial environment, people tend to choose B more often than what is optimal. This creates an environment where nudges for A can be effective.

3.2.3 The Problem of the Choice Architect

Assume without loss of generality that the Choice Architect prefers A over B. Hence, the Architect maximizes utility by maximizing the probability that the Decision Maker chooses A:

$$\max_{s_1,s_2} U = \max_{s_1,s_2} P(A) = \max_{s_1,s_2} \delta P(A|d = 1, \eta) + (1 - \delta)\rho. \tag{3.2.2}$$

The Choice Architect may target either the probability $P(A|d = 1, \eta)$ that expresses the direction and the strength of the inattentive decision shortcut or the probability of

[6] Without loss of generality, I break ties in favor of A.

[7] It also a theoretical possibility that no cue is found, in which case, neither A or B by an inattentive Decision Maker and this possibility is firmly part of the choice architecture. Opt-in schemes that automatically sign people in a program lead to a choice even if people do not actively choose to participate or not to participate. On the other hand, some choice opportunities are easily missed due to an architecture that, for example, hides the decision opportunity, making it common that neither A nor B is chosen. This means in general that $P(A|d = 1, \eta) + P(B|d = 1, \eta) \leq 1$. In my experimental setting, however, this concept is not important.

inattentiveness, δ. The two options correspond respectively to a System 1 nudge, s_1, and a System 2 nudge, s_2. Denote an active nudge with $s_i = 1$ and an inactive nudge with $s_i = 0$, for $i = 1, 2$. System 2 nudges aim to enhance conscious and attentive decision-making, while System 1 nudges rely on the opposite happening. Therefore, Choice Architects are assumed to select only one of the two nudges at a time, or neither, making the maximization subject to: $s_1 \cdot s_2 = 0$.

When a System 1 nudge is chosen, the probability of choosing A becomes then:

$$P(A|s_1) = \delta P(A|d = 1, s_1) + (1 - \delta)\rho. \tag{3.2.3}$$

The effect size depends on the System 1 nudge. The strongest System 1 nudge imaginable guarantees that all inattentive individuals choose A: $P(A|d = 1, s_1) = 1$. A weaker nudge has a correspondingly weaker effect. I focus on nudges that I consider to be 'strong' in the sense that they make more people choose A than what attentive people would choose on average: $P(A|d = 1, s_1) > \rho$. However, any System 1 nudge that satisfy $\rho \geq P(A|d = 1, s_1) > P(A|d = 1, \eta_0)$ is potentially useful for a policy maker wanting to change decision outcomes: I call these System 1 nudges 'weak'.

When a System 2 nudge is chosen, inattentiveness becomes $\delta(s_2)$. The probability of choosing A becomes:

$$P(A|s_2) = \delta(s_2)P(A|d = 1, \eta_0) + (1 - \delta(s_2))\rho. \tag{3.2.4}$$

A System 2 nudge is effective when inattentiveness is reduced such that $\delta(s_2) < \delta$, although counterproductive System 2 nudges are also possible. The strongest System 2 nudge imaginable cuts inattentiveness to zero, $\delta(s_2) = 0$, in which case all individuals decide attentively and a proportion ρ chooses A. Therefore, I focus on s_2 nudges for which $0 \leq \delta(s_2) < \delta$.

It is optimal for a Choice Architect to choose a System 1 nudge over a System 2 nudge if: $P(A|s_1) > P(A|s_2)$. Consider a rearranged version of this expression:

$$\delta \left[P(A|d = 1, s_1) - \rho\right] > \delta(s_2) \left[P(A|d = 1, \eta_0) - \rho\right] \tag{3.2.5}$$

This inequality is always true when I assume the following:

1. $P(A|d = 1, \eta_0) < \rho$; the original environment η_0 is such that fewer people on average choose A than they would if all chose attentively.

2. $P(A|d = 1, s_1) > \rho$; the System 1 nudge is 'strong' and causes more people to choose A than they would if all chose attentively.

3. $\delta(s_2) < \delta$; the System 2 nudge reduces inattention.

This is straightforward to see. All δ's are probabilities and thus either positive or zero. By the second assumption, the left hand side of inequality (3.2.5) is weakly positive and by the first assumption, the right hand side is weakly negative. A Choice Architect thus always prefers a 'strong' System 1 nudge to any System 2 nudge.[8]

Image concerns

Some Choice Architects might prefer one type of nudge over the other, or refrain from nudging altogether, because nudging may be considered to be manipulative in the sense that nudging can make people choose against their own interests. While System 2 nudges make Decision Makers, at least *in expectation, follow their rational preferences*, System 1 nudges potentially *mislead* Decision Makers to go against their interests. An image-concerned Choice Architect might dislike using System 1 nudges out of concern of being seen as manipulative. To capture this in a simple framework, I add these concerns as a linear and negative term to utility. Representing image concerns with parameter $\beta \geq 0$, the maximization problem, still subject to condition $s_1 \cdot s_2 = 0$, then becomes:

$$\max_{s_1,s_2} U = \max_{s_1,s_2} P(A) - \beta s_1 \tag{3.2.2'}$$

As a consequence, image concerns enter also inequality (3.2.5):

$$\delta \left[P(A|d=1,s_1) - \rho\right] - \beta s_1 > \delta(s_2) \left[P(A|d=1,\eta_0) - \rho\right]. \tag{3.2.5'}$$

Eq. (3.2.5') shows that if image concerns are sufficiently large, a Choice Architect will use a System 2 nudge instead of the more effective System 1 nudge. If the Choice Architect does not have image concerns, $\beta = 0$, inequality (3.2.5') simplifies back to (3.2.5), which means System 1 nudges are still more effective than System 2 nudges in making people choose A. Note that in this setup, System 2 nudges are not considered manipulative, but the opposite; they are thought of as active decision enhancers.

Choice Architects' other preferences

The Choice Architects may ground their decisions on preferences other than self-interest and image concerns. For example, it is possible that other-regarding preferences, such as altruism, guide some Architects, making them take into account what they believe to be best option for the Decision Makers. There is no reason to believe, however, that such preferences will interact with transparency. For this reason, they are not modeled explicitly.

[8] For 'weak' System 1 nudges the result is ambiguous and will depend on the relative strengths of the nudges and the underlying attention levels and shortcuts present in the initial environment.

3.2.4 Adding Transparency

Being transparent about the fact that the Choice Architects decide on the nudges can have at least three effects on the likelihood of a Decision Maker choosing A. First, transparency calls attention to the decision situation and therefore may reduce inattentive decision-making. Second, aware Decision Makers may want to punish the Choice Architects for the use of manipulative techniques by choosing the option not promoted by the nudge. This impulse is called *reactance* in the psychology literature. Third, revealing to the Decision Maker that the choice will affect another person's payoff might trigger general other-regarding preferences. I formally model only the first effect but discuss the latter two, and how they are accounted for in this study.

Transparency can also lead to the Decision Makers expecting nudges and the Choice Architects considering nudging to be manipulative when the Decision Makers are aware of it. These two effects may *license* the Choice Architects to use stronger methods of influence than what they would have chosen otherwise. I do not formally model this effect here, but take this into account in the experimental design, which I discuss in more detail later.

Effects on Decision Makers

First, consider System 1 nudges and transparency. Define $\delta - \tau$ as the level of inattention under transparency such that as a consequence $0 \leq \tau \leq \delta$. Thus the likelihood of choosing A under a System 1 nudge and transparency is given by:

$$P(A|s_1, \tau) = (\delta - \tau)[P(A|d = 1, s_1)] + (1 - \delta + \tau)\rho. \qquad (3.2.6)$$

This directly leads to the result that System 1 nudges have less of an effect under transparency: $P(A|s_1, \tau) < P(A|s_1)$. Transparency increases attentiveness and hence the System 1 nudge has fewer opportunities to influence decision-making.

Hypothesis 1: Transparency weakens the effect of System 1 nudges.

Second, consider System 2 nudges and transparency. Assume that $\delta(s_2, \tau) \ \leq \delta(s_2) < \delta$. This means that although both transparency and the System 2 nudge reduce inattentiveness, they can also crowd out each others' effects. The combined effect will however be at least as large as the effect of either one alone. Hence, I get that the probability of choosing A with a System 2 nudge and transparency becomes:

$$P(A|s_2, \tau) = \delta(s_2, \tau)P(A|d = 1, \eta_0) + (1 - \delta(s_2, \tau))\rho. \qquad (3.2.7)$$

It follows from $\delta(s_2, \tau) \leq \delta(s_2)$ that the effect is not reduced by transparency: $P(A|s_2, \tau) \geq P(A|s_2)$.

Hypothesis 2: Transparency does not weaken the effectiveness of System 2 nudges.

Effects on Choice Architects

A Choice Architect will use a System 1 nudge rather than a System 2 nudge if the benefits (less costs) from using the fast System 1 nudges outweigh those of the slow System 2 nudges:

$$(\delta - \tau)[P(A|d = 1, s_1) - \rho] - \beta(\tau)s_1 > \delta(s_2, \tau)[P(A|d = 1, \eta_0) - \rho] \tag{3.2.8}$$

Transparency may interact directly with the image concerns, $\beta(\tau)$, as there is now more audience to the decision, but this is not the only channel. Using notation $\delta(s_2, \tau^-) = \delta(s_2, \tau) - (\delta - \tau)$ and the fact that $\delta(s_2, \tau^-) \geq 0$, I can rewrite (3.2.8) to get a new expression for inequality (3.2.5'):

$$(\delta - \tau)[P(A|d = 1, s_1) - P(A|d = 1, \eta_0)] - \beta(\tau)s_1 > \delta(s_2, \tau^-)[P(A|d = 1, \eta_0) - \rho] \tag{3.2.9}$$

Transparency does not change the result that 'strong' System 1 nudges are more effective than any System 2 nudges in getting Decision Makers to choose A for two reasons. First, transparency crowds out the effectiveness of the System 2 nudge and diminishes its marginal impact. Second, unlike System 1 nudges, System 2 nudges are limited by individuals' true preferences. The left hand side of eq. (3.2.9) remains weakly negative. However, as the impact of System 1 nudges is also reduced by transparency, it means that for some image-concerned Choice Architects, the benefits might no longer be large enough to counteract the image costs, $\beta(\tau)s_1$, making the left hand side of the inequality (3.2.9) also negative, and potentially breaking this inequality.

Denote the lowest level of image concerns needed for the Choice Architect to prefer a System 2 nudge over the more powerful System 1 nudge by $\underline{\beta}$ and $\underline{\beta(\tau)}$ for the non-transparent and transparent case, respectively. With inequalities (3.2.5') and (3.2.9), we can derive expressions for each of them:

$$\underline{\beta} = \delta[P(A|d = 1, s_1) - \rho] - \delta(s_2)[P(A|d = 1, \eta_0) - \rho] \tag{3.2.10}$$

$$\underline{\beta(\tau)} = (\delta - \tau)[P(A|d = 1, s_1) - P(A|d = 1, \eta_0)] - \delta(s_2, \tau^-)[P(A|d = 1, \eta_0) - \rho] \tag{3.2.11}$$

Using our previous set of assumptions 1-3, we derive the result that: $\underline{\beta(\tau)} < \underline{\beta}$. Choice Architects with $\hat{\beta}$ such that $\underline{\beta(\tau)} \leq \hat{\beta} < \underline{\beta}$ thus use System 1 nudges in the non-transparent case, and System 2 nudges when there is transparency.

Hypothesis 3: Transparency is expected to reduce the use of System 1 nudges.

Hypothesis 4: Transparency is expected to increase the use of System 2 nudges.

3.3 Experimental Design

The predictions are tested with an online experiment. The software is coded on oTree (Chen et al., 2016) and the participants are recruited via Prolific.[9] I first enroll a set of Choice Architects to make their nudging decisions and subsequently implement these decisions on a set of Decision Makers. The main treatments are administered between subjects. To study transition patterns, all Choice Architects initially in the non-transparent treatment receive later on a delayed within-subject transparency treatment, as explained in more detail below.

The Prolific participant pool consists of more than 100 thousand participants located primarily in the United States and the United Kingdom (see Palan and Schitter, 2018). Prolific's participant pool is more diverse than the typical subject pool of university laboratories (Peer et al., 2017).[10]

3.3.1 An Overview of the Experiment

The Decision Makers do a series of real effort tasks. In the first round, they face simple piece-rate incentives, but for the second round, they get to choose a target. Specifically, each Decision Maker is asked to choose between two performance targets: a high target with high reward but more risk and a low target with lower reward but also considerably lower risk. The Choice Architect chooses how this question is to be presented to the Decision Maker. There are 3 options: a simple design (the benchmark); a design with a strong default that appeals to System 1; and a risk-benefit analysis design appealing to System 2. Table 3.3.1 lists the different formulations. Each formulation asks the Decision Maker to choose between the same two performance targets. The Choice Architect gains an additional bonus of five dollars if and only if the Decision Maker chooses the high target. The only way that the Architect can influence this choice is through the setting the question formulation.

The main treatment, transparency, alerts the Decision Makers to the fact that another participant, "Player B", has the power to choose the question formulation. To not use deception, I disclose both in the transparency treatment and its non-transparent control that the Decision Maker's actions impact the payoff of "Player B". The Choice Architects are

[9] https://www.prolific.co/

[10] Descriptive statistics and demographics collected in the experiment are summarized in Appendix Table 3.B.2.

also informed on what the Decision Maker knows. See Table 3.3.2 for the announcement texts.

The choice of a performance target was selected as the setup for a few reasons. First, the task is private and individualistic. A true nudge in this setting (i.e. without the incentives placed there by the experiment) is there a "paternalistic" nudge, as opposed to a "market" nudge, using Sunstein's (2016) classification.[11] Resistance and reactance, as discussed below, are expected to occur with paternalistic nudges more often than with market nudges, making paternalistic nudges an interesting setting to study the effects of transparency. Market nudges are complicated furthermore by the fact that they often target group behavior rather than individual behavior. Even if transparency neutralizes how a nudge exploits cognitive biases, the nudge might still be the only common signal to coordinate on with the group, making market nudges particularly difficult to neutralize in any setting, as is pointed out by Casal et al. (2019).

Secondly, there is no unambiguous optimal choice in this setting. What is best for the Decision Makers depends on personal factors such as ability and resilience. The Decision Makers are more informed about these factors than the Choice Architects. The Choice Architects, on the other hand, know more about the overall impact of this decision, in particular, that the Choice Architects are better off when the Decision Makers choose the riskier option. Similar situations happen in real life, for example, in entrepreneurship, innovation, and research and development, where individuals do not necessarily take on the socially optimal amount of risk. That the Decision Maker's optimal choice is not knowable may encourage Choice Architects' use of nudges in self-interest.

Third, the setting allows me to control the role of identity. Arad and Rubinstein (2018) find with some countries that attitudes are more negative when nudges are set by government compared to situation where the same nudge is set by an employer, also when people are in general agree with the goals of the nudge. In this study, a Choice Architect is identified only as "Player B", which does not provide much information about the identity. The Decision Makers do not learn much about the Architects. This is by design, as Altmann et al. (2013) show in their study that interest alignment and the informational advantages explain to a large extent whether nudges are followed or not. I collect the Decision Makers' attitudes towards the Choice Architects after they have made their choices, to control for these factors.

Choice Architects

Aside from basic instructions, Choice Architects receive a demonstration in general terms of how different question presentations impact choices. Namely, I tell them the results of

[11] Paternalistic nudges protect individuals from their own mistakes and market nudges protect individuals from market failures (f. ex. externalities, coordination problems, and prisoner's dilemmas).

Table 3.3.1: The three formulation options (nudges) that Choice Architects choose from

Nudge	Text:
Simple	Your earnings depend on a target. You can choose your target: ○ Receive 4 dollars if you answer at least X sums correctly, the same result as you had in Part 1. ○ Receive 6 dollars if you answer at least X + 2 sums correctly, two more than you did in Part 1.
Default	Your earnings depend on a target. You will receive 6 dollars if you answer at least X + 2 sums correctly, two more than you did in Part 1. Alternatively, you can choose to receive 4 dollars if you answer at least X sums correctly, the same as you did in Part 1. Choose one: ○ Switch to the target of X correct sums for 4 dollars. ○ Keep the target of X + 2 correct sums for 6 dollars. (This choice is pre-selected in the software)
Risk-Benefit	Sometimes taking risks is worth it! Before you answer the question below, please consider your real chances of success. Note that you can still choose as you wish and that responding to the following question is voluntary and does not affect your earnings in any way. Which option is the best for you? ○ The lower target is the best in terms of risks and benefits. ○ The higher target is the best in terms of risks and benefits. It is, of course, advisable to follow your own risk-benefit analysis Your earnings depend on a target. You can choose your target: ○ Receive 4 dollars if you answer at least X sums correctly, the same result as you had in Part 1. ○ Receive 6 dollars if you answer at least X + 2 sums correctly, two more than you did in Part 1.

Notes: X refers to the earlier performance in the same task in Part 1. If earlier performance was 0, the lower target is set to 1 instead of 0. The higher target is still set to 2 (that is, +2). With the default, the high target is already preselected by the program when participants enter the choice page. It is voluntary to answer the extra question with the risk-benefit analysis nudge.

Table 3.3.2: Transparency announcement and the non-transparent counterpart

	Announcement text
Transparent	There are 3 possible presentations of the next question. The options and their consequences are the same in all presentations. They are, however, worded and structured differently, and in a way that has been shown to affect people's choices. Another participant, Player B, chose which of the three presentations is shown to you. Depending on your choice in this part, Player B may also get some money. Player B has also seen this message.
Non-transparent	Depending on your choice in this part, another participant, Player B, may also get some money. Player B has also seen this message.

Notes: In the transparent case, the participants need to check a box to indicate they have read and understood the message. This is done in order to make sure all the participants receive the transparency treatment when assigned to it. The non-transparent text is not highlighted in any way and does not require checking a checkbox.

an earlier pilot, done in a different country and in a different language, that demonstrates how one presentation led to 25 percent and another led to 57 percent of the people choosing the intended target, without specifying what these presentations were. They then learn the relevant transparency or control announcement and choose one frame out of the three available to be shown to the Decision Maker (see Table 3.3.1). To guide Choice Architects on their decision making, I give them the following information with each nudge: with the simple, I tell that in a related pilot study on Prolific, 30% of participants chose the high target. With the default, I highlight that the high target is pre-selected by the software already when the players enter the page. The risk-benefit analysis framing states that it is expected that people use about twice as much time on the decision with this formulation than without.[12] The choice architecture options are presented to the Architects in a random order. The Choice Architects receives a five-dollar reward if the Decision Maker chooses the higher target. The Decision Makers are not aware of this reward scheme or of the fact that in the aggregate, risk-taking is socially optimal, and optimal for them conditional on there being a high chance that the higher target can be reached. After the choice architecture decision, I elicit the Architects' beliefs about the effectiveness and manipulativeness of each nudge.

The three frame options, 1) simple, 2) default and 3) risk-benefit analysis, represented in Table 3.3.1, have been selected so that each nudge operates primarily through one system (either System 1 or System 2), has a strong impact on decision making, promotes the same (high) target, keeps information constant, and is salient in that it would be obvious to those subjected to the frame what the nudge is and which option it is promoting. The lower target is intentionally listed first for each frame to invoke the order effect. It is common that people choose the item listed first in a list. It is thus expected that the simple formulation promotes the lower target, creating an initial environment η_0, in which the choice are suboptimal from the Choice Architect's point of view.

Licensing among Choice Architects

Transparency about nudging can also make Decision Makers expect manipulations and thus make nudging more acceptable in general. This expectation can *license* some Choice Architects to use stronger methods of influence than what they would have chosen otherwise. As the Decision Makers are aware of nudging, the action of nudging itself becomes less manipulative, which as such can also encourage nudging.

If licensing is common, we might want to distinguish between intensive and extensive margins of nudge using. The extensive margin looks at how many Choice Architects

[12] These sets of information are based on two small pilots ran on the Prolific platform before the main experiment. The two pilots comprised of 28 and 40 participants in the role of the Decision Maker, split across the 6 treatment cells, and 18 and 17 Choice Architects split across the 2 treatments.

use a nudge compared to those that do not use one at all. The intensive margin looks at what type of nudge is chosen when nudges are at use. Hypothesis 3 and 4 should still apply to the intensive margin, that is, a Choice Architect is expected to choose a System 2 nudge with higher probability under transparency than without transparency. This is conditional on the result that the Choice Architect uses nudges in the first place. The licensing effect, on the other hand, is expected to dominate the extensive margin, in essence, to increase the use of nudges in general when their use is transparent.

Delayed transparency treatment

To study the effects of transparency as it is introduced, all those Choice Architects that are assigned to the non-transparent control group receive the transparency treatment later in a delayed within-subject treatment. These participants repeat the Choice Architects' task , but on the second time, they are matched with Decision Makers that are in the transparency treatment. The task remains the same otherwise, including the rewarding scheme and the menu of nudges. The additional treatment aims at understanding better if transparency makes Choice Architects change the nudge that they use (intensive margin), or change whether they use nudges at all (extensive margin). Within-subject reactions to transparency reveal detailed information on the transitional patterns, making it possible to distinguish different kinds of direct reactions to transparency. This reveals specifically if Choice Architects switch from System 1 nudges to System 2 nudges with transparency, or if they start or stop using nudges. However, because of inertia, I expect the *within subjects* treatment effects to be weaker than those in the *between subject* treatment.

Decision Makers

For Decision Makers, the core of the experiment consists of two rounds of a real effort task. In the first round, the Decision Makers gather experience on the task and establish a performance level. Before they repeat the task in the second round, Decision Makers choose between two performance targets that are determined by the performance in the first round. Choice Architects decide how this question over the two performance targets is presented to the Decision Makers. We are interested in how the nudges in the question presentation affect the choice of the performance target in the transparent and the non-transparent cases.

I use the real effort task (RET) developed by Weber and Schram (2017). In each round, participants have 7 minutes to complete as many exercises as they wish. An exercise consists of two 10-by-10 matrices made of randomly generated numbers. The goal is to find the largest number in each matrix and then to sum the two together. This sum is the answer that the participants need to submit. Before the task starts, the participants get to test the task and the software on the instructions page. An example of this task can be found with the participant instructions in Appendix 3.C. On average, participants answer 5 summations correctly in 7 minutes.

The first time the Decision Makers face the RET, they receive 30 cents for every correct answer. In the second round, their payment depends on the chosen target. The low target rewards 4 dollars conditional on repeating earlier performance. The riskier high target rewards 6 dollars conditional on improving the earlier performance by 2 correct answers. If the chosen target is not reached, the participant receives 0 dollars. If he or she reaches the higher target but chose the lower target, they receive the lower reward of 4 dollars. In the case that their earlier performance was zero, the lower target is set to 1 correct answer instead of zero correct answers.

I collect data also on Decision Makers' risk preferences and confidence in being able to achieve different performance targets. The full timeline of the experiment is summarized in Table 3.3.3. Risk preferences are measured with the Bomb Risk Elicitation Task (BRET; Crosetto and Filippin, 2013) administered at the beginning of the experiment. This is followed by instructions for the real effort task (RET) and the first round of the RET with the piece-rate incentives. Afterwards, I elicit the Decision Makers' beliefs in being able to reach different target levels. To incentivize the elicitation, but to not confound it with the main interest of the study (the second round choice between the targets) I add a third round of the same task at the end of the experiment. This belief elicitation concerns the performance in this third round of the RET. I explain each new part in more detail below.

Table 3.3.3: Summary of the timeline for Decision Makers

Part	Description
Part 0	Risk preference elicitation task
Part 1	RET 1: 30 cents per correct answer
Part 2	Elicitation of confidence in reaching certain performance levels
Transparency/Control Announcement	
Choice	High or low target
Part 3	RET 2: with the chosen target
Part 4	RET 3: 30 cents per correct answer or based on the confidence task (Part 2)

Notes: RET: Real Effort Task. When introducing the RET for the first time, the participants are told that they will repeat the task twice more.

Risk preference elicitation task. BRET by Crosetto and Filippin (2013) consists of 100 boxes: 99 of the boxes contain 1 cent and 1 box contains a bomb. An individual chooses how many boxes to collect, $choice \in [0, 100]$. Then, the bomb is randomly positioned in one of the hundred boxes. If the choice is less than the bomb's position, the individual earns $choice \cdot 0.01$ dollars. If the choice is greater or equal to the bomb's position, the earnings are 0 dollars. The *risk neutral* action is to open half of the boxes, *choice* = 50. Results for this task are revealed in the end of the experiment.

RET 1. The first round of the real effort task introduces the participants to the task and measures their baseline performance levels. It rewards participants 30 cents for every correct answer given in the 7 minutes. Wrong answers have no effect on the payment. Afterwards, participants learn how many tasks they did correctly, how many they attempted, and how much they earned from this part.

Elicitation of confidence. To determine how confident the participant is in reaching particular targets, I ask whether they prefer receiving 3 dollars by reaching a performance target of X correct answers or by participating in a random lottery with Y% probability of winning. The targets (X) go from 1 to 10 correct answers and the probabilities (Y) are 20%, 40%, 60%, and 80%. In total, there are 40 target-probability combinations and thus 40 questions. There is a 50% likelihood that one of these 40 questions is picked to be the basis of payment for the third round of the real effort task. If this is the case, the person will be paid by her chosen method. With the remaining 50% likelihood, the participant earns 30 cents for every correct answer. Participants learn the randomly chosen payment method at the end of the experiment. When eliciting the confidence levels, I imposed some rationality requirements. A person may switch from a performance target to a lottery only once per a target level. This helps to interpret the results as the level of confidence in reaching that target. For example, if the person chooses the performance target at probability levels of 20% and 40%, but the lottery at 60% and 80%, we interpret that she is 40-60% confident in her ability to reach the target. Confidence must also be weakly decreasing over the target size, as in order to reach a higher target, one must have reached all of the lower ones before it. Participants were not able break these two rules when choosing between the targets and the lotteries.

RET 3. The third and last real effort task round is rewarded either with a 30 cent piece-rate or according to a randomly chosen confidence elicitation question. The method of payment is revealed only after the round has ended. This way, the participants are given incentives to accurately report their confidence levels and to perform well in the third RET independent of the answers given in the earlier confidence elicitation task.

Reactance and social preferences among Decision Makers

Decision Makers can react to transparency about nudging also aside the effects on attentiveness„ although the theoretical model above does not allow for this. One example is *reactance*, where the individual chooses the option not promoted by the nudge regardless of her true preferences (Sharon S. Brehm, 1966). If reactance is a common reaction to transparent nudging, we will observe that "B" or the lower target is chosen more often with transparent nudges than with the non-transparent ones. However, in a previous study by Arad and Rubinstein (2018), only a minority of people react in this way. Thus it might be difficult to observe reactance on top of (or separate to) other reactions. To control for this, I record decision makers' attitudes towards the Choice Architects in the experiment and their beliefs about the interests of the Choice Architect. This allows me to check if transparency changes these attitudes and beliefs, especially when nudges are at play.

On another hand, other-regarding preferences may come into play when the presence of Choice Architects is made salient to the Decision Makers. For example, if the Choice Architect is considered an ally, other regarding preferences might increase the choice of an option that the Decision Maker believes will increase the Choice Architect's utility. In the experiment, I make Decision Makers aware of the other player both in the treatment and in the control, thus reducing the potential importance of this effect.

Extra belief elicitation after the main study

At the very end of the experiment, I collect information on certain beliefs to understand better the motives behind the decisions. All of these questions are unincentivized.

With the Decision Makers, we are particularly interested in reactance. The Decision Makers are reminded of the question that they faced earlier and are then asked the following questions:

1. Which choice do you believe to have benefited Player B more? (The options are: Higher target, lower target, or I don't know.)

2. On a scale from 1 to 7, one meaning negative, four meaning neutral, and seven meaning positive, what are your feelings towards Player B?

3. On a scale from 1 to 7, one meaning dissatisfied, four meaning neutral, and seven meaning satisfied, how satisfied are you with your choice of target in Part 3?

4. On a scale of 1 to 7, one meaning not at all, four meaning I don't know, and seven meaning by a lot, how much do you feel that your choice was affected by the way the question was formulated?

With the Choice Architects, we are interested to see if transparency led to licensing or image concerns (or both) aside from possible reducing effectiveness of nudges. Therefore, each question formulation is reshown to the Choice Architects, who are then asked the following two questions:

1. What is the percentage of people that you expect to choose the higher target when the question is formulated as above?

2. On a scale from 1 to 10, one meaning 'Fully agree' and 10 meaning 'Fully disagree', how much do you agree with the statement: 'It is manipulative to formulate the question as above.'?

I collect these beliefs right after the first choice architecture decision to ensure comparability between those that make only one architecture decision and those that make two decisions. This means that Choice Architects in the extra treatment answer the questions after the non-transparent first round, and not after the second round where they get the delayed within-subject transparency treatment.

3.4 Results

The goal was to recruit 190 Choice Architects and 420 Decision Makers from the United States via Prolific.[13] In total, 190 Choice Architects and 445 Decision Makers were recruited. Both Choice Architects and Decision Makers were required to pass enough attention checks, that is, they were not allowed to fail more than 1 fair attention check.[14] The average earnings were 10.7 USD for Decision Makers for an estimated 35-minute task and 2.6 USD for Choice Architects for an estimated 7-minute task. Choice Architects were recruited among BSc degree holders due to the abstract nature of the task, and an approval rating of at least 90% on Prolific was required.[15] The study was pre-registered at AEA RCT Registry (Kujansuu, 2020).

In the analysis, the main statistical test is a non-parametric permutation t-test (PtT-test). This type of test is more high-powered than the standard t-test (Moir, 1998; Schram et al., 2019). The use of this test was not outlined the pre-analysis plan. For this reason, t-test results are also reported when their interpretation is different.

3.4.1 Choice Architects

The dataset consists of 190 Choice Architects. 95 of them first experience the non-transparent setting and then the transparent setting in an extra round of the experiment. The remaining 95 experience only the transparent setting once. Observable characteristics are evenly balanced across these two groups, indicating that random treatment assignment was successful. Table 3.B.1 in the Appendix 3.B summarizes the observable characteristics.

Primary outcomes

Figure 3.4.1 captures the main results for the Choice Architects. The left-hand panel reveals that 17% of the Architects choose the simple question formulation in the non-transparent case. A majority of the participants, 54%, choose the default nudge, while the second most popular choice is the risk-benefit analysis nudge, chosen by 29% of the participants. The differences are significant and persistent: we find the same patterns

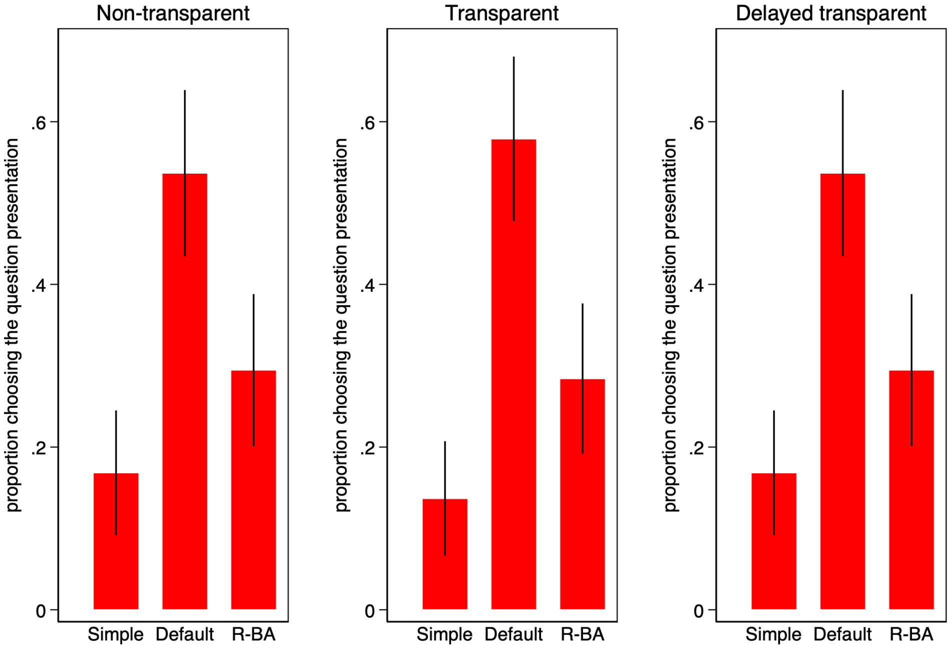
[13] See the power analysis in Appendix 3.A.

[14] See Appendix 3.B.3 for data on rejections and returned studies due to attention check failures.

[15] Participants submissions on the platform can be rejected if the participant does not follow the guidance sufficiently. Prolific allows to pre-screen people by the percentage of approved studies.

also in the transparent setting, as shown in the middle panel of Figure 3.4.1.[16] With transparency, 14% of the Architects choose the simple presentation, 3 percentage points less than in the non-transparent case, 58% choose the default nudge, up by 4 percentage points, and 28% choose the risk-benefit analysis nudge, down by 1 percentage point. None of these differences between the non-transparent and transparent cases are significant (chi-squared test: $p = 0.787$). Transparency thus has no effect on how the Choice Architects use nudges.

Figure 3.4.1: Nudging choices by Choice Architects



Notes: There are 95 Choice Architects in each treatment (non-transparent, transparent and delayed transparent). Simple is the baseline, default is a System 1 nudge and risk-benefit analysis (R-BA) is a System 2 nudge. The 95 participants in the non-transparent setting are the same as the 95 participants in the Delayed transparent setting.

The delayed transparency treatment captured in the right-hand panel of Figure 3.4.1 gives similar results. This treatment was administered as an additional second round for those that were originally assigned to the non-transparent setting, meaning that left-hand and right-hand panels represent the same pool of participants. 23% of the Choice Architects in this group choose the simple presentation, 6 percentage points more than in the non-transparent case, 46% choose the default nudge, 8 percentage points less than in the non-transparent case, and 31% choose the risk-benefit analysis nudge, a 2 percentage point increase compared to the non-transparent case. None of these differences are significant (chi-squared test: $p = 0.477$). Movements were the following. Out of the 16

[16] All relative risk ratios are significantly different at 0.05 level with the exception of the relative risk ratio between non-transparent simple and non-transparent risk-benefit analysis which is significant only marginally at $p = 0.074$.

people that chose simple in the first round, 50% of them reselected it in the second round, 31% switched to default and 19% switched to risk-benefit analysis. Out of the 51 people that chose the default in the first round, 63% stayed with the default, 18% chosen simple instead, and 20% chose the risk-benefit analysis nudge. Out of the 28 people who chose the risk-benefit analysis nudge in the first round, 58% of them chose the same nudge also in the second round, 18% switched to simple and 25% switched to default. There seems to be no clear reaction patterns – most people kept the same nudge as before, and quite equal amounts switched to the two alternatives – indicating that we do not find image concern or licensing patterns in the data.

Result 1: The most frequently chosen question presentation was the default nudge, which, considering the theoretical predictions, was expected to be the most impactful presentation.

Result 2: Transparency does not impact Choice Architects' use of nudges.

Secondary outcomes

To better understand the Choice Architects' decisions, I collected data on the participants' expectations and beliefs; how many people they expect to choose the high target with each architecture option and how manipulative they believe each nudge to be. Table 3.4.1 reports the average expectations by treatment and what nudge the Choice Architect chose earlier. Note that with the simple question presentation, people's average estimates are high compared to the information given to them that about 30% of the participants chose the high target in the simple setting in a related pilot, on Prolific. With the default and risk-benefit analysis nudges, people expect the chosen nudge to be the most effective one of the three. Transparency does not affect these evaluations. Among the people who chose the default nudge, the average expectation was that about 55-58% would choose the high target with the default. This is significantly higher than the other expectations, by a margin of 10-21 percentage points. Among the risk-benefit analysis choosers, the chosen nudge is expected to be 8-13 percentage points more effective than the other two options. Again, the differences are significant. This pattern breaks down with those that chose the simple question design: pooled across non-transparent and transparent cases, the average expectations are 49% for simple, 49% for default and 46% for risk-benefit analysis nudge, none significantly different from the other. This result of no difference indicates that either those choosing the simple formulation do not believe nudges have any effect, or they have a harder time analyzing the situation and what would be efficient, or that they use some other criteria in making their decision, for example, avoiding manipulation.

Average judgments of manipulativeness are reported in Table 3.4.2. Manipulativeness is rated on a scale from 1 to 10 to measure agreement with the statement "It is manipulative to formulate the question as above." For presentational purposes, I have inverted the

Table 3.4.1: Average estimates of effectiveness by nudge choice

		Expectation (%) for:		
Nudge choice:		Simple	Default	Risk-benefit analysis
Simple	Non-transparent	47	46	45
Simple	Transparent	52	53	47
Default	Non-transparent	42	58**	37
Default	Transparent	45**	55**	38**
R-B Analysis	Non-transparent	48	52	60**
R-B Analysis	Transparent	43	47	56*(*)

Notes: How many people are expected to choose the higher target with a given nudge? Simple is the baseline, default is the System 1 nudge and risk-benefit analysis is the System 2 nudge. ** indicates that the value is significantly different from the other two on the same row at 5% level, tested with PtT-tests and t-tests, and *(*) signifies marginal significance with PtT-tests and significance with the t-tests at 5% level of significance.

original scale such that the numbers can be interpreted intuitively: a high score means that the nudge is rated high in manipulativeness.

Table 3.4.2: Manipulativeness ratings by architecture choice

		Manipulativeness score for:		
Nudge choice:		Simple	Default	Risk-benefit analysis
Simple	Non-transparent	3.8**	5.2	5.9
Simple	Transparent	4.5	5.1	5.6
Default	Non-transparent	3.0**	6.1**	5.0**
Default	Transparent	3.5**	6.3**	5.3**
R-B Analysis	Non-transparent	2.9**	4.5**	5.7**
R-B Analysis	Transparent	4.8	5.2	5.5
Averages:				
Pooled	Non-transparent	3.1**	5.5	5.4
Pooled	Transparent	4.0**	5.8	5.4

Notes: The scale is from 1 to 10, by whether they agree with the statement "It is manipulative to formulate the question [with this nudge]." The higher the number, the more they agree that the nudge is manipulative. Simple is the baseline, default is a System 1 nudge and risk-benefit analysis (R-BA) is a System 2 nudge. ** indicates that the value is significantly different from the other two on the same row at 5% level, tested with PtT-tests and t-tests. In the pooled results, the rating for simple is significantly higher in the transparent case than in the non-transparent one ($p < 0.01$ with PtT-test).

The two first rows of Table 3.4.2 report the manipulativeness ratings among those who chose the simple question formulation. In the non-transparent case, the simple presentation gets a rating of 3.8, which is significantly lower than the ratings for default, 5.2, and the ratings for risk-benefit analysis, 5.9. In the transparent case, the ratings for the different formulations are not significantly different from each other, and in general the margins are smaller among those choosing the simple framing than among those using the nudges. These results support both the idea that those choosing the simple framing had a hard time differentiating the question formulations in terms of effectiveness and manipulativeness (in the transparent case), and that those choosing the simple might prefer less manipulative nudges (in the non-transparent setting). Of course, I cannot exclude the possibility that they use other criteria.

Among participants that chose one of the two nudges, people tended to choose the most manipulative nudge. The average scores for manipulativeness are 3.0 for simple, 6.1 for default, and 5.0 for risk-benefit analysis for those who chose the default nudge in the non-transparent case. The results are similar with transparency, and the differences are significant in both cases. Among those who chose the risk-benefit analysis in the non-transparent case, the ratings are 2.9 for the simple, 4.5 for the default, and 5.7 for the risk-benefit analysis. The ratings are different from each other in the non-transparent case, but the differences disappear with transparency. All in all, the nudging architects choose on average the most manipulative tool by their own judgment, indicating that Choice Architects do not shy away from potentially manipulative methods. Interestingly, I do not find evidence that the participants would consider slow System 2 nudges as less manipulative than fast System 1 nudges, only the simple question formulation is considered less manipulative. Therefore, image concerns, if they were to appear, would not move people towards the use of System 2 nudges but towards the simple question formulation.

Result 3: Those Architects that choose to nudge selected, on average, the most manipulative and most effective nudge based on their own evaluations.

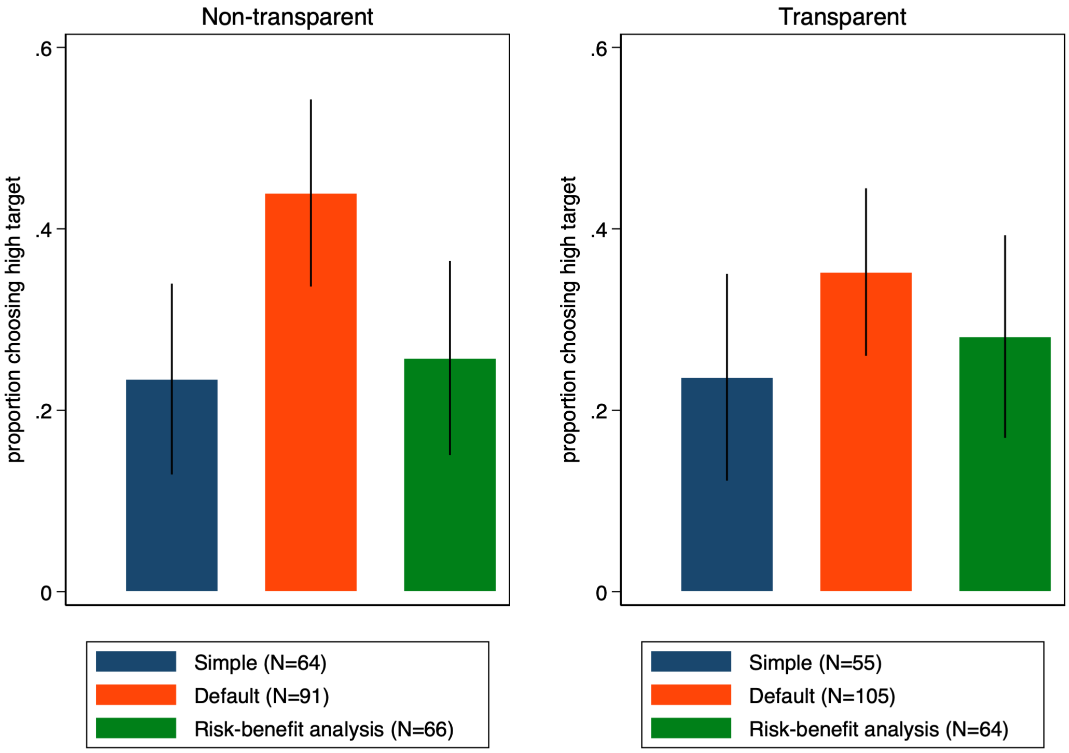
3.4.2 Decision Makers

The dataset consists of 445 Decision Makers, split across the 6 treatment-nudge combinations such that there are 55-105 observations per cell.[17] The data is balanced across age, income, labor market roles and family characteristics. It is not balanced in all treatment-nudge cells with respect to gender and Prolific specific experience. More information is provided in the Appendix Table 3.B.2.

I use the non-transparent simple question frame as the benchmark for two reasons. First, it is not clear how transparency alone affects individuals' decisions when there is no nudge in place and no good measure of true preferences, making the size and direction of the effect difficult to predict. Second, transparently doing 'nothing' is a somewhat empty action when it is not specified what could have been done instead, therefore making the transparent simple a strange special case.

[17] The number of observations per cell is based on the Choice Architects' earlier decisions that are not split evenly across the treatment cells. A major concern identified in the pre-analysis plan was that some cells would not have a sufficient amount of observations for statistical analysis. Therefore, the plan was to recruit 20-40 extra participants in each cell. For budget reasons, this plan was later adjusted downwards with the transparent nudges as they had large sizes to begin with. Only 6 and 8 extra were recruited to the transparent default and the risk-benefit analysis cells, respectively.

Figure 3.4.2: How many Decision Makers choose the high target?



Notes: The proportion of people choosing the high target for each question presentation, in non-transparent and transparent settings. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. There are 55-105 observations in each bin, the exact number for each bin is report below in the legend. The black lines represent the 95% confidence intervals for the mean.

Primary outcomes

The primary outcome of interest is the proportion of Decision Makers choosing the high target. The main results are reported in Figure 3.4.2. Looking at the left panel, we observe that 23% of the participants choose the high target in the baseline, that is, with the non-transparent simple question framing. The non-transparent default nudge increases this proportion by 21 percentage points to 44% in total. The effect is significant at the 1% level ($p = 0.007$, $n = 155$, one-sided PtT-test).[18] The non-transparent risk-benefit analysis nudge increases the choice of the high target by 3 percentage points. This effect is small and insignificant ($p = 0.463$, $n = 130$, one-sided PtT-test). With transparency, the effect of the default nudge shrinks to 12 percentage points and becomes only marginally significant, but neither is the drop in the default's effect significant.[19] The impact of the risk-benefit nudge is higher with transparency, 5 percentage points,

[18] One-sided tests are used for those hypotheses that have clear directional predictions, as specified in the pre-analysis plan.

[19] Comparing the transparent default to the benchmark gives a p-value of $p = 0.077$ ($n = 169$) with a one-sided PtT-test. Comparing transparent default to the non-transparent default gives $p = 0.135$ ($n = 196$), also with a one-sided PtT-test.

but it remains insignificant both in comparison to the benchmark and to the transparent risk-benefit analysis nudge.[20] To summarize, the default nudge targeting the fast System 1 thinking has a large effect on behavior, significantly increasing the take-up of the high target. This effect is weakened by transparency such that it is no longer significant at the 5% level. The risk-benefit analysis nudge targeting the slow System 2 thinking, on the other hand, does not have any significant effect on behavior in this setting nor is its impact affected by transparency.[21]

Result 4: The effect of the default is dampened by transparency, however, the difference is not significant.

Result 5: The effect of the risk-benefit nudge is insignificant in general. Transparency has no impact on this effect.

Heterogeneity in primary outcomes

This section, unlike the others, describes exploratory analysis that was not registered in the pre-analysis plan. It addresses the problem that the data does not generally seem to satisfy the model's assumption regarding the inferiority of the initial environment. The environment should be such that there is still room to nudge people towards the high target – however, based on the data, evidence of such room is limited. First, System 2 nudges do not lead to observable changes in behavior. The reason behind this might be that System 2 nudges are, in particular, limited by the true preferences that people hold. Second, these preferences are unobservable, but approachable. With the risk-benefit analysis nudge, I ask people which option they prefer in terms of risks and benefits. Only about 30% of the participants state that they prefer the high target over the low one ($n = 130$), indicating that there is little room for extra persuasion. With perfect hindsight, I can also check how many were able to reach the high target in the second round of the real effort task: 33% of the participants reach the it. The percentage is 52% for those who chose the high target, and it is 25% for those who chose the low target. Each number represents an upper limit for *persuasion.* It is difficult to convince people to voluntarily choose the high target above this level.

With the data collected in this experiment, especially on confidence levels and actual performance, I can identify groups of people who are over-confident, under-confident, correctly confident and correctly insecure and look if these groups respond differently to the nudges. I classify people as high ability if they reach the high target in the second or

[20] Comparison to the benchmark gives $p = 0.342$ ($n = 128$) with a one-sided PtT-test and a comparison with the non-transparent risk-benefit analysis gives $p = 0.924$ ($n = 130$) with a two-sided PtT-test, as the theory does not give a strong prediction.

[21] As a robustness check, I also regress the choice of the high target on the treatment-nudge interactions and all of the controls outlined in the pre-analysis plan. These regression results are reported in Appendix 3.B.5. The regression confirms the results described here.

third round of the real effort task. I include performance in the third round to account for reasonable risk taking – it is not always certain that one succeeds in a given round – as reaching the chosen target in the third round clearly indicates ability. Those not able to reach the high target in either round are classified as low ability.

Recall that participants complete an incentivized task where they estimate their own probability of reaching certain performance levels. Based on these estimates and the assumption of risk neutrality, I classify people as having high confidence if they are estimated to maximize earnings by choosing the high target. Similarly, I classify people as having low confidence if they are expected to maximize earnings by choosing the low target. Last, I define *overconfident* people as those who have high confidence but low ability, *underconfident* people as those with low confidence but high ability, *correctly confident* as those with high confidence and ability, and *correctly insecure* people as those with low confidence and ability. Note that self-reported confidence does not predict later success, in fact, the correlation between confidence and true ability to reach the high target is close to zero and insignificant. However, hindsight is not always available and hence the ability dimension has less policy relevance than the confidence dimension. For this reason, Appendix 3.B.2 reports the results only by confidence level.

Figures 3.4.3 and 3.4.4 report the results for the underconfident and overconfident individuals. Figure 3.4.3 shows that the underconfident people are relatively unaffected by the nudges. The number of observations is low for each bin, for which reason, I will only report the PtP test results. None of the differences are significant in Figure 3.4.3, $p > 0.384$. The contrast is stark when compared to the overconfident people, who seem to be sensitive to both default types, and to the transparent risk-benefit analysis nudge.[22] Interestingly, transparency on the risk-benefit analysis nudge seems to have the opposite effect for these two groups: the over-confident people choose the high target more frequently, while the underconfident people choose it less frequently, making each group worse off overall – the overconfident people should choose the low target and vice versa. However, neither effect is statistically significant ($p > 0.173$, two-sided PtT).

Figures 3.4.5 and 3.4.6 report the results for the people who holding accurate beliefs. Figure 3.4.5 shows how the correctly confident people seem to be affected by all of the nudges, regardless of transparency.[23] Figure 3.4.6 depicts the data for the correctly inse-

[22] The difference between non-transparent simple and non-transparent default is significant at $p = 0.011, n = 42$, between non-transparent simple and non-transparent risk-benefit analysis is insignificant at $p = 0.539, n = 32$, between non-transparent simple and transparent default is insignificant at $p = 0.102, n = 27$, and between non-transparent simple and transparent risk-benefit analysis is significant at $p = 0.024, n = 28$.

[23] The difference between non-transparent simple and non-transparent default is marginally significant at $p = 0.073, n = 38$, between non-transparent simple and non-transparent risk-benefit analysis is insignificant at $p = 0.120, n = 30$, between non-transparent simple and transparent default is significant at $p = 0.024, n = 41$, and between non-transparent simple and transparent risk-benefit analysis is marginally significant at $p = 0.089, n = 27$. Difference between transparent and non-transparent defaults is signif-

Figure 3.4.3: How many chose the high target, underconfident people

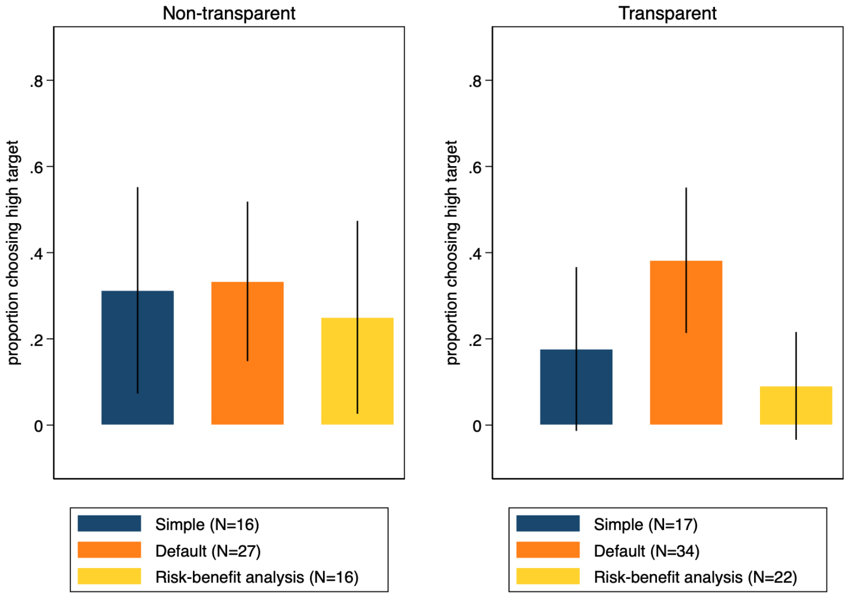
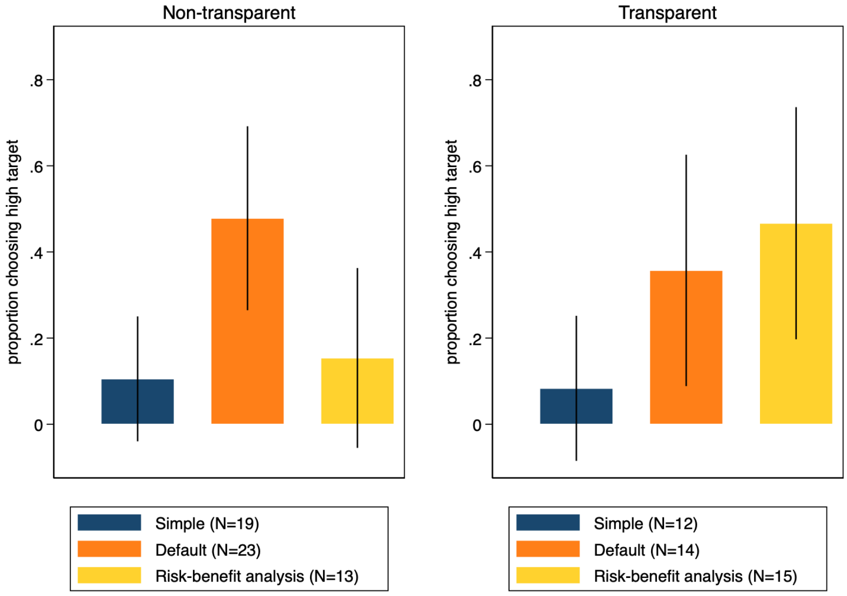


Figure 3.4.4: How many chose the high target, overconfident people



Notes: The proportion of people choosing the high target for each question presentation. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. The number of observations per each bin is reported in the legend, the black lines represent the 95% confidence level.

cure people. Nudges, excluding the non-transparent default, seem to have the opposite effect from the intended one: both risk-benefit analysis nudges and the transparent default

icant at $p = 0.819, n = 51$, similar to that between the transparent and non-transparent risk-benefit analyses, $p = 0.999, n = 29$.

Figure 3.4.5: How many chose the high target, correctly confident people

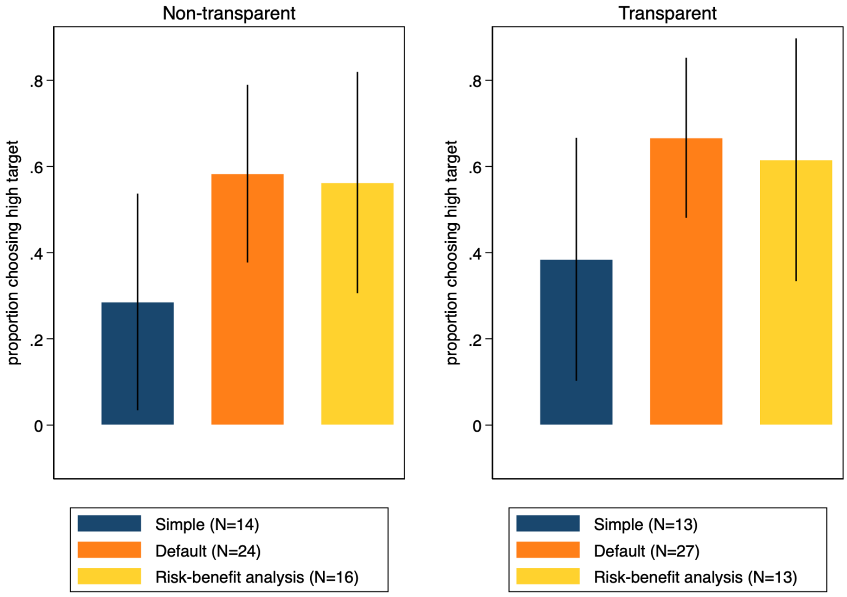
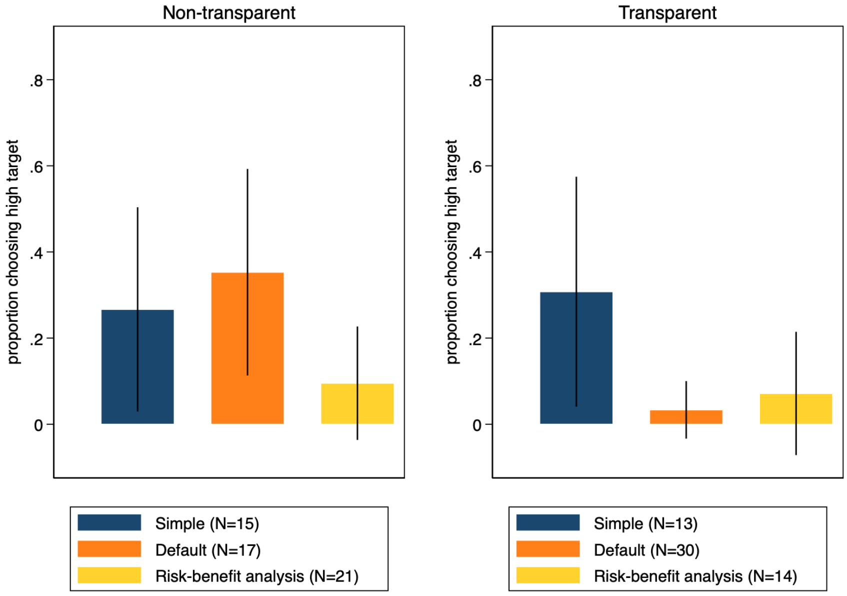


Figure 3.4.6: How many chose the high target, correctly insecure people



Notes: The proportion of people choosing the high target for each question presentation. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. The number of observations per each bin is reported in the legend, the black lines represent the 95% confidence level.

discourage people from choosing the high target.[24] The difference between the transparent and non-transparent defaults is significant at $p = 0.007, n = 47$, while it is not for the risk-benefit analysis, $p = 0.999, n = 35$. For those holding accurate beliefs about their

[24] All the differences with the non-transparent simple (benchmark) are insignificant, $p > 0.437$.

ability, nudges and transparency have positive effects on the decisions on average.

To summarize, I find that nudges and transparency have different effects for different confidence and ability levels. For those holding inaccurate beliefs about their ability (either overconfident or underconfident), transparency strengthens the impact of these inaccurate beliefs through the risk-benefit analysis nudge. However, these effects are not significant. Transparency reduces the effectiveness of the default nudge only among those individuals that hold correct beliefs about their low ability – these people behave more inconsistently when the question formulation is simple or the non-transparent default, while with the risk benefit analysis nudge and the transparent default, their behavior is most consistent. Transparency has no impact on the group of individuals that hold correct beliefs about their high ability.

Finally, I consider whether there are gender differences in how people respond to nudges and transparency. It appears that women's choices are more strongly affected than men's; in fact, women respond exactly in the ways that the model predicts. In contrast, men are on average unaffected by nudges and transparency; this appears to be attributable to (over)-confidence. Men who report low confidence still frequently choose the high target, a tendency that the nudges and transparency reduce. The treatments have the expected impact on men with high confidence, however, looking at all men together, the average effect is close to zero. Note, however, that no gender effects were anticipated by the model or by the pre-analysis plan. For this reason, I collected additional data to test whether they would be replicated. In particular, the plan was to replicate the strong results with women only. It turns out that the results did not replicate (highlighting the importance of preregistration). More information is provided in Appendix 3.B.3.

Result 6: The reduction in the effectiveness of System 1 nudges is driven by people making fewer inconsistent choices.

Result 7: The risk benefit analysis nudge increases the choice of the high target among people who report high confidence levels, while it has the opposite effect on people of low confidence. Conditioning on confidence level, the impact of the risk-benefit analysis nudge is not weakened by transparency.

Secondary outcomes

I first check whether the nudges worked as intended, through fast and slow thinking. It appears that the risk-benefit analysis nudge increases the time spent on the decision. Participants use considerably more time on the choice with the risk-benefit analysis framing than with either of the other two, as argued. Of course, some of this difference is explained by the fact that there is more text to read with the risk-benefit analysis nudge, however, the results suggest that the text length does not explain the increased time usage alone. More details are provided in Appendix 3.B.6.

Next, I consider what the participants report as the best choice for them, as was asked with the risk-benefit analysis nudge, and how closely people follow these reported preferences. More than 90% of the participants answer this question even if it does not affect payments directly. About two thirds prefer the lower target to the higher one, while about 30% prefer the higher target. The preference for the low target is followed more closely than the preference for the high target, indicating that many people who think the high target is better still choose the low target, perhaps due to risk aversion.

Third, I consider questions on beliefs, satisfaction and attitudes. The participants are reminded of the two performance targets and they are then asked, which option they believed to have benefited Player B (the Choice Architect) more. The answers are spread evenly across the three possible options: the higher target, the lower target, and 'I don't know'. In particular, with only one exception, beliefs of high target and low target are equal across nudges, the transparency treatment and gender. This means that people have a hard time interpreting the intentions behind the nudges. The participants were also asked to rate their attitude towards Player B, how satisfied they were with the performance target choice that they made and if they felt affected by how the question was presented. All responses are similar across nudges and transparency treatment. Thus, I do not find evidence of *reactance* in this setup – that negative attitudes towards the nudger would increase with transparency.

Last, neither transparency nor the nudges have a significant effect on the payoffs.

3.5 Conclusions

This project studies how transparency affects the use and effectiveness of nudges while differentiating between two kinds of nudges: fast System 1 nudges that provide quick decision shortcuts and slow System 2 nudges that encourage reflective thinking. By transparency, it becomes common knowledge that Choice Architects may influence Decision Makers' decisions by choosing how a question is presented to them. Transparency is expected to weaken and lessen the use of fast System 1 nudges, while the opposite is expected with the slow System 2 nudges.

The predictions of the model are tested in a framed field experiment, online. The results demonstrate that behavior can be influenced through System 1 and System 2 nudges, and that the impact of System 1 nudges in particular is weakened by transparency. The System 2 nudge, *risk-benefit analysis*, does not have an impact on the average choices. Looking at people different ability and confidence levels, I find, however, that the System 2 nudge does have an impact on the choices made. Transparency does not have a significant effect on the System 2 nudge.

I do not find support for the predictions concerning Choice Architects' behavior. The

Choice Architects do not react to transparency in the experiment. Instead, what I find is that the Architects commonly use the most effective and manipulative nudge as judged by themselves. I find no evidence that participants consider the risk-benefit analysis nudge to be less manipulative than the default nudge.

It is foreseeable that people become more knowledgeable about nudging and choice architecture. It does not matter if this happens through greater exposure to behavioral interventions or by political will that demands nudges to be used more transparently. What this paper shows is that transparency has the power to weaken some nudges, specifically those that rely on fast, automatic thinking processes. This is not necessarily concerning, however, as this study shows that this reduction in effectiveness is mainly driven by reductions in choices that appear to be mistakes. Moreover, transparency is not found to weaken nudges that promote more critical thinking. All in all, these results imply that although some of the earlier nudging successes might become smaller over time, transparency is not a threat to choice architecture.

References

Altmann, Steffen, Armin Falk, and Andreas Grunewald. 2013. "Incentives and information as driving forces of default effects."

Ambuehl, Sandro, B Douglas Bernheim, and Axel Ockenfels. 2021. "What Motivates Paternalism? An Experimental Study." *American Economic Review*, 111(3): 787–830.

Arad, Ayala, and Ariel Rubinstein. 2018. "The People's Perspective on Libertarian-Paternalistic Policies." *Journal of Law and Economics*, 61(May).

Bang, H. Min, Suzanne B. Shu, and Elke U. Weber. 2018. "The role of perceived effectiveness on the acceptability of choice architecture." *Behavioural Public Policy*, 1–21.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics." *Quarterly Journal of Economics*, 124(1): 51–104.

Blount, Sally, and Richard P. Larrick. 2000. "Framing the game: Examining frame choice in bargaining." *Organizational Behavior and Human Decision Processes*, 81(1): 43–71.

Bovens, Luc. 2009. "The Ethics of Nudge." In *Preference Change: Approaches from Philosophy, Economics and Psychology.* , ed. Till Grüne-Yanoff and S.0. Hansson, Chapter 10. Springer.

Brehm, Sharon S. 1966. *A Theory of Psychological Reactance.* New York:Academic Press.

Bruns, Hendrik, Elena Kantorowicz-Reznichenko, Katharina Klement, Marijane Luistro Jonsson, and Bilel Rahali. 2018. "Can nudges be transparent and yet effective?" *Journal of Economic Psychology*, 65: 41–59.

Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics*, 124(4): 1639–1674.

Casal, Sandro, Francesco Guala, and Luigi Mittone. 2019. "On the Transparency of Nudges: An On the Transparency of Nudges: An Experiment." *CEEL Working Paper*, 2-19.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. “oTree – An open-source platform for laboratory, online, and field experiments.” *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Crosetto, Paolo, and Antonio Filippin. 2013. “The "bomb" risk elicitation task.” *Journal of Risk and Uncertainty*, 47(1): 31–65.

Daniels, David P., and Julian J. Zlatev. 2019. “Choice architects reveal a bias toward positivity and certainty.” *Organizational Behavior and Human Decision Processes*, 151(May 2018): 132–149.

Ericson, Keith Marzilli. 2017. “On the interaction of memory and procrastination: Implications for reminders, deadlines, and empirical estimation.” *Journal of the European Economic Association*, 15(3): 692–719.

Felsen, Gidon, Noah Castelo, and Peter B. Reiner. 2013. “Decisional enhancement and autonomy: Public attitudes towards overt and covert nudges.” *Judgment and Decision Making*, 8(3): 202–213.

Grüne-Yanoff, Till. 2018. “Boosts vs. nudges from a welfarist perspective.” *Revue d'Economie Politique*, 128(2): 209–224.

Grüne-Yanoff, Till, and Ralph Hertwig. 2016. “Nudge Versus Boost: How Coherent are Policy and Theory?” *Minds and Machines*, 26(1-2): 149–183.

Hansen, Pelle Guldborg, and Andreas Maaløe Jespersen. 2013. “Nudge and the Manipulation of Choice.” *European Journal of Risk Regulation*, 4(1): 3–28.

Harrison, Glenn W., and John A. List. 2004. “Field experiments.” *Journal of Economic Literature*, 42(4): 1009–1055.

Holden, Stephen S., Natalina Zlatevska, and Chris Dubelaar. 2016. “Whether Smaller Plates Reduce Consumption Depends on Who's Serving and Who's Looking: A Meta-Analysis.” *Journal of the Association for Consumer Research*, 1(1): 134–146.

Johnson, E.J., and D. Goldstein. 2003. “Do Defaults Save Lives?” *Science*, 302(5649): 1338–1339.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Kroese, Floor M., David R. Marchiori, and Denise T.D. De Ridder. 2016. “Nudging healthy food choices: A field experiment at the train station.” *Journal of Public Health (United Kingdom)*, 38(2): e133–e137.

Kujansuu, Essi. 2020. "Choice Architecture and Transparency." *AEA RCT Registry*, #0006308.

Loewenstein, George, Cindy Bryce, David Hagmann, and Sachin Rajpal. 2015. "Warning: You are about to be nudged." *Behavioral Science & Policy*, 1(1): 35–42.

Löfgren, Åsa, and Katarina Nordblom. 2020. "A theoretical framework of decision making explaining the mechanisms of nudging." *Journal of Economic Behavior and Organization*, 174: 1–12.

McCrudden, Christopher, and Jeff King. 2016. "The dark side of nudging: the ethics, political economy, and law of libertarian paternalism." In *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging.* , ed. Alexandra Kemmerer, Christoph Möllers, Maximilian Steinbeis and Gerhard Wagner, 75–139. Hart and Nomos.

McKenzie, Craig R.M., Michael J. Liersch, and Stacey R. Finkelstein. 2006. "Recommendations implicit in policy defaults." *Psychological Science*, 17(5): 414–420.

Michaelsen, Patrik, Lars-olof Johansson, and Martin Hedesström. 2020. "Experiencing Nudges : Autonomy, Intrusion and Choice Satisfaction as Judged by People Themselves."

Moir, Robert. 1998. "A Monte Carlo analysis of the fisher randomization technique: Reviving randomization for experimental economists." *Experimental Economics*, 1(1): 87–100.

Moore, Don A., and Paul J. Healy. 2008. "The Trouble With Overconfidence." *Psychological Review*, 115(2): 502–517.

OECD. 2019. *Delivering Better Policies Through Behavioural Insights.*

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A subject pool for online experiments." *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Paunov, Yavor, Michela Wänke, and Tobias Vogel. 2018. "Transparency effects on policy compliance: disclosing how defaults work can enhance their effectiveness." *Behavioural Public Policy*, 1–22.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research." *Journal of Experimental Social Psychology*, 70: 153–163.

Pronin, Emily, Daniel Y. Lin, and Lee Ross. 2002. “The bias blind spot: Perceptions of bias in self versus others.” *Personality and Social Psychology Bulletin*, 28(3): 369–381.

Schram, Arthur, Jordi Brandts, and Klarita Gërxhani. 2019. “Social-status ranking: a hidden channel to gender inequality under competition.” *Experimental Economics*, 22(2): 396–418.

Steffel, Mary, Elanor F. Williams, and Ruth Pogacar. 2016. “Ethically deployed defaults: Transparency and consumer protection through disclosure and preference articulation.” *Journal of Marketing Research*, 53(5): 865–880.

Sunstein, Cass R. 2016. *The ethics of influence: Government in the age of behavioral science.* Cambridge University Press.

Sunstein, Cass R., Lucia A. Reisch, and Micha Kaiser. 2019. “Trusting nudges? Lessons from an international survey.” *Journal of European Public Policy*, 1763.

Thaler, Richard H. 2015. *Misbehaving: The making of behavioral economics.* W W Norton & Co.

Thaler, Richard H, and Cass R Sunstein. 2008. *Nudge.* Yale University Press.

Weber, Matthias, and Arthur Schram. 2017. “The Non-equivalence of Labour Market Taxes: A Real-effort Experiment.” *Economic Journal*, 127(604): 2187–2215.

Appendix 3.A Power Analysis

In a pilot done in Bologna BLESS laboratory with 124 participants, 25% of the participants among those facing the simple question formulation chose the high target. Nudges increased the take-up rate of the high target by 16-32 percentage points. The weakest nudge called 'recommendation' (16 percentage point increase) was replaced with the current 'risk-benefit analysis' due to inconsistencies in the time spending patterns and due to the overall weak impact. The 'risk-benefit analysis' framing was not pre-tested, but I expected a somewhat larger effect size. Therefore, I set minimum detectable effect size at 20 percentage points. With a one-sided test, $\alpha = 0.05$, $\beta = 0.2$, starting from a 25% take-up rate, I thus need 70 observations for each nudge (3) x treatment (2) combination. However, given that I am not in control of what the Choice Architects choose, I am not going to have equal numbers of observations in each nudge-treatment combination. Hence, I use a specific matching strategy described in the next section.

In another pilot, this time on the Choice Architects' behavior, 50% of the Choice Architects chose the 'simple' question formulation, 40% chose the 'default', and 10% chose the 'risk-benefit analysis' in the control. In the treatment, the numbers were 11%, 33%, and 56%, respectively. This sample was very small, only 19 participants altogether, and sophisticated (they were PhD students or professors interested in Experimental Social Sciences). Based on this pilot, we can expect to see sizeable changes in behavior. On one hand, a 7-point difference is too low to be detectable with any reasonable experimental sample size, on the other end, detecting a 40 percentage point change from a 10% starting level (as is the case with the risk-benefit analysis nudge in the small pilot)requires a sample size of about 20 observations per treatment cell. As Choice Architects experience only the main treatment dimension (2), their treatment structure is much lighter than that of the Decision Makers. Setting the minimum detectable effect size at 20% with a one-sided test, $\alpha = 0.05$, $\beta = 0.2$, starting from 15% take-up rate, I need 57 observations per treatment cell.

Appendix 3.B Extended Results

3.B.1 Characteristics and Treatment Assignment

To check that the treatment assignment was random, we want to study the descriptive statistics for those in the treatment and the control group to confirm that they are balanced. I start with the Choice Architects and their simple treatment structure between non-transparent and transparent setting. Table 3.B.1 summarizes the descriptive statistics collected by the study and a few variables provided by Prolific. None of the variables are significantly different between the treatments.

Table 3.B.1: Mean descriptive statistics by treatment assignment for Choice Architects

	Non-transp.	**Transp.**	**PtT-test**
Age	32.03	32.63	0.716
Gender (1: man, 2: woman, 3: other)	1.56	1.56	1.000
Income (measured with brackets)	$43605	$42132	0.767
Similar study experience	.084	.053	0.582
Prolific experience (brackets)	13.12	12.77	0.579
Labor market role (freq)			
Worker	56	48	
Student	16	15	
Other	2	3	
Unemployed	11	18	
Pensioner	3	1	
Employer or self-employed	7	10	
Total	95	95	(Chi2-test) 0.540
Lives with (freq)			
Parents	21	19	
Partner	31	30	
Family with kids	19	13	
Flatmates	10	13	
Alone	13	16	
Other	1	4	
Total	95	95	(Chi2-test) 0.587
Variables from Prolific			**PtT-test**
Number of approved studies	262.79	207.78	0.148
Number of rejected studies	0.94	1.08	0.617
Prolific Score (Share of approved studies)	99.82	99.64	0.197

Notes: For categorical variables, the table reports frequencies and the results of a Chi2 test measuring statistical differences between the two groups.

Next, we look at the Decision Makers, who are split across 6 treatment-nudge groups. Table 3.B.2 summarizes the descriptive statistics for the Decision Makers. The 6 groups are balanced except for gender and prolific experience, as is reported in the last column. For these two variables, we observe statistically significant differences across the groups. The non-transparent risk-benefit analysis group has more women than the other treatment groups. Average experience varies significantly between multiple groups. While the differences in experience matter slightly, more experienced people are less likely to choose the high target, when controlling for this variable, the results do not change. Gender and its effects are discussed at length in the main text.

What about attrition? Do people leave the study at equal rates in each treatment? Tables 3.B.3 and 3.B.4 show the numbers for Choice Architects and Decision Makers respective. We find no evidence of attrition bias. With the Choice Architects, 82% of the participants that start the study in the control treatment complete it. The equivalent

Table 3.B.2: Mean descriptive statistics by treatment assignment for Choice Architects

	C-Simp.	**C-Def.**	**C-R-BA**	**T-Simp.**	**T-Def.**	**T-R-BA**	**Balanced**
Age	32.11	31.16	30.76	33.64	32.27	31.69	Yes
Gender	1.55	1.44	1.62	1.56	1.50	1.48	No
Income	$37500	$43846	$43068	$40455	$44929	$38203	Yes
Similar experience	0.063	0.044	0.030	0.036	0.019	0.016	Yes
Prolific experience	13.80	11.96	13.00	11.98	13.01	12.83	No
Prolific variables							
Num. of approved st.	331.13	213.96	237.89	246.82	216.37	234.30	No
Num. of rejected st.	1.66	0.49	0.64	1.29	1.10	1.27	No
Prolific Score	99.58	99.92	99.86	99.44	99.72	99.64	No
Labor market role (freq)							**p-value**
Worker	38	46	37	26	54	36	
Student	12	21	12	9	18	13	
Other	5	0	4	3	3	1	
Unemployed	3	14	7	10	2	6	
Pensioner	1	5	1	1	2	1	
Employer or self-employed	5	5	5	6	7	7	
Total	64	91	66	55	105	64	0.349
Lives with (freq)							
Parents	15	27	16	16	30	17	
Partner	16	24	14	8	17	10	
Family with kids	15	15	18	13	33	14	
Flatmates	8	14	10	7	7	6	
Alone	9	11	7	10	16	16	
Other	1	0	1	1	2	1	
Total	64	91	66	55	105	64	0.608

Notes: C refers to control, that is, the non-transparent case and T refers to the treatment or the transparent case. Simp. is short for simple, Def. for default, R-BA for risk-benefit analysis. For continuous variables, CI states if the 95% confidence intervals of the means overlap or not. For categorical variables, the table reports frequencies and the results of a Chi2 test measuring statistical differences between the two groups.

number in the transparency treatment is 84%. Testing that the rates are independent of the treatment gives us the Chi2 statistics of $p = 0.780$ for Table 3.B.3. Equally for the Decision Makers, we find that 82-94% of the participants complete the study that they started. The Chi2 test gives $p = 0.293$ for Table 3.B.4, indicating that the treatments do not lead to differences in attrition rates.

The participants are required to pass enough attention checks to qualify for payment. The study was designed such that as soon as two attention checks were failed, the participant was informed about this. Most of these participants then left and "returned" the study.

3.B.2 Analysis by Confidence

Figures 3.B.1 and 3.B.2 split the sample between the low and high confidence individuals. Figure 3.B.1 shows the average choices for those with low confidence. With this group, the non-transparent default increases the take-up rate, while the transparent default does

Table 3.B.3: Attrition for Choice Architects

	Approved	Rejected	Returned	Timed-out	Approval rate
Non-transparency (control)	95	1	17	3	82%
Transparency (treatment)	95	0	15	3	84%

Table 3.B.4: Attrition for Decision Makers

	Approved	Rejected	Returned	Timed-out	Approval rate
Non-transparent Simple	64	3	2	0	93%
Non-transparent Default	91	3	3	0	94%
Non-transparent R-BA	66	5	5	4	83%
Transparent Simple	55	4	6	2	82%
Transparent Default	105	5	7	2	88%
Transparent R-BA	64	4	2	0	91%
Not assigned	0	3	60	11	

Notes: Control refers to the non-transparent setting, and Treatment to the transparent setting. R-BA is short for risk-benefit analysis nudge. In the experiment for Decision Makers, the participants are not assigned to a treatment until they reach page 17. Thus all of those who quite before page 17 are counted as *not assigned*.

the opposite. Neither effect is significant.[25] The difference between the default take-up rates is marginally significant with a t-test but not significant with the PtT-test.[26] The non-transparent and transparent risk-benefit analysis nudges have negative effects when positive ones were expected.[27] Figure 3.B.2 shows the average choices among those that reported high confidence in being able to reach the high target. Among these people, the non-transparent and transparent defaults have about the same effect, significantly encouraging the choice of the high target.[28] The risk-benefit analysis nudge encourages the choice of the high target, and the effect is significant in both transparency conditions.[29] Notice that with this System 2 nudge, transparency has the opposite impact on the low confidence and high confidence people, moving the average choice towards the one judged optimal for the subgroup based on their self-reported confidence levels.

[25] Comparing the non-transparent default to the non-transparent simple baseline gives a p-value of 0.416 with a one-sided PtT-test ($n = 75$). Comparing the transparent default to the non-transparent simple baseline gives a p-value of 0.846 in a one-sided PtT-test ($n = 95$).

[26] The p-values are $p = 0.081$ with a one-sided t-test, $n = 108$, but the PtT-test results in $p = 0.119$.

[27] We fail to reject the alternative hypothesis that these nudges have a negative effect: comparing the non-transparent risk-benefit analysis nudge to the benchmark gives $p = 0.943$ ($n = 68$) and the transparent comparison gives $p = 0.996$ ($n = 67$), again with one-sided PtT-tests. The difference between the two risk-benefit analysis nudges is not significant: $p = 0.495$, $n = 73$, two-sided PtT-test.

[28] Comparing the non-transparent default with the benchmark of non-transparent simple gives a $p = 0.001$ with a one-sided PtT-test ($n = 80$). Comparing transparent default with the same benchmark gives $p = 0.001$ with a one-sided PtT-test ($n = 74$). The difference between the two defaults is not significant: $p = 0.682$ with a one-sided PtT-test, $n = 88$.

[29] Compared to the non-transparent simple benchmark, the take-up rate with non-transparent risk-benefit analysis nudge is marginally different at $p = 0.073$ ($n = 62$) with a one-sided PtT-test (with a t-test this difference is significant, $p = 0.042$), and the transparent risk-benefit analysis is significantly different at $p = 0.005$ ($n = 61$), also one-sided PtT-test. The difference between the two is not significant: $p = 0.3432$ ($n = 57$) two-sided test.

Figure 3.B.1: How many chose the high target, low confidence

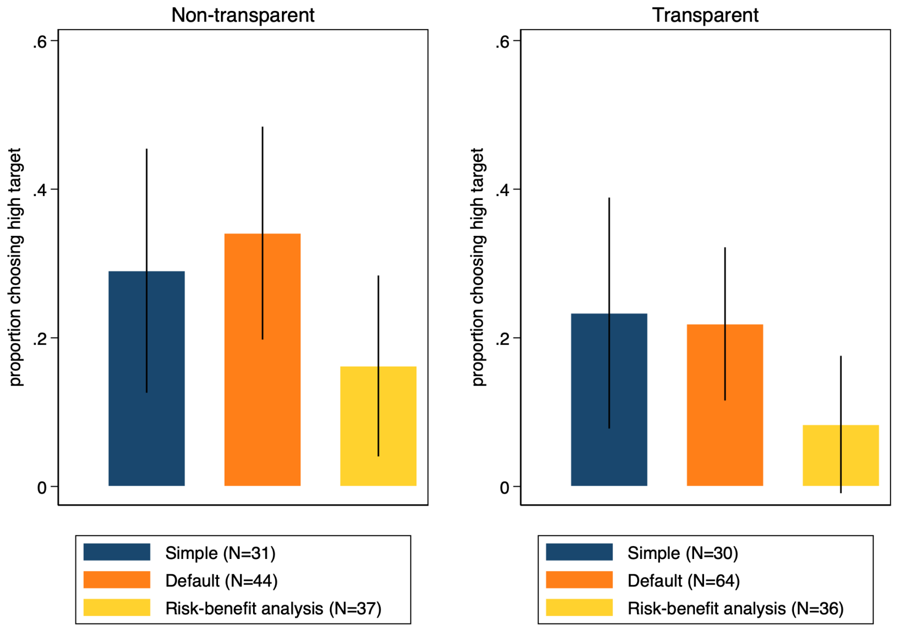
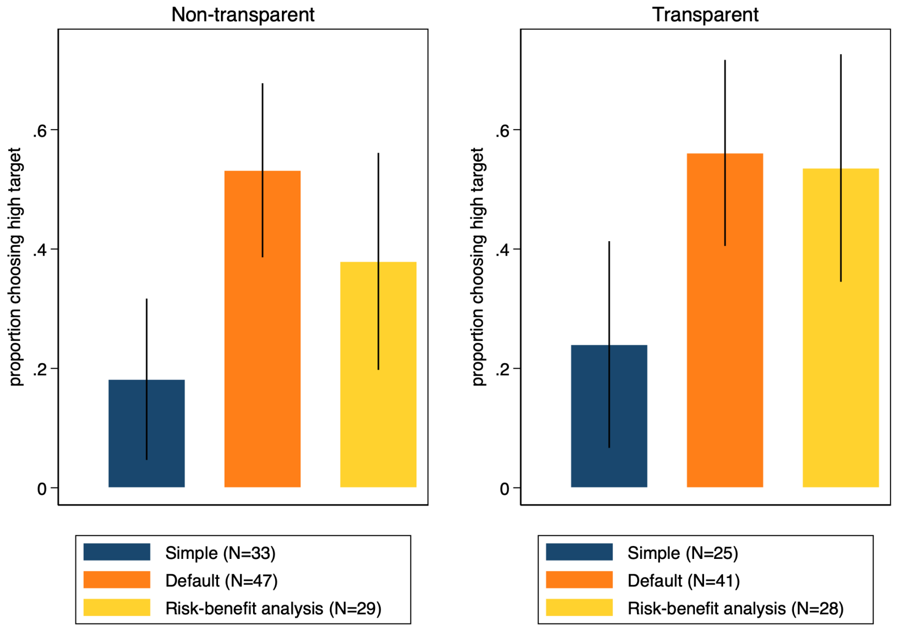


Figure 3.B.2: How many chose the high target, high confidence



Notes: The proportion of people choosing the high target for each question presentation. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. The number of observations per each bin is reported in the legend, the black lines represent the 95% confidence level.

Split this way, the data shows a pattern where consistency between stated confidence and choices is increased by both transparency and the risk-benefit analysis nudge. The default nudge, on the other hand, is compatible with high inconsistency between the self-reported confidence and the subsequent choices, but only in the non-transparent case. The effectiveness of the default nudge is smaller with transparency (although only with marginal significance here), due to the fact that the people of low confidence do not follow the transparent default nudge against their own beliefs.

3.B.3 Analysis by Gender

Figures 3.B.3 and 3.B.4 report the results for women and for men, respectively. Gender is self-reported.[30] Women choose the higher target much less frequently in the benchmark than men. In the non-transparent simple case, 10% of the women chose the high target, while for men the percentage is as high as 38%. This difference is significant ($p = .0278$, $n = 62$, two-sided PtT-test), and can be interpreted as women under-choosing the high target while men over-choose it against the backdrop that approximately only 30% of both men and women report that they believe high target is better in terms of risks and benefits and roughly the same proportions choose it with the risk-benefit analysis.[31] This means that the initial environment is as theorized but only with women.

I discuss the effects of the nudges on women first. The non-transparent default nudge encourages more women to choose the high target. The 36 percentage points increase is large and statistically significant ($p = 0.001$, $n = 66$, one-sided PtT-test). The non-transparent risk-benefit analysis nudge increases the take-up rate by 14 percentage points, however, this effect is not significant.[32] Transparency reduces the impact of the default nudge to 12 percentage points. This effect is no longer significant ($p = 0.133$, $n = 81$, one-sided test in comparison to the non-transparent simple) and it is significantly different from that of the non-transparent default ($p = 0.017$, $n = 85$, one-sided PtT).[33] The slow risk-benefit analysis nudge has a larger impact in the transparent case. The 22 percentage point increase is statistically significant ($p = 0.033$, $n = 59$, one-sided PtT) but not significantly different from the effect of the non-transparent risk-benefit analysis nudge ($p = 0.688$, $n = 65$, two-sided PtT).

With men, neither nudges nor transparency has a consistent effect on the choices made. The non-transparent default nudge increases the take-up rate negligibly by 6 percentage

[30] Participants self-report their gender both for this study and and for the recruitment service provider Prolific in general. The latter can also be used to screen participants. Among the 445 participants, 430 report their gender consistently across these two datasets. Out of the 15 remaining, 2 seem to report also their age and study experience inconsistently, for which reason, I exclude these individuals from the analysis that control for individual characteristics, as I cannot rely on the data. 7 out of the remaining 13 do not wish to report their gender or report it as 'other' when given the opportunity, and 6 report it as man and as woman in the respective datasets. Including these 13 individuals or excluding them does not fundamentally change the subsequent results.

[31] Participants report which target they find better in terms of risks and benefits with the risk-benefit analysis nudge before making the choice. See more in Table 3.B.7 in the Appendix.

[32] The difference is not significant with the one-sided PtT-test, $p = 0.102$, $n = 68$. It is marginally significant with a traditional t-test ($p = 0.059$, one-sided).

[33] With traditional t-test the difference between non-transparent simple and the transparent default is marginally significant: $p = 0.079$.

Figure 3.B.3: How many women chose the high target, by treatment and nudge

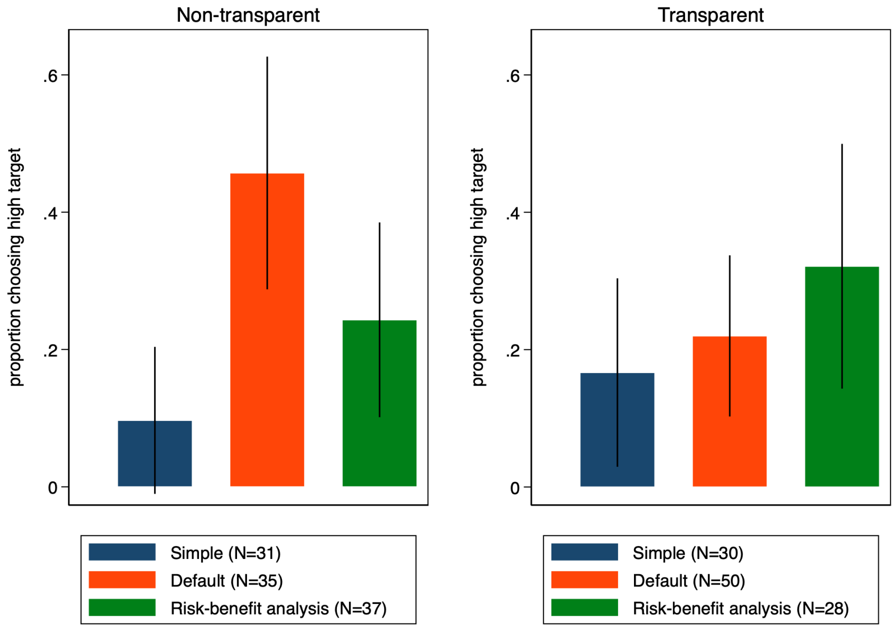
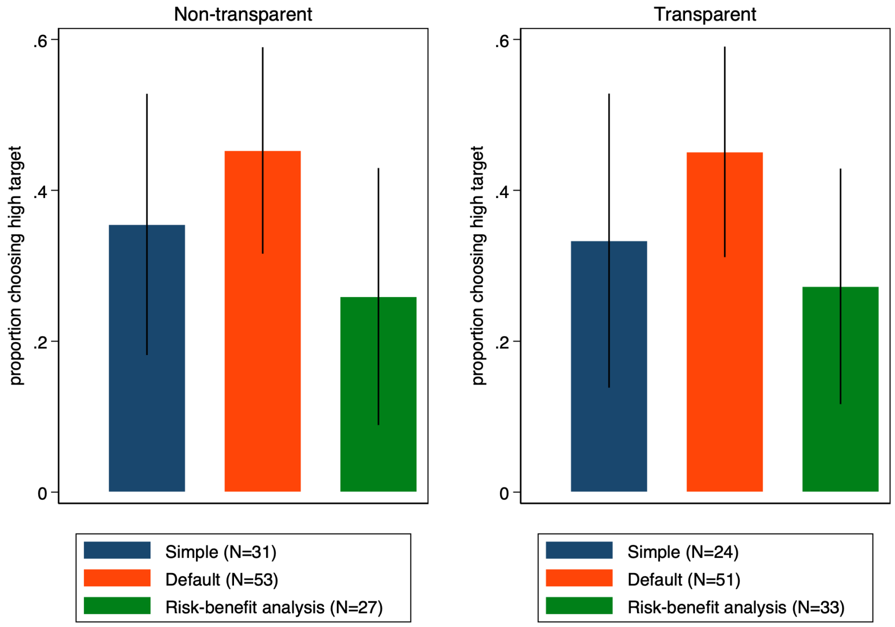


Figure 3.B.4: How many men chose the high target, by treatment and nudge



Notes: The proportion of people choosing the high target for each question presentation. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. The number of observations per each bin is reported in the legend, the black lines represent the 95% confidence level.

points.[34] The risk benefit nudge, on the other hand, has a negative effect. It decreases the take-up rate by 13 percentage points, but again, this effect is insignificant.[35] The

[34] This is insignificant with $p = 0.379$ ($n = 81$) in a one-sided PtT test.

[35] One-sided test expecting an increase gives the following test results: $p = 0.894$, $n = 57$. As a side note, a two-sided PtT-test gives a p-value of $p = 0.507$, $n = 56$, meaning that the drop is not significant.

transparent nudges have effects similar to the non-transparent ones. Transparent default nudge increases take-up by 5 percentage points, and transparent risk-benefit analysis nudge decreases take-up by 10 percentage points. Neither is significantly different from its non-transparent counterpart.[36]

Why are men's choices on the average unaffected by the nudges and transparency? We can exclude simple explanations of risk aversion and preferences, as they are largely similar between men and women.[37] There is some evidence that differences in confidence can explain some of the pattern. Men perform slightly better than women, by about half a task per 7 minutes in the first real effort task.[38] As the targets are set against previous performance, this actually advantages those that perform worse. Still, when estimating their chances of repeating the performance of the first round later, men give a higher estimate than women. This would indicate that women attribute their performance more to good luck than personal ability and that men are more overconfident.[39] When overconfidence is coupled with underestimating the task difficulty (called *overestimation* by Moore and Healy (2008)), it leads to a situation where the higher target is chosen more than what is optimal. This situation is likely to be acerbated by inattentiveness.

3.B.4 Replication of the Results on Women

As the differences between men and women were not anticipated in the the pre-analysis plan, a new pre-analysis plan was was made anticipating the same results on a new sample of women. The plan was to recruit female 150-180 Decision Makers, and 178 women were recruited in the end.

Figure 3.B.5 reports the results for this extra session. The non-transparent default

[36] Comparing transparent default to the baseline gives $p = 0.430$, $n = 78$, one-sided PtT. Comparing transparent risk-benefit analysis to the baseline gives $p = 0.864$, $n = 61$, one-sided PtT. Comparing non-transparent and transparent defaults gives $p = 0.521$, $n = 101$, one-sided PtT, and non-transparent and transparent risk-benefit analysis $p > 0.999$, $n = 59$, two-sided PtT.

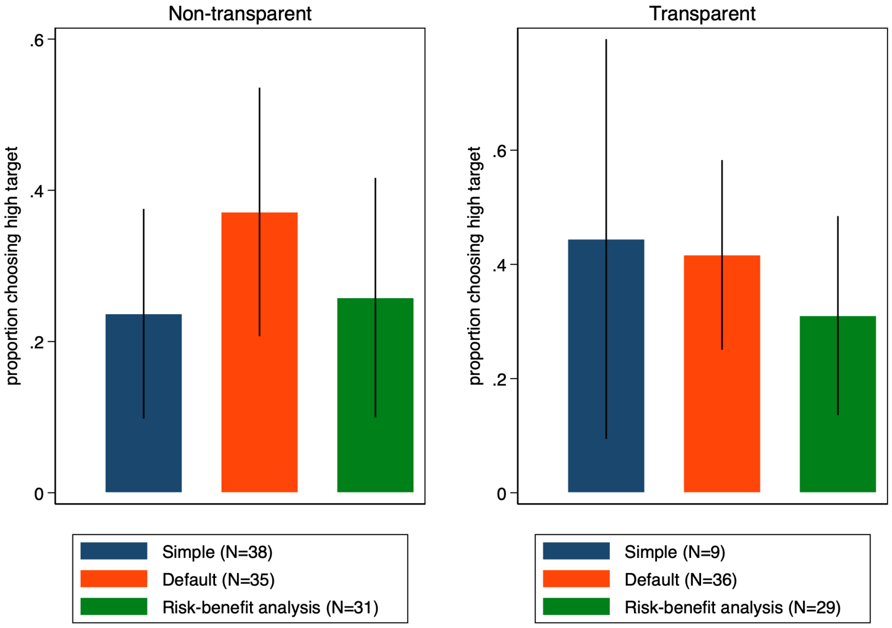
[37] One explanation is that women are more risk averse and thus shy away from the high target, while men do not. However, based on the BRET risk-aversion measure, men and women in this sample do not have different risk preferences. Both are on average risk averse. Score of 50 is risk neutral. Men's average score is 37 ($n = 219$), while women have an average score of 35 ($n = 211$). The difference is not significant ($p = 0.691$, PtT-test). Another explanation is that men and women have different preferences for the task itself, specifically that men prefer more difficult tasks. However, as has been mentioned before, state preferences are similar between men and women.

[38] Men's average performance is 5.6 correct tasks in the 7 minutes while women's average is 5.1 correct tasks. The difference is marginally significant with $p = 0.069$ and persists also in the last round. Results in the second round are not comparable as men and women choose different targets.

[39] Men ($n = 211$) are on average 62% confident that they can repeat previous result, while women ($n = 208$) are 55% confident; the difference is significant at 5% level ($p = 0.028$). Furthermore, as a group, women's belief in the ability of reaching targets is negatively and significantly correlated with the target size. Among men, the correlation is weak and insignificant. Correlation coefficient for women between belief and target size is -0.223 ($p = 0.001$) for the lower target (repeating performance) and -0.232 ($p = 0.001$) for the higher target that requires improvement by 2. The correlation coefficients for men are -0.115 ($p = 0.095$) for low target and -0.088 ($p = 0.225$) for high target.

increases the take-up of the high target by about 13 percentage point which is not significant ($p = 0.163$, $n = 73$, one-sided PtT). The transparent default increases take-up by 18 percentage point, which is marginally significant ($p = 0.083$, $n = 74$, one-sided PtT). The difference between the two is not significant, in fact, the effect is opposite of what was expected: transparency increases the point estimate by 4 percentage points ($p = 0.649$, $n = 71$, one-sided PtT). The risk-benefit analysis nudge has a small 2 percentage point effect when non-transparent ($p = 0.528$, $n = 69$, one-sided PtT), and a 7 percentage point effect when transparent ($p = 0.347$, $n = 76$, one-sided PtT), neither significant. The difference between the two types of risk-benefit analysis are also insignificant ($p = 0.877$, $n = 60$, two-sided PtT). Hence, we do not replicate the effects that we previously recorded with women: 1) transparency does not weaken the System 1 nudge here, and 2) risk-benefit analysis does not have a positive impact on women's choices as previously recorded.

Figure 3.B.5: Second experiment, women's choices



Notes: The proportion of people choosing the high target for each question presentation. Simple is the baseline, default is a fast System 1 nudge and risk-benefit analysis is a slow System 2 nudge. The number of observations per each bin is reported in the legend, the black lines represent the 95% confidence level.

3.B.5 Regression

To check the robustness of my results, I regress the choice of the high target on the interaction of the transparency treatment and the question formulation (simple, default or risk-benefit analysis) and the following controls:

- risk preference (number of bombs selected in the BRET)
- confidence at the lower target
- confidence at the higher target
- tasks correct in the first round
- tasks attempted in the first round
- tasks correct in the tutorial (instruction page)
- tasks attempted in the tutorial (instruction page)
- gender
- age
- income
- similar experience (in terms of studies)
- prolific experience
- household type (with family, with partner, with flatmate(s), with parents, alone, other)
- employment status (employee, employer or self-employed, student, pensioner, unemployed, other)

The regression results are reported in Table 3.B.5. We find the results with the default nudge to be quite robust. The impact of the System 1 nudge is significant with the pooled data and with women, as is the case with the simple t-tests reported earlier. The impact is no longer significant when there is transparency, compared to the benchmark but neither is it significantly different from the impact of the nudge without transparency. With the basic tests, we got the result that the effort of the default nudge in the treatment group was marginally significantly different from the effect in the control group. In general, the control variables explain some of the variation and the effect sizes found with the regression are smaller than those found with the simple comparisons across the groups.

Table 3.B.5: Regressions of choice for all data, for women, and for men

ALL	Coefficient	(Standard error)		90% confidence interval
Simple x Control (Constant)	0.173	(0.155)		[-0.082, 0.429]
Simple x Treatment	-0.003	(0.082)		[-0.139, 0.132]
Default x Control	0.176	(0.074)	***	[0.055, 0.298]
Default x Treatment	0.113	(0.072)		[-0.005, 0.231]
R-B Analysis x Control	0.019	(0.078)		[-0.109, 0.147]
R-B Analysis x Treatment	0.011	(0.079)		[-0.118, 0.141]
Controls	YES			
N	443			
R-squared	0.1776			
Adj R-squared	0.1262			

WOMEN	Coefficient	(Standard error)		90% confidence interval
Simple x Control (Constant)	0.093	(0.216)		[-0.264, 0.450]
Simple x Treatment	0.061	(0.106)		[-0.114, 0.235]
Default x Control	0.373	(0.104)	***	[0.200, 0.545]
Default x Treatment	0.166	(0.097)	**	[0.006, 0.326]
R-B Analysis x Control	0.168	(0.102)		[-0.001, 0.338]
R-B Analysis x Treatment	0.159	(0.108)		[-0.020, 0.338]
Controls	YES			
N	211			
R-squared	0.2761			
Adj R-squared	0.1738			

MEN	Coefficient	(Standard error)		90% confidence interval
Simple x Control (Constant)	0.140	(0.255)		[-0.281, 0.562]
Simple x Treatment	-0.014	(0.134)		[-0.236, 0.208]
Default x Control	0.036	(0.114)		[-0.152, 0.223]
Default x Treatment	0.090	(0.115)		[-0.101, 0.280]
R-B Analysis x Control	-0.121	(0.130)		[-0.336, 0.094]
R-B Analysis x Treatment	-0.125	(0.124)		[-0.330, 0.079]
Controls	YES			
N	213			
R-squared	0.2074			
Adj R-squared	0.0966			

Notes: Control refers to the non-transparent setting, and Treatment to the transparent setting. R-B Analysis is short for risk-benefit analysis nudge.

With women, I also observe that the results with the default nudge are robust. The estimated effects sizes are slight larger than in the simple comparisons earlier. Looking at the 90% confidence interval, we learn that the default nudge has a significantly different effect in treatment compared to that in control: transparency more than halves the proportion of people that choose the high target with default. The results with the risk-benefit analysis nudge are not robust: based on the regression, we no longer find this nudge to change behavior significantly (only with marginal significance, $0.05 < p < 0.10$). No difference is found between the risk-benefit analysis estimates for the non-transparent and the transparent cases.

With men, we repeat the no effect-result.

3.B.6 Secondary Variables

To confirm that the nudges worked as intended, we check the time spent on the page where the performance target is chosen. The prediction is that people use more time with the risk-benefit analysis nudge than with the two other formulations. Some of this effect is mechanical as there is more text to be read. Some of it is explained by the fact that participants need more time analyzing the question.[40] Table 3.B.6 reports the mean and median seconds spent on the choice by each nudge. The median is a better indicator of average behavior in online environments due to the fact that participants are free to take breaks at any moment, thus creating large outliers. On average, people spent considerably more time on the question with the risk-benefit analysis nudge, and this effect is significant.

Table 3.B.6: Time spent making the decision between the high and the low target

	Simple	Default	R-B Analysis	Simple vs R-BA	Default vs R-BA
Non-transparent: mean	49 s	47 s	71 s	p = 0.054	p =0.005
Non-transparent: median	24 s	39 s	53 s	p < 0.001	p = 0.010
obs.	64	91	66		
Transparent: mean	45 s	58 s	102 s	p < 0.001	p = 0.015
Transparent: median	29 s	38 s	57 s	p < 0.001	p < 0.001
obs.	55	105	64		

Notes: Time spent (in seconds) on the program's page asking people to choose between the high target, by treatment and nudge. Simple is the baseline, default is the System 1 nudge and risk-benefit analysis is the System 2 nudge. Difference in means are tested with PtT-tests, differences in medians are tested with continuity corrected Pearson Chi2 median tests.

How did the answers to the risk-benefit analysis question affect the later choice? Table 3.B.7 reports the risk-benefit analysis results and whether people followed them, together and separately for men and women. A large proportion of the participants came to the conclusion that the lower target is better, this is about two thirds of the participants both in the non-transparent and the transparent case (64% and 67%). The question was left unanswered only by a few participants (9% in the non-transparent and 3% in the transparent case), meaning the high target was rated best by the remaining 27% and

[40] There are very few individuals ($n = 8$) that do not answer the risk-benefit analysis question. The median time among these 8 individuals is 41 seconds on the choice page, while the median among those who answered the question ($n = 122$) is 55 seconds, suggesting that the extra time needed is quite evenly split between reading and answering. The difference is (marginally) significant with Pearson chi2 test giving $p = 0.044$ (continuity corrected test giving $p = 0.099$).

Table 3.B.7: Risk-benefit analysis and choice of target

ALL	Low target	High target	Did not answer	Total
Non-transparent: preference	64%	27%	9%	66
Followed preference	83%	44%	Chose high: 33%	
Transparent: preference	67%	30%	3%	64
Followed preference	88%	63%	Chose high: 50%	

WOMEN	Low target	High target	Did not answer	Total
Non-transparent: preference	70%	19%	11%	37
Followed preference	85%	57%	Chose high: 25%	
Transparent: preference	64%	32%	4%	28
Followed preference	89%	78%	Chose high: 0%	

MEN	Low target	High target	Did not answer	Total
Non-transparent: preference	55%	37%	7%	27
Followed preference	80%	30%	Chose high: 50%	
Transparent: preference	69%	28%	3%	31
Followed preference	87%	56%	Chose high: 100%	

Notes: If nudged with the risk-benefit analysis, the participants could state the lower target or the higher target as the best option in terms risks and benefits, or not answer the question at all.

30% of the participants. The numbers are similar in treated and in control, but also between men and women.[41] People follow their own preference more tightly when they had identified the low target as the best option as compared to the high target.[42] Hence, we find no evidence that men and women carry considerably different preferences or that they follow them differently.

I collected Decision Makers' beliefs on which option benefited Player B, the Choice Architect, more. The results are summarized in Table 3.B.8, and in general, the beliefs are spread quite evenly across the three options: *higher target*, *lower target*, and *I don't know*. About equal numbers of people believe the lower target to have benefited Player B and the higher target to have benefited player B (38% and 34%). There are no significant differences between these two beliefs broken by gender, transparency and nudge, except for transparent default with women, where women believed the target to have been the lower one. This result indicates that it was not easy for the participants to judge the goals of the nudges. This is in line with the results from previous research finding that people underestimate the impact of nudges and have a hard time estimating their effects Bang et al. (2018); Pronin et al. (2002).

[41] On average 66% of men prefer the low target ($n = 57$), while the number is 73% for women ($n = 60$), $p = 0.574$. 70% of men follow their stated preference, while 82% of women do it. The difference is not however significant $p = 0.211$.)

[42] Out of the 35 participants that through the higher target was best in terms of risks and benefits, only 54% chose the high target. Out of the 82 who preferred the lower target, 85% also chose it. The difference is significant: $p < 0.001$.

Table 3.B.8: Beliefs: which option was believed to benefit "Player B" (the Choice Architect)

		I don't know	Higher Target	Lower Target
Non-transparent	Simple	34%	27%	39%
Non-transparent	Default	26%	36%	37%
Non-transparent	R-B Analysis	29%	32%	52%
Transparent	Simple	35%	33%	33%
Transparent	Default	28%	36%	36%
Transparent	R-B Analysis	19%	39%	42%

Notes: In the end of the study, the participants were asked:"Which choice do you believe to have benefited Player B [i.e., the Choice Architect] more?", and given 3 options: the lower target, the higher target, and "I don't know".

The last set of secondary variables are concerned with attitudes towards Player B, satisfaction with the target chosen, and sense of being affected by the question formulation. These results can be found in Table 3.B.9. No differences stand out between those in the treated and those in control, nor among those subjected to the different nudges, or between genders.

Table 3.B.9: Liking "Player B", satisfaction, affected

		Liking Player B	Satisfaction	Affected
Non-transparent	Simple	4.4	4.9	4.0
Non-transparent	Default	4.6	4.7	4.1
Non-transparent	R-B Analysis	4.4	4.7	4.0
Transparent	Simple	4.6	5.1	3.9
Transparent	Default	4.5	5.1	4.0
Transparent	R-B Analysis	4.3	4.9	3.9

Notes: In the end of the study, the participants were asked the following questions. Liking Player B: "on a scale from 1 to 7, one meaning negative, four meaning neutral, and seven meaning positive, what are your feelings towards Player B [i.e., the Choice Architect]?"; Satisfaction: "On a scale from 1 to 7, one meaning dissatisfied, four meaning neutral, and seven meaning satisfied, how satisfied are you with your choice of target in Part 3?"; and Affected: "On a scale of 1 to 7, one meaning not at all, four meaning "I don't know", and seven meaning by a lot, how much do you feel that your choice was affected by the way the question was formulated?".

Finally, to touch on the welfare effects of the interventions, we look into whether nudges led to higher or lower earnings among the Decision Makers subjected to the nudges. Table 3.B.10 summarizes the average payment from the real effort task for which the participants had to chose the target. Average payments are reported by treatment and nudge, together and separately for men and women. None of the differences observed were significant, which indicates that at least no harm was done with the nudges.

Table 3.B.10: Average Decision Maker payoff by Treatment and Nudge

ALL	Simple	Default	R-B Analysis
Non-transparent	3.0	3.0	2.8
Transparent	3.1	3.4	2.9

WOMEN	Simple	Default	R-B Analysis
Non-transparent	2.8	3.2	2.6
Transparent	2.8	3.2	2.8

MEN	Simple	Default	R-B Analysis
Non-transparent	3.2	2.8	3.1
Transparent	3.4	3.6	2.9

Notes: The payoff from this part could be either 0 dollars, 4 dollars, or 6 dollars, depending on which target was chosen and if it was reached.

Appendix 3.C Instructions to the Participants

The full code for the program and instructions can be found on my GitHub account. Add link.

3.C.1 Choice Architects

Overview of the study

There are two kinds of players in this study. You are Player B, and you will be paired with a Player A. Your earnings depend on a choice that your paired Player A makes.

Player A participants will do a task where we measure their performance. Before doing the task, they must choose between a high and a low performance target. If Player A chooses the high target, you will be rewarded a bonus of 5 dollars, otherwise you will not receive a bonus.

Your task is to decide how this choice over the performance targets is presented to Player A. There are three (3) different versions of this question. You decide which one of these three formulations is shown to Player A.

Demonstration

From previous research, we know that the way a question is presented affects how it is later answered. For example, in a recent related study (in a different country and in a different language), players also had to choose between a low and a high target. This choice was presented in 4 different ways, however, the options and the outcomes were the same.

The "take-up rate" refers to the share of people that chose the high target. Each of the 4 presentations led to a different take-up rate. You can see the take-up rates from this study in the table below.

Presentation A led to 25% of the participants choosing the high target. This is one in every four people, and it was the lowest rate recorded. The highest take-up rate was achieved with Presentation D, where 57% chose the high target. Hence, the take-up rate was more than doubled with respect to Presentation A, leading more than half of the participants to choose the high target.

Presentations (ordered and named by take-up rate)	take-up rate	number of participants
Presentation A	25%	32
Presentation B	41%	32
Presentation C	53%	32
Presentation D	57%	28

[Attention question 1] What percentage of people chose the higher target when they saw Presentation B?

For data collection purposes, some decisions by B Players will be implemented for multiple A Players. You will not be informed if this will involve your decision. If your decision is used multiple times, we will randomly select one Player A to determine your payoffs. This means that you should make your choices assuming there is only one Player A affected by your decision.

Click "Next" to read the more detailed instructions.

Instructions

Your earnings depend on the choice that another participant, Player A, makes between a high and a low performance target. **If Player A chooses the high target, you will be rewarded a bonus of 5 dollars, otherwise you will not receive a bonus. Your task will is to choose between three ways of asking Player A which performance target he or she would prefer.**

Before seeing the chosen question and making the decision, Player A will read the following information:
[**Control**] Depending on your choice in the next part, another participant, Player B, may also get some money. Player B has also seen this message.
[**Treatment**] There are 3 possible presentations of the next question. The options and their consequences are the same in all presentations. They are, however, worded and structured differently, and in a way that has been shown to affect people's choices. Another participant, Player B, chose which of the three presentations is shown to you. Depending on your choices in the next part, Player B may also get some money. Player B has also seen this message.

This is all the information that Player A will learn about Player B (you).

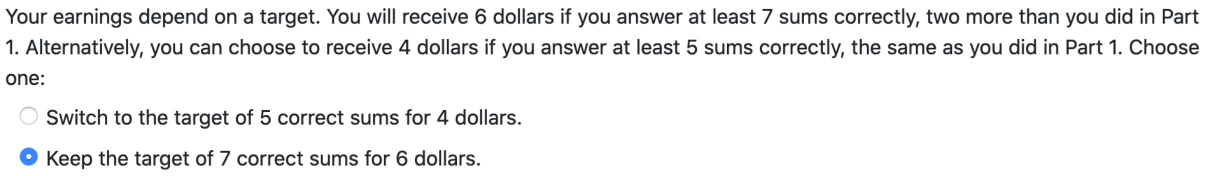
When answering this question, Player A is about to repeat a 7-minute task. Player A's reward depends on the chosen target and whether it is reached. If Player A fails to reach the chosen target, Player A earns nothing from this task.

Your choice

The different presentations are listed in a random order below. Your task is to choose one of them to be shown to Player A. Note that in their essence, the question and the options available are the same in all three presentations. Furthermore, note that the targets are set relative to previous performance and thus the exact number will depend on how well the individual did earlier in the same task. For simplicity, we use the example targets of 5 and 7 sums in the pictures below.

Choose one of the following question presentations to be shown to Player A:

Presentation 1: Note that the high target is already pre-selected when Player A sees this question.

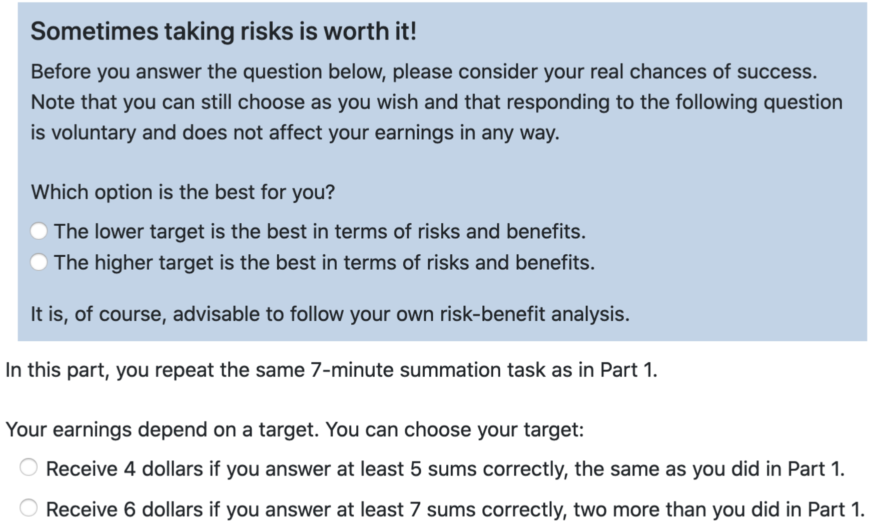


Your earnings depend on a target. You will receive 6 dollars if you answer at least 7 sums correctly, two more than you did in Part 1. Alternatively, you can choose to receive 4 dollars if you answer at least 5 sums correctly, the same as you did in Part 1. Choose one:

○ Switch to the target of 5 correct sums for 4 dollars.

● Keep the target of 7 correct sums for 6 dollars.

Presentation 2: It is expected that people spend twice as much time answering the question with this presentation than with the others.



Sometimes taking risks is worth it!

Before you answer the question below, please consider your real chances of success. Note that you can still choose as you wish and that responding to the following question is voluntary and does not affect your earnings in any way.

Which option is the best for you?

○ The lower target is the best in terms of risks and benefits.
○ The higher target is the best in terms of risks and benefits.

It is, of course, advisable to follow your own risk-benefit analysis.

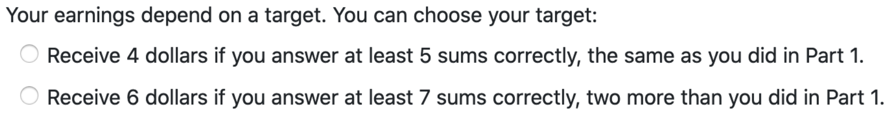
In this part, you repeat the same 7-minute summation task as in Part 1.

Your earnings depend on a target. You can choose your target:

○ Receive 4 dollars if you answer at least 5 sums correctly, the same as you did in Part 1.

○ Receive 6 dollars if you answer at least 7 sums correctly, two more than you did in Part 1.

Presentation 3: About 30% of people chose the high target with this formulation on an earlier Prolific study.



Your earnings depend on a target. You can choose your target:

○ Receive 4 dollars if you answer at least 5 sums correctly, the same as you did in Part 1.

○ Receive 6 dollars if you answer at least 7 sums correctly, two more than you did in Part 1.

[Attention Check 2] It is important that you pay attention to this study. Please, uncheck this box. By deselecting it, we know you are paying attention.

Submit your answer by clicking "Next".

3.C.2 Decision Makers

The structure of the study

This study consists of 5 consecutive parts and 3 of these tasks are very similar. You will receive specific instructions separately for each part. Using the full-screen mode for your browser is highly recommended. Click "Next" to receive the instructions for Part 0.

Part 0: Bomb Task

01	02	03	04	05	06	07	08	09	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

On the screen you see a field composed of 100 numbered boxes. Behind one of these boxes a time bomb is hidden; the remaining 99 boxes are empty. You do not know where the time bomb is. You only know that it can be in any place with equal probability.

Your task is to choose how many boxes to collect. Boxes will be collected in numerical order. So you will be asked to choose a number between 1 and 100.

At the end of the experiment, we will randomly determine the number of the box containing the time bomb by the computer randomly drawing a number between 1 and 100. If you happen to have collected the box in which the time bomb is located – i.e., if your chosen number is greater than, or equal to, the drawn number – you will earn zero. If the time bomb is located in a box that you did not collect – i.e., if your chosen number is smaller than the drawn number – you will earn an amount in dollar equivalent to the number you have chosen in dollar cents. In other words, each collected empty box is worth 1 cent. Your decision:

[Input box] How many boxes do you want to collect?

Part 1: First Series of Summations

In Part 1, you will have 7 minutes to solve a series of summation problems. You earn 30 cents for every correct answer.

Instructions to the series of summations

With each series of summations, the task is to find the largest number in each box and then to sum these numbers together. Each box contains 100 two-digit numbers that have been randomly generated by a computer.

Below, you see an example task. The largest number in Box 1 on the left is 87 and the largest number in Box 2 on the right is 58. They sum up to 87 + 58 = 145, thus 145 is the correct answer to be submitted. Enter the answer in the input box below. To submit, click "Submit" or press the Enter-key on your keyboard. Once you have submitted an answer, a new set of numbers appears immediately. You also receive information regarding whether the answer was correct or, in case it was not, what the correct answer would have been. Underneath you find counters for correct answers and attempts.

Box 1

19	47	42	30	65	50	11	32	43	19
38	24	84	64	13	77	28	37	53	54
16	31	56	54	36	19	30	13	71	56
43	84	66	56	75	37	82	81	47	27
64	61	33	30	76	38	55	65	17	47
35	42	10	69	58	25	49	27	48	41
82	31	61	78	59	60	38	10	39	66
76	64	46	87	21	83	59	73	57	20
73	85	27	74	75	85	13	56	86	30
28	85	21	77	43	35	33	35	29	55

Box 2

54	31	17	44	42	23	46	39	53	10
17	41	34	46	56	55	52	37	29	10
44	11	20	31	15	34	40	38	12	30
31	52	53	10	54	24	31	51	55	47
17	30	53	35	10	49	22	11	24	13
14	20	31	26	42	13	14	12	23	32
14	15	40	57	41	57	21	50	47	19
23	21	12	28	18	26	55	15	19	46
58	32	26	36	16	23	36	58	44	12
26	33	10	23	37	22	39	55	24	35

[Input box] Your answer:
Correct answers so far: 0
Attempts: 0

Feel free to solve another task (the old task is immediately replaced with a new task once you submit your answer), and also try what happens when a wrong answer is submitted.

To the next question, which asks you for your favorite number, you must write the answer salmon.
(Attention check) After having read the instruction above, what would you say is your favorite number between 1 and 100?

You will repeat this same 7-minute series of summations also in Parts 3 and 4 of this study. When you are ready, click “Next”. The task will open on the next page and the timer for the 7 minutes will start immediately.

Part 2: Confidence: Target or Lottery?

Overview

You will repeat the 7-minute series of summations task that you just completed both in Part 3 and Part 4. The goal of Part 2 is to determine how many correct answers you expect to be able to give in **the last 7-minute series of summations** in **Part 4**. For this purpose, we ask you to make 40 choices between different lotteries and performance targets, which will reveal how confident you are about reaching certain performance levels. **The more accurately you estimate your chances of reaching different levels of correct answers, the more likely you will earn 3 dollars.**

Instructions

Part 4, the third series of summation, is rewarded with one of two methods chosen at random by the computer. The method is selected after the summation task of Part 4, at the end of the experiment.

- With 50% probability, **you earn 30 cents for every sum correctly inserted** in Part 4, just like in Part 1.
- With 50% probability, **one of the 40 questions below is randomly chosen as the basis of payment for Part 4.**

Each of the 40 questions is a choice between a performance target and a random lottery. If one question is selected to be the basis for rewarding, the reward will depend on what you chose in this part:
If your choice was **"target"**, you earn 3 dollars only if you have reached that particular target. Otherwise you get 0 dollars.
if your choice was **"lottery"**, you have a chance to win 3 dollars with that particular probability that the lottery offers. If you do not win, you get 0 dollars.

You can think of the lottery as extracting one ball at random from a concealed jar that contains 5 balls altogether.

- **20%** chance of winning corresponds to **one** winning ball among the **5 balls**.
- **40%** chance of winning corresponds to **two** winning balls among the **5 balls**.
- **60%** chance of winning corresponds to **three** winning balls among the **5 balls**.
- **80%** chance of winning corresponds to **four** winning balls among the **5 balls**.

An example (not compulsory to answer)

How do you prefer to have a chance at getting 3 dollars? Choose between:

- A target of 4 correct summations
- A lottery with a 60% chance of winning

If you choose the lottery, you indicate that you believe you have less than 60% probability of reaching the result of 4 correct summations or more. If you choose the target, you indicate that you believe that with a probability of 60% or higher you can reach a result of 4 correct summations or more.

Thus, if the computer chooses this question to be the basis of the payment, and you selected the target, you earn 3 dollars if you have submitted 4 or more answers correctly in Part 4. If you selected the lottery, you participate in a random lottery in which you have a 60% chance of winning the prize of 3 dollars.

Rules

When answering these questions, we impose two rules. First, you can switch from "target" to "lottery" at most once per target level. This way we can interpret your answer as how confident you are in achieving this particular target. Second, being confident in reaching a particular level of performance, you must be at least as confident in reaching all the lower levels before this one. In order to reach a higher target, you must fulfill all the lower ones first. Hence, confidence is not allowed to increase with the size of the target.

The answering form is coded in such a way that you cannot break these two rules. When these rules apply to other questions, they are auto-filled by the same two rules. Note that your answers will be submitted only once you press "Next" in the bottom of the page, where you will be also provided with a summary of your choices.

You must answer the next question about sports by writing your favorite color in the box.

(Attention check) Based on the text above, what would you say is your favorite sport?

In the first series of summations (Part 1) you answered 0 sums correctly and attempted 0 in total.

1. How do you prefer to have a chance at getting 3 dollars? Choose between:

- A target of 1 correct summation
- A lottery with a 20% chance of winning

etc.

40. How do you prefer to have a chance at getting 3 dollars? Choose between:

- A target of 10 correct summations
- A lottery with a 80% chance of winning

Based on your responses you are NA confident you can reach the target of 10.

To summarize, based on your responses above:
You are NA percent confident you can reach the target of 1.
You are NA percent confident you can reach the target of 2.
You are NA percent confident you can reach the target of 3.
You are NA percent confident you can reach the target of 4.
You are NA percent confident you can reach the target of 5.
You are NA percent confident you can reach the target of 6.
You are NA percent confident you can reach the target of 7.
You are NA percent confident you can reach the target of 8.
You are NA percent confident you can reach the target of 9.
You are NA percent confident you can reach the target of 10.

Confirm by clicking "Next".

Part 2: Confidence: Target or Lottery

Thank you, your choices have been registered. Instructions for Part 3 start on the next page.
[**Control:**] Depending on your choices in the next part, Player B may also get some money. Player B has also seen this message.
[**Treatment:**] **Please, note!** There are 3 possible presentations of the next question. The options and their consequences are the same in all presentations. They are, however, worded and structured differently and in a way that has been shown to affect people's choices. Another participant, Player B, chose which of the three presentations is shown to you. Depending on your choices in the next part, Player B may also get some money. Player B has also seen this message.
(Check box) I have read and understood the message.

Instructions for Part 3

[If risk-benefit analysis] Sometimes taking risks is worth it!

Before you answer the question below, please consider your real chances of success. Note that you can still choose as you wish and that responding to the following question is voluntary and does not affect your earnings in any way.

Which option is the best for you?

- The lower target is the best in terms of risks and benefits.

- The higher target is the best in terms of risks and benefits.

It is, of course, advisable to follow your own risk-benefit analysis. [**If risk-benefit analysis part ends**]

In this part, you repeat the same 7-minute summation task as in Part 1. Your earnings depend on a target. You can choose your target:

- Receive 4 dollars if you answer at least X sum correctly, the same as you did in Part 1.
- Receive 6 dollars if you answer at least X+2 sums correctly, two more than you did in Part 1.

[**If default, instead:**] Your earnings depend on a target. You will receive 6 dollars if you answer at least X+2 sums correctly, two more than you did in Part 1. Alternatively, you can choose to receive 4 dollars if you answer at least X sums correctly, the same as you did in Part 1. Choose one:

- Switch to the target of X correct sums for 4 dollars
- Keep the target of X+2 correct sums for 6 dollars

[**Common to all:**] If you do not reach the chosen target, your earnings will be 0 dollars for this part. If you reach the higher target without having selected it, you will receive the lower reward of 4 dollars. When you are ready, click "Next". The task will open on the next page and the timer for the 7 minutes will start immediately.

Part 4 - third series of summations

There are two ways that you can be rewarded for Part 4:

- With 50% probability, you earn 30 cents for every correct answer, just like in Part 1.
- With 50% probability, one of the 40 questions that you answered in Part 2 is randomly chosen as the basis of the reward. Depending on your choice, you will have a chance to get a 3 dollar reward either by winning a lottery or by reaching a performance target.

The method will be chosen at random and communicated in the end of the experiment, after Part 4. Please write down what is the probability that you will earn 30 cents for every correct answer:
(Attention check:) When you are ready, click "Next". The task will open on the next page and the timer for the 7 minutes will start immediately.