The Role of Incentive Design in Firms

Shaping Employee and Customer Behavior With Non-Monetary and Monetary Incentives

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2022

vorgelegt
von

Thomas Vogt, M.Sc.

aus
Bonn

Referent:	Prof. Ulrich W. Thonemann, Ph.D.
Koreferent:	Prof. Dr. Dirk Sliwka
Tag der Promotion:	20.12.2022

To my family

Acknowledgments

I am grateful for the support and help of many people during my time at the chair and the time of my dissertation.

First, I want to thank my supervisors. Thank you Ulrich W. Thonemann for giving me the opportunity to write this dissertation and for supporting and guiding me along the way. I learned a lot not only about research and teaching but also about management and leadership in general. Thank you for teaching and challenging me and for giving me the freedom for my personal and professional development. Thank you Dirk Sliwka for co-supervising my dissertation and your support in the past years. Thank you both for the inspiring research discussions and your help in the analysis and the writing of this dissertation.
I also thank Nicolas Fugger for chairing the examination board.

Second, I would like to give special thanks to the following friends and colleagues from my past eight years at the chair. Thank you Sabrina and Cedric for your friendship and support since the beginning in 2015 - I will always look back at our time at the chair, our study abroad, vacations, conference trips, and nice evenings together. Thank you Sabrina for being my office mate, for always having an open ear, and helping me with everything.
Thank you Andreas for the great fun together, the many Pandemic board game sessions, and for being a good teacher. Thank you Anna-Lena for our joint sustainability project. Thank you Simon for the nice office time together and the many soccer games we played. Thank you Stephanie for all your administrative support and extra effort and for being the glue of the department. Thank you Tobias for your extra time and effort, the enthusiastic discussions about various (research) topics, and supervising my master thesis. Thank you Wolfram for telling me about the Ph.D. Fast Track Program and the open student assistant position at the chair.

Third, I want to say a big thank you to all the other friends and colleagues, who accompanied my time at the chair and the time of my dissertation: Jakob Alfitian, Michael Becker-Peth, Sidney Block, Jose Benedicto Duhaylongsod, Daniel Gass, Theresa Gattermann-Itschert, Christiane Haubitz, Silke Jütte, Lea Kaftan, David Könen, Chantale Köster, Max Pflitsch, Daniel Piersig, Laura Maria Poreschack, Margarita Protopappa-Sieke, Christian Rählmann, Robin Reiners, Anne Schönhofen, Sebastian Siuda, Jeffrey Starck, Michael Völkel, Felix Wagener, Dominik Walzner, Moritz Wegener, Monika Westphal, Ilya Zarubin, Roman Zeiß, and Yingshuai Zhao.

Fourth, I thank my dear friends from my bachelor and master studies for the numerous Kasalla and Tonnen Raus events, trips, and cycling tours.

Last but above all, I would like to thank my family - I am grateful for your love and support in general and in particular during my dissertation. Thank you Mama and Papa for laying the foundation for the person I am today, for giving me opportunities that many people do not get, and for supporting me ever since. Thank you Benedikt for introducing me to the world of Academia already during my bachelor studies, for your inspiration and endless support, and for all our discussions about research. Thank you Christine and Moritz for your open ears and support over the years. Thank you Paula for always being there for me, for listening, understanding, and supporting me through the ups and downs of this dissertation - I am truly happy to have you. Thanks to you six for the moments we had in the past and the many in the future to come!

Thank you all for the great time and support!

Contents

List of Figures

List of Tables

List of Abbreviations

AE	Always Expiring
AEARCTR	American Economic Association's registry for randomized controlled trials
CO	Crowding Out
C-SEB	Center for Social and Economic Behavior (of the University of Cologne)
D	Dummy Treatment
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
ECU	experimental currency unit
HIT	Human Intelligence Tasks
MTurk	Amazon's Mechanical Turk
ND	No Dummy Treatment
NE	Never Expiring
OLS	ordinary least squares
PS	Price Sensitive
TD	Transparent Dummy Treatment

Chapter 1

Introduction

”Consider a thought experiment: You meet an attractive person, and in due time you tell that person, “I like you very much and would like to have sex with you.” Alternatively, consider the same situation, but now you say, “I like you very much and would like to have sex with you, and, to sweeten the deal, I’m also willing to pay you $20!” Only a certain kind of economist would expect your partner to be happier in the second scenario. However, offering $20 worth of (unconditional) flowers might indeed make the desired partner happier.”

Uri Gneezy, Stephan Meier, and Pedro Rey-Biel

2011. When and Why Incentives (Don’t) Work to Modify Behavior. Journal of Economic Perspectives, 25(4): p.201

1.1 Motivation

According to standard economic theory, employees and customers are two of the most important pillars of the success of firms. Employees are the labor input of firms and their performance is a major driver of the production efficiency and thereby of the quantity, quality, and cost of the output of firms. Customer buying behavior defines how much output firms can sell at a certain price, need to stock, or must discard. Consequently, employee performance and customer buying behavior determine not only the short-term performance but also the long-term market position and survival of firms. Not surprisingly, Clarkson (1995, p.106) defines employees and customers as one of the ”primary stakeholder group[s] [...] without whose continuing participation the corporation cannot survive as a going concern”.

The principal-agent model describes that firms face moral hazard of customers and employees due to information asymmetry and misaligned incentives (see for example Holmström 1979): Firms benefit from certain actions taken by self-interested (that is own utility maximizing) employees and customers, but cannot fully observe these actions. Moreover, actions that maximize the utility of employees or customers do not necessarily maximize the profit of the respective firms. Thus, there is the possibility that employees and customers act against the interest of the respective firms.

To overcome moral hazard problems, firms can provide employees and customers with incentives that their actions contribute to the output of the firms. There is an extensive theoretical literature on the optimal design of incentive schemes in firms (see for an overview Köszegi 2014 and Schmidt 2017). There is also a rich empirical literature that studies how incentive schemes affect employee performance (see for an overview Prendergast 1999, Condly et al. 2003, Bandiera et al. 2011, List and Rasul 2011 and Levitt and Neckermann 2014) and customer behavior (see exemplary White et al. 2019).

Indeed, firms excessively use incentives to align employee and customer behavior with their own interests: For instance, a study estimates that firms in the United States spend about 40 and 32 billion dollars alone on incentives such as award points, gift cards, trips, events, or merchandise for their employees and customers, respectively in 2022 (Garlick 2022).

It can broadly be distinguished between two types of incentives: Non-monetary and monetary (see for example Deci and Ryan 1985, Gneezy et al. 2011, Bowles and Polanía-Reyes 2012). The literature provides ample evidence that either incentive type can be effective in directing individuals' behavior towards a behavior desired by firms. However, the effectiveness of an incentive depends on its specific design and the setting where it is employed – see as an illustrative example the introductory quote above. For example, some studies find a positive influence of introducing monetary incentives (see for example Ariely et al. 2009), others discuss a negative (Titmuss 1970) or non-monotonic impact (Gneezy and Rustichini 2000). Looking at the interplay of monetary and non-monetary incentives paints a similar complex picture. The underlying argument is that depending on the situation, extrinsic rewards can crowd out, that is negatively affect intrinsic motivation or other drivers of behavior such as social preferences, image motivation or individuals' norms (Bénabou and Tirole 2003, 2006, Gneezy et al. 2011). The literature also reports heterogeneous reactions to each incentive type or the combination of both (see as an example for the latter case Mellström and Johannesson 2008). Thus, whether or not (a combination of) incentives work, is not always clear ex ante.

The syntheses of the literature reveals that the specific design of an incentive scheme determines how employees and customers react to it. Due to the importance of employees and customers for the success of firms, the design of their incentives affect the short-term performance as well as long-term market position and survival of firms. Accordingly, a better understanding of how non-monetary and monetary incentives affect employee and customer behavior is key to increase the performance of firms and to design more efficient markets. This dissertation aims to contribute to this understanding by studying examples of how non-monetary and monetary incentives and the combination of both affect employee and customer behavior.

1.2 Outline

We now present the structure of this dissertation. Each of the three main Chapters 2 to 4 represent independent research projects. We study examples of how the design of incentives induces employees (Chapters 2 and 3) and customers (Chapter 4) to contribute to firm output. In Chapters 2 and 3, we analyze how the design of rating scales of performance appraisals can incentivize employees to perform better. In Chapter 4, we study how the design of sales promotions can incentivize customers to buy earlier expiring items. We provide further analyses as well as experimental instructions and screenshots in the appendices of each chapter.

While we focus on different incentive settings and mechanisms, we use the same methodological approach – controlled (online and laboratory) experiments – to causally identify the effect of interest in each research project. For the research reported in Chapter 2, we ran two online field experiments on Amazon's Mechanical Turk (MTurk), a crowd-sourcing marketplace run by Amazon.com, Inc (Buhrmester et al. 2011, Horton et al. 2011, Lee et al. 2018). We chose the online setting since this is a natural environment of the gig-economy where short-term employer-employee relations are common. Moreover, it allowed us to recruit 2,344 subjects by investing a reasonable amount of time and effort that would have not been possible in the same manner in an offline field experiment. For the research reported in Chapter 3, we ran a laboratory experiment that comprised 16 experimental sessions at the Cologne Laboratory of Economic Research at the University of Cologne. We chose the laboratory experiment instead of an online experiment due to the experimental design that involved interactions between subjects. For the research reported in Chapter 4, we ran an online experiment on MTurk. We chose an online experiment instead of a laboratory experiment since the design did not involve interactions between subjects and running laboratory experiments was not possible due to the COVID-19 pandemic. Additionally, an online experiment most closely mirrors an online shopping setting. In total, we analyze data of 3,358 subjects for all three research projects. In what follows, we shortly summarize each chapter:

In **Chapters 2** and **3**, we investigate how the incentive design choice of firms affects employee performance. Many firms assess employee performance on predefined rating scales but rarely use the lowest rating categories. Following the approach of behavioral economic engineering (Bolton and Ockenfels 2012), we study an unused low rating category in performance appraisals as an incentive design choice. Employing an unused low rating category is – by design – a non-monetary incentive design tool since it does not involve and hence affect monetary payments. However, if individuals believe that the additional category is used, they face perceived higher incentives since low performance might result in lower personal performance ratings (non-monetary incentive) and lower monetary payments (monetary incentive).

In **Chapter 2**, we investigate how an unused low rating category in performance appraisals affects employee performance in short-term employer-employee relations. We ran two field experiments where we hired individuals on MTurk to digitize handwritten class grades for our university department. Individuals worked over two periods and received private rank feedback. In the baseline treatment individuals saw the rating scale that was used to evaluate performance. In the other treatments individuals saw a rating scale with an additional – but unused – low rating category. We varied whether individuals were informed or not that the low rating was unused. In our experiments reflecting short-term employer-employee relations, only low performing subjects showed higher performance when they did not learn that the additional rating category was unused. Contrarily, other individuals showed lower performance when they did not learn that the category was unused. Overall, an unused low rating category did not raise average performance, irrespective of whether individuals learned or not that the category was unused. Our results suggest that individuals consider not only their monetary incentives and personal performance ratings but also the design and kindness of rating scales in short-term employer-employee relations. Monetary incentives and personal performance ratings seem to have a stronger performance effect on low performing individuals than on other individuals.

While conducting the research of Chapter 2, I received valuable input from Prof. Dr. Dirk Sliwka, Dr. Tobias Stangl, and Prof. Ulrich W. Thonemann, Ph.D.. A similar version of Chapter 2 is published as a working paper by Vogt (2021).

In **Chapter 3**, we examine how an unused low rating category in performance appraisals affects employee performance in long-term employer-employee relations when employees receive multiple performance ratings. We ran a real effort laboratory experiment where individuals worked over six periods and received private rank feedback in each period. In the baseline treatment, individuals saw the actual rating scale used to evaluate performance. In two other treatments, individuals saw an additional low rating category that was never used. In the treatment where individuals were not informed about the non-usage of the low category, performance was about 20% higher than in the baseline without the additional low rating category. In line with results of Chapter 2, this effect evolved over time and low performing individuals reacted stronger than the remaining individuals. In the treatment where individuals were informed about the non-usage of the low category, performance was not significantly higher than in the baseline. Our findings suggest that individuals value their monetary incentives and personal performance ratings more than the kindness of rating scales in long-term employer-employee relations. Moreover, we find that (perceived) higher incentives in presence of an unused low rating category seem to drive the performance results rather than reciprocal reactions to higher personal performance ratings. In accordance with the results of Chapter 2, this mechanism seems to be more pronounced for low performers as they increase performance stronger than the remaining individuals.

Chapter 3 is joint work with Prof. Dr. Dirk Sliwka and Prof. Ulrich W. Thonemann, Ph.D.. I designed, implemented, ran, and analyzed the experiment as well as wrote all sections of the chapter. Both co-authors gave input for the design and analysis of the experiment, and line edited the chapter.

In **Chapter 4**, we investigate how the incentive design choice of firms affects customer buying behavior. We analyze how firms can design sales promotions using non-monetary and monetary incentives to shape customer buying behavior. To reduce food waste at retailers, we analyze how sustainability messages (non-monetary incentive) and price discounts (monetary incentive) affect purchases of earlier expiring items. We ran an online experiment in which individuals chose between earlier expiring and longer lasting items. We find that both price discounts and sustainability messages increased purchases of earlier expiring items. Moreover, we find heterogeneous reactions to either incentive across different customer types. Some individuals switched from buying longer lasting items to earlier expiring items or vice versa while others did not change their buying behavior when receiving price discounts. Similarly, some individuals switched from buying longer lasting items to earlier expiring items while others did not when seeing a sustainability message. Our results suggest that retailers who recognize how either incentive type affects buying behavior of different customer types can offer a respective incentive only when an increase of purchases of earlier expiring items can be expected.

Chapter 4 is joint work with Dr. Anna-Lena Sachs and Prof. Ulrich W. Thonemann, Ph.D.. I formulated the model, designed, implemented, ran, and analyzed the experiment as well as wrote most of the sections of this chapter. Both co-authors gave input for the model formulation, design and analysis of the experiment, and line edited the chapter.

In **Chapter 5**, we conclude by summarizing and critically reviewing the key results of this dissertation. We also propose directions for future research.

1.3 Contribution

This dissertation contributes to the experimental literature on the effect of non-monetary and monetary incentives and the combination of both on employee and customer behavior. We add to the fields of behavioral management science, behavioral economics, and behavioral economic engineering. In the following, we shortly summarize the main contribution of each chapter. More details about the contribution can be found in the respective chapters and in Chapter 5.

In **Chapters 2 and 3**, we contribute to the literature on subjective performance evaluations and the experimental literature on the effect of feedback on performance. We also add to the research on reciprocity in employer-employee relations and rank preferences. Research in psychology and economics associated unused low rating categories in performance appraisals with rating biases of supervisors who give too lenient ratings. We, however, investigate the phenomenon of unused low rating categories in performance appraisals as an intentional incentive design choice of firms.

To evaluate the performance effect of unused low rating categories in short-term and long-term employer-employee relations, we designed two online field experiments and a laboratory experiment, respectively. We developed a novel real effort task for the research on short-term employer-employee relations in Chapter 2: Digitizing handwritten exam grades from artificially created exam cover sheets (see Appendix 2.A for more details). This is a simple and tedious task that demands low cognitive effort. It primarily requires attention and can be solved without prior knowledge while potential learning effects can be neglected. For the research on long-term employer-employee relations in Chapter 3 we use the real effort task by Berger et al. (2013).

Our results indicate that employees not only care about the specific feedback, which they receive for their performance but also about the choice of the scale on which this feedback is given. Employing an unused low rating category in long-term employer-employee relations seems to be an effective design choice: It raised performance by 20% when individuals did not know that the category was unused. We also find that the (perceived) higher incentives in presence of an unused low rating category seem to be the dominating performance driver. This mechanism seems to be stronger for low performers as they increase performance stronger than other individuals. Moreover, the results of both chapters show that incentive mechanisms affect individual behavior differently dependent on whether individuals are exposed to them repeatedly and can react dynamically. Thus, insights about a short-term performance effect of incentive mechanisms do not necessarily allow conclusions about respective long-term influences.

In **Chapter 4**, we contribute to the literature on customer buying behavior and food waste reduction in retailing by improving the understanding of customer behavior when buying perishable products. Research so far focused on price discounts, while we analyze a novel approach for food waste reduction by studying sustainability messages to incentivize purchases of earlier expiring items. We add to the research on social preferences and the crowding out effect of monetary incentives.

To test the influence of sustainability messages and price discounts on customer buying behavior, we designed an online experiment. We also developed a behavioral model that predicts overall and individual buying reactions to sustainability messages and price discounts.

The results are consistent with the model and show that displaying sustainability messages and offering price discounts in retailing can induce individuals to purchase earlier expiring items. Moreover, as predicted by the behavioral model, we observe that the reactions to a sustainability message and price discounts varied between types of customers. We hence conclude that targeted promotions using sustainability message and price discounts for selected groups of customers seem to be more effective than untargeted policies that address all customers equally.

Overall, this dissertation provides evidence that not only the actual provision but also the perception of non-monetary and monetary incentives can affect employee and customer behavior. Moreover, in our settings the (perceived) combination of non-monetary with monetary incentives induced behavior that increases the output of firms. Furthermore, we document heterogeneous reactions to both incentive types and their combination, and observe that these reactions depended on whether or not individuals were exposed to the incentives repeatedly. Thus, an incentive designer may consider the composition of the workforce and customers and whether or not they are repeatedly exposed to an incentive scheme.

Chapter 2

On Rating Scales in Performance Appraisals: Performance Effect of a Dummy Rating Category in Short-Term Employer-Employee Relations

We investigate unused low rating categories in performance appraisals as an incentive design choice in short-term employer-employee relations. The literature proposes that unused low rating categories trigger what we term an Incentive, Evaluation and Kindness-of-the-Scale Effect. We explore how these effects impact performance in short-term employer-employee relations in two field experiments on Amazon's Mechanical Turk. Subjects worked on a real effort task over two periods and received private rank feedback from a computer. The computer rated performance using three categories. In the baseline treatment subjects saw the rating scale with the actual three rating categories. In the other treatments subjects saw an additional fourth – but unused – low rating category. Depending on the treatment, subjects were informed or not that the additional low category was unused. We do not find evidence that an unused low rating category increases average performance in short-term employer-employee relations, independent of whether individuals were informed or not that this category is unused. When individuals did not learn that the additional rating category was unused low performers worked more while the remaining individuals worked less. Our results indicate that individuals do not only consider their monetary incentives and personal performance rating but also pay attention to the kindness of a rating scale in short-term employer-employee relations. Low performers seem to focus more on their monetary incentives and personal performance rating than other individuals.

2.1 Introduction

About 46% of non-managers and 60% of managers receive performance based payments (U.S. Census Bureau 2015). In such payment schemes, companies usually define rating scales to evaluate employees' performance and determine payments accordingly.

Studies find that employees almost never rank in lower rating categories (see for example Ockenfels et al. 2015, Frederiksen et al. 2017) and thus that lower rating categories are often unused in performance appraisals.

The literature in psychology and economics considers unused rating categories in performance appraisals as rating biases of supervisors who assign too lenient ratings (see for example Landy and Farr 1980 and Prendergast 1999). However, most of the companies do not prevent unused rating categories by for example requiring supervisors to rank predefined percentages of employees in each rating category (Holland 2006). While low rating categories are often unused, they are not removed from the scale of possible evaluations.

Motivated by the observation that many firms employ scales where the lowest categories are unused, we investigate unused low rating categories as an incentive design choice. We examine whether the presence of an unused low rating category in performance appraisals increases performance in short-term employer-employee relations. We refer to an unused low rating category as "dummy category".

When individuals believe that a dummy category is actually used, economic reasoning and tournament theory suggest that it triggers an *Incentive Effect* that raises performance. Employees have higher incentives to perform in the presence of a dummy category as low performance may result in lower performance ratings and monetary payments. In this light, Berger et al. (2013) show that forcing supervisors to use lower rating categories increases performance. Moreover, following Lazear and Rosen (1981), a dummy category increases incentives when the payment scheme resembles a tournament since it increases the (perceived) prize spread of possible payments.

A broad stream of literature demonstrates that individuals reciprocate behavior of others by rewarding favors and penalizing unkindness (Rabin 1993, Fehr et al. 1993, 1997, Fehr and Rockenbach 2003, Falk et al. 2008). Employees receive more generous ratings in the form of higher relative ratings when they believe that the dummy category is actually used. Accordingly, a dummy category may trigger positively reciprocal reactions that raise performance (see for example Ockenfels et al. 2015, Sebald and Walzl 2014). We refer to this as *Evaluation Effect*. However, such a dummy category may also be seen as unkind and signal bad intentions of the employer as an additional punishment option is introduced (Bowles and Polanía-Reyes 2012). This, in turn, may trigger negatively reciprocal reactions that reduce performance (Levine 1998, Dufwenberg and Kirchsteiger 2004). We refer to this as *negative Kindness-of-the-Scale Effect*.

When employees know that the category is unused, incentives and performance ratings of employees are the same with and without a dummy category. However, a transparent dummy category may signal kindness and good faith of employers transmitting that they intentionally do not use an available punishment option. This, in turn, may induce positively reciprocal reactions and increase performance (*positive Kindness-of-the-Scale Effect*).

We tested how a dummy category affects performance in short-term employer-employee relations in two field studies in the online labor market of Amazon's Mechanical Turk (MTurk; Horton et al. 2011). In Study I, we investigated the potential reciprocal reactions to a dummy category when subjects believe that the category is used. More specifically, we tested whether the *Evaluation Effect* raises performance and hence overcompensates the *negative Kindness-of-the-Scale Effect*. Therefore, we excluded the *Incentive Effect* of a dummy category by design. In Study II, we examined the total performance effect of a dummy category and hence the joint effect of potential reciprocal reactions and the *Incentive Effect*.

More specifically, we tested whether a dummy category raises performance and hence whether the *Incentive Effect* and *Evaluation Effect* jointly raise performance and thus overcompensate the *negative Kindness-of-the-Scale Effect.* Moreover, we examined whether a transparent dummy category and thus the *positive Kindness-of-the-Scale Effect* raises performance.

Both studies followed the same protocol. As a university department, we hired subjects to digitize handwritten class grades but did not disclose that this task was an experiment. Subjects worked twice in two consecutive weeks. We used week one only to rank and provide feedback to subjects in week two. For their work in week one, subjects received a bonus payment based on relative performance. We explained the incentive mechanism of the bonus payment, but did not reveal the rating scale such that week one was identical across treatments. Accordingly, we analyze the treatment effect only in week two. In week two, subjects saw their individual performance rating and resulting bonus payment for week one before they worked again on the same task. To evaluate how a dummy category affects subjects' well-being and the perception of rating scales, we asked how satisfied they were with their performance rating and how kind they perceived their rating scale.

The computer rated performance using three categories but subjects either saw three or four rating categories, depending on the treatment. In treatment No Dummy (ND), subjects saw the actual three rating categories used by the computer. In treatment Dummy (D), subjects saw an additional fourth category that was never used. We did not inform that the additional category was never used.

In Study I, we tested whether the potential positive *Evaluation Effect* raises performance and hence is stronger than the potential *negative Kindness-of-the-Scale Effect.* To exclude the potential *Incentive Effect*, subjects did not receive a rating or bonus payment but the same fixed payment in both treatments ND and D in week two. Thus, only the rating (scale) shown for week one varied between treatments as either 3 or 4 rating categories were displayed in the performance appraisal. Accordingly, differences between treatments in week two can be attributed to the rating (scale) seen for week one.

A dummy category did not raise performance in week two of Study I: Average performance and performance across rating categories did not differ significantly between treatments ND and D. Subjects did not report higher satisfaction with their individual rating when seeing a dummy category in treatment D. They did, however, evaluate the rating scale in treatment D as being less kind.

In Study II, we tested the total performance effect of a dummy category. We analyzed whether a dummy category raises performance when subjects believe that the dummy category is used and hence whether the *Evaluation Effect* and *Incentive Effect* are stronger than the *negative Kindness-of-the-Scale Effect*. Subjects received a bonus payment based on relative performance and an additional relative performance rating also for week two: They learned that the rating scale of week one was also used for week two. Thus, not only the rating (scale) shown for week one but also the incentive scheme in week two varied between treatments D and ND. Subjects in treatment D faced an additional low rating category, which increased the (perceived) prize spread of the bonus tournament compared to treatment ND. Accordingly, differences between these treatments in week two may not only be influenced by the rating (scale) seen for week one but also by the anticipation of and hence incentive induced by the performance rating and payment for week two.

A dummy category did not raise average performance in week two of Study II either. We do, however, observe opposing performance effects of a dummy category. Subjects receiving the lowest rating worked significantly more while those receiving higher ratings worked significantly less in treatment D. As in Study I, subjects did not report different rating satisfaction between treatments. Moreover, the rating scale in treatment D was perceived as less kind – however, only from those receiving the lowest rating.

In addition, we analyzed whether a dummy category raises performance when subjects are informed that the category is unused (*positive Kindness-of-the-Scale Effect*). Therefore, we ran an additional treatment Transparent Dummy (TD): Subjects saw an additional unused low rating categories but were informed that this rating category was unused. The communicated number of rating categories used and the prize spread were equivalent in treatments ND and TD.

Also a transparent dummy category did not raise performance in week two of Study II: Average performance and performance across rating categories did not differ significantly between treatments ND and TD. However, subjects who did not receive the lowest rating category evaluated the rating scale in treatment TD as being more kind.

Our results suggest two main insights: (1) A dummy category did not raise average performance in short-term employer-employee relations. While low performers showed higher performance, the remaining subjects showed lower performance when they did not know that the dummy category was unused. (2) It seems that individuals do not only consider their monetary incentives and personal performance rating but also pay attention to the design and kindness of a rating scale in short-term employer-employee relations. The results also indicate that low performers seem to focus more on their personal performance ratings and monetary incentives than other individuals.

This chapter is closest to Chapter 3 where we investigate how a dummy category affects performance in a setting where employees experience multiple ratings and can react dynamically. The results of our investigation of Chapter 3 are consistent with the findings of this chapter: We do not see significant performance differences in the first working period and the performance differences that evolve over time are driven by low performers who increase performance stronger than other individuals.

The remainder of this chapter is structured as follows. In Section 2.2, we develop our hypotheses. In Sections 2.3 and 2.4, we present Study I and II, respectively. In Section 2.5, we conclude.

2.2 Literature and Hypotheses Development

Research on reciprocity in employer-employee relations, which originated from the theory of gift exchange (Adams 1963, Akerlof 1982), shows that employees reciprocate kind and punish unkind behavior (see for instance Akerlof and Yellen 1988, 1990, Fehr et al. 1993, 1997, Charness 2004, Chung and Narayandas 2017). Falk et al. (2008) show that intentions of the gift-giver are crucial to provoke reciprocal reactions (see also Falk and Ichino 2006, Kube et al. 2012).

Deploying a dummy category may increase performance by triggering positively reciprocal responses when subjects believe the category is actually used: Ceteris paribus, employees receive more generous feedback in the form of relatively higher ratings: For example, a rating in the category 2 of 3 (top 66%) becomes a rating in the category 2 of 4 (top 75%). This may shift employees' reference point which, in turn, increases positively reciprocal reactions among those who receive high ratings or reduce negatively reciprocal reactions of those whose ratings fall short of their expectations (for an analysis of reciprocal reactions to performance ratings see for instance Ockenfels et al. 2015). Bol (2011, p.1555) points out: "The behavioral perspective expects leniency bias to positively affect perceived fairness by increasing the congruence between the rating the employee thinks s/he deserves and the rating s/he actually receives". Following the argument of Ellingsen and Johannesson (2007), more generous ratings positively influence employees if they see them as a sign of employer appreciation. Moreover, receiving higher relative ratings may also make employees happier (Parducci 1965) and consequently motivate higher performance (Oswald et al. 2015). We refer to a positive performance effect of awarding higher relative ratings in the presence of a dummy category as *Evaluation Effect*.

But there may also be a negatively reciprocal performance effect when individuals believe that the dummy category is used. Bowles and Polanía-Reyes (2012, p.388) argue that incentive schemes transmit information about the type and intentions of the incentive designer. If the specific incentive scheme signals an employer's "bad" intentions, employees may punish this. Adding a low rating category to the rating scale may be judged as being unkind and signal "bad news" about the employer's type or intention since an additional punishment option is introduced. In turn, this can induce negative reactions (Fehr and Rockenbach 2003). We refer to a negative performance effect of employing a less kind rating scale in the presence of a dummy category as *negative Kindness-of-the-Scale Effect*.

We expected individuals to focus more on their personal performance rating than on the overall kindness of a rating scale. Therefore, we hypothesized the positive *Evaluation Effect* to be stronger than the *negative Kindness-of-the-Scale Effect* when individuals are not informed that the dummy category is unused. We preregistered:

Hypothesis 1a: Evaluation Effect *If incentives are held constant, average performance is higher in the presence of a dummy category when individuals are not informed that the dummy category is unused than if there is no dummy category.*

Simple economic reasoning suggests that a dummy category raises performance when employees believe that the category is actually used. Employees have higher incentives to work as the additional low rating category penalizes low performance more. A key result in tournament theory is that higher prize spreads should induce higher performance (see for example Lazear and Rosen 1981). Since adding a dummy category increases the (perceived) prize spread, a dummy category should raise performance. We refer to a positive performance effect of higher incentives in the presence of a dummy category as *Incentive Effect*.

We expected individuals to focus more on their personal performance rating and monetary incentives than on the overall kindness of a rating scale. Therefore, we conjectured the positive *Incentive* and *Evaluation Effect* of a dummy category to be stronger than the *negative Kindness-of-the-Scale Effect* when individuals are not informed that the category is unused. Consequently, we preregistered:

Hypothesis 1b: Evaluation & Incentive Effect *Average performance is higher in the presence of a dummy category when individuals are not informed that the dummy category is unused than if there is no dummy category.*

When individuals know that the dummy category is unused it may trigger positively reciprocal reactions. Ratings and incentives do not change if individuals know that the dummy category is never used. However, the transparency of not using the lowest category may transmit "good news" about employers as it signals that they refrained from using an available punishment option. Following the reasoning outlined above, this might signal kindness and good faith of employers and in turn increase performance. We refer to a positive performance effect of employing a more kind rating scale in the presence of a transparent dummy category as *positive Kindness-of-the-Scale Effect.* We hence preregistered:

Hypothesis 2: Positive Kindness-of-the-Scale Effect *Average performance is higher in the presence of a dummy category when individuals are informed that the dummy category is unused than if there is no dummy category.*

We expected that the *Evaluation* and *Incentive Effect* are stronger than the *Positive Kindness-of-the-Scale Effect* since we expected that individuals focus more on their personal performance rating and monetary incentives. Accordingly, we preregistered:

Hypothesis 3: Evaluation & Incentive Effect II *Average performance is higher in the presence of a dummy category when individuals are not informed that the dummy category is unused than when they are informed that the dummy category is unused.*

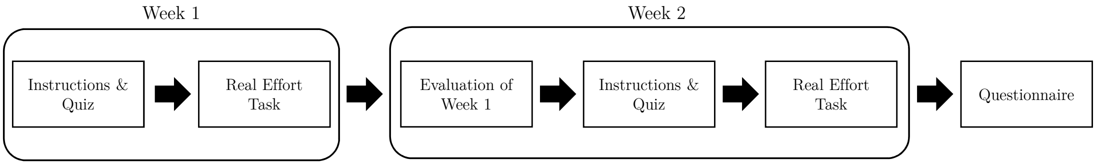
The literature suggests that those receiving the lowest possible performance rating react stronger to an additional low – but unused – rating category. Studies on performance feedback report a de-motivating performance effect of receiving negative feedback in the form of low ratings, see for example Barankay (2011, 2012) or Gill et al. (2019). If individuals are not aware that the category is unused, employing a dummy category avoids giving harsh negative feedback to those with the lowest rating. In the same light – if subjects know that the dummy category is unused – we expect those with the lowest rating to perceive the kind act of not using a punishment option stronger. Moreover, one might argue that the proposed *Incentive Effect* is stronger or – even more conservative – only present for those with the lowest rating as they have the highest probability to be ranked in that category. We refer to a stronger performance increase of those with the lowest rating when seeing a (transparent) dummy category as *Last Place Effect*. We hence preregistered:

Hypothesis 4: Last Place Effect *The performance increase in the presence of a dummy category is stronger for those with the lowest performance rating than for those with higher performance ratings.*

2.3 Study I: Performance Effect of a Dummy Category When Incentives are Held Constant

In Study I, we tested the reciprocal responses to a dummy category when subjects are not informed that the dummy category is unused. We tested Hypothesis 1a and investigated whether the potentially positive performance effect of giving more generous feedback (*Evaluation Effect*) outweigh the potentially negative performance effect of employing a less kind rating scale (*negative Kindness-of-the-Scale Effect*). Therefore, we excluded the potential *Incentive Effect* of a dummy category by holding incentives constant between treatments. We also analyzed

Figure 2.1: Experimental Procedure



whether the performance effect was stronger for those with the lowest performance rating (Hypothesis 4).

2.3.1 Experimental Design

Overview As a university department, we recruited subjects from Amazon's Mechanical Turk online labor market to digitize handwritten grades. Our design is in accordance with standard ethical guidelines but we did not disclose that the task was an experiment. Subjects worked in two consecutive weeks. We paid a bonus payment based on relative performance for week one. We explained that subjects receive higher bonus payments, the higher they rank relative to their peers. However, we did not explain details about the rating scale such that week one was identical across treatments. Subjects received private performance feedback for week one before they worked again in week two. In treatment Dummy, we displayed a dummy category in the performance evaluation of week one. In treatment No Dummy, we did not display a dummy category. For their work in week two, subjects did not receive a performance dependent payment or performance rating in order to prevent the potential *Incentive Effect* from different incentives between treatments. We sent out a questionnaire on the kindness of the rating scales and demographics after week two (see Appendix 2.G). Figure 2.1 shows the experimental procedure. See Appendices 2.D & 2.E for screenshots of week one and two.

Experimental Details We used week one only to provide performance ratings in week two. Therefore, it was kept the same across treatments to avoid treatment specific performance effect that distort ratings between treatments. Subjects worked for 20 minutes digitizing grades from scanned exam cover sheets. See Appendix 2.A for details on the real effort task. Subjects learned that they received a bonus based on relative performance in addition to a fixed wage. However, they did not learn the rating scale or respective number of rating categories. Performance was defined as the number of correctly entered cover sheets; a cover sheet was evaluated as entered correctly if all grades were entered correctly. Subjects had to pass a quiz on the task and payment structure to be able to work.

We used week two to test the performance effect of a dummy category. Subjects were invited via e-mail to work again. They could work on the task between Monday and Friday. Upon entering the task, subjects received private performance ratings for their work in week one. Depending on the treatment, subjects saw or saw not a dummy category in the rating scale. They were then asked how satisfied they were with their individual rating. The incentive scheme of week two was explained in the instructions afterwards: We paid the same fixed wage in both treatments but no performance dependent bonus to eliminate a performance effect due to different incentive schemes. Hence, subjects did not receive rank feedback for week two. Subjects had to pass a quiz on the task and payment structure to work again. Working time was not restricted.

The computer rated performance in week one using three rating categories in both treatments. Ratings were based on relative performance and followed the same procedure in all treatments such that only categories one to three were actually used. Category one was used for the highest performing subjects and category three to the lowest performing subjects. Subjects were not informed about the specific details of the rating procedure.

Treatment Variation Subjects either saw three or four rating categories, depending on the treatment. Figure 2.2 depicts exemplary the scale subjects saw when receiving the rating ”Grade 3” across treatments. In treatment “No Dummy”

Figure 2.2: Performance Rating When Receiving Grade 3 Across Treatments

	Grade	1	2	3
No Dummy	Bonus	$2.00	$1.50	$1.00
	% of workers	30%	40%	30%

	Grade	1	2	3	4
Dummy	Bonus	$2.00	$1.50	$1.00	$0.00
	% of workers	30%	40%	30%	

(ND) subjects saw the actual three-point rating scale. In treatment "Dummy" (D), an additional fourth rating category was displayed at the bottom. Subjects were not informed that the additional category was unused. We randomly assigned subjects to either treatment D or ND stratifying assignment based on the performance in week one. In treatment ND, subjects learned that 30% of the ratings were given in the top category – Grade 1 –, 40% in the middle category – Grade 2 –, and 30% in the lowest category – Grade 3. To avoid deception in treatment D, subjects learned that 30% of the ratings were given in category 3 and 4 that is, Grade 3 and 4, respectively.

Experimental Protocol and Subject Pool We recruited subjects on Amazon's Mechanical Turk (MTurk) online labor market using the service of TurkPrime (Litman et al. 2017). The experiment was conducted online with Qualtrics and a self-developed Javascript.

Over the past decade, MTurk has received increased attention of researchers as platform to conduct scientific experiments: See for example Horton (2010), Barankay (2011), and the reference in Horton et al. (2011). On Amazon's platform, employers can post job offers, so called Human Intelligence Tasks (HITs), to a workforce of at least eighty-five thousand US workers active during the time of our study (Robinson et al. 2019). For a more detailed description of the marketplace see Ipeirotis (2010) or Paolacci et al. (2010).

We ran our treatments in March 2018. We recruited subjects on Monday and Tuesday. No subject participated in more than one treatment. Subjects were invited via e-mail to work again on Monday the week after. To increase the likelihood of returning in week two, subjects could work on the task all week (Monday-Friday) as well as pause and return to the task later. We only recruited residents of the United States and required workers to have had completed at least 100 HITs with an approval rate of at least 90% to ensure that subjects were familiar with MTurk, avoid complications arising from difficulties in understanding the English task instructions, and to prevent performance noise due to different time-zones. Note that theses sampling restrictions still allow a sufficient total population size (Robinson et al. 2019) and thus worker non-naïveté (Chandler et al. 2014) cannot be a problem in our study.

Selective attrition is not a concern in our study. We avoided selection in the return rate, as treatment details were revealed only after subjects returned in week two. However, when returned, subjects could drop-out after receiving their performance rating. Moreover, we excluded subjects from the experiment who failed the quiz or worked on a device without sufficient screen resolution. Additionally, subjects had the choice to answer the questionnaire as it was sent out after working in week two. We check selective attrition for the afore mentioned cases. There are no statistically significant differences between treatments neither for the drop-out rates ($\chi^2(1) = 0.27$ $p = .60$), the screen-out rates ($\chi^2(1) = 0.47$ $p = .49$) or the questionnaire-return rates ($\chi^2(1) = 0.05$ $p = .82$). Our study hence does not suffer from selective attrition.

946 subjects completed week two. Of those who answered the questionnaire, 58% were female, the median age was 35. The median educational level was a bachelor's degree and the median income class ranged from $30,001 to $40,000. See Table 2.3 in the Appendix 2.B for detailed sample demographics. Earnings ranged between $5.50 and $8.50 depending on the bonus payment. The median experiment duration was 46.67 minutes (sum of week one and two). This results in a median hourly wage of $7.55, which is above median earnings on MTurk and above the federal minimum wage in the United States.

2.3.2 Results

We first analyze how a dummy category affected performance. We then examine questionnaire data to explore how it affected rating satisfaction and the perceived kindness of a rating scale.

Performance Effect of a Dummy Category

We hypothesized that average performance is higher in treatment Dummy (D) compared to treatment No Dummy (ND) since we expected individuals to focus more on their personal performance rating than on the kindness of a rating scale (Hypothesis 1a: *Evaluation Effect*).

We analyze performance in week two as week one was the same across treatments. Subjects did not receive a performance rating for week two and incentives did not differ between treatments in week two. Thereby, we excluded any potential *Incentive Effect* of a dummy category. Hence, performance can only be affected by reciprocal responses induced by the additional rating category shown.

Contrary to our hypothesis, a dummy category did not increase average performance: The coefficient of the treatment indicator "Dummy Category" is insignificant and negative in our regression analysis shown in column (1) of Table 2.1. We report OLS regressions and control for week one performance to capture individual performance differences that can still occur despite our sampling procedure. The results are robust to using tobit regressions and controlling for the day, as well as time of day subjects worked (Appendix 2.C.1). Thus, we do not find support that the *Evaluation Effect* is stronger than the *Kindness-of-the-Scale Effect* when subjects receive one performance rating. It seems that the two opposing reciprocal performance effects offset each other.

We next analyze how a dummy category affected the performance of those with the lowest performance rating. We hypothesized that the performance effect

Table 2.1: Performance Effect of a Dummy Category When Incentives are Held Constant

Dependent Variable: Number of Cover Sheets Entered Correctly	ND vs. D (1)	 (2)
Dummy Category	-3.77 (2.78)	-4.57 (3.55)
Dummy Category#Grade 3 in t-1		2.98 (5.19)
Grade 3 in t-1		-2.26 (5.18)
Pre-round Performance	0.57*** (0.05)	0.57*** (0.08)
Constant	15.86*** (3.38)	16.96*** (6.51)
Observations	946	946

Notes: Ordinary least squares regressions on individual output are performed. D, Dummy Treatment; ND, No Dummy Treatment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

is stronger for these subjects since being rated second last in treatment D as compared to last in treatment ND may trigger a stronger reaction (Hypothesis 4: *Last Place Effect*).

Grade 3 was the lowest possible rating. In our setting, grades were equivalent to performance ranks in the pre-round: Subjects receiving grade 3 belonged to the lowest 30% of the performance distribution, grade 1 and 2 belonged to the top 70%.

Contrary to our hypothesis, also low performers did not work more when seeing a dummy category: The interaction term of receiving grade 3 with the treatment indicator "Dummy Category" in column (2) of Table 2.1 is insignificant. The results are robust to using tobit regressions and controlling for day and time of day (see Appendix 2.C.1). Hence, we also do not observe a *Last Place Effect* when subjects receive only one performance rating.

Effect of a Dummy Category on Rating Satisfaction

We test whether individuals in treatment D were more satisfied with their rating (grade) than those awarded with the same category in treatment ND.

When receiving their performance rating in week two, we asked subjects – on the same screen – how satisfied they were with their rating. The scale ranged from 1 (not satisfied at all) to 7 (extremely satisfied). The payment scheme of the second part could not influence satisfaction as incentives did not differ between treatments and were communicated after the rating was shown.

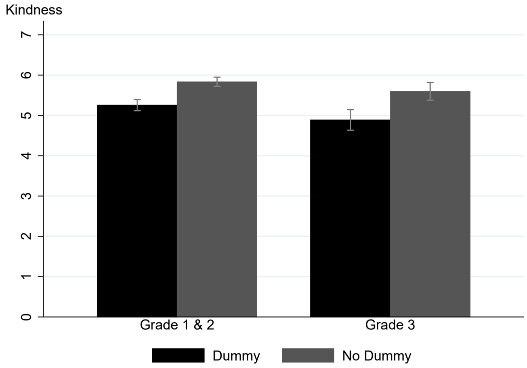
Subjects across treatments did not report different rating satisfaction when receiving grade 1, 2, or 3 (see Table 2.6 in Appendix 2.C.2 for the *p*-values of Wilcoxon rank-sum tests comparing the satisfaction levels across grades). Hence, in case of one evaluation, seeing a dummy category did not increase rating satisfaction – in line with the results comparing performance between treatments.

Effect of a Dummy Category on the Perceived Kindness of a Rating Scale

To investigate whether a dummy category conveys "bad news" about an employer, we analyze survey data obtained in the post-trial questionnaire (see Appendix 2.G). We showed subjects the rating scale of their treatment again and asked how kind they perceived it. The scale ranged from 1 (very unkind) to 7 (very kind).

If a dummy category was interpreted as bad news, subjects should evaluate the rating scales of treatment Dummy (D) as less kind. We test this by comparing subjects' kindness evaluations between treatments ND and D. Figure 2.3 shows the mean kindness evaluations between treatments. We differentiate between subjects receiving the top rating categories grade 1 and 2 (on the left) and the lowest rating category grade 3 (on the right hand side) to analyze whether performance effect differs across performance classes.

Figure 2.3: Evaluation of the Kindness of a Rating Scale Across Treatments I

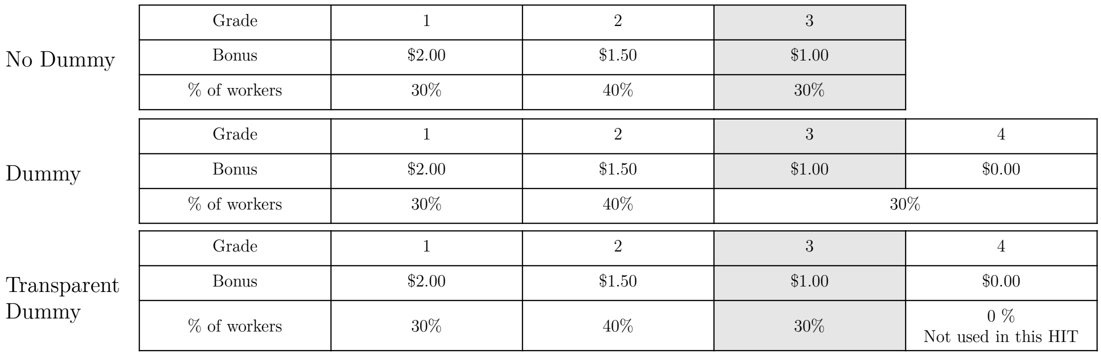


The additional low rating category was interpreted as "bad" news. Across performance classes, subjects evaluated the rating scale in treatment D – when they did not know that the additional rating category was unused – as being less kind (Wilcoxon rank-sum test, two-sided, $p = .000$ and $p = .000$, respectively).

2.4 Study II: Performance Effect of a Dummy Category

In Study II, we investigated the total performance effect of a dummy category. We tested whether a dummy category raises performance when individuals believe that the category is used (Hypothesis 1b) and hence whether the potentially positive performance effect of higher incentives (*Incentive Effect*) and more generous feedback (*Evaluation Effect*) outweigh the potentially negative performance effect of employing a less kind rating scale (*negative Kindness-of-the-Scale Effect*). In addition, we tested whether a dummy category raises performance when individuals are informed that the dummy category is unused (Hypothesis 2) and hence whether a potentially more kind rating scale raises performance (*positive Kindness-of-the-Scale Effect*). We also examined whether performance was higher in the treatment where individuals were not informed that the dummy category is unused than in the treatment where they were informed (Hypothesis 3). In all three analyses, we

Figure 2.4: Performance Rating When Receiving Grade 3 Across Treatments



	Grade	1	2	3
No Dummy	Bonus	$2.00	$1.50	$1.00
	% of workers	30%	40%	30%

	Grade	1	2	3	4
Dummy	Bonus	$2.00	$1.50	$1.00	$0.00
	% of workers	30%	40%	30%	

	Grade	1	2	3	4
Transparent Dummy	Bonus	$2.00	$1.50	$1.00	$0.00
	% of workers	30%	40%	30%	0 % Not used in this HIT

tested if the performance effect was stronger for those with the lowest performance rating (Hypothesis 4).

2.4.1 Experimental Design

Compared to Study I, subjects received a bonus payment and hence performance rating also for their performance in week two. Study II followed the protocol of Study I and everything was the same except for the payment scheme in week two. Subjects learned that the rating scale of week one was also used for determining the rating and payment in week two. Accordingly, not only the rating (scale) shown for week one but also the anticipation of and the incentives induced by the rating in week two can affect performance in week two. See Appendices 2.D - 2.F for screenshots of week one and week two.

To test the effect of a dummy category when individuals know that the category is unused, we ran an additional treatment ”Transparent Dummy” (TD). In the new treatment subjects also saw four rating categories but learned that the fourth rating category was unused. Note that treatment ND and treatment TD have the same prize spread and the communicated number of rating categories in use is equivalent. Figure 2.4 shows exemplary the scale subjects saw when receiving the rating ”Grade 3” across treatments. We randomly assigned subjects to either treatment D, ND, or TD stratifying assignment based on the performance in week one.

We recruited subjects on Amazon's Mechanical Turk (MTurk) in June 2018. We excluded participants of Study I and no subject took part in more than one treatment. A questionnaire was sent out to all subjects two weeks after week two.

Selective attrition is not a concern in our study. We avoided selection in the return rate as treatment details were revealed only after subjects returned in week two. We check selective attrition for the drop-out, screen-out, and questionnaire return rates. There are no statistically significant differences across treatments neither for the drop-out rates ($\chi^2(2) = 0.21$ $p = .90$), the screen-out rates ($\chi^2(2) = 1.78$ $p = .41$), or the questionnaire-return rates ($\chi^2(2) = 4.63$ $p = .10$).

1,398 subjects completed week two. Of those who answered the questionnaire 61% were female, the median age was 34. The median educational level was a bachelor's degree and the median income class ranged from \$30,001 to \$40,000. See Table 2.3 in the Appendix 2.B for detailed sample demographics. Earnings ranged between \$5.50 and \$8.50 depending on the bonus payment. The median experiment duration was 51.47 minutes resulting in a median hourly wage of \$8.74, which is substantially above median earnings on MTurk and above the federal minimum wage in the United States.

2.4.2 Results

We first analyze the performance effect of a dummy category. We then examine whether a dummy category affected individual rating satisfaction or the perceived kindness of a rating scale.

Performance Effect of a Dummy Category

In all treatments, incentive schemes in week two resembled a tournament in which participants competed for bonus payments. Subjects in treatment ND

Table 2.2: Performance Effect of a Dummy Category

Dependent Variable:	ND vs. D		ND vs. TD		D vs. TD	
Number of Cover Sheets Entered Correctly	(1)	(2)	(3)	(4)	(5)	(6)
Dummy Category	-3.69	-7.51*			-4.54	-5.50
	(3.31)	(4.05)			(2.93)	(3.51)
Dummy Category#Grade 3 in t-1		15.32**				3.68
		(6.62)				(6.30)
Transparent Dummy Category			0.90	-2.05		
			(3.35)	(4.01)		
Transparent Dummy Category#Grade 3 in t-1				11.56		
				(7.07)		
Grade 3 in t-1		-14.48**		-15.11**		-0.14
		(6.83)		(6.64)		(6.08)
Pre-round Performance	0.95***	0.88***	0.96***	0.87***	0.92***	0.94***
	(0.05)	(0.08)	(0.05)	(0.08)	(0.05)	(0.07)
Constant	26.05***	34.47***	25.32***	35.49***	29.15***	28.02***
	(4.26)	(7.59)	(4.15)	(7.14)	(3.92)	(6.39)
Observations	928	928	934	934	934	934

Notes: Ordinary least squares regressions on individual output are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

and TD faced a three prize tournament while those in treatment D entered a (perceived) four prize tournament. A conjecture of the incentive literature is that the dummy category in treatment D – where subjects did not know that it is unused – increases performance (*Incentive Effect*). Compared to Study I, behavior in these experimental conditions may hence not only be influenced by a different rating (scale) shown (*Evaluation & negative Kindness-of-the-Scale Effect*) but also by different communicated incentives (*Incentive Effect*).

Table 2.2 shows our regression results. We investigate the performance effect in week two as week one was the same across treatments. We report OLS regressions and control for week one performance to capture individual performance differences that can still occur within the degrees of freedom of our sampling procedure. The results are robust to controlling for the day and time of day subjects worked as well as performing tobit regressions (see Appendix 2.C.1).

First, we compare treatment D – where individuals were not informed that the dummy category was unused – to treatment ND. We expected the positive performance effect of higher relative ratings (*Evaluation Effect*) and higher incentives

(*Incentive Effect*) to be stronger than the potentially negative performance effect of a less kind rating scale in the presence of a dummy category (*negative Kindness-of-the-Scale Effect*). We thus hypothesized that average performance in treatment D is higher than in treatment ND (Hypothesis 1b: *Evaluation & Incentive Effect*).

The results do not support our hypothesis. Average performance in treatment D was not higher than in treatment ND: The coefficient of the treatment indicator "Dummy Category" is insignificant and negative in column (1) of Table 2.2. Thus, we do not find support for a stronger *Evaluation* and *Incentive Effect* when subjects receive two ratings. Instead, it seems that the negative performance effect of a less kind rating scale offsets the positive performance effect of higher ratings and higher incentives.

We next analyze the performance effect of the dummy category in treatment D on subjects receiving the lowest rating (grade 3). The underlying hypothesis is that both *Incentive* and *Evaluation Effect* are stronger for those with the lowest performance rating (Hypothesis 4: *Last Place Effect*).

The results support the hypothesis of a *Last Place Effect* in the presence of a dummy category. Subjects who received grade 3 in treatment D worked significantly more than subjects with the same rating in treatment ND: The interaction term of the treatment indicator "Dummy Category" with receiving a grade 3 for the performance in week one is significant in column (2) of Table 2.2. Interestingly, subjects receiving grade 1 or 2 worked significantly less in treatment D indicated by the significant negative treatment indicator "Dummy Category" in column (2) of Table 2.2. This indicates that – when individuals receive two ratings – a dummy category has a positive performance effect only on those with the lowest performance rating. Accordingly, the *Evaluation* and *Incentive Effect* seem to have a stronger performance effect for these individuals.

Second, we compare treatment TD – where individuals did know that the dummy category was unused – to treatment ND. Incentives did not differ between these treatments. However, a transparent dummy category might signal kindness of the employer and induce positively reciprocal reactions that increase performance. We hence, hypothesized that average performance in treatment TD is higher than in treatment ND (Hypothesis 2: *Positive Kindness-of-the-Scale Effect*).

Contrary to our hypothesis, we do not find support for a *positive Kindness-of-the-Scale Effect* when subjects receive two ratings. Average performance in treatment TD was not higher than in treatment ND: The coefficient of the treatment indicator "Transparent Dummy Category" is statistically and economically insignificant in column (3) of Table 2.2.

Comparing performance between those receiving the lowest rating, we do not find support for a *Last Place Effect* in presence of a transparent dummy category (Hypothesis 4: *Last Place Effect*). The interaction term of the treatment indicator "Transparent Dummy Category" with grade 3 is insignificant in column (4) of Table 2.2.

Third, we compare performance in treatment D with performance in treatment TD. We hypothesized that average performance is higher in treatment D than in treatment TD (Hypothesis 3: *Evaluation & Incentive Effect II*) since we expected individuals to focus more on their personal performance rating and monetary incentives than on the kindness of a rating scale.

We do not find support that average performance is higher in the presence of a dummy category as compared to a transparent dummy category. Average performance was not significantly different between treatments D and TD: The coefficient of the treatment indicator "Transparent Dummy Category" is insignificant and negative in column (5) of Table 2.2.

We do not find support for a *Last Place Effect* (Hypothesis 4) since there are also no significant differences comparing performance between those with the lowest performance rating. The interaction term of the treatment indicator "Dummy Category" with grade 3 is insignificant in column (6) of Table 2.2.

Effect of a Dummy Category on Rating Satisfaction

We analyze questionnaire data to investigate whether higher relative ratings in treatment Dummy increase subjects' rating satisfaction. When subjects received their performance rating in week two, we asked them how satisfied they were with their rating. The scale ranged from 1 (not satisfied at all) to 7 (extremely satisfied).

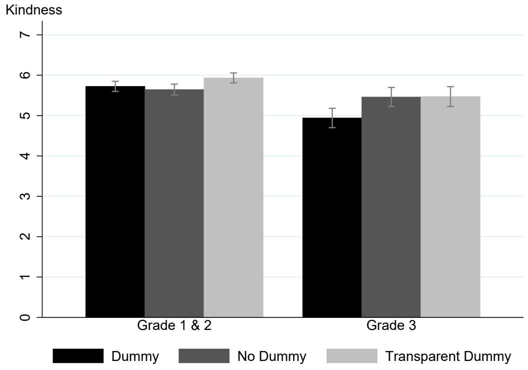
As in Study I, higher relative ratings did not increase rating satisfaction in Study II, which is also in line with the results of the performance analysis. Subjects across treatments did not report different rating satisfaction when receiving grade 1, 2, or 3 (see Table 2.6 in Appendix 2.C.2 for the *p*-values of Wilcoxon rank-sum tests comparing the satisfaction across grades). Note that the different payment schemes across treatments can not influence responses as they were communicated afterwards.

Effect of a Dummy Category on the Perceived Kindness of a Rating Scale

To analyze whether the presence of a dummy category affects the perceived kindness of a rating scale, we report survey data obtained in the post-trial questionnaire (see Appendix 2.G). We showed subjects the rating scale of their treatment again and asked how kind they perceived it on a scale from 1 (very unkind) to 7 (very kind). Increasing values of the score reflect higher kindness levels. Figure 2.5 shows the kindness evaluations of subjects receiving grade 1 and 2 on the left and grade 3 on the right hand side. See Table 2.7 in Appendix 2.C.2 for the *p*-values of Wilcoxon rank-sum tests comparing the kindness across treatments.

If individuals do not know that the additional rating category is unused, a dummy category introduces an additional punishment option and thereby may signal

Figure 2.5: Evaluation of the Kindness of a Rating Scale Across Treatments II



an employers unkindness and bad intentions. The literature suggests – and we also observed in Study I – that individuals perceive the rating scale in treatment Dummy (D) as being less kind compared to the rating scale used in treatment No Dummy (ND).

We find again support that a dummy category is interpreted as "bad" news. Subjects receiving grade 3 – the lowest rating – evaluated the rating scale in treatment D as being less kind than the scale in treatment ND (Wilcoxon rank-sum test, two-sided, $p = .000$). Subjects receiving grade 1 and 2, however, did not perceive the rating scale in treatment D as being less kind than in treatment ND (Wilcoxon rank-sum test, two-sided, $p = .544$).

If individuals do know that the additional rating category is unused, employing a transparent dummy category may signal kindness and good intentions of employers as they refrain from using an available punishment option. Therefore, the literature suggests that subjects perceive the scale in treatment Transparent Dummy (TD) as being more kind compared to the rating scale used in treatment D and ND.

We find support that employing a transparent dummy category is interpreted as being kind. Compared to treatment D, subjects across performance classes evaluated the scale in treatment TD as more kind (Wilcoxon rank-sum test, two-sided, $p = .011$ $p = .002$ comparing grade 1&2 and grade 3, respectively). Compared to treatment ND, only subjects receiving grade 1 and 2 evaluated the scale in treatment TD as more kind (Wilcoxon rank-sum test, two-sided, $p = .003$ $p = .782$ comparing grade 1&2 and grade 3, respectively).

2.5 Conclusion

We studied the performance effect of a dummy category in short-term employer-employee relations. Contrary to our hypotheses, a dummy category did not increase average performance in our setting – independent of whether subjects were informed or not that the additional category is unused.

We expected that more generous ratings (*Evaluation Effect*) and higher incentives (*Incentive Effect*) in the presence of a dummy category increase average performance and thus offset a potential negative performance effect arising from employing a less kind rating scale (*negative Kindness-of-the-Scale Effect*).

We did not see an increase of average performance in the presence of a dummy category. It seems that the opposing performance effects of a dummy category outweigh each other on the aggregate level. In this light, we observe that subjects with the lowest performance rating increased performance in the presence of a dummy category while those receiving higher ratings reduced performance: The *Incentive Effect* and *Evaluation Effect* of a dummy category seem to be stronger for those with the lowest performance rating than for those receiving higher ratings. Moreover, we found indication that subjects perceive the rating scale in treatment Dummy as being less kind.

We expected that employing a transparent dummy category increases average performance by triggering positively reciprocal reactions to employing a more kind rating scale (*positive Kindness-of-the-Scale Effect*).

We did not see a performance increase in the presence of a transparent dummy category either. Interestingly, we found indication that subjects perceive the rating scale in treatment Transparent Dummy as being more kind. However, these kindness perceptions did not translate into a performance effect.

Overall, we find that a dummy category raised performance only for those with the lowest performance rating but we do not find evidence that a dummy category increases average performance in short-term employer-employee relations. It seems that individuals also pay attention to the kindness of rating scales in addition to their personal performance rating and monetary incentives when they receive one or two performance ratings. Moreover, the results suggest that personal performance ratings and monetary incentives have a stronger performance effect on low performing individuals than on other individuals.

In practice, employees usually work over multiple periods and receive multiple consecutive performance ratings over a longer time period. It is thus a very important question which of the effects – *Incentive, Evaluation*, or *Kindness-of-the-Scale Effect* – prevails or whether they cancel out in the long-run. We investigate this in the following Chapter 3.

2.6 Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for financial support through the research unit “Design & Behavior” [FOR 1371; projects TP3 and TP5]. The studies are preregistered under AEARCTR-0002736 and AEARCTR-0003029 (Study I and Study II, respectively). Please note that the change in the authorship is due to the fact that most of the work was done by Thomas Vogt. Please see Appendices 2.D - 2.G for the experimental instructions and screenshots.

Appendix of Chapter 2

2.A A Novel Real Effort Task

The working screen of the field experiment is shown in Figure 2.13 in Appendix 2.D. We provided sets of artificially created exam cover sheets that contained a table of six handwritten grades. Employees' task was to enter the grades displayed at the top into the entry fields on the bottom of the screen.

We chose the appearance of the real effort task in line with a job (1.) Amazon's Mechanical Turk (MTurk) workers are used to, (2.) a university department actually does, and (3.) which is reasonably outsourced. The task is very similar to the jobs otherwise found in the MTurk marketplace in terms of the type of work, difficulty, and time required. It is also common practice for university departments to digitize class results that were initially marked using pen and paper. To make it plausible that the class results had not been digitized yet, we labeled the artificially created exams with exam dates of the year 2004.

The real effort task is very tedious, simple, and demands low cognitive effort. It mainly requires attention and can be solved without prior knowledge. It induces positive effort costs and potential learning effects can be neglected. Due to the nature of the task, we assume that intrinsic motivation does not play a role in our setting.

We assured that the individual grading on a cover sheet as well as the overall class grading followed a reasonable distribution. Furthermore, grades were written from the same person to eliminate difficulties due to different handwriting. We created an initial set of 200 cover sheets. Based on the initial set we created new sets by changing the layout of the cover sheets – the exam date and number of pages of the exams – leaving the handwritten grades untouched. Note that thereby the actual grades on the exam cover sheets were identical across sets. We randomly varied which set was displayed and assured that every subject saw a specific version only once. In addition, the display order of individual cover sheets within one set was randomly varied for each subject. In Study I, subjects could enter up to 200, in Study II up to 400 cover sheets in week two.

We also paid attention to reducing noise in performance due to different technical skills and varying technical conditions. We did so by restricting the input of grades to capital letters and did not allow any special character other than ”+” and ”-”. Moreover, we required a screen resolution of at least 1200 (width) x 700 (height) such that none of the subjects had to scroll while transferring the grades.

2.B Sample Demographics

Table 2.3: Sample Demographics Study I & II

Demographics	Percentage	
	Study I (N=838)	Study II (N=1,339)
Age[1]	38.21 (12.05)	36.16 (11.36)
Female	58.00	61.24
Highest level of education		
Less than High school degree	0.00	0.60
High school graduate	6.80	8.51
Vocational/technical school	6.56	5.15
Some college	31.74	28.23
Bachelor’s degree	41.89	42.49
Master’s degree	10.38	12.40
Doctoral degree	0.84	1.12
Advanced professional degree (JD, MD, MBA, etc.)	1.79	1.49
Employment status		
Working (paid employee)	67.30	65.42
Working (self-employed)	16.23	16.36
Not working	14.44	15.46
Other	2.03	2.76
Annual income from all sources before taxes		
$10,000 or less	15.39	12.85
$10,001 to $20,000	10.38	10.68
$20,001 to $30,000	9.90	15.24
$30,001 to $40,000	14.08	12.70
$40,001 to $50,000	12.41	11.73
$50,001 to $60,000	12.05	11.80
$60,001 to $70,000	5.85	8.14
$70,001 to $80,000	7.76	7.54
Over $80,000	12.17	9.33

Note: [1] Mean in years (standard deviation)

2.C Further Analyses

2.C.1 Robustness Checks of Regressions in Table 2.1 & Table 2.2

Table 2.4: Effect of a Dummy Category on Individual Performance II

Dependent Variable:	Study I		Study II					
Number of Cover Sheets Entered Correctly	ND vs. D		ND vs. D		ND vs. TD		D vs. TD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dummy Category	-4.17 (2.92)	-4.37 (3.67)	-3.44 (3.35)	-7.57* (4.07)			-4.62 (2.95)	-5.61 (3.52)
Dummy Category#Grade 3 in t-1		0.75 (5.61)		16.70** (6.81)				3.78 (6.37)
Transparent Dummy Category					1.27 (3.37)	-1.98 (4.02)		
Transparent Dummy Category#Grade 3 in t-1						12.86* (7.26)		
Grade 3 in t-1		-1.50 (5.42)		-15.54** (6.97)		-16.21** (6.79)		-0.13 (6.10)
Pre-round Performance	0.61*** (0.05)	0.59*** (0.09)	0.97*** (0.06)	0.90*** (0.09)	0.98*** (0.05)	0.88*** (0.08)	0.93*** (0.05)	0.95*** (0.07)
Constant	12.57*** (3.63)	13.73** (6.80)	24.38*** (4.39)	33.28*** (7.64)	23.62*** (4.27)	34.33*** (7.19)	28.49*** (3.97)	27.32*** (6.42)
Observations	946	946	928	928	934	934	934	934

Notes: Tobit regressions on individual output are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

Table 2.5: Effect of a Dummy Category on Individual Performance III

Dependent Variable:	Study I		Study II					
Number of Cover Sheets Entered Correctly	ND vs. D		ND vs. D		ND vs. TD		D vs. TD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dummy Category	-3.73 (2.81)	-4.67 (3.59)	-3.16 (3.38)	-6.70 (4.12)			-3.46 (2.98)	-4.45 (3.57)
Dummy Category#Grade 3 in t-1		3.50 (5.32)		14.41** (6.74)				3.79 (6.36)
Transparent Dummy Category					1.34 (3.32)	-0.59 (3.97)		
Transparent Dummy Category#Grade 3 in t-1						7.56 (7.10)		
Grade 3 in t-1		-2.40 (5.31)		-14.26** (6.78)		-12.71* (6.55)		-1.22 (6.13)
Pre-round Performance	0.56*** (0.05)	0.56*** (0.08)	0.94*** (0.06)	0.87*** (0.08)	0.95*** (0.05)	0.86*** (0.08)	0.92*** (0.05)	0.93*** (0.07)
Constant	-9.48*** (1.63)	-6.88 (7.35)	75.21 (46.69)	84.42* (46.34)	87.54** (36.72)	98.11*** (36.77)	26.78** (11.78)	26.63** (12.81)
Observations	946	946	928	928	934	934	934	934
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: OLS regressions on individual output are performed. Session dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

2.C.2 Effect of a Dummy Category on Rating Satisfaction and the Perceived Kindness of a Rating Scale

Table 2.6: Tests Comparing Rating Satisfaction Across Treatments

	Rating Satisfaction in Study I			Rating Satisfaction in Study II		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
D vs. ND	.607	.332	.296	.752	.574	.862
D vs. TD	-	-	-	.511	.821	.315
ND vs. TD	-	-	-	.734	.756	.232

Note: We report p-values of two-sided Wilcoxon rank-sum tests.

Table 2.7: Tests Comparing the Evaluation of Kindness of Own Rating Scale Across Treatments

	Scale Kindness in Study I		Scale Kindness in Study II	
	Grade 1 & 2	Grade 3	Grade 1 & 2	Grade 3
D vs. ND	.000	.000	.544	.004
D vs. TD	-	-	.011	.002
ND vs. TD	-	-	.003	.782

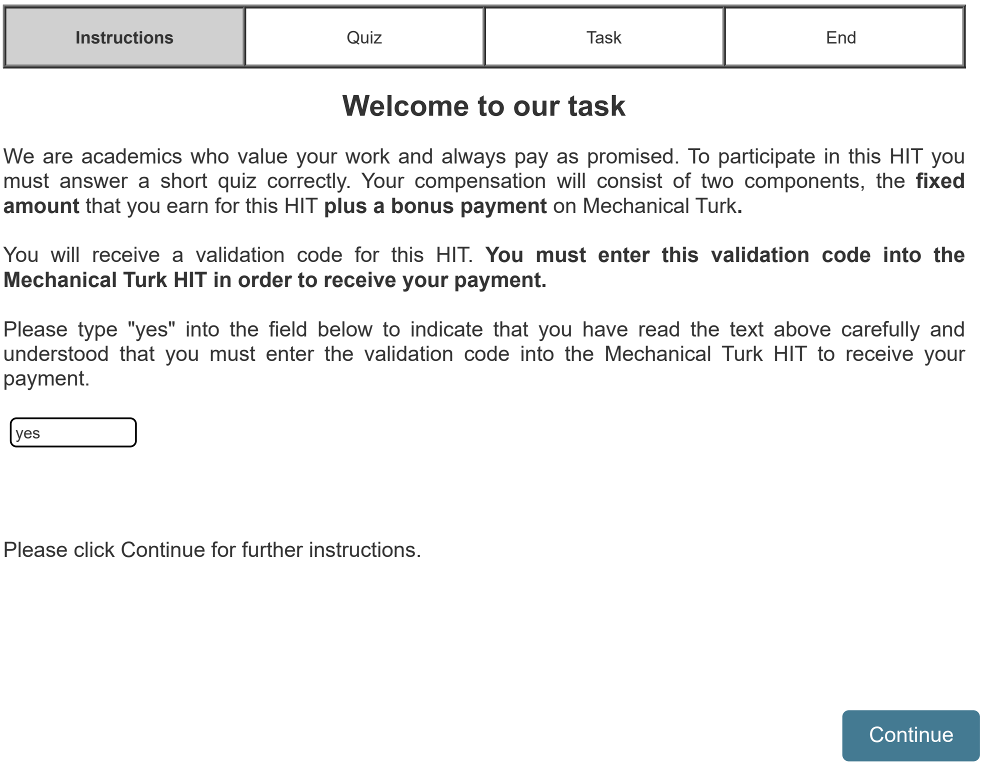
Note: We report p-values of two-sided Wilcoxon rank-sum tests.

2.D Screenshots of Week one of Study I and II

In the following, we present screenshots of the first week of Study I and II. The screens were identical in Study I and II and across treatments. The correct answers to the questions of the quiz are selected on the respective screenshot (Figure 2.9). We show a randomly generated example of the validation and task screen.

2.D.1 Welcome Screen and Instructions

Figure 2.6: Welcome Screen



Welcome to our task

We are academics who value your work and always pay as promised. To participate in this HIT you must answer a short quiz correctly. Your compensation will consist of two components, the **fixed amount** that you earn for this HIT **plus a bonus payment** on Mechanical Turk**.**

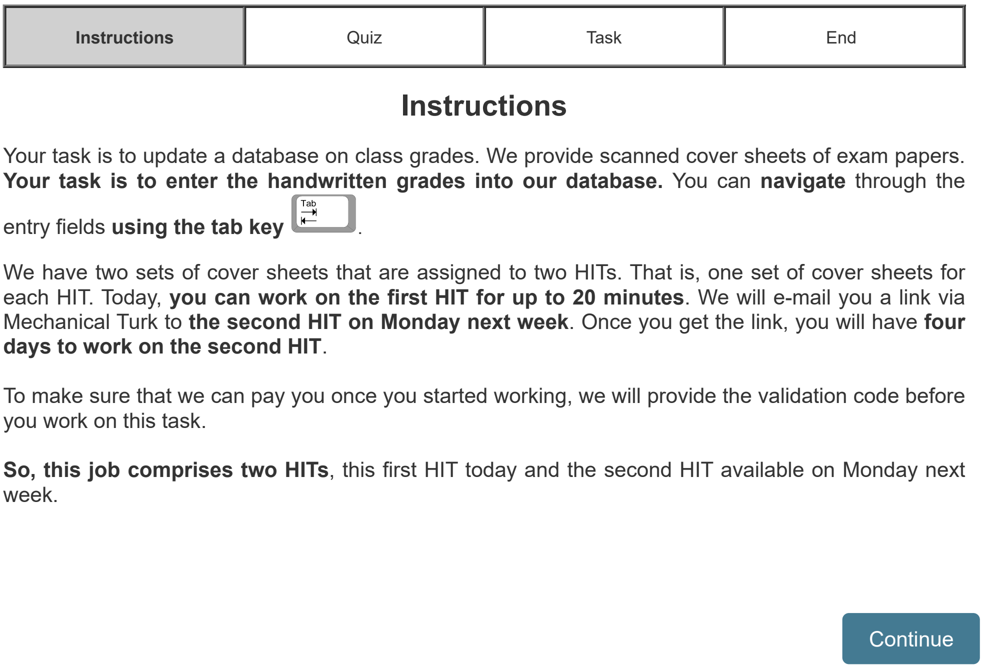
You will receive a validation code for this HIT. **You must enter this validation code into the Mechanical Turk HIT in order to receive your payment.**

Please type "yes" into the field below to indicate that you have read the text above carefully and understood that you must enter the validation code into the Mechanical Turk HIT to receive your payment.

yes

Please click Continue for further instructions.

Figure 2.7: Instructions I



Instructions

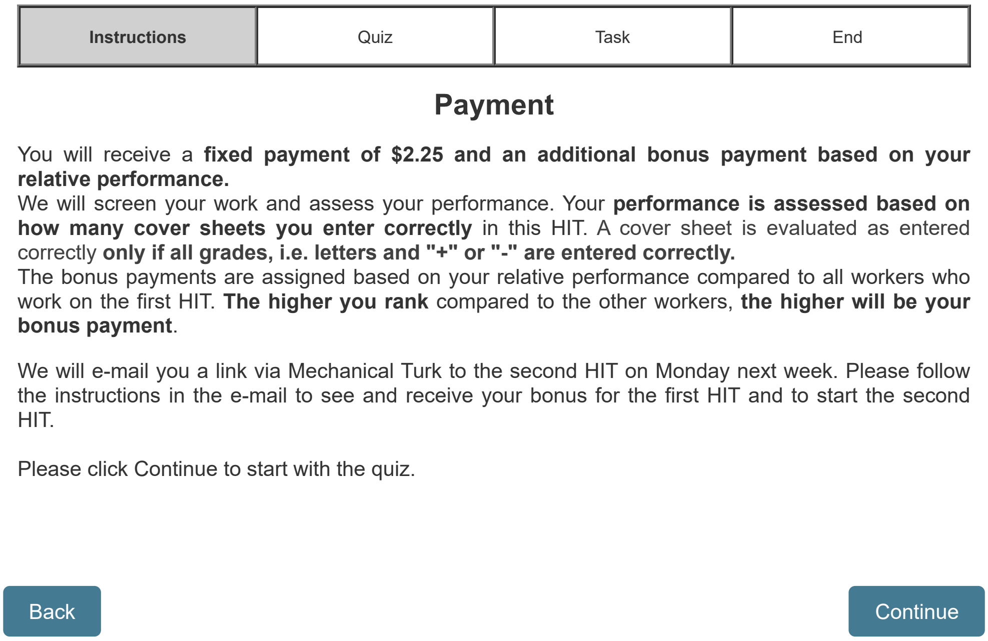
Your task is to update a database on class grades. We provide scanned cover sheets of exam papers. **Your task is to enter the handwritten grades into our database.** You can **navigate** through the entry fields **using the tab key** Tab.

We have two sets of cover sheets that are assigned to two HITs. That is, one set of cover sheets for each HIT. Today, **you can work on the first HIT for up to 20 minutes**. We will e-mail you a link via Mechanical Turk to **the second HIT on Monday next week**. Once you get the link, you will have **four days to work on the second HIT**.

To make sure that we can pay you once you started working, we will provide the validation code before you work on this task.

So, this job comprises two HITs, this first HIT today and the second HIT available on Monday next week.

Figure 2.8: Instructions II



Payment

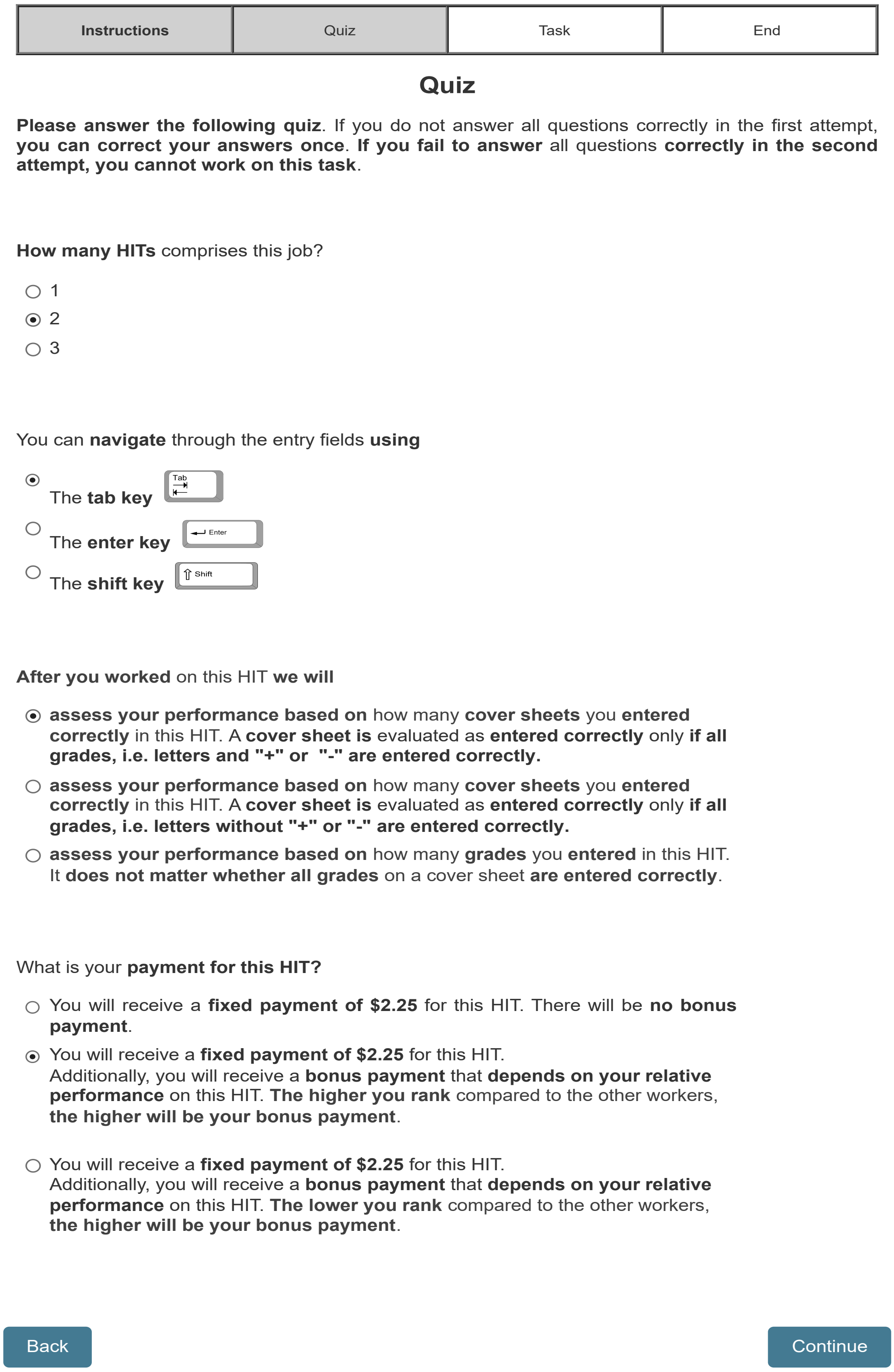
You will receive a **fixed payment of $2.25 and an additional bonus payment based on your relative performance.**
We will screen your work and assess your performance. Your **performance is assessed based on how many cover sheets you enter correctly** in this HIT. A cover sheet is evaluated as entered correctly **only if all grades, i.e. letters and "+" or "-" are entered correctly.**
The bonus payments are assigned based on your relative performance compared to all workers who work on the first HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.

We will e-mail you a link via Mechanical Turk to the second HIT on Monday next week. Please follow the instructions in the e-mail to see and receive your bonus for the first HIT and to start the second HIT.

Please click Continue to start with the quiz.

2.D.2 Quiz

Figure 2.9: Quiz Questions



Quiz

Please answer the following quiz. If you do not answer all questions correctly in the first attempt, **you can correct your answers once**. **If you fail to answer** all questions **correctly in the second attempt, you cannot work on this task**.

How many HITs comprises this job?

○ 1
● 2
○ 3

You can **navigate** through the entry fields **using**

● The **tab key**
○ The **enter key**
○ The **shift key**

After you worked on this HIT **we will**

● **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters and "+" or "-" are entered correctly.**

○ **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters without "+" or "-" are entered correctly.**

○ **assess your performance based on** how many **grades** you **entered** in this HIT. It **does not matter whether all grades** on a cover sheet **are entered correctly**.

What is your **payment for this HIT?**

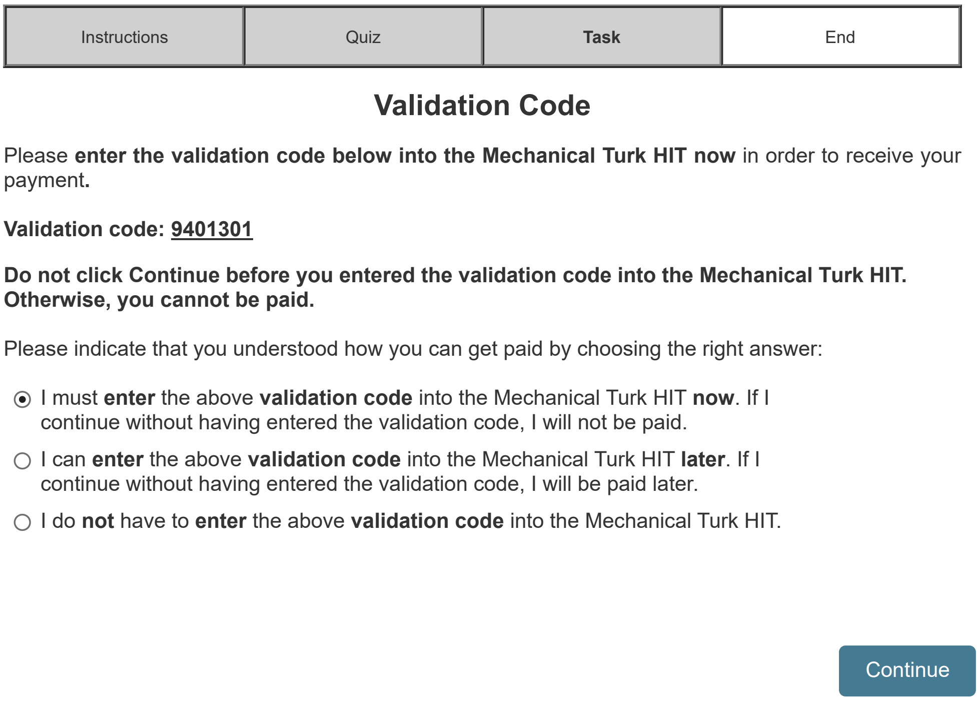
○ You will receive a **fixed payment of $2.25** for this HIT. There will be **no bonus payment**.

● You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.

○ You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The lower you rank** compared to the other workers, **the higher will be your bonus payment**.

2.D.3 Main Part

Figure 2.10: Validation Code I



Validation Code

Please **enter the validation code below into the Mechanical Turk HIT now** in order to receive your payment**.**

Validation code: 9401301

Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

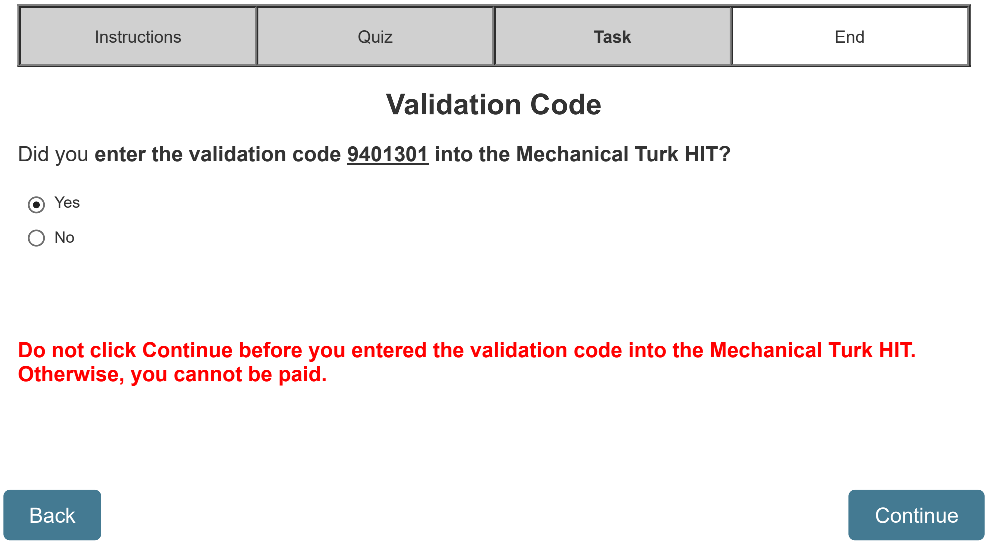
Please indicate that you understood how you can get paid by choosing the right answer:

● I must **enter** the above **validation code** into the Mechanical Turk HIT **now**. If I continue without having entered the validation code, I will not be paid.

○ I can **enter** the above **validation code** into the Mechanical Turk HIT **later**. If I continue without having entered the validation code, I will be paid later.

○ I do **not** have to **enter** the above **validation code** into the Mechanical Turk HIT.

Figure 2.11: Validation Code II



Validation Code

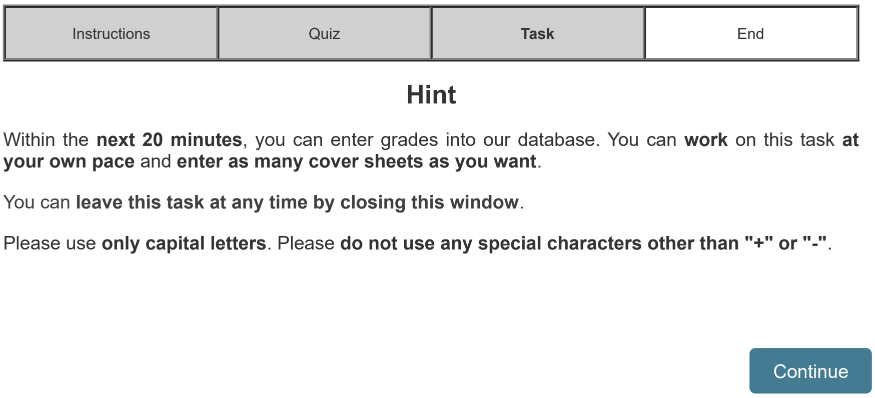
Did you **enter the validation code 9401301 into the Mechanical Turk HIT?**

● Yes

○ No

Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

Figure 2.12: Hint



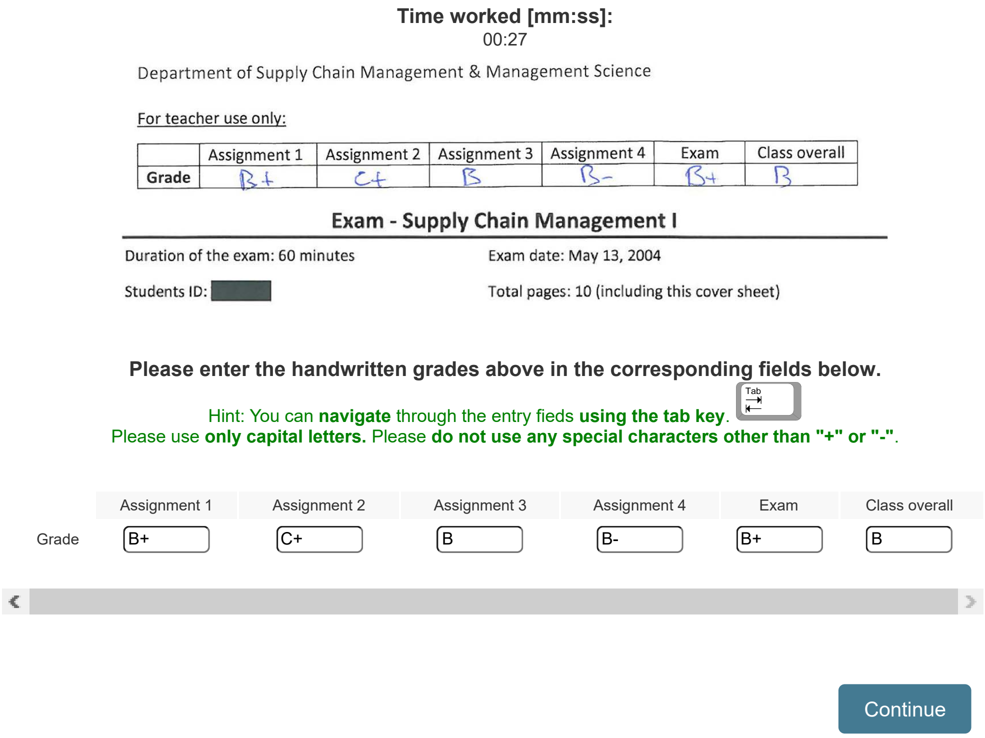
Hint

Within the **next 20 minutes**, you can enter grades into our database. You can **work** on this task **at your own pace** and **enter as many cover sheets as you want**.

You can **leave this task at any time by closing this window**.

Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

Figure 2.13: Task Screen



Time worked [mm:ss]:
00:27

Department of Supply Chain Management & Management Science

For teacher use only:

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B+	C+	B	B-	B+	B

Exam - Supply Chain Management I

Duration of the exam: 60 minutes

Exam date: May 13, 2004

Students ID: [redacted]

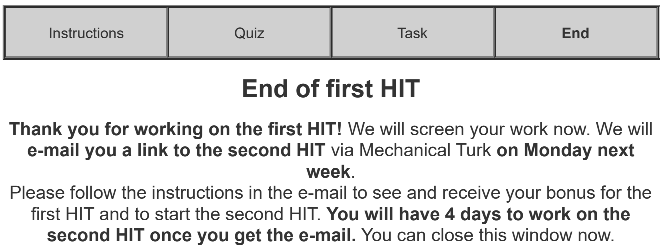
Total pages: 10 (including this cover sheet)

Please enter the handwritten grades above in the corresponding fields below.

Hint: You can **navigate** through the entry fieds **using the tab key**.
Please use **only capital letters.** Please **do not use any special characters other than "+" or "-"**.

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B+	C+	B	B-	B+	B

Figure 2.14: Final Screen



End of first HIT

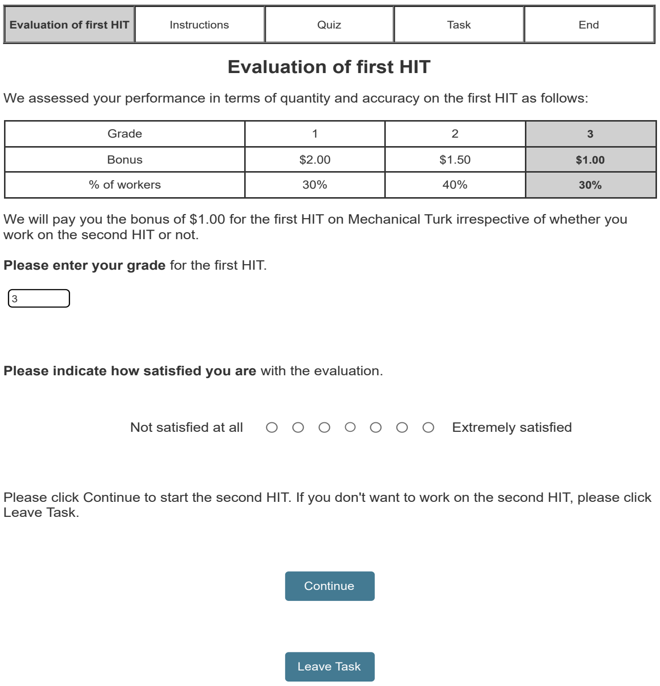
Thank you for working on the first HIT! We will screen your work now. We will **e-mail you a link to the second HIT** via Mechanical Turk **on Monday next week**.
Please follow the instructions in the e-mail to see and receive your bonus for the first HIT and to start the second HIT. **You will have 4 days to work on the second HIT once you get the e-mail.** You can close this window now.

2.E Screenshots of Week two of Study I

In the following, we present screenshots of the second week of Study I. Exemplary, we show the feedback screen of subjects who received grade 3 for their work in week one (Figures 2.15 and 2.16). If subjects received grade 1 or 2, the respective other cells of the table were greyed. In treatment No Dummy, a 3-point scale was displayed on the feedback screen (see Figure 2.15). In treatment Dummy, a 4-point scale was displayed on the feedback screen (see Figure 2.16). This was the only difference between treatments. After seeing their evaluation, subjects could leave the experiment by clicking the button "Leave Task" on the feedback screen. The correct answer to the quiz question is selected on the respective screenshot (Figure 2.18). After the quiz followed the identical validation code screens as in week one of Study I and II (see Figures 2.10 and 2.11). The task screen was also identical to the one of week one of Study I and II (see Figure 2.13). After working for 20 minutes, a message was displayed on top of the task screen as shown in Figure 2.20. We show a randomly generated example of the evaluation, validation, and task screen.

2.E.1 Evaluation of Performance in Week one

Figure 2.15: Evaluation of Performance in Week one in No Dummy Treatment



Evaluation of first HIT

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	**3**
Bonus	$2.00	$1.50	**$1.00**
% of workers	30%	40%	**30%**

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

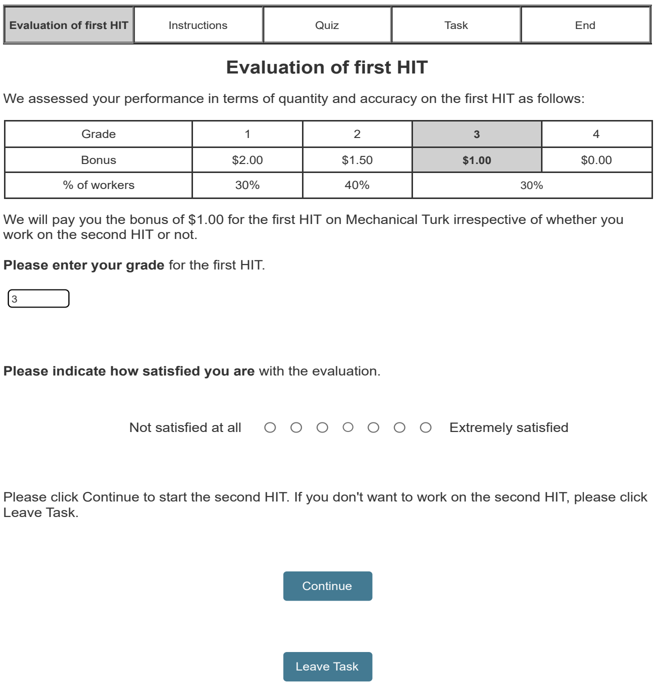
3

Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ○ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Figure 2.16: Evaluation of Performance in Week one in Dummy Treatment



Evaluation of first HIT

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	**3**	4
Bonus	$2.00	$1.50	**$1.00**	$0.00
% of workers	30%	40%	30%	

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

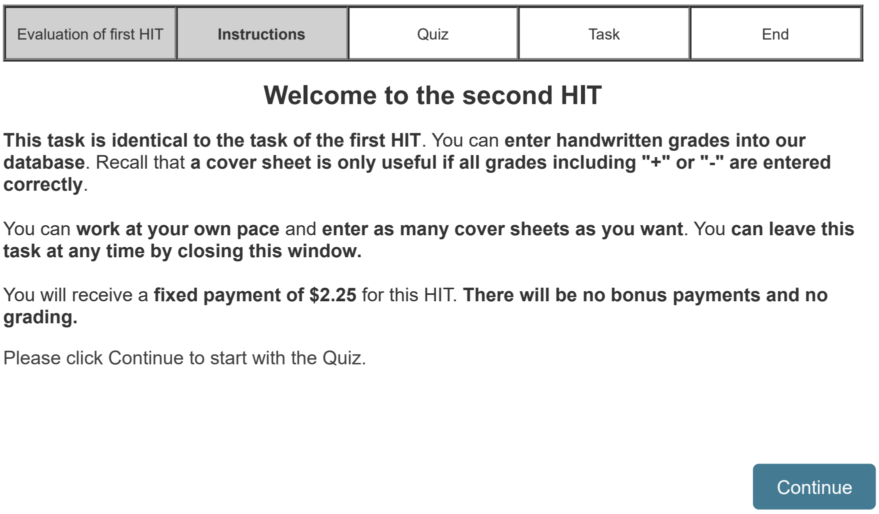
Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ○ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

2.E.2 Welcome Screen and Instructions

Figure 2.17: Welcome Screen and Instructions



Welcome to the second HIT

This task is identical to the task of the first HIT. You can **enter handwritten grades into our database**. Recall that **a cover sheet is only useful if all grades including "+" or "-" are entered correctly**.

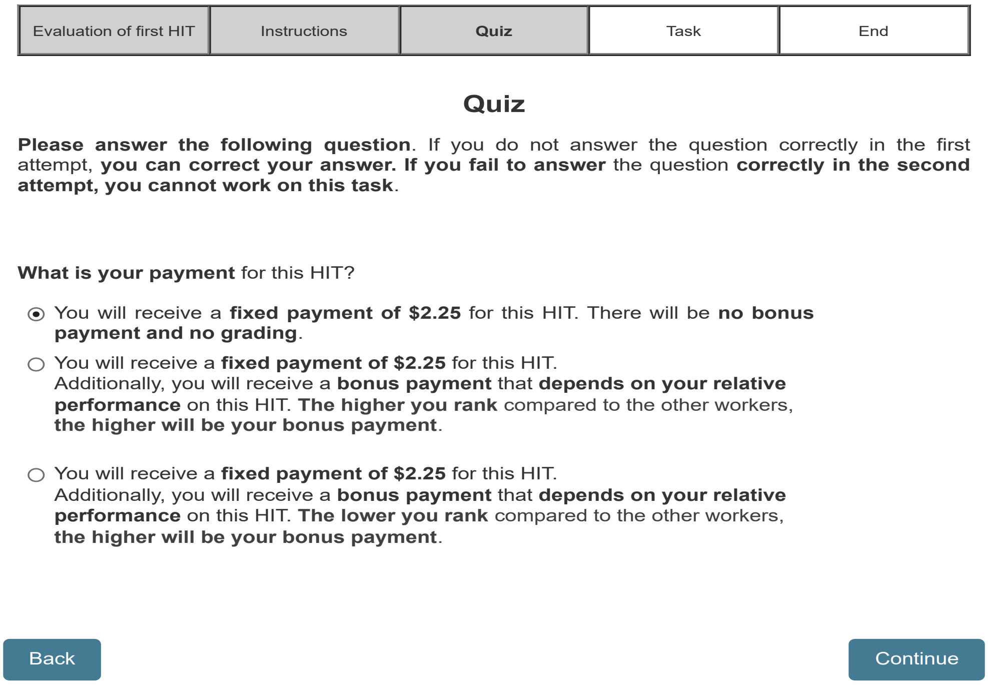
You can **work at your own pace** and **enter as many cover sheets as you want**. You **can leave this task at any time by closing this window.**

You will receive a **fixed payment of $2.25** for this HIT. **There will be no bonus payments and no grading.**

Please click Continue to start with the Quiz.

2.E.3 Quiz

Figure 2.18: Quiz Question



Quiz

Please answer the following question. If you do not answer the question correctly in the first attempt, **you can correct your answer. If you fail to answer** the question **correctly in the second attempt, you cannot work on this task**.

What is your payment for this HIT?

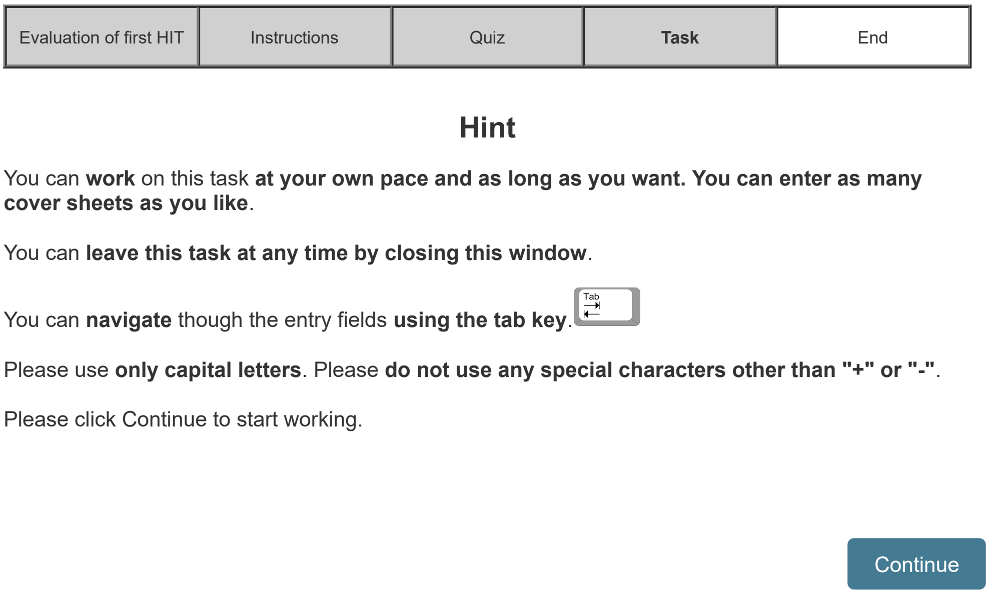
● You will receive a **fixed payment of $2.25** for this HIT. There will be **no bonus payment and no grading**.

○ You will receive a **fixed payment of $2.25** for this HIT.
Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.

○ You will receive a **fixed payment of $2.25** for this HIT.
Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The lower you rank** compared to the other workers, **the higher will be your bonus payment**.

2.E.4 Main Part

Figure 2.19: Hint



Hint

You can **work** on this task **at your own pace and as long as you want. You can enter as many cover sheets as you like**.

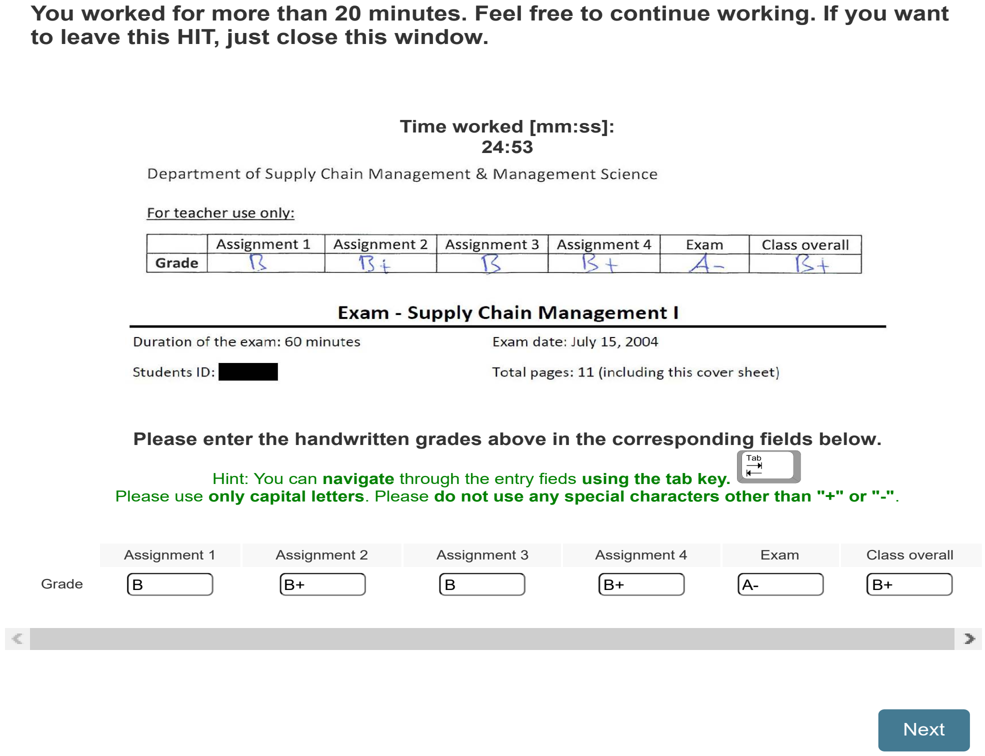
You can **leave this task at any time by closing this window**.

You can **navigate** though the entry fields **using the tab key**.

Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

Please click Continue to start working.

Figure 2.20: Final Screen: Task Screen After 20 Minutes



You worked for more than 20 minutes. Feel free to continue working. If you want to leave this HIT, just close this window.

Time worked [mm:ss]:
24:53

Department of Supply Chain Management & Management Science

For teacher use only:

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B	B+	B	B+	A-	B+

Exam - Supply Chain Management I

Duration of the exam: 60 minutes

Exam date: July 15, 2004

Students ID: [redacted]

Total pages: 11 (including this cover sheet)

Please enter the handwritten grades above in the corresponding fields below.

Hint: You can **navigate** through the entry fieds **using the tab key.**
Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

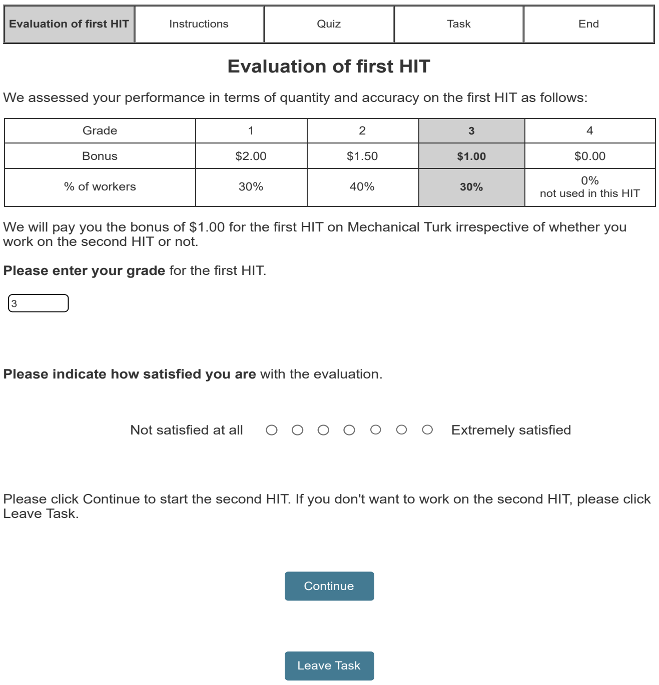
	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B	B+	B	B+	A-	B+

2.F Screenshots of Week two of Study II

In the following, we present screenshots of the second week of Study II. Exemplary, we show the feedback screen of subjects who received grade 3 for their work in week one. If subjects received grade 1 or 2, the respective other cells of the table were greyed. In treatment No Dummy, a 3-point scale was displayed on the feedback screen as shown in the previous subsection in Figure 2.15. In treatment Dummy, a 4-point scale was displayed on the feedback screen as shown in the previous subsection in Figure 2.16. In treatment Transparent Dummy, we displayed a 4-point scale on the feedback screen as shown in Figure 2.21. After seeing their evaluation, subjects could leave the experiment by clicking the button "Leave Task" on the feedback screen. We show the instructions of treatment Transparent Dummy in Figure 2.22. In the instructions of treatments No Dummy and Dummy, we displayed the rating scales of treatment No Dummy (Figure 2.15) and treatment Dummy (Figure 2.16), respectively. The text of the instructions and the quiz questions were identical across treatments. The correct answers to the questions of the quiz in treatment Transparent Dummy are selected on the respective screenshot (Figure 2.23). The correct answer to the second question in the quiz of treatments Dummy and No Dummy was "The top 30% receive Grade 1, the next 40% receive Grade 2, the worst 30% receive Grade 3.". The screens displaying the validation code, hint, task, and final screen were identical to the respective screens of Study I.

2.F.1 Evaluation of Performance in Week one

Figure 2.21: Evaluation of Performance in Week one in Transparent Dummy Treatment



Evaluation of first HIT

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	**3**	4
Bonus	$2.00	$1.50	**$1.00**	$0.00
% of workers	30%	40%	**30%**	0% not used in this HIT

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

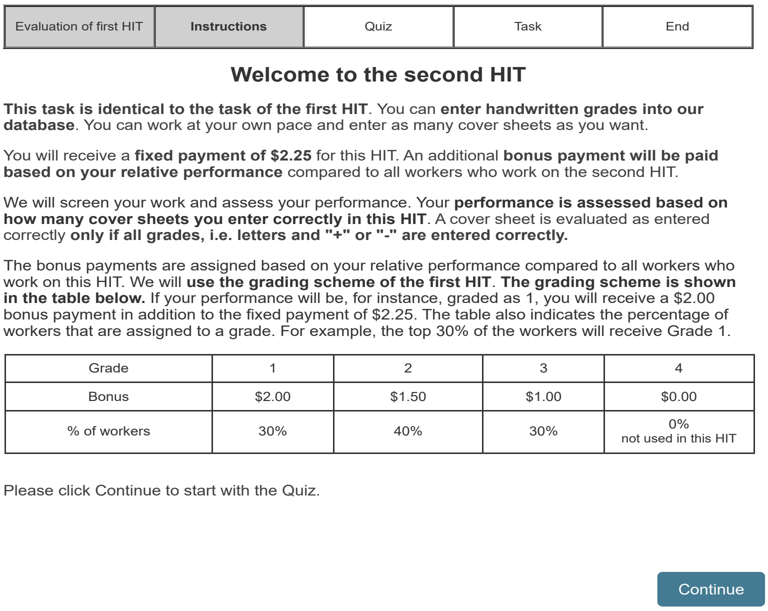
Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ○ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

2.F.2 Welcome Screen and Instructions

Figure 2.22: Welcome Screen and Instructions in Transparent Dummy Treatment



Welcome to the second HIT

This task is identical to the task of the first HIT. You can **enter handwritten grades into our database**. You can work at your own pace and enter as many cover sheets as you want.

You will receive a **fixed payment of $2.25** for this HIT. An additional **bonus payment will be paid based on your relative performance** compared to all workers who work on the second HIT.

We will screen your work and assess your performance. Your **performance is assessed based on how many cover sheets you enter correctly in this HIT**. A cover sheet is evaluated as entered correctly **only if all grades, i.e. letters and "+" or "-" are entered correctly.**

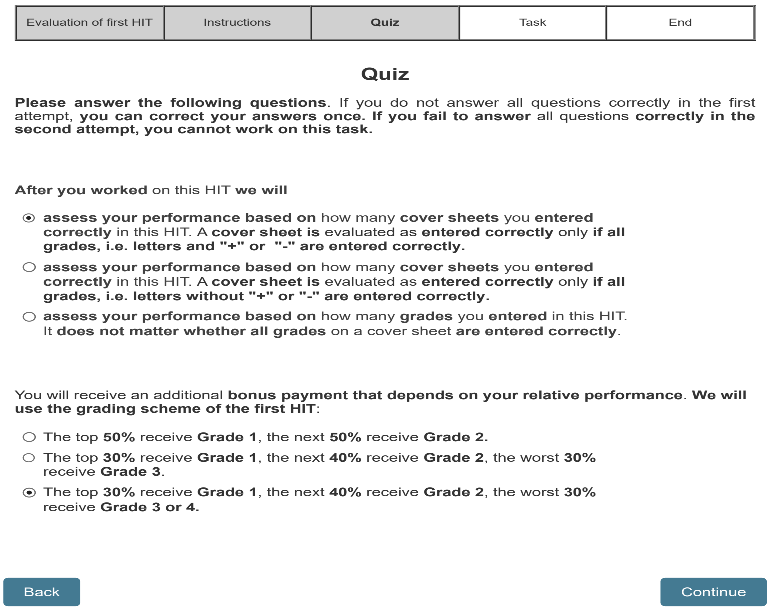
The bonus payments are assigned based on your relative performance compared to all workers who work on this HIT. We will **use the grading scheme of the first HIT**. **The grading scheme is shown in the table below.** If your performance will be, for instance, graded as 1, you will receive a $2.00 bonus payment in addition to the fixed payment of $2.25. The table also indicates the percentage of workers that are assigned to a grade. For example, the top 30% of the workers will receive Grade 1.

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	0% not used in this HIT

Please click Continue to start with the Quiz.

2.F.3 Quiz

Figure 2.23: Quiz Questions



Quiz

Please answer the following questions. If you do not answer all questions correctly in the first attempt, **you can correct your answers once. If you fail to answer** all questions **correctly in the second attempt, you cannot work on this task.**

After you worked on this HIT **we will**

● **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters and "+" or "-" are entered correctly.**

○ **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters without "+" or "-" are entered correctly.**

○ **assess your performance based on** how many **grades** you **entered** in this HIT. It **does not matter whether all grades** on a cover sheet **are entered correctly**.

You will receive an additional **bonus payment that depends on your relative performance**. **We will use the grading scheme of the first HIT**:

○ The top **50%** receive **Grade 1**, the next **50%** receive **Grade 2.**

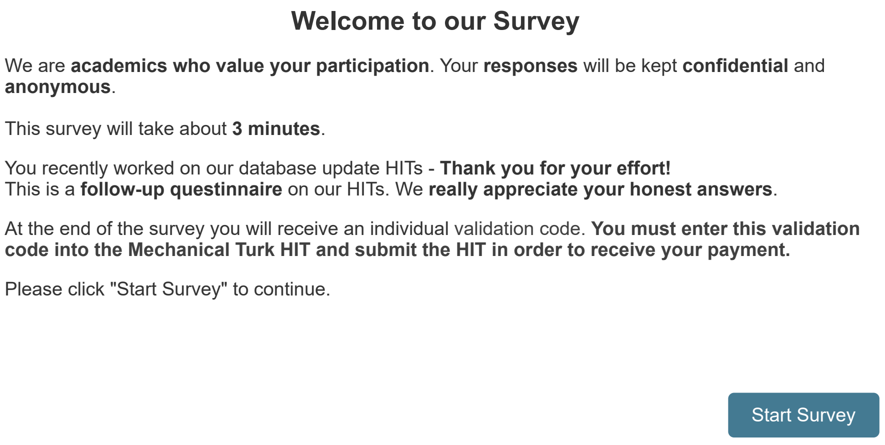
○ The top **30%** receive **Grade 1**, the next **40%** receive **Grade 2**, the worst **30%** receive **Grade 3**.

● The top **30%** receive **Grade 1**, the next **40%** receive **Grade 2**, the worst **30%** receive **Grade 3 or 4.**

2.G Screenshots of the Questionnaire of Study I and II

In the following, we present screenshots of the questionnaire. Exemplary, we show the rating scale as shown in treatment Transparent Dummy in Figure 2.25. In treatment No Dummy, a 3-point scale was displayed on the evaluation screen as shown in Figure 2.15. In treatment Dummy, a 4-point scale was displayed on the evaluation screen as shown in Figure 2.16.

Figure 2.24: Welcome Screen



Welcome to our Survey

We are **academics who value your participation**. Your **responses** will be kept **confidential** and **anonymous**.

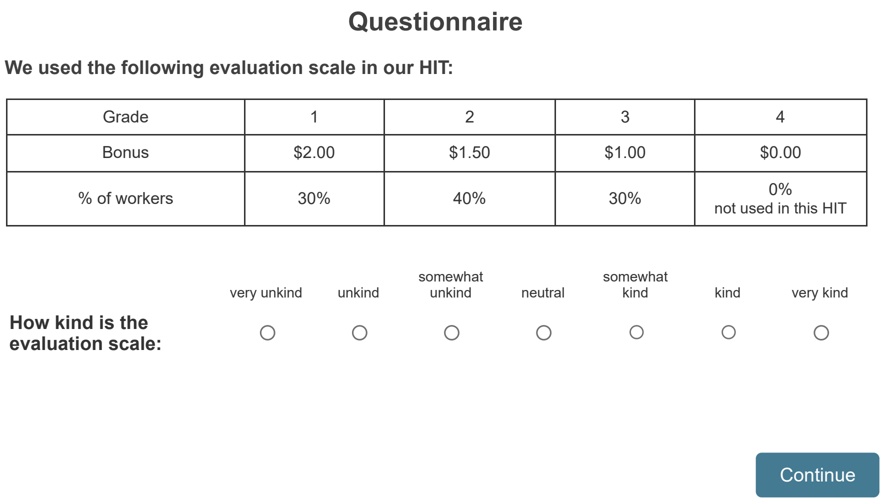
This survey will take about **3 minutes**.

You recently worked on our database update HITs - **Thank you for your effort!**
This is a **follow-up questinnaire** on our HITs. We **really appreciate your honest answers**.

At the end of the survey you will receive an individual validation code. **You must enter this validation code into the Mechanical Turk HIT and submit the HIT in order to receive your payment.**

Please click "Start Survey" to continue.

Figure 2.25: Evaluation of the Kindness of the Rating Scale in Transparent Dummy Treatment



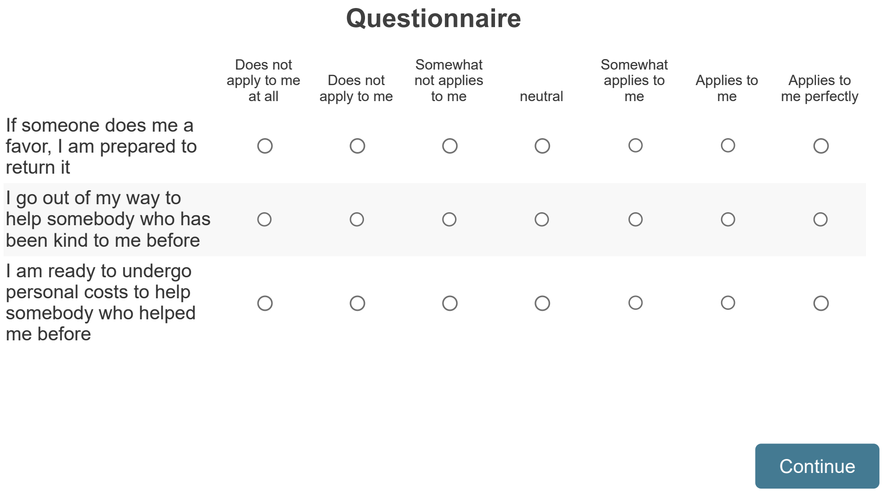
Questionnaire

We used the following evaluation scale in our HIT:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	0% not used in this HIT

	very unkind	unkind	somewhat unkind	neutral	somewhat kind	kind	very kind
How kind is the evaluation scale:	○	○	○	○	○	○	○

Figure 2.26: Questions on Positive Reciprocity

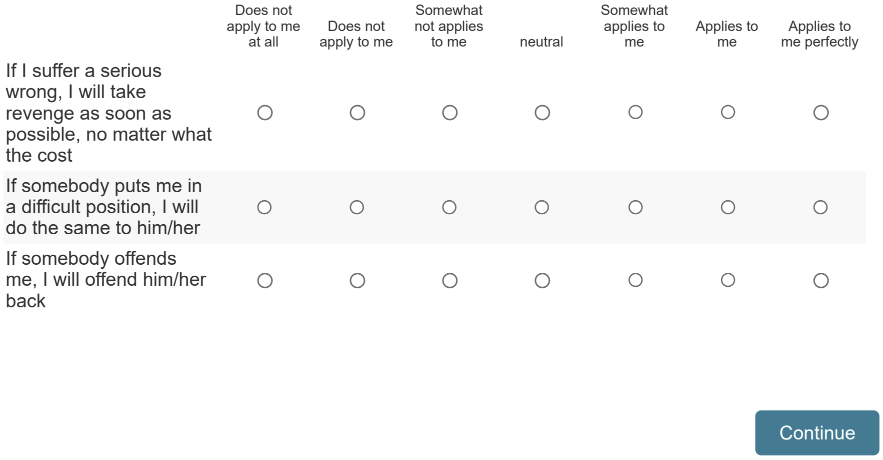


Questionnaire

	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If someone does me a favor, I am prepared to return it	○	○	○	○	○	○	○
I go out of my way to help somebody who has been kind to me before	○	○	○	○	○	○	○
I am ready to undergo personal costs to help somebody who helped me before	○	○	○	○	○	○	○

Continue

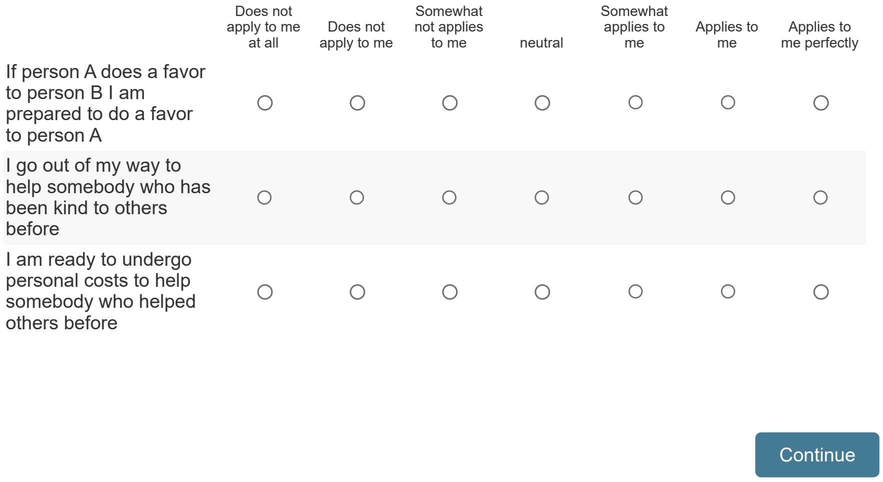
Figure 2.27: Questions on Negative Reciprocity



	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost	○	○	○	○	○	○	○
If somebody puts me in a difficult position, I will do the same to him/her	○	○	○	○	○	○	○
If somebody offends me, I will offend him/her back	○	○	○	○	○	○	○

Continue

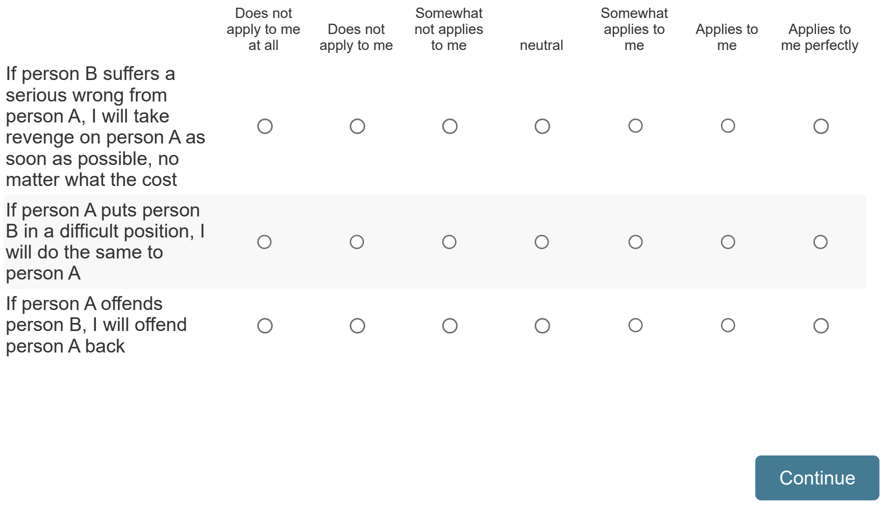
Figure 2.28: Questions on Positive Indirect Reciprocity



	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If person A does a favor to person B I am prepared to do a favor to person A	○	○	○	○	○	○	○
I go out of my way to help somebody who has been kind to others before	○	○	○	○	○	○	○
I am ready to undergo personal costs to help somebody who helped others before	○	○	○	○	○	○	○

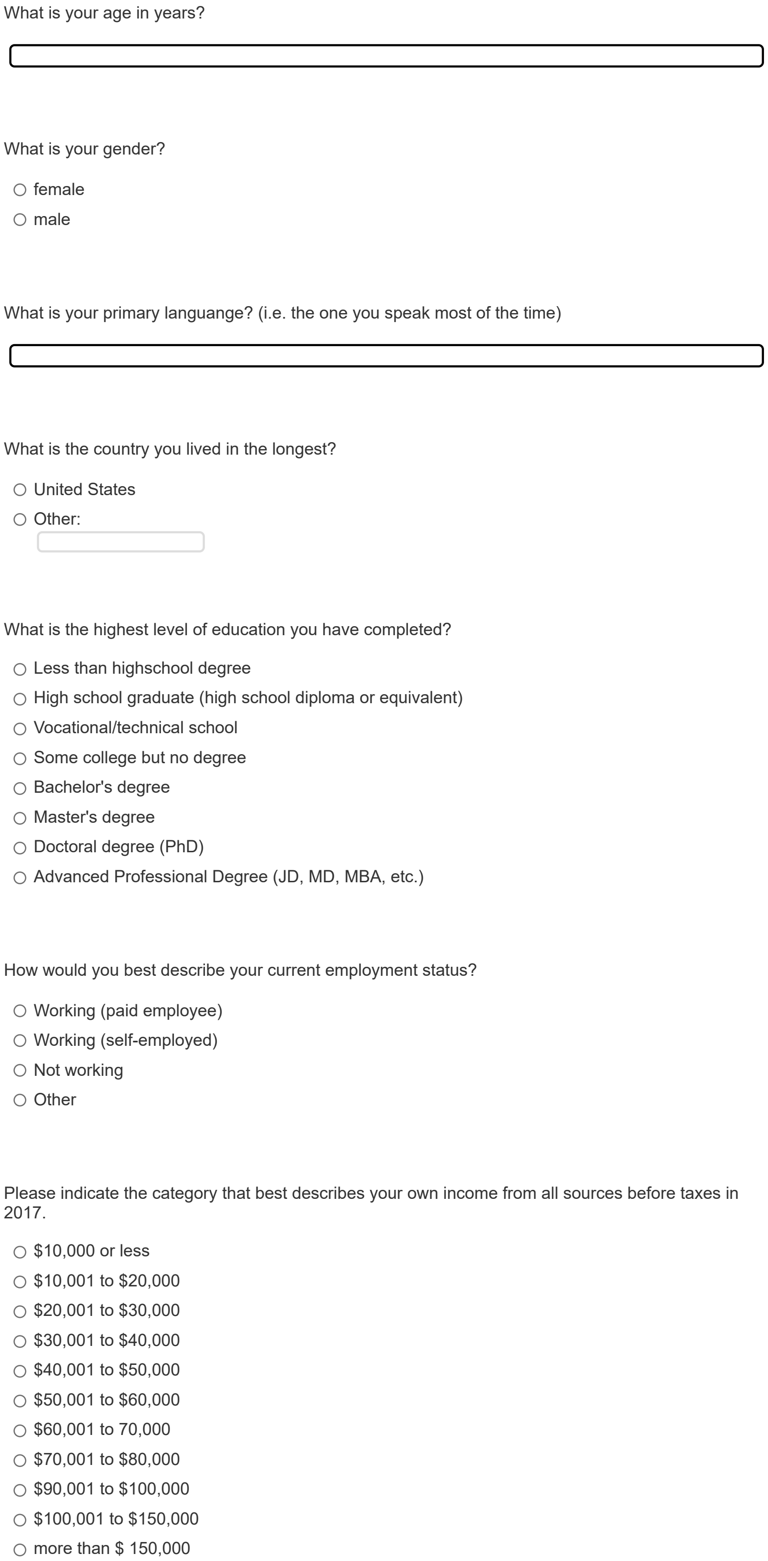
Continue

Figure 2.29: Questions on Negative Indirect Reciprocity



	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If person B suffers a serious wrong from person A, I will take revenge on person A as soon as possible, no matter what the cost	○	○	○	○	○	○	○
If person A puts person B in a difficult position, I will do the same to person A	○	○	○	○	○	○	○
If person A offends person B, I will offend person A back	○	○	○	○	○	○	○

Figure 2.30: Demographic Questions



What is your age in years?

What is your gender?

○ female
○ male

What is your primary languange? (i.e. the one you speak most of the time)

What is the country you lived in the longest?

○ United States
○ Other:

What is the highest level of education you have completed?

○ Less than highschool degree
○ High school graduate (high school diploma or equivalent)
○ Vocational/technical school
○ Some college but no degree
○ Bachelor's degree
○ Master's degree
○ Doctoral degree (PhD)
○ Advanced Professional Degree (JD, MD, MBA, etc.)

How would you best describe your current employment status?

○ Working (paid employee)
○ Working (self-employed)
○ Not working
○ Other

Please indicate the category that best describes your own income from all sources before taxes in 2017.

○ $10,000 or less
○ $10,001 to $20,000
○ $20,001 to $30,000
○ $30,001 to $40,000
○ $40,001 to $50,000
○ $50,001 to $60,000
○ $60,001 to 70,000
○ $70,001 to $80,000
○ $90,001 to $100,000
○ $100,001 to $150,000
○ more than $ 150,000

Chapter 3

On Rating Scales in Performance Appraisals: Performance Effect of a Dummy Rating Category in Long-Term Employer-Employee Relations

Firms use rating scales to assess employee performance. The lowest rating categories are often rarely used in practice, which triggers the question whether it has value to include them at all in the scales. We investigated in a real effort laboratory experiment how an unused low rating category in performance appraisals affects performance. In the experiment subjects worked over six periods. Their performance was rated on a three point rating scale and the rating determined a bonus payment. In the baseline treatment, subjects saw the actual three point rating scale. In another treatment, subjects saw an additional fourth low rating category that was never used but they were not informed about the non-usage. We find that adding the unused low category increased performance by about 20%. This effect vanished in a treatment where subjects were aware that the lowest rating category was never used.

3.1 Introduction

Companies use performance appraisals to incentivize employees. The Management and Organizational Practices Survey of the U.S. Census Bureau, for instance, reports that 46.47% of non-managers and 59.74% of managers receive performance based bonus payments (U.S. Census Bureau 2015). Typically, firms use predefined rating scales (for example from 1-5) to determine bonus payments.

However, low rating categories often remain unused in performance appraisals resulting in compressed ratings. Research on subjective performance evaluations associates unused rating categories with rating biases where supervisors assign too lenient ratings (see Landy and Farr 1980, for a classical contribution in psychology or Prendergast 1999, for a survey from an economics perspective).

Interestingly, the majority of companies do not prevent compressed ratings (Holland 2006) but instead seem to be confident with employing unused rating categories: "[...] Even though most organizations report systems with five levels, generally only three levels are used [...]. It is common for 60 to 70% of an organization's workforce to be rated in the top two performance levels. [...] Skewed performance distributions not only exist, but are common." (Bretz et al. 1992, p.333). Frederiksen et al. (2017), for instance, investigate performance ratings in six large companies and find, with only one exception, that the lowest rating category is assigned to less than 0.2% of employees. Similarly, in the multinational company studied by Ockenfels et al. (2015) the lowest rating category (out of 5) is assigned to 0.1% of employees.

The key question we address in this chapter is whether the inclusion of low rating categories that are in fact never used should be an intentional incentive design choice. We refer to an unused low rating category henceforth as "dummy category".

The literature suggests that a dummy category raises performance as long as employees believe the category may actually be used. According to simple economic reasoning, individuals face higher incentives to work as the additional rating category punishes low performance more. In the language of tournament theory, the use of a low rating category raises the prize spread increasing performance incentives (Lazear and Rosen 1981). Berger et al. (2013) show in laboratory experiments that raters tend to be too lenient and forcing them to assign low performance ratings can raise performance. Behavioral economic research suggests a further potential benefit of adding a low rating category. A substantial body of research has shown that people exhibit preferences for reciprocity (Rabin 1993, Fehr and Rockenbach 2003, Dufwenberg and Kirchsteiger 2004, Falk and Ichino 2006, Falk et al. 2008) and reward kind and punish unkind acts of others. As indeed shown by Sebald and Walzl (2014), Ockenfels et al. (2015), or Bellemare and Sebald (2019), subjective performance evaluations trigger reciprocal reactions by those who have been evaluated towards those who have evaluated their performance. A given rating may appear more generous when a dummy category is included in the scale and trigger stronger positively reciprocal reactions for those with higher ratings or reduce negatively reciprocal reactions among those receiving a lower rating. In other words, the use of a low rating category may shift the reference point in relation to which workers evaluate their ratings.

However, a dummy category may also have negative consequences if employees think that the category is used. The introduction of a low rating category itself may be seen as an additional punishment option and hence signal unkindness and bad intentions of the employer. This may induce negatively reciprocal reactions (Levine 1998, Ellingsen and Johannesson 2007, Bowles and Polanía-Reyes 2012) and reduce performance.

On the other hand, if workers know that the additional category is unused, such a transparent dummy category may signal kindness and good faith of the employers transmitting that they intentionally do not use an available punishment option. This, in turn, may induce positively reciprocal reactions that raise performance.

Chapter 2 (Vogt 2021) reports results that are in line with the idea that a dummy category may trigger negatively reciprocal and a transparent dummy positively reciprocal reactions. However, in Chapter 2 we only analyze short-term employer-employee relations, while we now analyze behavior over longer time frames, where workers experience multiple ratings and can react dynamically.

The aim of this chapter is to investigate the performance effect of having a dummy category if individuals receive repeated ratings over time. To study this question, we ran a real effort laboratory experiment.

The experiment involved subjects in the role of "employers" and "employees". Employers initially determined the rating scale used to evaluate their employees – and thus determined the treatment. That is, they chose between a rating scale that consisted of either three or four rating categories. Employers benefited from higher work effort of their employees as their payment depended on their employees' performance. Employees then worked on a real effort task over six periods. After each period, employees received performance ratings that determined bonus payments.

Employees only learned the scale and not the specific procedure by which the ratings were assigned within the scale. Importantly, ratings were then conducted by the computer and subjects received a rating based on the same predetermined absolute performance thresholds irrespective of the treatment. Therefore, only three rating categories were used irrespective of the chosen scale. This allows a clean ceteris paribus comparison as any treatment differences must be driven by the employees' perceptions induced by the choice of the rating scale.

We consider three treatments. Treatment No Dummy (ND) serves as a baseline, where subjects saw the actual three-point rating scale. In treatment Dummy (D), subjects saw an additional fourth category that was never used. Subjects were not informed that the additional category was never used. In treatment Transparent Dummy (TD), subjects also saw four rating categories but were informed that the fourth rating category was never used. Note that the communicated number of rating categories in use and the bonus spread were equivalent in treatments ND and TD.

Our main result is that the use of an unused low rating category raised performance significantly: Total performance was 20.92% higher in treatment D compared to treatment ND. This performance effect evolved over time: While performance did not differ in the first period, over time subjects worked increasingly more in treatment D than in treatment ND. However, these performance differences were not present in the treatment TD where subjects knew that the low rating was in fact never used.

We do not find evidence that the performance increase was driven by reciprocal subjects. Therefore, we conclude that it could not be reciprocal reactions to more generous ratings that increased performance in the presence of a dummy category. We thus argue that it was the threat of a potential low rating that caused these performance gains – which as we show were driven by the low performers.

We also elicited the employees' perceptions about the "kindness" of the rating scales in two different ways. First, we asked subjects after the last working period to evaluate the scale that their employer implemented (without knowledge of the other scales). Second, we presented subjects the other scales and asked them to evaluate them. When knowing all rating scales, subjects evaluated the rating scale in treatment D as being less kind than the scales in treatment ND and TD. But this was not the case when they only knew their own rating scale.

Beyond the literature on subjective performance evaluations discussed in the above, our results contribute to the experimental literature on the effect of feedback on performance (for a recent overview see Villeval 2020, for further experimental evidence see, for instance, Azmat and Iriberri 2010, Barankay 2012, Tran and Zeckhauser 2012, Gill et al. 2019, or Hoffmann and Thommes 2020). As our results show, not only the specific feedback matters that an employee receives for her or his performance but also the choice of the scale on which this feedback is given.

he remainder of this chapter is structured as follows. In Section 3.2, we introduce our laboratory study and state our hypotheses. In Section 3.3, we present our results. In Section 3.4, we conclude and discuss the managerial implications of our findings.

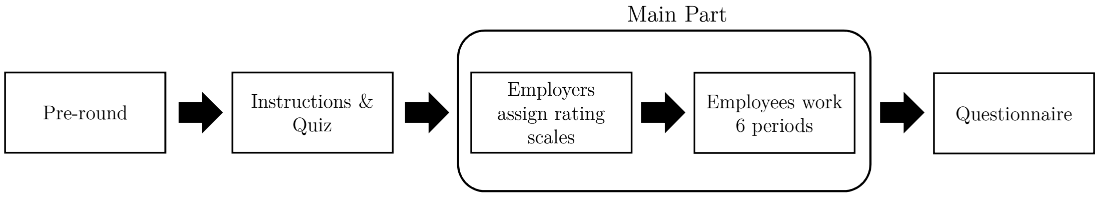
3.2 Experimental Design and Hypotheses Development

3.2.1 Experimental Design

Overview We conducted a laboratory experiment that involved the role of "employees" and "employers". In all treatments, subjects in the role of employees were randomly matched into groups with one subject in the role of employer. Employees worked on a real effort task and received a performance dependent bonus. Employers benefited from high work effort of employees as their payment depended on their employees' performance. In the first stage, employers determined whether three or four categories were shown in the rating scale presented to their employees. In the second stage, employees worked on a real effort task over six periods receiving private performance ratings given by the computer after each period. Ratings were based on the same predetermined absolute performance thresholds irrespective of the treatment. However, dependent on employer's scale choice, employees saw a fourth, unused low rating category in their rating scale. Employees only learned their own rating scale but did not see other rating scales that the employers could have chosen.[1] A questionnaire section followed after

[1] Most commonly, employees in firms learn the details of the performance appraisal scheme and hence their own rating scale only after they started their job and are not made aware of the potential alternatives considered by the employer.

Figure 3.1: Experimental Procedure



the main part. Figure 3.1 shows the experimental procedure. The Appendix 3.C contains the experimental instructions and screenshots.

Experimental Details We assessed individuals' task ability in a pre-round where all subjects worked 2.5 minutes on the real-effort task that was also used in the main part. Subjects had to count the number "7" in blocks of randomly generated numbers as used in Berger et al. (2013). Performance was measured by the number of "points" earned. Subjects received +2 points for each correct answer and -0.5 points for each wrong answer. Subjects received 10 (euro) cents per point in the pre-round. Subjects could take a "time-out" by pushing a button that locked their screen for 20 seconds during which subjects could not work. For each time-out taken, subjects received 8 (euro) cents, representing potential opportunity costs of not continuing working.

Based on the pre-round results, best performing subjects became employers reflecting that employers are usually more productive than their employees in firms. The remaining subjects became employees and were randomly assigned to an employer. Using stratified sampling, we ensured similar ex-ante performance across employee groups and an evenly distributed number of employees across employers. Matching was anonymous and participants did not receive information on the identity of other subjects. Decisions were anonymous, communication was not permitted and also not observed. The assigned roles and the group matching were kept constant during the experiment.

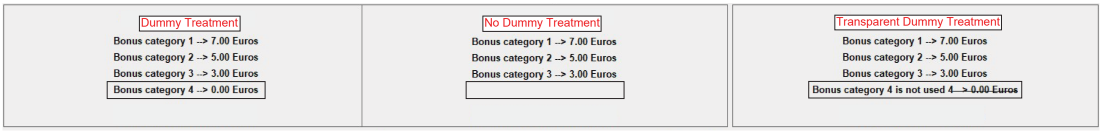
After the pre-round, subjects received feedback on the number of correct and false answers, time-outs taken, points, and money earned. Before the main part, subjects had to pass a comprehension quiz and learned their role (employer or employee).

In the main part employers first determined whether three or four categories were shown in the rating scales to their employees when receiving performance ratings. Subsequently, employees learned the rating scale chosen by their employer. Then employees worked on the real effort task for six periods, each lasting 2.5 minutes. As in the pre-round, subjects could take a time-out receiving 25 (euro) cents for it. After each period, the computer rated employees' performance on a three-point scale using the same absolute performance thresholds across treatments. The performance rating determined employees (potential) payment for the respective period. One period was randomly selected as payment relevant at the end of the experiment. Employees received 7 euros, 5 euros, or 3 euros, when being rated as a "Grade 1, 2, or 3", respectively. An employer received 2.5 (euro) cents per point earned by an individual employee. Employees learned their personal performance rating and resulting payment, their number of time-outs taken and resulting payment as well as the total period payment. Employees also saw the individual contribution to the employer's payment.[2] A questionnaire section followed after the main part.

Treatment Variation In all treatments, the computer followed the same procedure and rated employees' performance on a three-point scale using predetermined, absolute performance thresholds. We based the performance thresholds for assigning bonus categories on the performance distribution of Berger et al. (2013), such that about 30% of the subjects received grade 1, 40% grade 2, and 30% grade 3. Subjects did not learn the specific details of the rating procedure to reflect the situation in firms that employees usually do not know exactly how their performance will be evaluated.

[2] Note that information on employers' surplus is crucial for employees reciprocal reactions as shown in Hennig-Schmidt et al. (2010).

Figure 3.2: Rating Scales Across Treatments



Dummy Treatment
Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 --> 0.00 Euros

No Dummy Treatment
Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros

Transparent Dummy Treatment
Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~

We analyze three treatments: In treatment No Dummy (ND) subjects saw the actual three-point rating scale used by the computer. In treatment Dummy (D), an additional, unused low rating category was displayed and subjects hence saw a four-point scale. The scale shown in treatment Transparent Dummy (TD) was similar to the one in treatment D. However, the non-usage of the added rating category was disclosed. Figure 3.2 shows the rating scales displayed across treatments.

We introduced the three treatments ND, D, and TD with the following mechanism: Three equally sized employee groups were randomly labeled A, B, and C and randomly matched to the corresponding employer. The three groups had similar pre-round performance as we stratified group assignment based on pre-round performance. Employers saw three different rating scales as described in the treatments ND, D and, TD above. Employers had to assign each of the three rating scales to one of the three employee groups. Other information on the employee groups than the group labels A, B, C were not given. We thereby guaranteed that employers randomly assigned all rating scales to employee groups and ensured exogenous treatment variation. To assure balanced treatment assignment, we forced the selection of all rating scales to avoid that employers choose only the most attractive one(s) as observed in a related context of contract choice by Fehr et al. (2007). We thereby assured that all three treatment conditions were run simultaneously within each session. Additionally, we guaranteed stratified sampling to avoid session effects between treatments.

Experimental Protocol and Subject Pool We recruited a total of 468 subjects from the Cologne Laboratory of Economic Research at the University of

Cologne in 16 sessions in October 2019 via the Online Recruitment System for Economic Experiments (Greiner 2004). Sessions were evenly distributed across the day (morning, noon, and after-noon) and followed the same protocol. The experiment was programmed and conducted using z-Tree (Fischbacher 2007). All subjects participated only once. 61% of the participants were female, the median age was 25. 70% of the subjects worked 10 hours a week (median) receiving a median hourly wage of 10 euro. 96% were students with the majority studying at the faculty of Management, Economics and Social Sciences and being enrolled in a Master's program. See Table 3.5 in the Appendix 3.A for further sample demographics. Subject received a median payment of 11 euros, including a participation fee of 4 euros. With a median experiment time of 64 minutes, the median hourly wage was 10.31 euros.

3.2.2 Hypotheses Development

The key hypothesis at the outset was that the dummy category raises performance due to two effects. First, it should generate a standard *Incentive Effect*: When workers believe that low performance may now trigger a lower rating they should have stronger incentives to work harder. To see this, consider for instance a simple moral hazard problem where an agent with utility function $u(w) - c(a)$ – where $u(w)$ is increasing and $c(a)$ is increasing and convex – chooses an effort level a and obtains wage w. Suppose for simplicity that the effort determines the perceived likelihood of obtaining a high rather than a low rating such that $w \in L, H$ and $\Pr(w = H) = a$. Then by applying the implicit function theorem to the first order condition of the agent's optimization problem we have that $\frac{\partial a}{\partial L} = \frac{-u'(L)}{c''(a)} < 0$. Hence, decreasing the lower rating should increase efforts from this perspective.

Moreover, there may be what we term an *Evaluation Effect* as the dummy category shifts the reference standard for the evaluation. A given rating may appear more generous relative to the lowest rating category and this may trigger positively reciprocal reactions. Hence, these two channels suggest the hypothesis that performance in treatment Dummy is highest and exceeds the performance in treatment Transparent Dummy and that in treatment No Dummy.

However, the results in Chapter 2 (Vogt 2021) indicated a countervailing *Kindness-of-the-Scale Effect* as the overall rating scale may be perceived as harsher when an additional low rating category is displayed – potentially triggering negatively reciprocal reactions to the choice of the scale itself. We expected this effect also to be prevalent in our laboratory experiment where in contrast to the previous study, subjects in the role of employees worked for subjects in the role of employers, and employees knew that employers had chosen the respective scale in use.[3]

To the extent that the *Kindness-of-the-Scale Effect* dominates, this would suggest the opposing hypothesis that performance is highest under the most "generous" scale, i.e. in the Transparent Dummy treatment, followed by the No Dummy and then the Dummy treatment. We preregistered our hypotheses accordingly. However, we also preregistered the hypothesis that the *Kindness-of-the-Scale Effect* will fade in importance over time relative to the *Incentive Effect* and *Evaluation Effect* as potential initial reciprocal reactions to the choice of the scale decrease over time as subjects get used to the scale. Gneezy and List (2006) or Sliwka and Werner (2017), for instance, show that wage increases trigger effort increases in the short-term, but these reciprocal reactions then fade over time.

3.3 Results

In Section 3.3.1, we first analyze the effect of a dummy category on total individual performance and study how this effect evolved over time. To explore the underlying performance drivers in more detail, we then investigate whether the performance effect differs between high and low performers and between strongly and weakly reciprocal subjects. In Section 3.3.2, we analyze how subjects perceived the

[3] Please note that we would expect even stronger reciprocal reactions to the scale choice or evaluations if employees knew reference scales or employers faced a costly choice of a rating scale. However, we argue that these structures are not common in firms.

Table 3.1: Effect of a Dummy Category on Total Individual Performance

Dependent Variable:	Total Number of Points			Log of Total Number of Points		
	D vs. ND (1)	D vs. TD (2)	TD vs. ND (3)	D vs. ND (4)	D vs. TD (5)	TD vs. ND (6)
Dummy Category	13.94*** (3.72)	10.48*** (3.68)		0.19*** (0.05)	0.16*** (0.05)	
Transparent Dummy Category			3.39 (4.20)			0.03 (0.06)
Pre-round Number of Points	4.63*** (0.41)	4.46*** (0.34)	4.76*** (0.43)	0.05*** (0.01)	0.05*** (0.01)	0.06*** (0.01)
Constant	30.27*** (7.52)	36.44*** (6.89)	28.12*** (7.87)	3.70*** (0.12)	3.70*** (0.13)	3.56*** (0.13)
Observations	293	290	289	293	290	289

Notes: Ordinary least squares regressions on total individual performance are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* p < 0.10, ** p < 0.05, *** p <0.01, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

kindness of rating scales across treatments.

3.3.1 Performance Effect of a Dummy Category

Total Individual Performance We start our analysis by investigating the effect of a dummy category on total individual performance. We regress total individual performance over six periods – one data point per individual – on a treatment dummy. We control for pre-round performance to capture individual performance differences that can still occur despite our sampling procedure. Our results do not qualitatively change when we exclude the pre-round control, include session dummies, or use non-parametric Wilcoxon rank-sum tests to compare performance (see Appendix 3.B.1).

Table 3.1 reports our main results. In columns (1)-(3) we use absolute points as performance measure, in columns (4)-(6) we show regression results using the log of points as performance measure. We pairwise compare treatment Dummy (D) with No Dummy (ND) in columns (1) and (4); treatment D with Transparent Dummy (TD) in columns (2) and (5), and treatment ND with TD in (3) and (6). Total individual performance was significantly higher in treatment D – where subjects were not aware that the additional rating category was never used – as compared to treatment ND and TD. Absolute performance differences amounted to 13.94 and 10.48 points (columns (1) and (2)), translating into a performance increase of 20.92% and 17.35%, respectively (columns (4) and (5)).[4] In Appendix 3.B.1 we also analyze other output measures: Subjects solved significantly more number blocks and took significantly less time-outs in treatment D than in treatment ND and TD while the number of mistakes made did not differ between treatments. Performance in treatment ND and TD did not differ significantly (columns (3) and (6)).

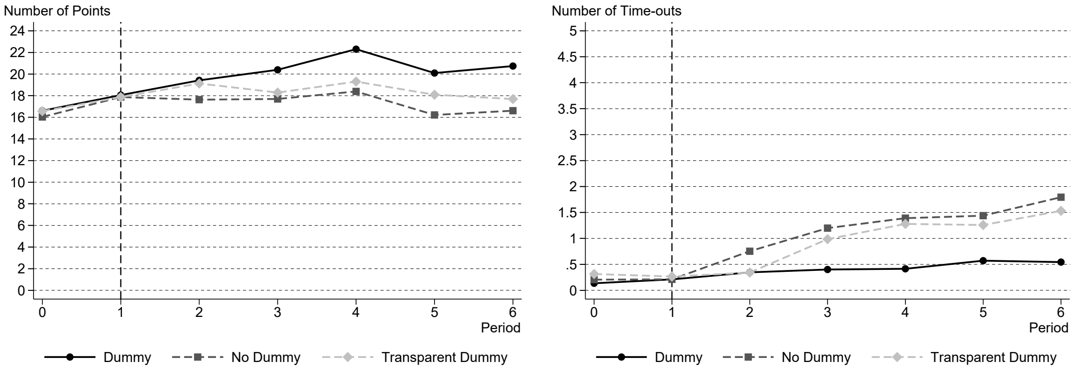
To summarize, a dummy category raised performance significantly. Contrary to our hypothesis, we do not find support for a *Kindness-of-the-Scale Effect* on performance even though the experimental setting emphasizes the scale choice of the employer. Instead, the results indicate a strong *Incentive* and *Evaluation Effect* in the presence of a dummy category.

Individual Performance and Time-outs Taken Over Time We now examine how the observed treatment effect evolved over time. The left graph in Figure 3.3 depicts average individual performance over time across treatments. Average performance was – by design – not different across treatments in the pre-round (period 0). Also in the first period – after the treatment intervention – performance was not significantly different between treatments (see Table 3.2 for a regression analysis below). This mirrors the results from Chapter 2 (Vogt 2021) where we studied a short-term setting and did not see significant differences in total performance across conditions.

Over time, however, average performance in treatment D increased while performance stayed rather flat in the ND and TD group. Note that performance differences evolved faster between D and ND as compared to D and TD where performance was also identical in period 2.

[4] Note that the coefficients in the log specifications of .19 and .16 are equivalent to an increase of $e^{0.19} = 1.2092$ and $e^{0.16} = 1.1735$.

Figure 3.3: Distribution of Individual Performance and Time-outs Over Time Across Treatments



(a) Average Performance Across Treatments **(b)** Average Time-outs Across Treatments

The right graph in Figure 3.3 plots the average number of time-outs taken over time across treatments. The evolution of time-outs taken reversely parallels the evolution of performance differences across treatments. Average number of time-outs taken increased more strongly over time in treatment ND and TD as compared treatment D.

The visual indication is confirmed by the regression analysis shown in Table 3.2. We regress individual period performance (columns (1) and (2)), individual period effort (i.e. the number of blocks worked on in columns (3) and (4)) and individual period time-outs taken (columns (5) and (6)) – six observations per subject – on a treatment dummy estimating pooled OLS regressions. To analyze whether the treatment effect differed between periods, we include interaction terms of the treatment indicator with the working periods 2-6 (period 1 is the baseline). We control for individual differences using pre-round performance, effort, and time-outs taken as well as time trends by including period dummies in all specifications. Our results are robust to controlling for session effects (see Appendix 3.B.2). We pairwise compare the effect in treatment D with ND in columns (1), (3), and (5) and with TD in columns (2), (4), and (6).

The point estimate of the treatment indicator "Dummy Category" is small and insignificant in all specifications confirming that there were no sizeable differences

Table 3.2: Effect of a Dummy Category on Individual Performance, Effort Provision, and Time-outs Taken Over Time I

Dependent Variable:	Number of Points		Number of Blocks		Number of Time-outs	
	D vs. ND (1)	D vs. TD (2)	D vs. ND (3)	D vs. TD (4)	D vs. ND (5)	D vs. TD (6)
Dummy Category	-0.26 (0.51)	0.16 (0.53)	-0.11 (0.21)	0.00 (0.22)	0.05 (0.09)	0.05 (0.09)
Dummy Category#Period 2	1.60** (0.62)	0.10 (0.61)	0.52* (0.29)	-0.03 (0.24)	-0.41*** (0.15)	0.06 (0.13)
Dummy Category#Period 3	2.51*** (0.74)	1.92** (0.74)	1.34*** (0.38)	1.04*** (0.36)	-0.80*** (0.21)	-0.53*** (0.20)
Dummy Category#Period 4	3.72*** (0.84)	2.81*** (0.84)	1.73*** (0.45)	1.68*** (0.44)	-0.97*** (0.23)	-0.81*** (0.22)
Dummy Category#Period 5	3.69*** (0.84)	1.82** (0.83)	1.70*** (0.45)	1.33*** (0.45)	-0.87*** (0.24)	-0.63*** (0.24)
Dummy Category#Period 6	3.94*** (0.85)	2.87*** (0.84)	2.42*** (0.50)	1.91*** (0.49)	-1.25*** (0.26)	-0.93*** (0.26)
Pre-round Number of Points	0.77*** (0.07)	0.74*** (0.06)				
Pre-round Number of Blocks			0.80*** (0.06)	0.84*** (0.06)		
Pre-round Number of Time-outs					0.67*** (0.16)	0.60*** (0.11)
Constant	5.51*** (1.16)	5.55*** (1.04)	3.23*** (0.62)	2.69*** (0.61)	0.07 (0.07)	0.08 (0.07)
Observations	1758	1740	1758	1740	1758	1740
Individuals	293	290	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Pooled ordinary least squares regressions on individual performance are performed. Period dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

in performance, effort or time-outs taken in the first period. However, comparing treatment D with ND in columns (1), (3), and (5), the point estimates of the interaction terms of the treatment indicator with the remaining periods 2-6 are all significant. Moreover, the magnitude of the point estimates increases from for example 1.60 in period 2 to 3.94 in period 6 (column (1)). Comparing treatment D with TD (columns (2), (4), and (6)), we observe a similar pattern with the exclusion that output did not differ significantly in period 2. Hence, we observe that over time subjects achieved more points, solved more number blocks, and took less time-outs in treatment D than in treatment ND and TD.

We thus find support for the hypothesis that any potential *Kindness-of-the-Scale Effect* fades over time observing that the positive performance effect of a dummy category became dominant over time.

Moreover, we observe that performance differences can be partly explained by time-outs taken: Over the course of the working periods, subjects in treatment D continued working while subjects in treatment ND and TD exploited the time-out option more often maximizing their own pay-out and free time.

In the following, we analyze whether the observed performance increase in the presence of a dummy category can be attributed to perceived higher incentives (*Incentive Effect*), reciprocal reactions to higher relative ratings (*Evaluation Effect*), or both. Therefore, we first explore whether the performance effect is stronger for low performers. We then examine whether the performance effect differs between strongly and weakly reciprocal subjects.

Individual Performance Across Performance Quantiles Low performers have the highest probability of "falling" into the dummy category and hence have the highest incentive to increase performance. Accordingly, a stronger performance increase among low performers would support the hypothesis that performance effect is driven by the *Incentive Effect.* We test this conjecture by analyzing quantile regressions estimating the effect of a dummy category on period performance for the *.25*-, *.50*- , and *.75*-Quantile (see Table 3.3). The results do

Table 3.3: Effect of a Dummy Category on Individual Performance Across Performance Quantiles

Dependent Variable:	*.25*-Quantile		*.50*-Quantile		*.75*-Quantile	
Number of Points	D vs. ND	D vs. TD	D vs. ND	D vs. TD	D vs. ND	D vs. TD
	(1)	(2)	(3)	(4)	(5)	(6)
Dummy Category	2.13**	2.13**	1.35***	0.84	0.95**	0.49
	(0.85)	(0.89)	(0.52)	(0.52)	(0.48)	(0.52)
Percentile of Pre-round Number of Points	14.58***	14.53***	12.82***	11.13***	11.42***	9.62***
	(1.54)	(1.65)	(0.93)	(0.98)	(1.05)	(0.96)
Constant	6.36***	5.97***	11.24***	11.91***	14.31***	15.61***
	(1.03)	(1.39)	(0.59)	(0.77)	(0.68)	(0.64)
Observations	1758	1740	1758	1740	1758	1740
Individuals	293	290	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Quantile Regressions on individual output are performed. Period Dummies are included.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

not change if we include session dummies or estimate the effect on total individual performance (see Appendix 3.B.3). In columns (1), (3), and (5) of Table 3.3 we compare the effect in treatment D with ND, in the other columns treatment D with TD. We control for individual differences by including subjects' pre-round performance percentile.

The estimated treatment effect was highest for low performers. The higher the performance quantile, the smaller the point estimate of the treatment indicator "Dummy Category": Subjects in treatment Dummy worked about 2 points more per period at the *.25*-Quantile (columns (1) and (2) of Table 3.3), while performance differences at the *.75*-Quantile were less than 1 point. Moreover, comparing treatment D with TD, we observe that the point estimate is only significant for the *.25*-Quantile. This suggests, that the treatment effect was particularly present for low performers. See Appendix 3.B.3 for a visualization of this result in Figures 3.5 and 3.6 and an additional regression analyses as robustness check.

Performance Effect by Reciprocity Score We next analyze whether we can attribute the observed performance increase in treatment D not only to higher communicated incentives but also to behavioral responses to more generous ratings. A given rating may appear more generous relative to the lowest rating category in the presence of a dummy category. This, in turn, may induce reciprocal reactions resulting in higher subsequent performance (*Evaluation Effect*). If the observed performance effect was indeed partially reciprocal reactions, strongly reciprocal subjects should increase performance to a stronger extent than weakly reciprocal subjects. To test this conjecture, we extend our analysis of total performance (Table 3.1) and our analysis of performance across performance quantiles (Table 3.3) by an interaction term *Dummy Category#Standardized Reciprocity Score*.[5] The results are shown in Table 3.4 and in Table 3.14 in the Appendix 3.B.4. As the insignificant interaction term in all specifications shows, we find no evidence

[5] See Appendix 3.B.4 for details on subjects' reciprocity score.

Table 3.4: Effect of a Dummy Category on Total Individual Performance Depending on Reciprocity

Dependent Variable: Total Number of Points	D vs. ND (1)	D vs. TD (2)	D vs. ND (3)	D vs. TD (4)
Dummy Category	14.17***	10.40***	14.15***	10.65***
	(3.76)	(3.73)	(3.71)	(3.77)
Dummy Category#Standardized Reciprocity Score	3.93	3.37	5.47	3.07
	(4.33)	(3.46)	(4.45)	(3.66)
Standardized Reciprocity Score	-3.48	-2.90	-4.38	-2.50
	(3.84)	(2.83)	(3.85)	(3.08)
Pre-round Number of Points	4.59***	4.47***	4.61***	4.55***
	(0.41)	(0.35)	(0.41)	(0.35)
Constant	30.53***	36.36***	36.13***	31.05***
	(7.56)	(6.91)	(8.55)	(9.13)
Observations	293	290	293	290
Session Dummies	No	No	Yes	Yes

Notes: Ordinary least squares regressions on total individual performance are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

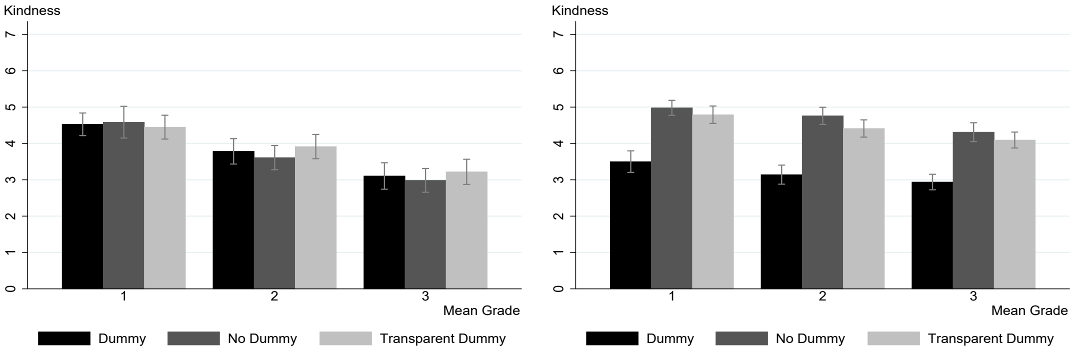
that performance differed substantially between strongly and weakly reciprocal subjects.

As a result, we do not find evidence for an *Evaluation Effect* as in Chapter 2 (Vogt 2021). We hence attribute the performance increase in the presence of a dummy category to an *Incentive Effect* in our dynamic setting.

3.3.2 Effect of a Dummy Category on the Perceived Kindness of a Rating Scale

We next analyze whether subjects perceived the kindness of rating scales differently across treatments. We asked multiple questions on the kindness of a rating scale, kindness of an employer, and trust level of an employer using 7-point likert scales. Increasing values reflect higher perceived levels of kindness or trust. To evaluate the kindness of the rating scale, we asked for example whether subjects agree to the statements "the performance rating scale was fairly designed" or "the performance rating scale allowed kind evaluations". See Appendix 3.B.5 for more details on the questions and the respective scores. We present the effect on the

Figure 3.4: Evaluation of the Kindness of a Rating Scale Across Treatments



(a) Evaluation Without Reference Scales

(b) Evaluation With Reference Scales

kindness of a rating scale. The results are similar for the perceived kindness and trust level of an employer (Appendix 3.B.5).

The left graph in Figure 3.4 plots the mean kindness evaluation of subjects' own rating scale when they did not know the rating scales of the other treatments. We differentiate by the mean grade received over all six periods. The evaluation of the kindness of one's own rating scale was not significantly different across treatments within each grade class. In Appendix 3.B.5 we provide the p-values of Wilcoxon rank-sum tests comparing the evaluations across grade classes. Consistent with the observed performance effect, we do not find support for a *Kindness-of-the-Scale Effect* if subjects receive performance ratings over multiple periods.

The right graph in Figure 3.4 shows the mean evaluation of rating scales when subjects were asked to evaluate the rating scales of the other treatments. We differentiate by the mean grade received over all six periods. The rating scale used in treatment D was evaluated as being less kind than the scales used in treatment ND and TD (Wilcoxon rank-sum test, two sided, $p < 0.000$). Evaluations of the rating scales shown in ND and TD did not differ (Wilcoxon rank-sum test, two sided, $p = 0.6647$). This observation also holds across grade classes, see Appendix 3.B.5 for p-values of Wilcoxon rank-sum tests comparing the evaluations across grade classes. Subjects perceived the rating scale of treatment D – where subjects did not know that the additional rating category was never used – as "bad news" when they knew reference rating scales. Compared to the evaluation of rating scales without reference scales, subjects evaluated the rating scales of treatment ND and TD as more kind and the scale shown in treatment D as less kind.

The results can be explained by a reference point effect in the choice of a rating scale itself: Without prior reference standard subjects do not perceive a rating scale with a lower rating category as less kind. But they start to do so when apparently more generous scales become salient.

We can see this shift in evaluation comparing individual evaluation of one's own and the other rating scales: We control for the individual kindness evaluation of one's own rating scale by subtracting the initial evaluation of one's own rating scale from the evaluation of the other rating scales. Thus, negative values indicate a lower, positive values reflect a higher evaluation of the other rating scale. Independent of the mean grade, we observe negative values for the rating scale in treatment D and positive values for treatment ND and TD – see Figure 3.7 and Table 3.16 in Appendix 3.B.5.

We also observe that the rating scale in treatment D was evaluated as "bad news" in the answers to the question "which of the three ratings scales would you have chosen for yourself?": Subjects rarely chose the rating scale of treatment D (14.91% of all employees). In contrast, 58.48% of the employees chose the rating scale without and 26.61% chose the scale with the transparent additional low category (Table 3.17 in Appendix 3.B.5). This phenomenon is also present comparing the scale choice within treatments. Interestingly, also 68.75% of the employers did not choose the rating scale of treatment dummy for themselves – even though they received periodic feedback on the performance of each employee group and hence observed that employees in treatment D performed better.

3.4 Conclusion

We analyzed the performance effect of employing a dummy category in performance appraisals. Our main result is that the use of an – in fact then unused – low rating category increased performance by about 20% when subjects were not aware that the additional rating category was never used.

Our results indicate that it may be beneficial to keep low rating categories in performance appraisals even when in fact they are hardly ever used as is reported in many firms. While performance across treatments did not differ in the first period, seeing an additional low rating category over time raised performance – as long as subjects were not aware that this category was never used. We do not find that reciprocal reactions to perceived higher relative ratings in the presence of a dummy category affect performance (*Evaluation Effect*). We hence attribute the performance increase to (perceived) higher incentives in the presence of a dummy category (*Incentive Effect*). In line with this reasoning, we observe that the effect was driven by low performers. Over time, any potential negative behavioral consequence of employing a less kind rating scale was thus outweighed by (perceived) higher incentives in the presence of a dummy category. In line with the performance results, subjects perceived the rating scales equally kind across treatments when they only knew the scale chosen by their own employer and thus had no clear reference point for the evaluation of the scale itself. Subjects thus apparently focus on their personal performance ratings and incentives rather than on the kindness of a rating scale when they receive ratings repeatedly over time.[6]

Our findings have several implications for the design of performance ratings in organizations. First, our treatments No Dummy and Transparent Dummy show that when there is no additional rating category or employees know that the respective low category is unused, performance is significantly lower – both,

[6] Recall that individuals indeed perceived the scale in treatment Dummy as being less kind when they knew the rating scales of the other treatments.

economically and statistically. Hence, firms that employ unused rating categories in existing appraisal systems, should not disclose the non-usage of the respective categories. Moreover, when designing new appraisal systems, firms may frame the rating scale intentionally by adding a low rating category which in fact will never be used rather than employing a shorter scale.

Second, the evaluation of kindness of the overall rating scale depends on employees' knowledge about the design of other rating scales (say in the market or about previously employed scale). Without the comparison to other rating scales, subjects did not perceive the unused low rating category in treatment Dummy as unkind. But with the comparison to other rating scales, subjects perceived the added rating category as less kind. Moreover, our results from the dynamic interaction show, most likely any such *Kindness-of-the-Scale Effect* may vanish over time as employees accommodate to the used scale and the *Incentive Effect* predominates.

Third, the above suggests that the *Kindness-of-the-Scale Effect* would be stronger if individuals knew reference rating scales. However, since in practice employees are typically not exposed to salient comparisons between different rating scales, firms' common practice of having unused low rating categories indeed seems a sensible design choice.

3.5 Acknowledgements

This work was supported by the Reinhard Selten Institute and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy - EXC 2126/1–390838866; and the Center for Social and Economic Behavior (C-SEB) of the University of Cologne. This study is preregistered under AEARCTR-0004385. Institutional Review Board approval was obtained from the University of Cologne. Please see Appendix 3.C for the experimental instructions and screenshots.

Appendix of Chapter 3

3.A Sample Demographics

Table 3.5: Sample Demographics

Demographics	Percentage (N=468)
Age[1]	25.38 (5.61)
Female	60.90
Desired degree	
Bachelor's degree	4.06
Master's degree	59.61
State examination	36.11
Not studying	0.21
Employment status	
Working	70.30
Not working	29.70
Subject	
Economics	9.62
Business Administration	18.38
Social Science	4.27
Political Science	1.07
Information Systems	2.35
Health Economics	2.35
Vocational School Teacher Training	11.97
Economics and Social Sciences	0.85
Psychology	1.92
Medicine	4.27
Other	38.89
Not studying	4.06

Note: [1] Mean in years (standard deviation)

3.B Further Analyses

3.B.1 Effect of a Dummy Category on Total Individual Output

Total individual performance was significantly higher in treatment Dummy – where subjects did not know that the additional rating category was unused – as compared to treatment No Dummy and Transparent Dummy (Wilcoxon rank-sum test, one-sided, $p = .000$ and $p = .062$).

Table 3.6: Effect of a Dummy Category on Total Individual Performance II

Dependent Variable:	Total Number of Points			Log of Total Number of Points		
	D vs. ND (1)	D vs. TD (2)	TD vs. ND (3)	D vs. ND (4)	D vs. TD (5)	TD vs. ND (6)
Dummy Category	16.58***	10.64**		0.22***	0.16***	
	(4.56)	(4.53)		(0.06)	(0.06)	
Transparent Dummy Category			5.94			0.06
			(5.05)			(0.07)
Constant	104.43***	110.37***	104.43***	4.52***	4.58***	4.52***
	(3.59)	(3.55)	(3.59)	(0.05)	(0.05)	(0.05)
Observations	293	290	289	293	290	289

Notes: Ordinary least squares regressions on total individual performance are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* p < 0.10, ** p < 0.05, *** p <0.01, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

Table 3.7: Effect of a Dummy Category on Total Individual Performance III

Dependent Variable:	Total Number of Points			Log of Total Number of Points		
	D vs. ND (1)	D vs. TD (2)	TD vs. ND (3)	D vs. ND (4)	D vs. TD (5)	TD vs. ND (6)
Dummy Category	13.90***	10.73***		0.19***	0.17***	
	(3.66)	(3.72)		(0.05)	(0.05)	
Transparent Dummy Category			3.14			0.02
			(4.22)			(0.06)
Pre-round Number of Points	4.66***	4.55***	4.77***	0.05***	0.05***	0.06***
	(0.41)	(0.34)	(0.40)	(0.01)	(0.01)	(0.01)
Constant	35.73***	31.05***	34.44***	3.75***	3.65***	3.64***
	(8.41)	(9.18)	(9.64)	(0.14)	(0.15)	(0.15)
Observations	293	290	289	293	290	289
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Ordinary least squares regressions on total individual performance are performed. Session dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* p < 0.10, ** p < 0.05, *** p <0.01, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

Table 3.8: Effect of a Dummy Category on Total Effort Provision, Time-outs, and Mistakes I

Dependent Variable:	Total Number of Blocks			Total Number of Time-outs			Total Number of False Blocks		
	D vs. ND (1)	D vs. TD (2)	TD vs. ND (3)	D vs. ND (4)	D vs. TD (5)	TD vs. ND (6)	D vs. ND (7)	D vs. TD (8)	TD vs. ND (9)
Dummy Category	7.05***	5.95***		-4.02***	-2.53***		-0.19	-0.11	
	(1.99)	(1.91)		(0.92)	(0.87)		(0.96)	(0.90)	
Transparent Dummy Category			0.93			-1.56			-0.16
			(2.33)			(1.09)			(0.94)
Pre-round Number of Blocks	4.78***	5.04***	5.37***						
	(0.38)	(0.34)	(0.42)						
Pre-round Number of Time-outs				4.03***	3.62***	3.96***			
				(0.93)	(0.65)	(0.61)			
Pre-round Number of False Blocks							1.97***	1.70***	2.11***
							(0.36)	(0.30)	(0.36)
Constant	20.56***	19.10***	14.86***	5.96***	4.52***	5.97***	9.21***	9.46***	9.01***
	(4.06)	(4.13)	(4.41)	(0.83)	(0.76)	(0.82)	(0.78)	(0.72)	(0.76)
Observations	293	290	289	293	290	289	293	290	289

Notes: Ordinary least squares regressions on total individual performance are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* p < 0.10, ** p < 0.05, *** p <0.01, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

Table 3.9: Effect of a Dummy Category on Total Effort Provision, Time-outs, and Mistakes II

Dependent Variable:	Total Number of Blocks			Total Number of Time-outs			Total Number of False Blocks		
	D vs. ND	D vs. TD	TD vs. ND	D vs. ND	D vs. TD	TD vs. ND	D vs. ND	D vs. TD	TD vs. ND
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dummy Category	7.07***	6.03***		-4.01***	-2.56***		-0.17	-0.22	
	(1.97)	(1.94)		(0.92)	(0.87)		(0.95)	(0.91)	
Transparent Dummy Category			0.93			-1.56			-0.16
			(2.33)			(1.09)			(0.94)
Pre-round Number of Blocks	4.70***	5.11***	5.37***						
	(0.37)	(0.35)	(0.42)						
Pre-round Number of Time-outs				4.26***	3.60***	3.96***			
				(0.92)	(0.66)	(0.61)			
Pre-round Number of False Blocks							1.94***	1.53***	2.11***
							(0.37)	(0.30)	(0.36)
Constant	23.40***	16.43***	14.86***	3.66***	4.26**	5.97***	9.99***	11.69***	9.01***
	(4.99)	(6.09)	(4.41)	(1.17)	(1.80)	(0.82)	(1.77)	(1.88)	(0.76)
Observations	293	290	289	293	290	289	293	290	289
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Ordinary least squares regressions on total individual performance are performed. Session dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

3.B.2 Effect of a Dummy Category on Individual Performance, Effort Provision, and Time-outs Taken Over Time

Table 3.10: Effect of a Dummy Category on Individual Performance, Effort Provision, and Time-outs Taken Over Time II

Dependent Variable:	Number of Points		Number of Blocks		Number of Time-outs	
	D vs. ND (1)	D vs. TD (2)	D vs. ND (3)	D vs. TD (4)	D vs. ND (5)	D vs. TD (6)
Dummy Category	-0.26 (0.51)	0.20 (0.54)	-0.11 (0.22)	0.02 (0.23)	0.05 (0.09)	0.05 (0.10)
Dummy Category#Period 2	1.60** (0.63)	0.10 (0.62)	0.52* (0.29)	-0.03 (0.24)	-0.41*** (0.15)	0.06 (0.13)
Dummy Category#Period 3	2.51*** (0.74)	1.92** (0.75)	1.34*** (0.38)	1.04*** (0.37)	-0.80*** (0.21)	-0.53*** (0.20)
Dummy Category#Period 4	3.72*** (0.84)	2.81*** (0.84)	1.73*** (0.45)	1.68*** (0.44)	-0.97*** (0.23)	-0.81*** (0.22)
Dummy Category#Period 5	3.69*** (0.85)	1.82** (0.84)	1.70*** (0.46)	1.33*** (0.45)	-0.87*** (0.24)	-0.63*** (0.24)
Dummy Category#Period 6	3.94*** (0.86)	2.87*** (0.85)	2.42*** (0.50)	1.91*** (0.49)	-1.25*** (0.26)	-0.93*** (0.26)
Pre-round Number of Points	0.78*** (0.07)	0.76*** (0.06)				
Pre-round Number of Blocks			0.78*** (0.06)	0.85*** (0.06)		
Pre-round Number of Time-outs					0.71*** (0.15)	0.60*** (0.11)
Constant	6.42*** (1.33)	4.65*** (1.42)	3.70*** (0.79)	2.24** (0.93)	-0.31* (0.19)	0.03 (0.27)
Observations	1758	1740	1758	1740	1758	1740
Individuals	293	290	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Pooled ordinary least squares regressions on individual performance are performed. Period and session dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

3.B.3 Effect of a Dummy Category Across Performance Quantiles

As a robustness check, we run two regressions to investigate whether the observed performance effects differ across performance quantiles.

First, we report quantile regressions estimating the effect of a dummy category on individual period performance controlling for session effects in Table 3.11. In columns (1), (3), and (5) we compare effects in treatment D with ND, in the remaining columns treatment D with TD. The results replicate the regression analysis of individual period performance without controlling for session effects in the main text. The magnitude of the point estimates decreases with the performance quantile. The results are robust to regressing the effects of a dummy category on total individual performance (see Table 3.12).

Second, as treatment specifics were revealed only after the pre-round, we use subjects' pre-round performance percentile as an unbiased ability measure. We estimate a pooled OLS model regressing period performance on a treatment dummy and a treatment dummy interacted with subjects' pre-round performance percentile (see Table 3.13). In columns (1) and (3) we compare treatment D with ND, in columns (2) and (4) we compare treatment D with TD. The positive significant coefficient of the treatment dummy estimates that a dummy category increased performance of the lowest performers (performance percentile= 0) by 4.03 and 2.45 units, respectively. The negative interaction term, however, suggests that the effect of a dummy category decreased with subjects' initial performance percentile. For example, the estimated effect for the best performing subject is only $4.03 - 3.31 = 0.72$ and $2.45 - 1.25 = 1.20$ units. Figures 3.5 & 3.6 visualize this finding. We plot average performance over time differentiating subjects according to their pre-round performance quartile. The higher the initial performance quartile, the lower the treatment differences. Hence, we observe that performance effects were stronger for and hence driven by low performers.

Table 3.11: Effect of a Dummy Category on Individual Performance Across Performance Quantiles II

Dependent Variable:	*.25*-Quantile		*.50*-Quantile		*.75*-Quantile	
Number of Points	D vs. ND	D vs. TD	D vs. ND	D vs. TD	D vs. ND	D vs. TD
	(1)	(2)	(3)	(4)	(5)	(6)
Dummy Category	2.45***	2.40**	1.39***	1.10**	0.94**	0.39
	(0.79)	(0.96)	(0.53)	(0.53)	(0.47)	(0.55)
Percentile of Pre-round Number of Points	14.50***	14.78***	12.90***	11.31***	11.63***	10.27***
	(1.39)	(1.81)	(0.96)	(1.05)	(1.09)	(1.05)
Constant	6.42***	4.97**	11.70***	10.90***	14.37***	14.92***
	(1.69)	(2.04)	(1.21)	(1.28)	(1.14)	(1.09)
Observations	1758	1740	1758	1740	1758	1740
Individuals	293	290	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Quantile Regressions on individual performance are performed. Period and session dummies are included.

* $p < 0.10$, ** $p < 0.05$, *** $p <0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

Table 3.12: Effect of a Dummy Category on Total Individual Performance Across Performance Quantiles

Dependent Variable:	*.25*-Quantile		*.50*-Quantile		*.75*-Quantile	
Number of Points	D vs. ND	D vs. TD	D vs. ND	D vs. TD	D vs. ND	D vs. TD
	(1)	(2)	(3)	(4)	(5)	(6)
Dummy Category	16.97***	13.29**	9.22**	3.44	6.87**	4.31
	(6.54)	(5.91)	(4.23)	(4.04)	(3.35)	(3.75)
Percentile of Pre-round Number of Points	87.03***	93.28***	80.77***	69.48***	67.42***	52.08***
	(8.41)	(9.19)	(7.27)	(7.32)	(6.16)	(7.96)
Constant	41.25***	42.07***	71.36***	82.11***	95.08***	107.30***
	(7.21)	(8.96)	(5.92)	(6.11)	(4.42)	(6.52)
Observations	293	290	293	290	293	290

Notes: Quantile Regressions on total individual performance are performed. Period dummies are included.

* $p < 0.10$, ** $p < 0.05$, *** $p <0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

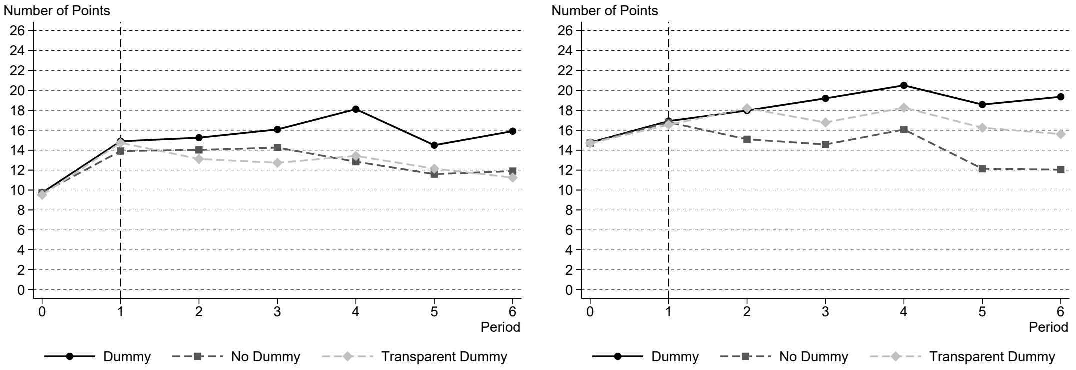
Table 3.13: Effect of a Dummy Category on Individual Performance Depending on Pre-round Performance

Dependent Variable:	D vs. ND	D vs. TD	D vs. ND	D vs. TD
Number of Points	(1)	(2)	(3)	(4)
Dummy Category	4.03***	2.45*	4.13***	2.34*
	(1.34)	(1.43)	(1.27)	(1.41)
Dummy Category#Percentile of Pre-round Number of Points	-3.31	-1.25	-3.50*	-0.96
	(2.10)	(2.12)	(1.98)	(2.12)
Percentile of Pre-round Number of Points	14.78***	12.72***	14.99***	12.85***
	(1.63)	(1.65)	(1.53)	(1.64)
Constant	9.24***	10.32***	9.68***	9.34***
	(0.93)	(1.05)	(1.34)	(1.55)
Observations	1758	1740	1758	1740
Individuals	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes
Session Dummies	No	No	Yes	Yes

Notes: Pooled ordinary least squares regressions on individual performance are performed. Period dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p <0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

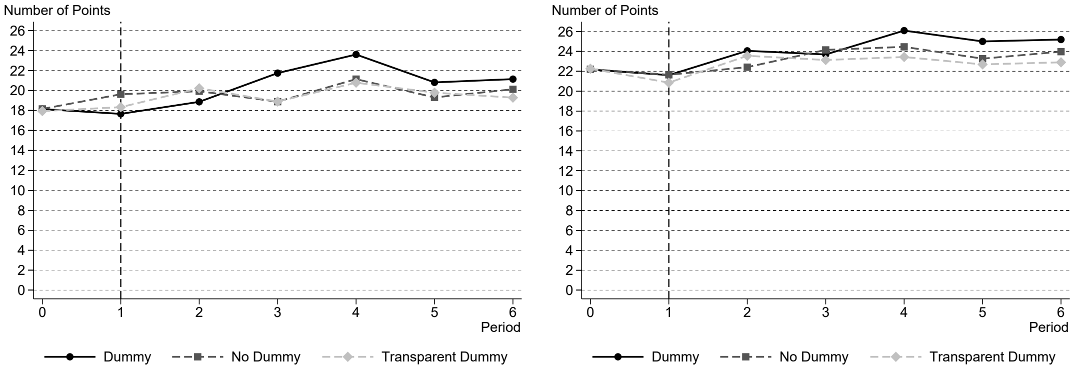
Figure 3.5: Distribution of Individual Performance Over Time Across Treatments: 1^{st} & 2^{nd} Performance Quartile



(a) Subjects Belonging to 1^{st} Quartile

(b) Subjects Belonging to 2^{nd} Quartile

Figure 3.6: Distribution of Individual Performance Over Time Across Treatments: 3^{rd} & 4^{th} Performance Quartile



(a) Subjects Belonging to 3^{rd} Quartile

(b) Subjects Belonging to 4^{th} Quartile

3.B.4 Effect of a Dummy Category Across the Reciprocity Score

We obtained individual reciprocity scores asking the six standardized 7-point likert questions of the German Socio-Economic Panel (Infratest Sozialforschung 2012) at the conclusion of the experiment. Increasing values reflect a higher score on the respective reciprocity dimension. In our setting it is not clear whether the positive or negative reciprocity dimension is the driving personality trait:

Subjects could repay better ratings (positive dimension) or retaliate worse ratings (negative dimension). Accordingly, we use a single reciprocity score indicating how high someone scores on both dimensions by taking the mean of the six individual answers for each subject.

Table 3.14: Effect of a Dummy Category on Individual Performance Depending on Reciprocity II

Dependent Variable:	*.25*-Quantile		*.50*-Quantile		*.75*-Quantile	
Number of Points	D vs. ND (1)	D vs. TD (2)	D vs. ND (3)	D vs. TD (4)	D vs. ND (5)	D vs. TD (6)
Dummy Category	2.22^{***} (0.80)	2.15^{**} (0.89)	1.41^{***} (0.52)	0.87 (0.53)	0.98^{*} (0.53)	0.59 (0.57)
Dummy Category#Standardized Reciprocity Score	0.30 (0.93)	0.56 (0.72)	0.62 (0.58)	0.53 (0.58)	-0.00 (0.52)	0.08 (0.65)
Standardized Reciprocity Score	-0.03 (0.85)	-0.40 (0.62)	-0.53 (0.46)	-0.40 (0.48)	-0.10 (0.43)	-0.18 (0.59)
Percentile of Pre-round Number of Points	14.66^{***} (1.47)	14.58^{***} (1.63)	12.88^{***} (0.91)	11.24^{***} (1.00)	11.48^{***} (1.07)	9.96^{***} (1.01)
Constant	6.34^{***} (1.00)	6.08^{***} (1.37)	10.97^{***} (0.56)	11.80^{***} (0.77)	14.23^{***} (0.71)	15.40^{***} (0.65)
Observations	1758	1740	1758	1740	1758	1740
Individuals	293	290	293	290	293	290
Period Dummies	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Quantile Regressions on individual performance are performed. Period dummies are included.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.

3.B.5 Effect of a Dummy Category on the Perceived Kindness of a Rating Scale

After the working periods, subjects first evaluated how kind they perceived their own rating scale as well as how kind and trusting they perceived their employer. We then revealed the rating scales used in the other two treatments. Subsequently, we asked subjects which rating scale they would have chosen for themselves. Afterwards, subjects evaluated the kindness of the two other rating scales – which they had not seen when evaluating their own rating scale – answering the same questions as for their own rating scale. See Appendix 3.C for all questions of the questionnaire.

We asked to evaluate positive and negative statements on three dimensions – kindness of a rating scale, kindness of an employer, trust level of an employer – using a 7-point likert scale. For example, we asked 10 questions on the kindness of the scale such as whether subjects agree to the statements ”the performance rating scale allowed generous evaluations” or ”the performance rating scale did not allow fair evaluations”. To evaluate the kindness and trust level of an employer, we asked for example whether the ”employer did not have good intentions” or whether the ”employer showed trust”. We re-coded the negative statements, such that increasing values reflect higher perceived levels of kindness and trust, respectively. For every subject we define a single score for each dimension by taking the mean score of all answers to a dimension.

Table 3.15: Tests Comparing the Kindness and Trust Evaluations Across Treatments

	Kindness of Scale			Kindness of Employer			Trust Level of Employer		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
D vs. ND	.838	.867	.099	.829	.766	.544	.291	.174	.185
D vs. TD	.961	.734	.924	.969	.171	.758	.800	.923	.892
ND vs. TD	.826	.507	.107	.913	.093	.378	.487	.297	.241

Note: We report p-values of two-sided Wilcoxon rank-sum tests.

Figure 3.7: Relative Evaluation of the Kindness of Other Rating Scales

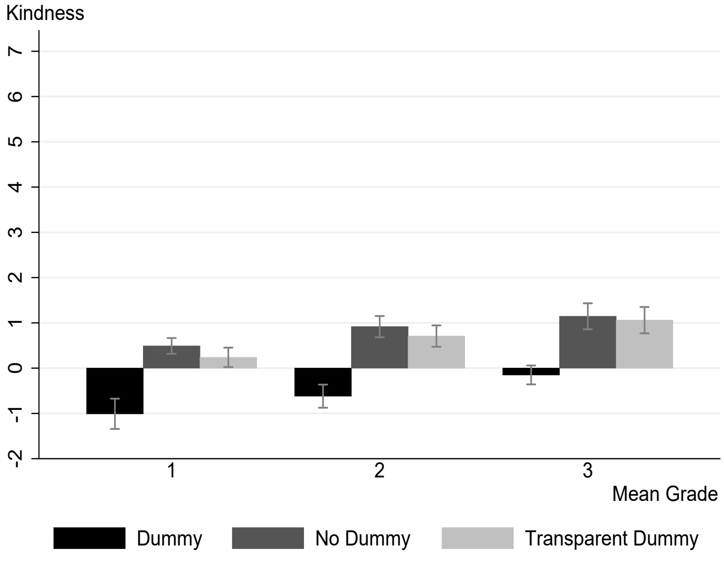


Table 3.16: Tests Comparing the Evaluation of Kindness of Own and Other Rating Scales Across Treatments

		Kindness of Scale	
Absolute/Relative	Grade 1	Grade 2	Grade 3
D vs. ND	.000 / .000	.000 / .000	.001 / .000
D vs. TD	.001 / .000	.000 / .000	.001 / .000
ND vs. TD	.275 / .115	.404 / .283	.221 / .366

Note: We report p-values of two-sided Wilcoxon rank-sum tests.

Table 3.17: The Choice of the Rating Scale for Oneself by Treatment

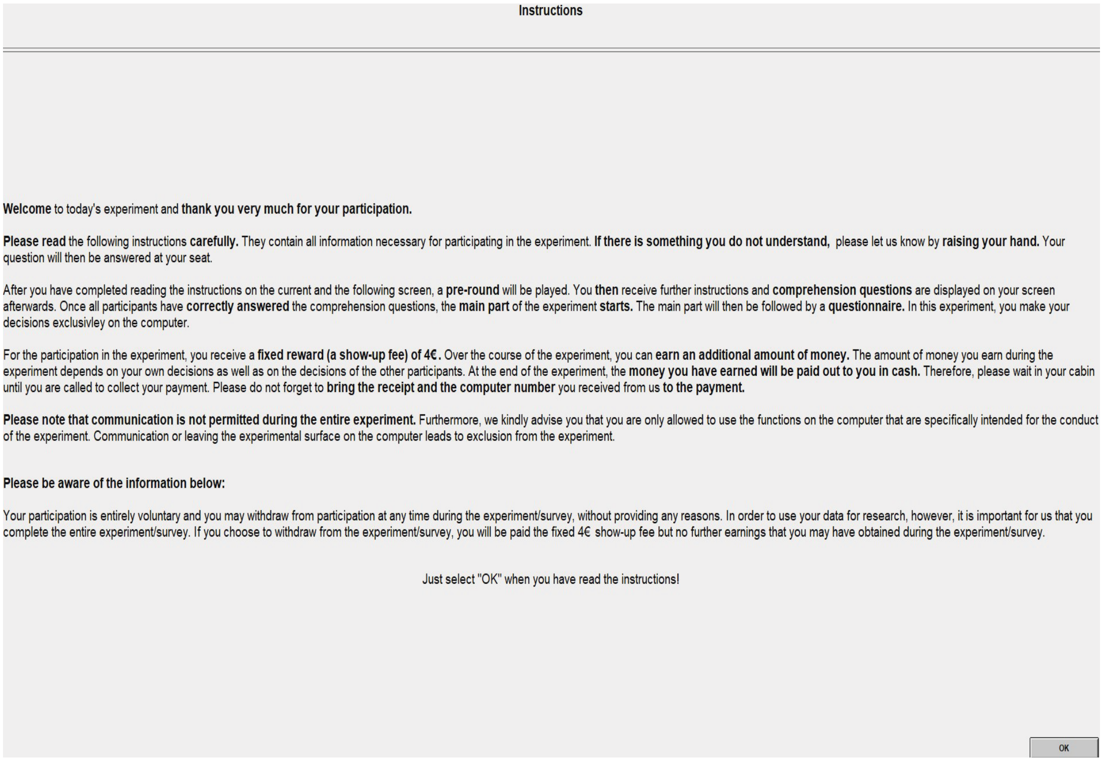
		Scale Chosen For Oneself		
Treatment	Dummy	No Dummy	Transparent Dummy	Total
Dummy	44	77	26	147
No Dummy	13	114	19	146
Transparent Dummy	8	64	71	143
Total	65	255	116	436

3.C Screenshots and Instructions of the Experiment

The experiment language was German. We translated the original German text into English for the following screenshots and instructions of the experiment. It was a computerized experiment except for the instructions that subjects received on paper after the pre-round and before the quiz (see Figures 3.15-3.17). The quiz questions were identical across treatments and subject roles. The correct answers to the quiz questions are given on the respective screenshots (see Figures 3.18-3.20). A questionnaire followed after the main part. The position of the rating scales (left, middle, right) and order of possible answers (top, middle, bottom) were randomly varied for each subject on the screen displayed in Figure 3.29. The position of the rating scale (left, middle, right) was randomly varied for each subject on the screens displayed in Figures 3.30 and 3.31. We show a randomly generated example of the pre-round and the main part. On the screens seen by subjects in the role of employees in the main part and questionnaire, we show the scale displayed in the Transparent Dummy treatment exemplarily. In treatment No Dummy and Dummy, we displayed the respective scales as shown for example in Figure 3.22.

3.C.1 Welcome Screen, Instructions, and Pre-round

Figure 3.8: Welcome Screen



Instructions

Welcome to today's experiment and **thank you very much for your participation.**

Please read the following instructions **carefully.** They contain all information necessary for participating in the experiment. **If there is something you do not understand,** please let us know by **raising your hand.** Your question will then be answered at your seat.

After you have completed reading the instructions on the current and the following screen, a **pre-round** will be played. You **then** receive further instructions and **comprehension questions** are displayed on your screen afterwards. Once all participants have **correctly answered** the comprehension questions, the **main part** of the experiment **starts.** The main part will then be followed by a **questionnaire.** In this experiment, you make your decisions exclusivley on the computer.

For the participation in the experiment, you receive a **fixed reward (a show-up fee) of 4€.** Over the course of the experiment, you can **earn an additional amount of money.** The amount of money you earn during the experiment depends on your own decisions as well as on the decisions of the other participants. At the end of the experiment, the **money you have earned will be paid out to you in cash.** Therefore, please wait in your cabin until you are called to collect your payment. Please do not forget to **bring the receipt and the computer number** you received from us **to the payment.**

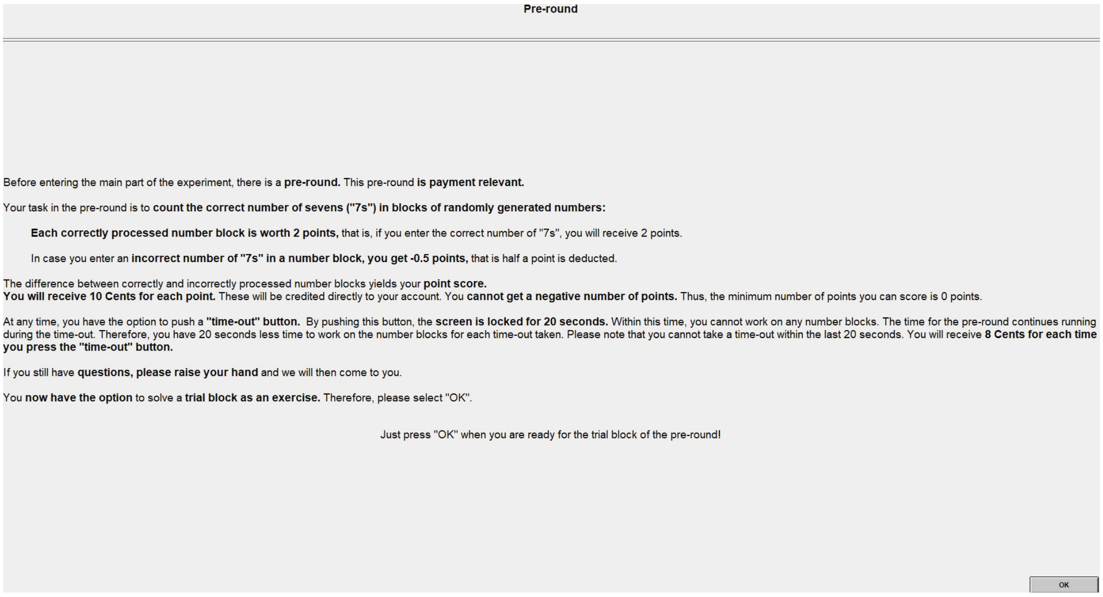
Please note that communication is not permitted during the entire experiment. Furthermore, we kindly advise you that you are only allowed to use the functions on the computer that are specifically intended for the conduct of the experiment. Communication or leaving the experimental surface on the computer leads to exclusion from the experiment.

Please be aware of the information below:

Your participation is entirely voluntary and you may withdraw from participation at any time during the experiment/survey, without providing any reasons. In order to use your data for research, however, it is important for us that you complete the entire experiment/survey. If you choose to withdraw from the experiment/survey, you will be paid the fixed 4€ show-up fee but no further earnings that you may have obtained during the experiment/survey.

Just select "OK" when you have read the instructions!

Figure 3.9: On Screen Instructions



Pre-round

Before entering the main part of the experiment, there is a **pre-round.** This pre-round **is payment relevant.**

Your task in the pre-round is to **count the correct number of sevens ("7s") in blocks of randomly generated numbers:**

Each correctly processed number block is worth 2 points, that is, if you enter the correct number of "7s", you will receive 2 points.

In case you enter an **incorrect number of "7s" in a number block, you get -0.5 points,** that is half a point is deducted.

The difference between correctly and incorrectly processed number blocks yields your **point score.**
You will receive 10 Cents for each point. These will be credited directly to your account. You **cannot get a negative number of points.** Thus, the minimum number of points you can score is 0 points.

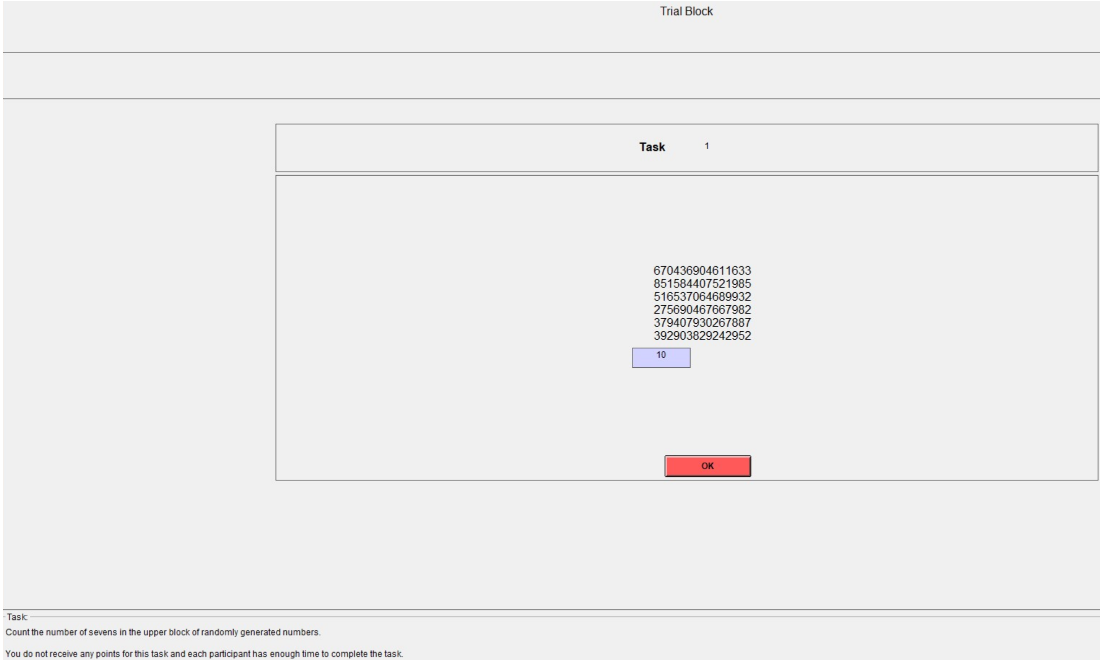
At any time, you have the option to push a **"time-out" button.** By pushing this button, the **screen is locked for 20 seconds.** Within this time, you cannot work on any number blocks. The time for the pre-round continues running during the time-out. Therefore, you have 20 seconds less time to work on the number blocks for each time-out taken. Please note that you cannot take a time-out within the last 20 seconds. You will receive **8 Cents for each time you press the "time-out" button.**

If you still have **questions, please raise your hand** and we will then come to you.

You **now have the option** to solve a **trial block as an exercise.** Therefore, please select "OK".

Just press "OK" when you are ready for the trial block of the pre-round!

Figure 3.10: Trial Block



Trial Block

Task 1

```
670436904611633
851584407521985
516537064689932
275690467667982
379407930267887
392903829242952
```

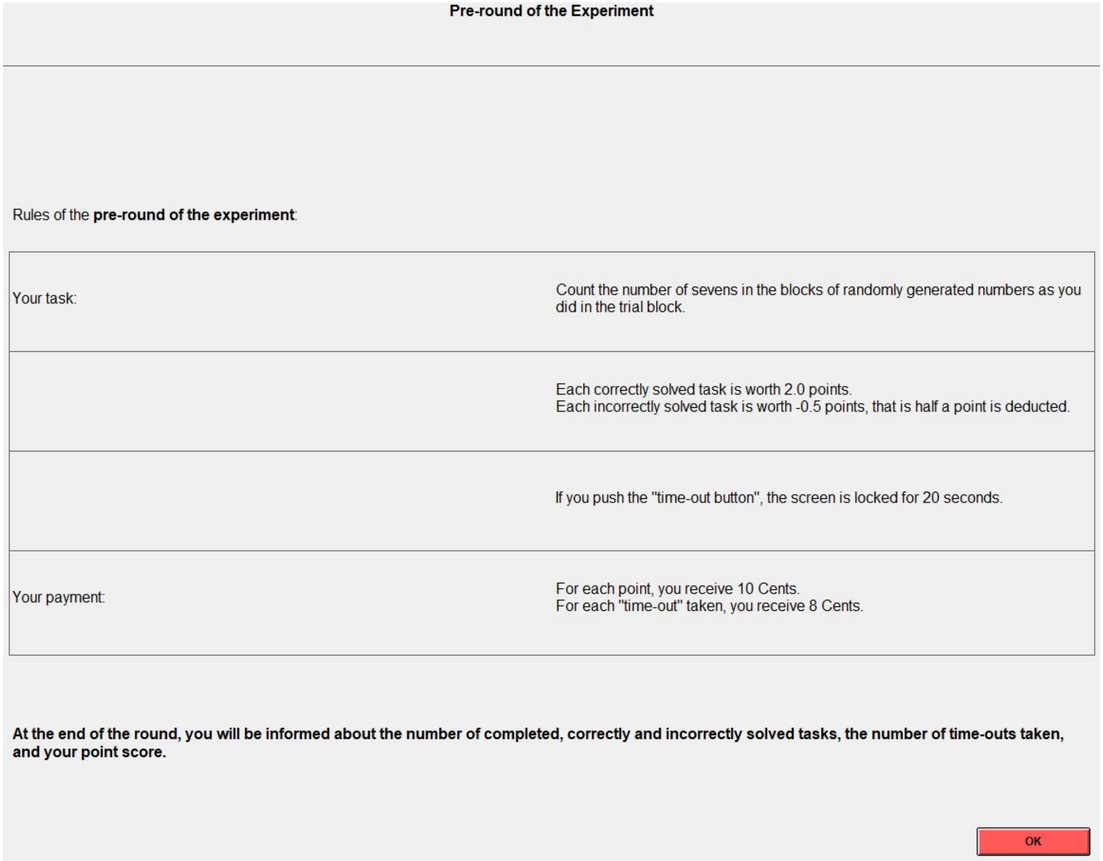
10

Task:

Count the number of sevens in the upper block of randomly generated numbers.

You do not receive any points for this task and each participant has enough time to complete the task.

Figure 3.11: Pre-round Instructions



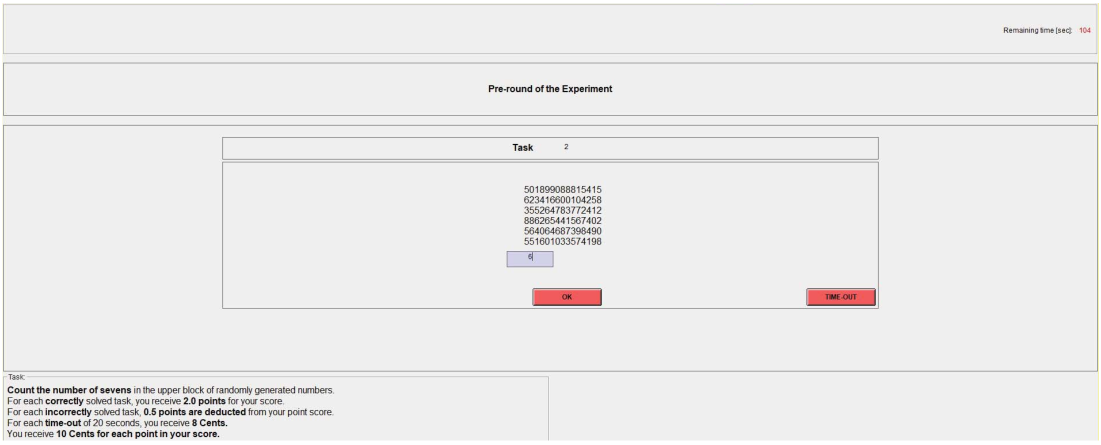
Pre-round of the Experiment

Rules of the **pre-round of the experiment**:

Your task:	Count the number of sevens in the blocks of randomly generated numbers as you did in the trial block.
	Each correctly solved task is worth 2.0 points. Each incorrectly solved task is worth -0.5 points, that is half a point is deducted.
	If you push the "time-out button", the screen is locked for 20 seconds.
Your payment:	For each point, you receive 10 Cents. For each "time-out" taken, you receive 8 Cents.

At the end of the round, you will be informed about the number of completed, correctly and incorrectly solved tasks, the number of time-outs taken, and your point score.

Figure 3.12: Pre-round Task Screen I



Remaining time [sec]: 104

Pre-round of the Experiment

Task 2

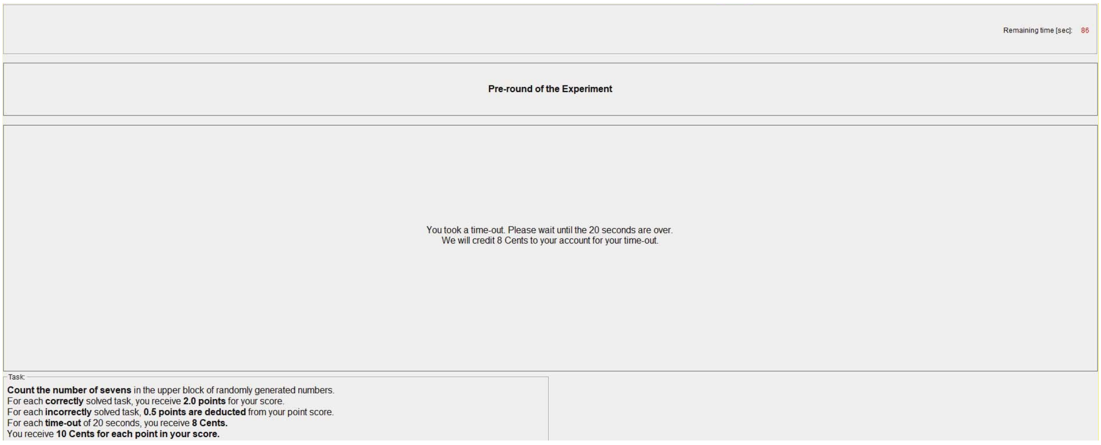
501899088815415
623416600104258
355264783772412
886265441567402
564064687398490
551601033574198

6

Task:

Count the number of sevens in the upper block of randomly generated numbers.
For each **correctly** solved task, you receive **2.0 points** for your score.
For each **incorrectly** solved task, **0.5 points are deducted** from your point score.
For each **time-out** of 20 seconds, you receive **8 Cents.**
You receive **10 Cents for each point in your score.**

Figure 3.13: Pre-round Task Screen II



Remaining time [sec]: 86

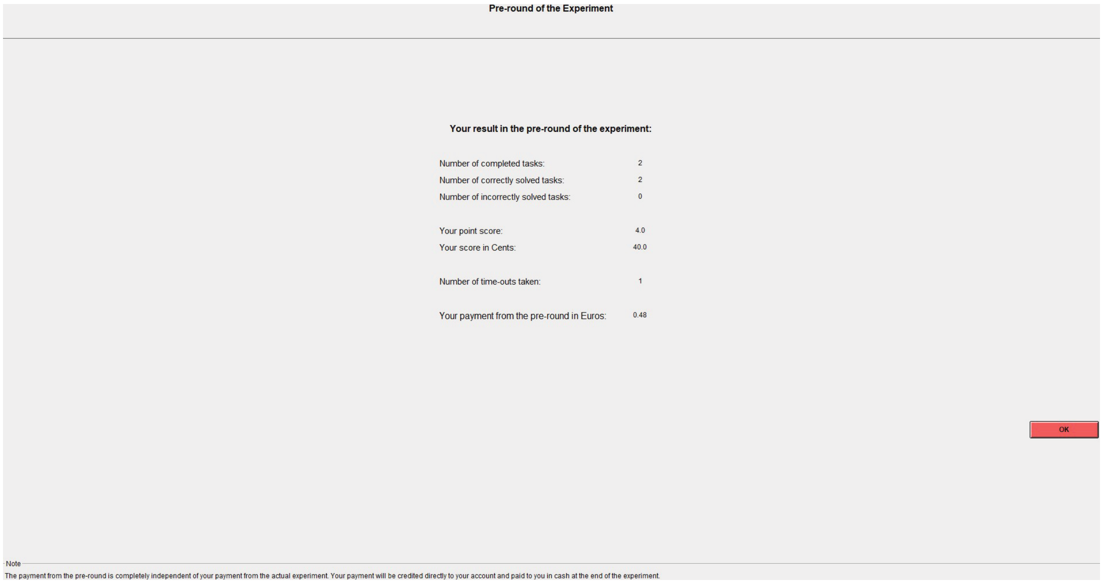
Pre-round of the Experiment

You took a time-out. Please wait until the 20 seconds are over.
We will credit 8 Cents to your account for your time-out.

Task:

Count the number of sevens in the upper block of randomly generated numbers.
For each **correctly** solved task, you receive **2.0 points** for your score.
For each **incorrectly** solved task, **0.5 points are deducted** from your point score.
For each **time-out** of 20 seconds, you receive **8 Cents.**
You receive **10 Cents for each point in your score.**

Figure 3.14: Pre-round Task Feedback



Pre-round of the Experiment

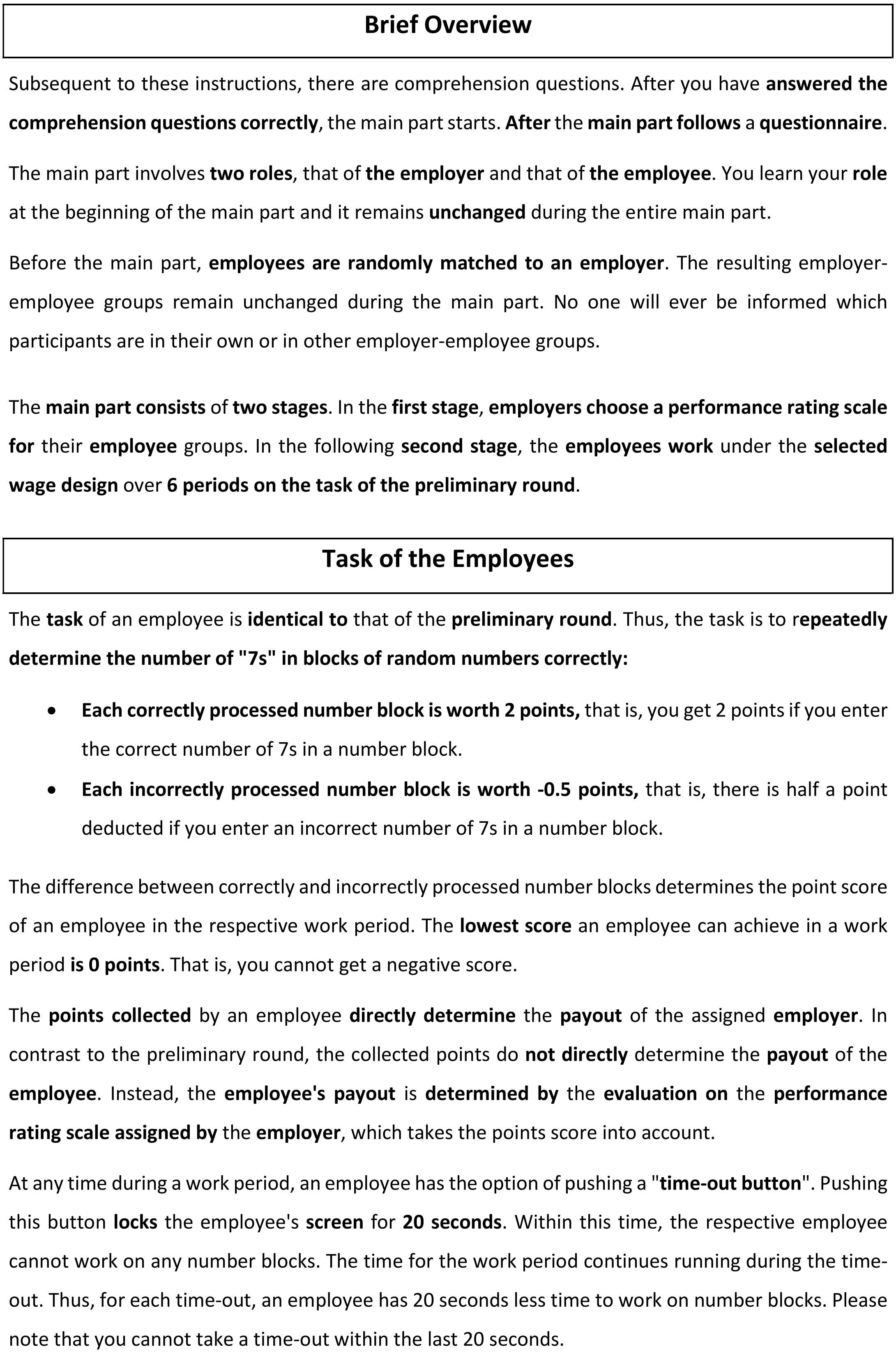
Your result in the pre-round of the experiment:

Number of completed tasks:	2
Number of correctly solved tasks:	2
Number of incorrectly solved tasks:	0
Your point score:	4.0
Your score in Cents:	40.0
Number of time-outs taken:	1
Your payment from the pre-round in Euros:	0.48

Note

The payment from the pre-round is completely independent of your payment from the actual experiment. Your payment will be credited directly to your account and paid to you in cash at the end of the experiment.

Figure 3.15: Paper Instructions I



Brief Overview

Subsequent to these instructions, there are comprehension questions. After you have **answered the comprehension questions correctly**, the main part starts. **After** the **main part follows** a **questionnaire**.

The main part involves **two roles**, that of **the employer** and that of **the employee**. You learn your **role** at the beginning of the main part and it remains **unchanged** during the entire main part.

Before the main part, **employees are randomly matched to an employer**. The resulting employer-employee groups remain unchanged during the main part. No one will ever be informed which participants are in their own or in other employer-employee groups.

The **main part consists** of **two stages**. In the **first stage**, **employers choose a performance rating scale for** their **employee** groups. In the following **second stage**, the **employees work** under the **selected wage design** over **6 periods on the task of the preliminary round**.

Task of the Employees

The **task** of an employee is **identical to** that of the **preliminary round**. Thus, the task is to r**epeatedly determine the number of "7s" in blocks of random numbers correctly:**

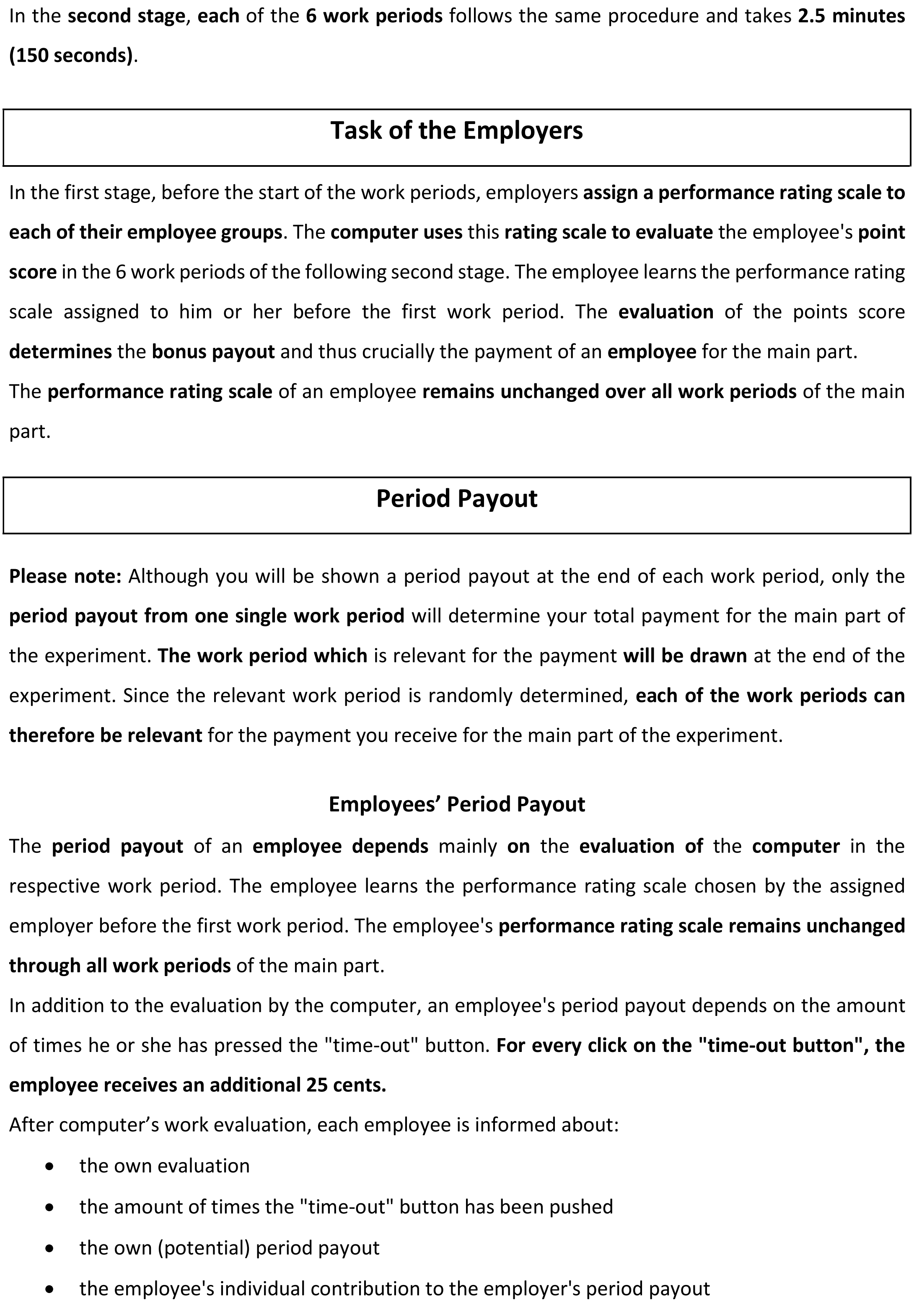
- **Each correctly processed number block is worth 2 points,** that is, you get 2 points if you enter the correct number of 7s in a number block.
- **Each incorrectly processed number block is worth -0.5 points,** that is, there is half a point deducted if you enter an incorrect number of 7s in a number block.

The difference between correctly and incorrectly processed number blocks determines the point score of an employee in the respective work period. The **lowest score** an employee can achieve in a work period **is 0 points**. That is, you cannot get a negative score.

The **points collected** by an employee **directly determine** the **payout** of the assigned **employer**. In contrast to the preliminary round, the collected points do **not directly** determine the **payout** of the **employee**. Instead, the **employee's payout** is **determined by** the **evaluation on** the **performance rating scale assigned by** the **employer**, which takes the points score into account.

At any time during a work period, an employee has the option of pushing a "**time-out button**". Pushing this button **locks** the employee's **screen** for **20 seconds**. Within this time, the respective employee cannot work on any number blocks. The time for the work period continues running during the time-out. Thus, for each time-out, an employee has 20 seconds less time to work on number blocks. Please note that you cannot take a time-out within the last 20 seconds.

Figure 3.16: Paper Instructions II



In the **second stage**, **each** of the **6 work periods** follows the same procedure and takes **2.5 minutes (150 seconds)**.

Task of the Employers

In the first stage, before the start of the work periods, employers **assign a performance rating scale to each of their employee groups**. The **computer uses** this **rating scale to evaluate** the employee's **point score** in the 6 work periods of the following second stage. The employee learns the performance rating scale assigned to him or her before the first work period. The **evaluation** of the points score **determines** the **bonus payout** and thus crucially the payment of an **employee** for the main part.

The **performance rating scale** of an employee **remains unchanged over all work periods** of the main part.

Period Payout

Please note: Although you will be shown a period payout at the end of each work period, only the **period payout from one single work period** will determine your total payment for the main part of the experiment. **The work period which** is relevant for the payment **will be drawn** at the end of the experiment. Since the relevant work period is randomly determined, **each of the work periods can therefore be relevant** for the payment you receive for the main part of the experiment.

Employees' Period Payout

The **period payout** of an **employee depends** mainly **on** the **evaluation of** the **computer** in the respective work period. The employee learns the performance rating scale chosen by the assigned employer before the first work period. The employee's **performance rating scale remains unchanged through all work periods** of the main part.

In addition to the evaluation by the computer, an employee's period payout depends on the amount of times he or she has pressed the "time-out" button. **For every click on the "time-out button", the employee receives an additional 25 cents.**

After computer's work evaluation, each employee is informed about:

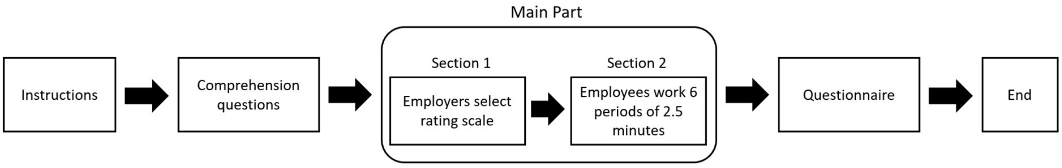
- the own evaluation
- the amount of times the "time-out" button has been pushed
- the own (potential) period payout
- the employee's individual contribution to the employer's period payout

Figure 3.17: Paper Instructions III

Employers' Period Payout

The employer's period payout is solely determined by the points scored by the assigned employees in the respective work period: **The employer receives 2.5 cents for each point collected by an employee.**

Summary



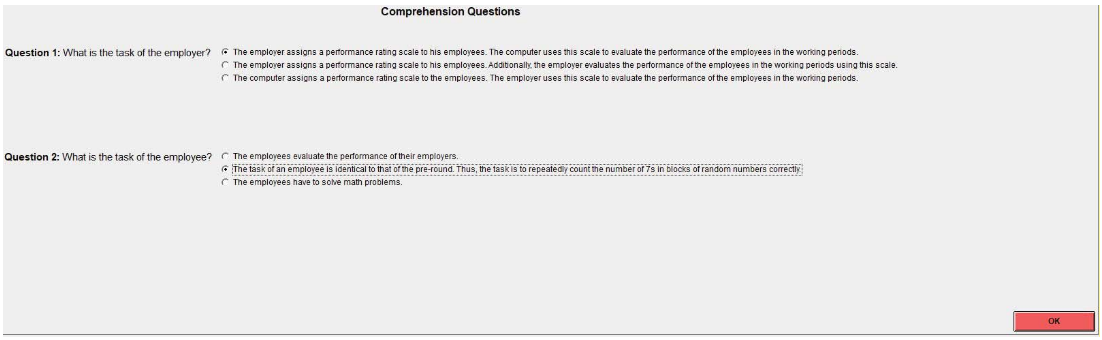
Please keep in mind that **your total payment** for today's experiment **is composed of** the **fixed attendance reward,** the **payout of** the **preliminary round** and the **payout of the randomly selected period in the main part**.

Please bring your **receipt** and **computer number** for the cash **payment** at the **end** of the experiment.

If you are **ready** to answer the **comprehension questions**, please **click** on the **"Ok " button** on the **screen**.

3.C.2 Quiz

Figure 3.18: Quiz Questions I



Comprehension Questions

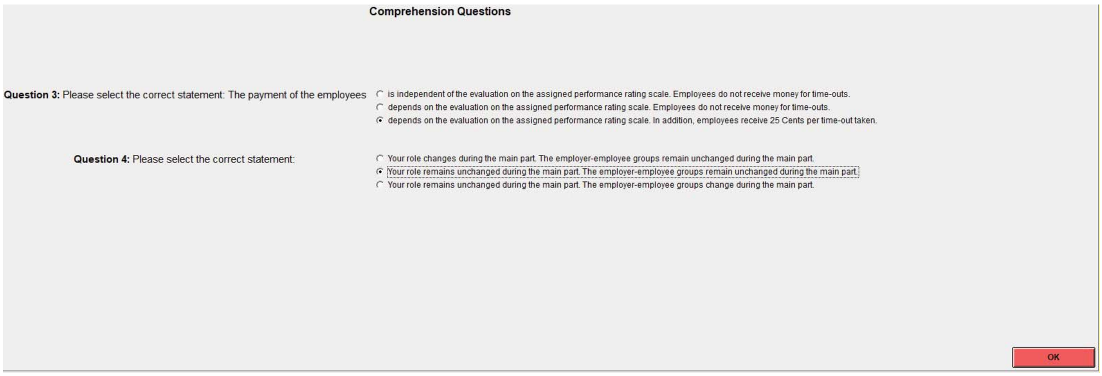
Question 1: What is the task of the employer?

● The employer assigns a performance rating scale to his employees. The computer uses this scale to evaluate the performance of the employees in the working periods.
○ The employer assigns a performance rating scale to his employees. Additionally, the employer evaluates the performance of the employees in the working periods using this scale.
○ The computer assigns a performance rating scale to the employees. The employer uses this scale to evaluate the performance of the employees in the working periods.

Question 2: What is the task of the employee?

○ The employees evaluate the performance of their employers.
● The task of an employee is identical to that of the pre-round. Thus, the task is to repeatedly count the number of 7s in blocks of random numbers correctly.
○ The employees have to solve math problems.

Figure 3.19: Quiz Questions II



Comprehension Questions

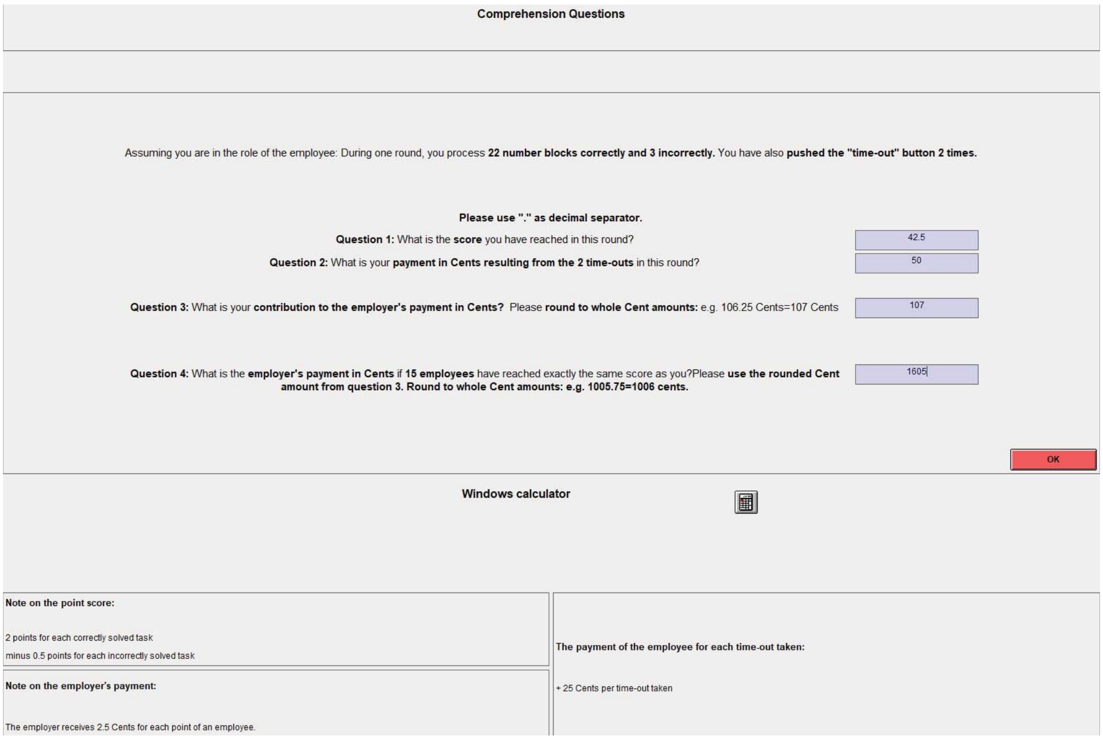
Question 3: Please select the correct statement: The payment of the employees

○ is independent of the evaluation on the assigned performance rating scale. Employees do not receive money for time-outs.
○ depends on the evaluation on the assigned performance rating scale. Employees do not receive money for time-outs.
● depends on the evaluation on the assigned performance rating scale. In addition, employees receive 25 Cents per time-out taken.

Question 4: Please select the correct statement:

○ Your role changes during the main part. The employer-employee groups remain unchanged during the main part.
● Your role remains unchanged during the main part. The employer-employee groups remain unchanged during the main part.
○ Your role remains unchanged during the main part. The employer-employee groups change during the main part.

Figure 3.20: Quiz Question III



Comprehension Questions

Assuming you are in the role of the employee: During one round, you process **22 number blocks correctly and 3 incorrectly.** You have also **pushed the "time-out" button 2 times.**

Please use "." as decimal separator.

Question 1: What is the **score** you have reached in this round? 42.5

Question 2: What is your **payment in Cents resulting from the 2 time-outs** in this round? 50

Question 3: What is your **contribution to the employer's payment in Cents?** Please **round to whole Cent amounts:** e.g. 106.25 Cents=107 Cents 107

Question 4: What is the **employer's payment in Cents** if **15 employees** have reached exactly the same score as you?Please **use the rounded Cent amount from question 3. Round to whole Cent amounts: e.g. 1005.75=1006 cents.** 1605

Note on the point score:

2 points for each correctly solved task
minus 0.5 points for each incorrectly solved task

Note on the employer's payment:

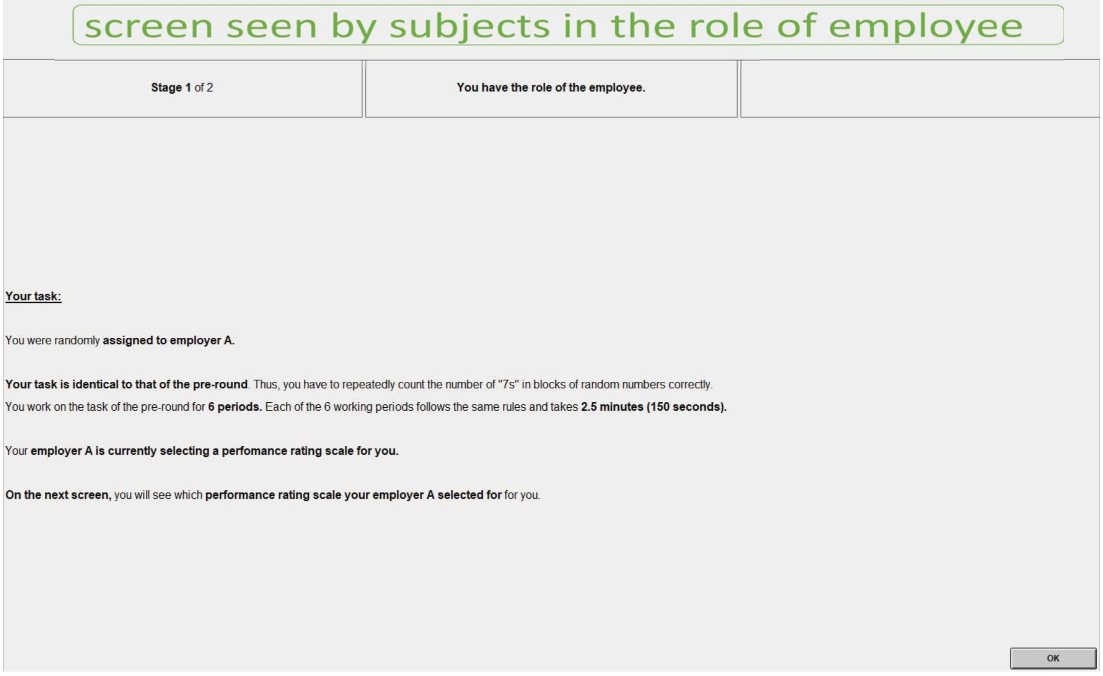
The employer receives 2.5 Cents for each point of an employee.

The payment of the employee for each time-out taken:

+ 25 Cents per time-out taken

3.C.3 Main Part

Figure 3.21: Role Assignment Screen of Subjects in the Role of Employee



screen seen by subjects in the role of employee

Stage 1 of 2

You have the role of the employee.

Your task:

You were randomly **assigned to employer A.**

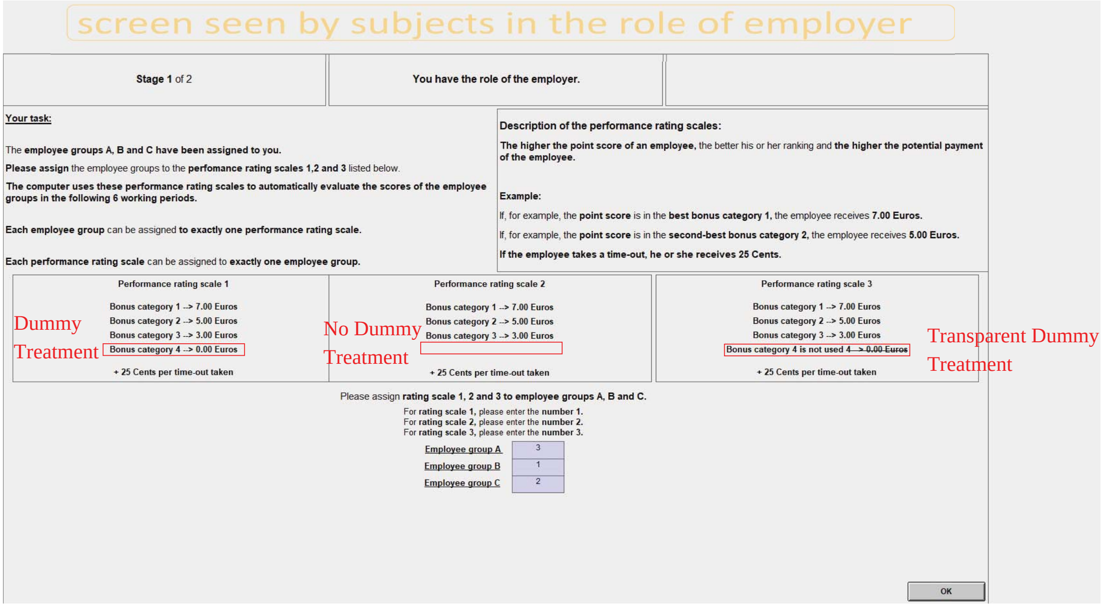
Your task is identical to that of the pre-round. Thus, you have to repeatedly count the number of "7s" in blocks of random numbers correctly.

You work on the task of the pre-round for **6 periods.** Each of the 6 working periods follows the same rules and takes **2.5 minutes (150 seconds).**

Your **employer A is currently selecting a perfomance rating scale for you.**

On the next screen, you will see which **performance rating scale your employer A selected for** for you.

Figure 3.22: Role Assignment & Task Screen of Subjects in the Role of Employer



screen seen by subjects in the role of employer

Stage 1 of 2

You have the role of the employer.

Your task:

The **employee groups A, B and C have been assigned to you.**

Please assign the employee groups to the **perfomance rating scales 1,2 and 3** listed below.

The computer uses these performance rating scales to automatically evaluate the scores of the employee groups in the following 6 working periods.

Each employee group can be assigned **to exactly one performance rating scale.**

Each performance rating scale can be assigned to **exactly one employee group.**

Description of the performance rating scales:

The higher the point score of an employee, the better his or her ranking and **the higher the potential payment of the employee.**

Example:

If, for example, the **point score** is in the **best bonus category 1,** the employee receives **7.00 Euros.**

If, for example, the **point score** is in the **second-best bonus category 2,** the employee receives **5.00 Euros.**

If the employee takes a time-out, he or she receives 25 Cents.

Performance rating scale 1	Performance rating scale 2	Performance rating scale 3
Bonus category 1 --> 7.00 Euros	Bonus category 1 --> 7.00 Euros	Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros	Bonus category 2 --> 5.00 Euros	Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros	Bonus category 3 --> 3.00 Euros	Bonus category 3 --> 3.00 Euros
Bonus category 4 --> 0.00 Euros		Bonus category 4 is not used ~~4 --> 0.00 Euros~~
+ 25 Cents per time-out taken	+ 25 Cents per time-out taken	+ 25 Cents per time-out taken

Dummy Treatment

No Dummy Treatment

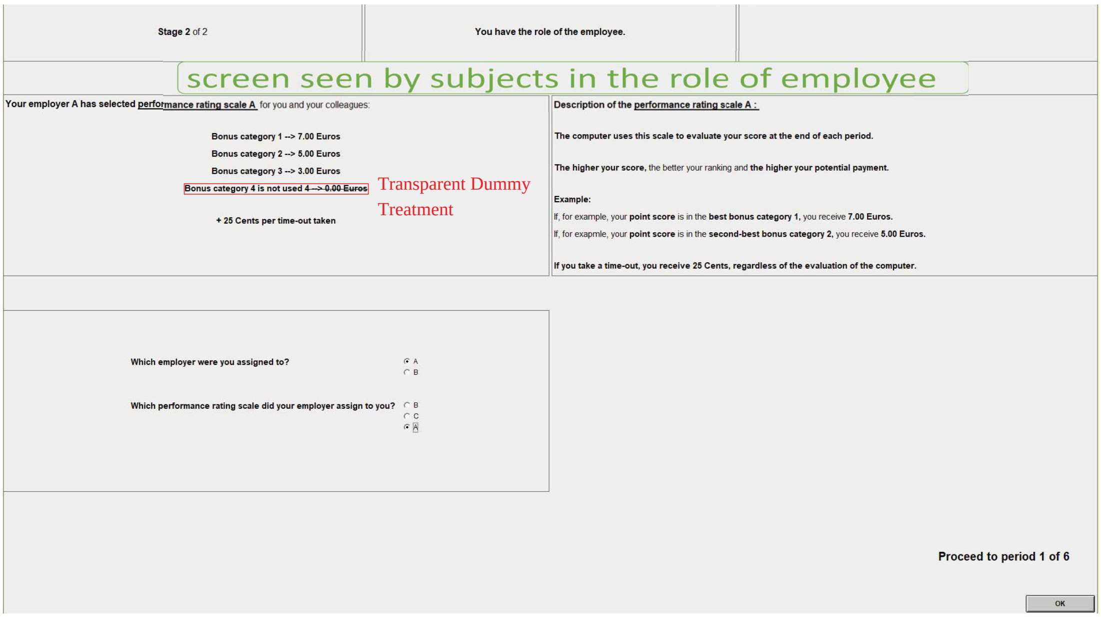
Transparent Dummy Treatment

Please assign **rating scale 1, 2 and 3 to employee groups A, B and C.**

For **rating scale 1,** please enter the **number 1.**
For **rating scale 2,** please enter the **number 2.**
For **rating scale 3,** please enter the **number 3.**

Employee group A	3
Employee group B	1
Employee group C	2

Figure 3.23: Scale Assignment Screen of Subjects in the Role of Employee



Stage 2 of 2

You have the role of the employee.

screen seen by subjects in the role of employee

Your employer A has selected performance rating scale A for you and your colleagues:

Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~

+ 25 Cents per time-out taken

Transparent Dummy Treatment

Description of the performance rating scale A :

The computer uses this scale to evaluate your score at the end of each period.

The higher your score, the better your ranking and **the higher your potential payment.**

Example:
If, for example, your **point score** is in the **best bonus category 1,** you receive **7.00 Euros.**
If, for exapmle, your **point score** is in the **second-best bonus category 2,** you receive **5.00 Euros.**

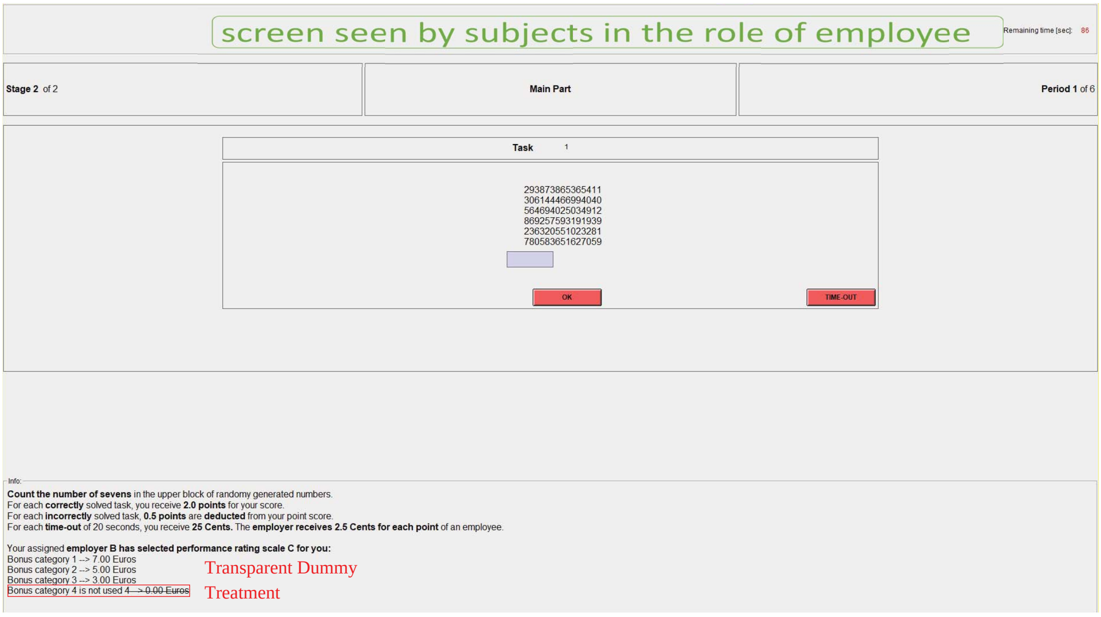
If you take a time-out, you receive 25 Cents, regardless of the evaluation of the computer.

Which employer were you assigned to? ● A ○ B

Which performance rating scale did your employer assign to you? ○ B ○ C ● A

Proceed to period 1 of 6

Figure 3.24: Task Screen of Subjects in the Role of Employee



screen seen by subjects in the role of employee

Remaining time [sec]: 86

Stage 2 of 2

Main Part

Period 1 of 6

Task 1

```
293873865365411
306144466994040
564694025034912
869257593191939
236320551023281
780583651627059
```

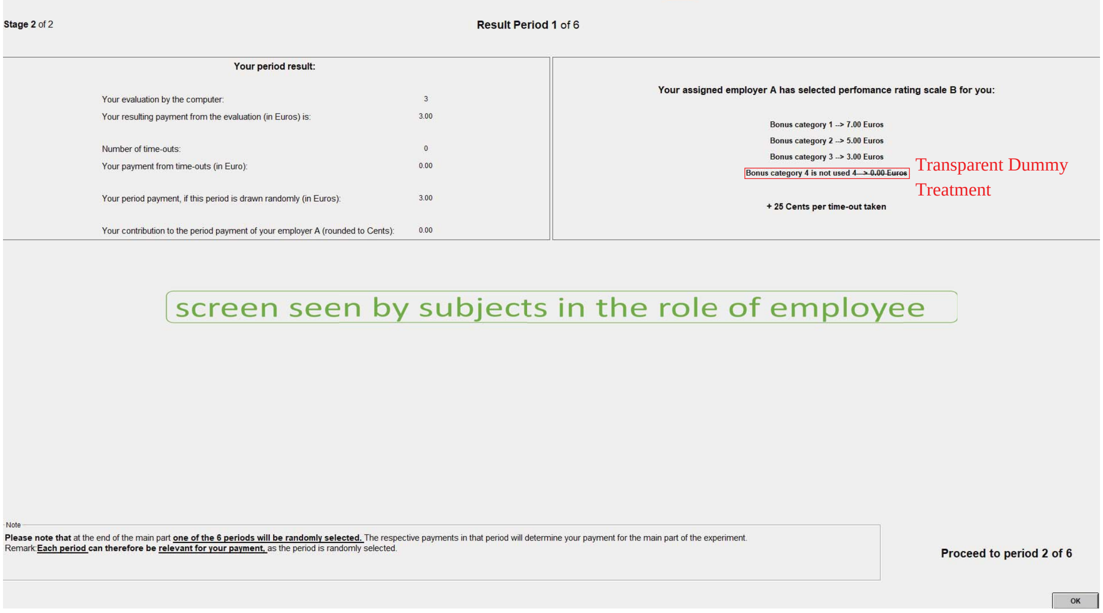
Info:

Count the number of sevens in the upper block of randomy generated numbers.
For each **correctly** solved task, you receive **2.0 points** for your score.
For each **incorrectly** solved task, **0.5 points** are **deducted** from your point score.
For each **time-out** of 20 seconds, you receive **25 Cents.** The **employer receives 2.5 Cents for each point** of an employee.

Your assigned **employer B has selected performance rating scale C for you:**
Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~

Transparent Dummy Treatment

Figure 3.25: Performance Feedback Screen of Subjects in the Role of Employee in Transparent Dummy Treatment



Stage 2 of 2

Result Period 1 of 6

Your period result:

Your evaluation by the computer:	3
Your resulting payment from the evaluation (in Euros) is:	3.00
Number of time-outs:	0
Your payment from time-outs (in Euro):	0.00
Your period payment, if this period is drawn randomly (in Euros):	3.00
Your contribution to the period payment of your employer A (rounded to Cents):	0.00

Your assigned employer A has selected perfomance rating scale B for you:

Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~

+ 25 Cents per time-out taken

Transparent Dummy Treatment

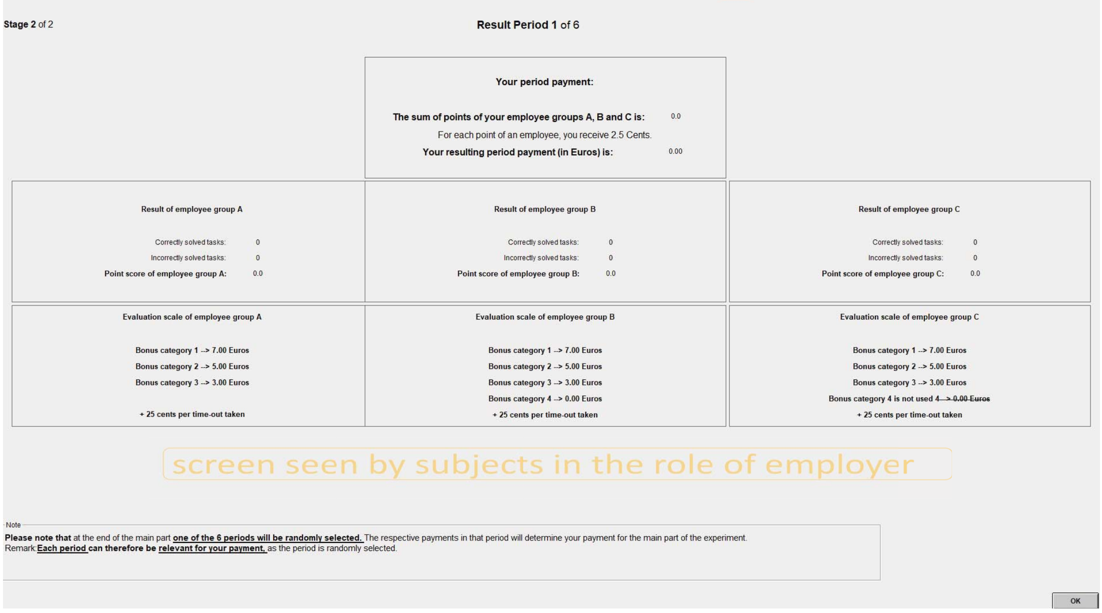
screen seen by subjects in the role of employee

Note

Please note that at the end of the main part **one of the 6 periods will be randomly selected.** The respective payments in that period will determine your payment for the main part of the experiment.
Remark:**Each period can therefore be relevant for your payment,** as the period is randomly selected.

Proceed to period 2 of 6

Figure 3.26: Performance Feedback Screen of Subjects in the Role of Employer



Stage 2 of 2

Result Period 1 of 6

Your period payment:

The sum of points of your employee groups A, B and C is: 0.0
For each point of an employee, you receive 2.5 Cents.
Your resulting period payment (in Euros) is: 0.00

Result of employee group A		**Result of employee group B**		**Result of employee group C**	
Correctly solved tasks:	0	Correctly solved tasks:	0	Correctly solved tasks:	0
Incorrectly solved tasks:	0	Incorrectly solved tasks:	0	Incorrectly solved tasks:	0
Point score of employee group A:	0.0	**Point score of employee group B:**	0.0	**Point score of employee group C:**	0.0

Evaluation scale of employee group A	**Evaluation scale of employee group B**	**Evaluation scale of employee group C**
Bonus category 1 --> 7.00 Euros	**Bonus category 1 --> 7.00 Euros**	**Bonus category 1 --> 7.00 Euros**
Bonus category 2 --> 5.00 Euros	**Bonus category 2 --> 5.00 Euros**	**Bonus category 2 --> 5.00 Euros**
Bonus category 3 --> 3.00 Euros	**Bonus category 3 --> 3.00 Euros**	**Bonus category 3 --> 3.00 Euros**
	Bonus category 4 --> 0.00 Euros	**Bonus category 4 is not used ~~4 --> 0.00 Euros~~**
+ 25 cents per time-out taken	**+ 25 cents per time-out taken**	**+ 25 cents per time-out taken**

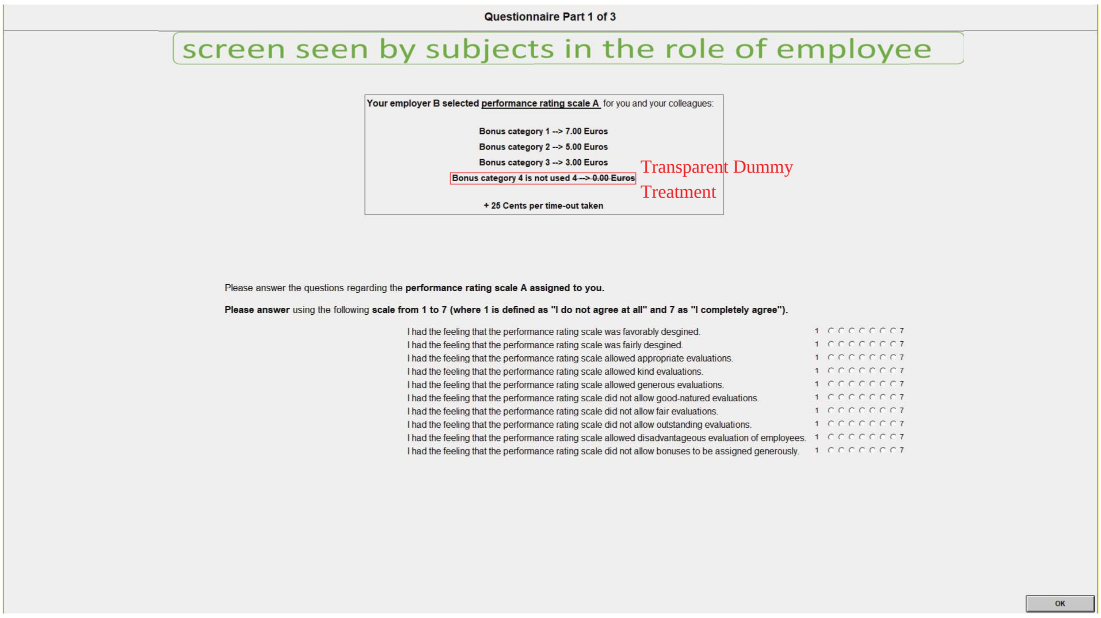
screen seen by subjects in the role of employer

Note

Please note that at the end of the main part **one of the 6 periods will be randomly selected.** The respective payments in that period will determine your payment for the main part of the experiment.
Remark:**Each period can therefore be relevant for your payment,** as the period is randomly selected.

3.C.4 Questionnaire

Figure 3.27: Evaluation of the Kindness of the Rating Scale Screen of Subjects in the Role of Employee in Transparent Dummy Treatment



Questionnaire Part 1 of 3

screen seen by subjects in the role of employee

Your employer B selected **performance rating scale A** for you and your colleagues:

Bonus category 1 --> 7.00 Euros

Bonus category 2 --> 5.00 Euros

Bonus category 3 --> 3.00 Euros

Bonus category 4 is not used ~~4 --> 0.00 Euros~~

+ 25 Cents per time-out taken

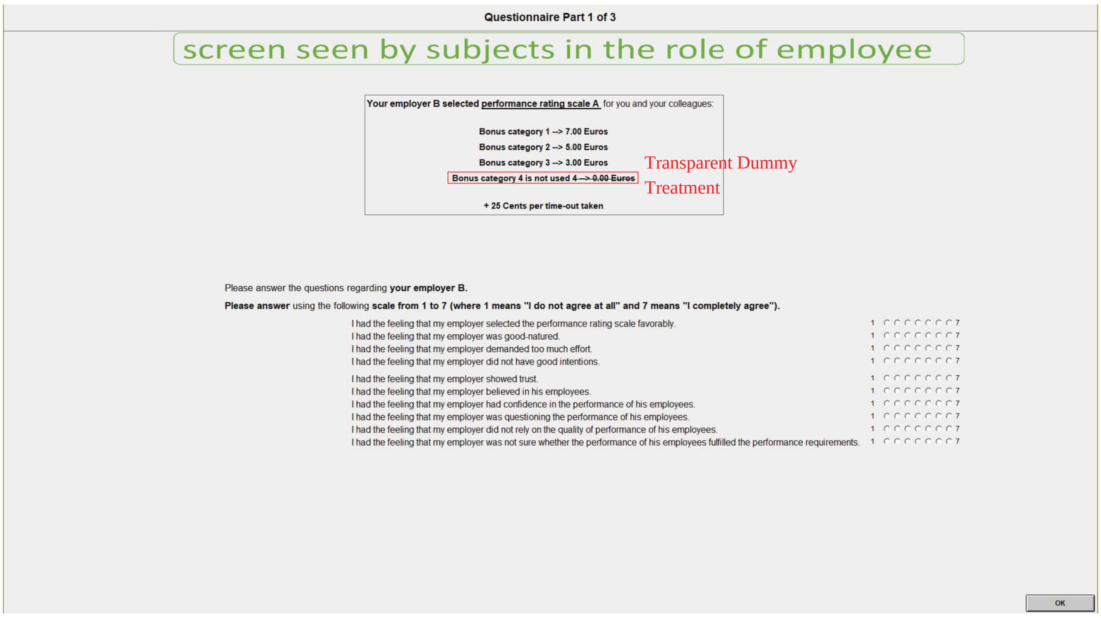
Transparent Dummy Treatment

Please answer the questions regarding the **performance rating scale A assigned to you.**

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I had the feeling that the performance rating scale was favorably desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale was fairly desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale allowed appropriate evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale allowed kind evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale allowed generous evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale did not allow good-natured evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale did not allow fair evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale did not allow outstanding evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale allowed disadvantageous evaluation of employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that the performance rating scale did not allow bonuses to be assigned generously. 1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.28: Evaluation of the Kindness of the Employer Screen of Subjects in the Role of Employee in Transparent Dummy Treatment



Questionnaire Part 1 of 3

screen seen by subjects in the role of employee

Your employer B selected **performance rating scale A** for you and your colleagues:

Bonus category 1 --> 7.00 Euros

Bonus category 2 --> 5.00 Euros

Bonus category 3 --> 3.00 Euros

Bonus category 4 is not used ~~4 --> 0.00 Euros~~

+ 25 Cents per time-out taken

Transparent Dummy Treatment

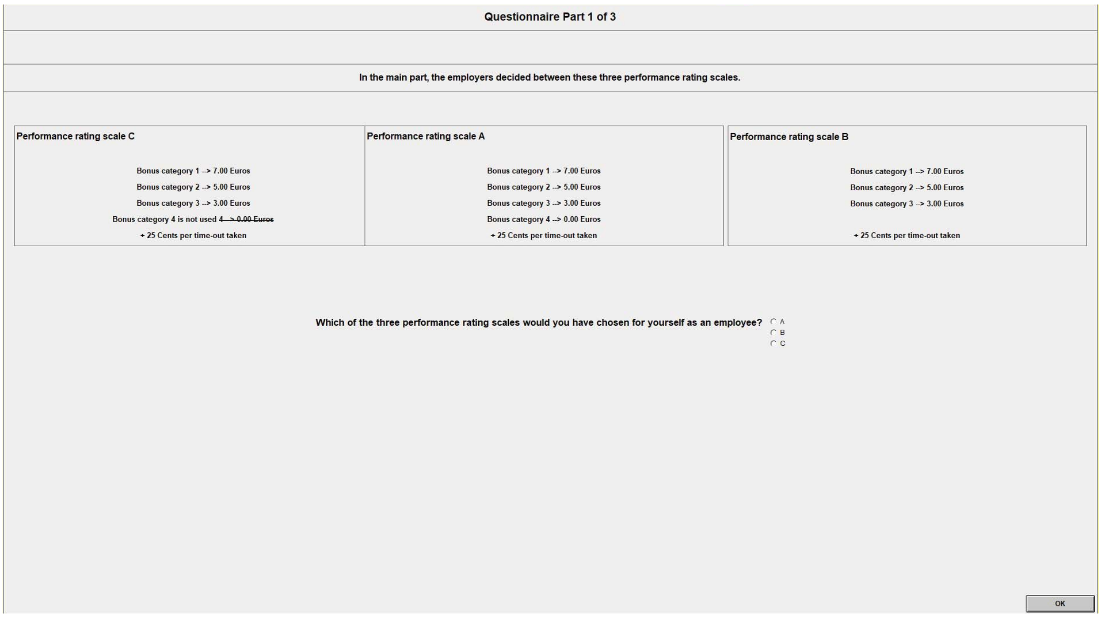
Please answer the questions regarding **your employer B.**

Please answer using the following **scale from 1 to 7 (where 1 means "I do not agree at all" and 7 means "I completely agree").**

I had the feeling that my employer selected the performance rating scale favorably. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer was good-natured. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer demanded too much effort. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer did not have good intentions. 1 ○ ○ ○ ○ ○ ○ ○ 7

I had the feeling that my employer showed trust. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer believed in his employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer had confidence in the performance of his employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer was questioning the performance of his employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer did not rely on the quality of performance of his employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
I had the feeling that my employer was not sure whether the performance of his employees fulfilled the performance requirements. 1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.29: Question on Scale Choice



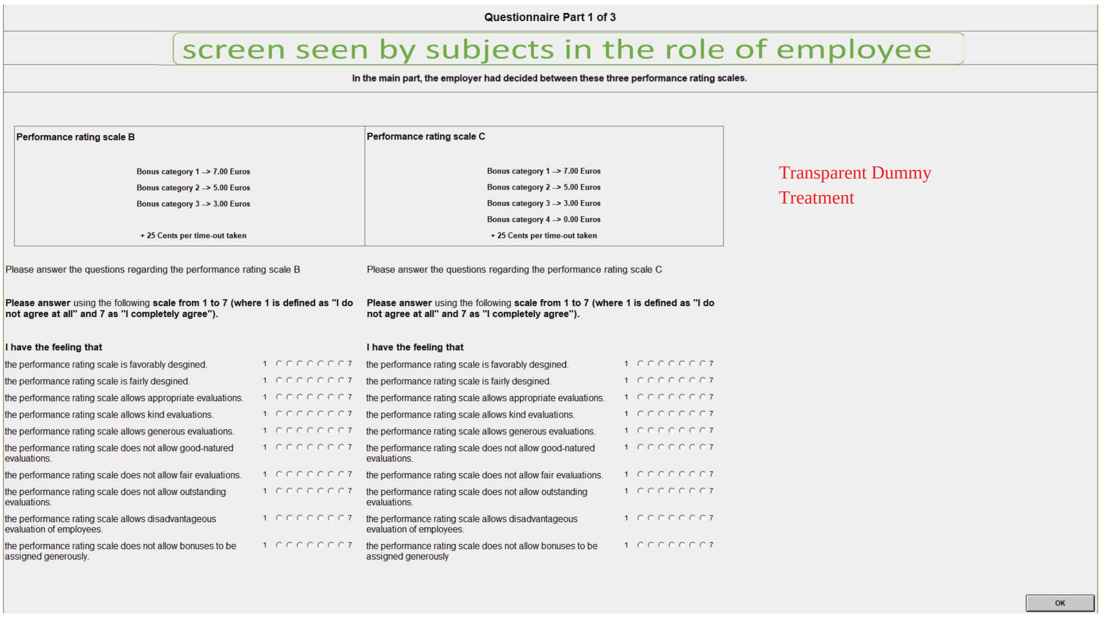
Questionnaire Part 1 of 3

In the main part, the employers decided between these three performance rating scales.

Performance rating scale C	Performance rating scale A	Performance rating scale B
Bonus category 1 --> 7.00 Euros	Bonus category 1 --> 7.00 Euros	Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros	Bonus category 2 --> 5.00 Euros	Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros	Bonus category 3 --> 3.00 Euros	Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~	Bonus category 4 --> 0.00 Euros	
+ 25 Cents per time-out taken	+ 25 Cents per time-out taken	+ 25 Cents per time-out taken

Which of the three performance rating scales would you have chosen for yourself as an employee? ○ A ○ B ○ C

Figure 3.30: Evaluation of the Kindness of Other Rating Scales Screen of Subjects in the Role of Employee in Transparent Dummy Treatment



Questionnaire Part 1 of 3

screen seen by subjects in the role of employee

In the main part, the employer had decided between these three performance rating scales.

Performance rating scale B	Performance rating scale C
Bonus category 1 --> 7.00 Euros	Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros	Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros	Bonus category 3 --> 3.00 Euros
	Bonus category 4 --> 0.00 Euros
+ 25 Cents per time-out taken	+ 25 Cents per time-out taken

Transparent Dummy Treatment

Please answer the questions regarding the performance rating scale B

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I have the feeling that

the performance rating scale is favorably desgined.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale is fairly desgined.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows appropriate evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows kind evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows generous evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow good-natured evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow fair evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow outstanding evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows disadvantageous evaluation of employees.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow bonuses to be assigned generously.	1 ○ ○ ○ ○ ○ ○ ○ 7

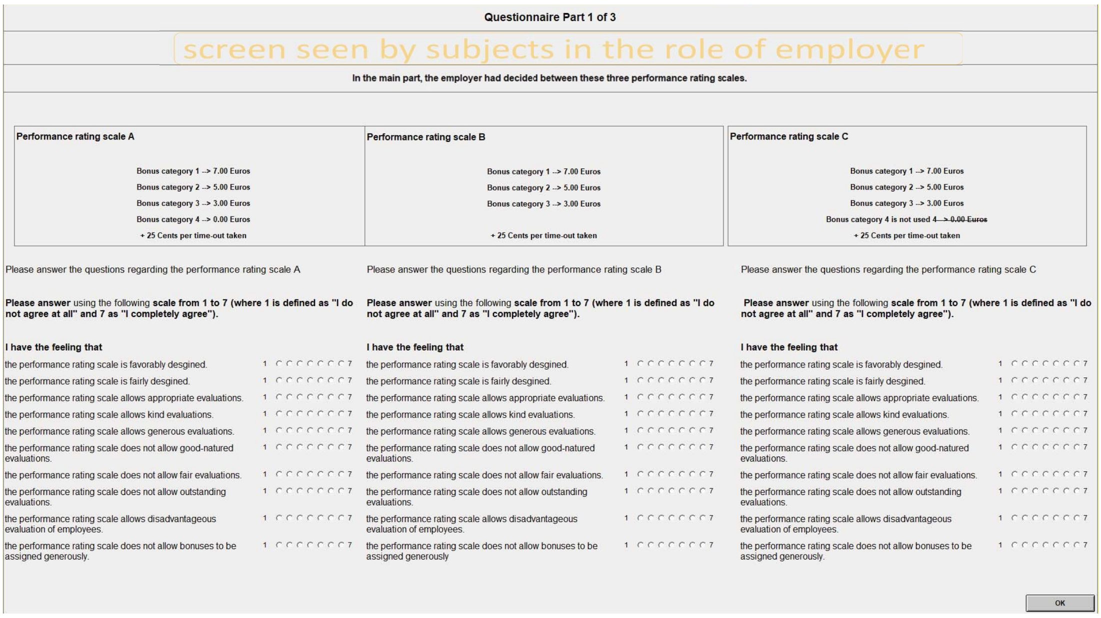
Please answer the questions regarding the performance rating scale C

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I have the feeling that

the performance rating scale is favorably desgined.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale is fairly desgined.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows appropriate evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows kind evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows generous evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow good-natured evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow fair evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow outstanding evaluations.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows disadvantageous evaluation of employees.	1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow bonuses to be assigned generously	1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.31: Evaluation of the Kindness of Other Rating Scales Screen of Subjects in the Role of Employer



Questionnaire Part 1 of 3

screen seen by subjects in the role of employer

In the main part, the employer had decided between these three performance rating scales.

Performance rating scale A

Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 --> 0.00 Euros
+ 25 Cents per time-out taken

Performance rating scale B

Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
+ 25 Cents per time-out taken

Performance rating scale C

Bonus category 1 --> 7.00 Euros
Bonus category 2 --> 5.00 Euros
Bonus category 3 --> 3.00 Euros
Bonus category 4 is not used ~~4 --> 0.00 Euros~~
+ 25 Cents per time-out taken

Please answer the questions regarding the performance rating scale A

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I have the feeling that

the performance rating scale is favorably desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale is fairly desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows appropriate evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows kind evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows generous evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow good-natured evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow fair evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow outstanding evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows disadvantageous evaluation of employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow bonuses to be assigned generously. 1 ○ ○ ○ ○ ○ ○ ○ 7

Please answer the questions regarding the performance rating scale B

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I have the feeling that

the performance rating scale is favorably desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale is fairly desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows appropriate evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows kind evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows generous evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow good-natured evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow fair evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow outstanding evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows disadvantageous evaluation of employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow bonuses to be assigned generously 1 ○ ○ ○ ○ ○ ○ ○ 7

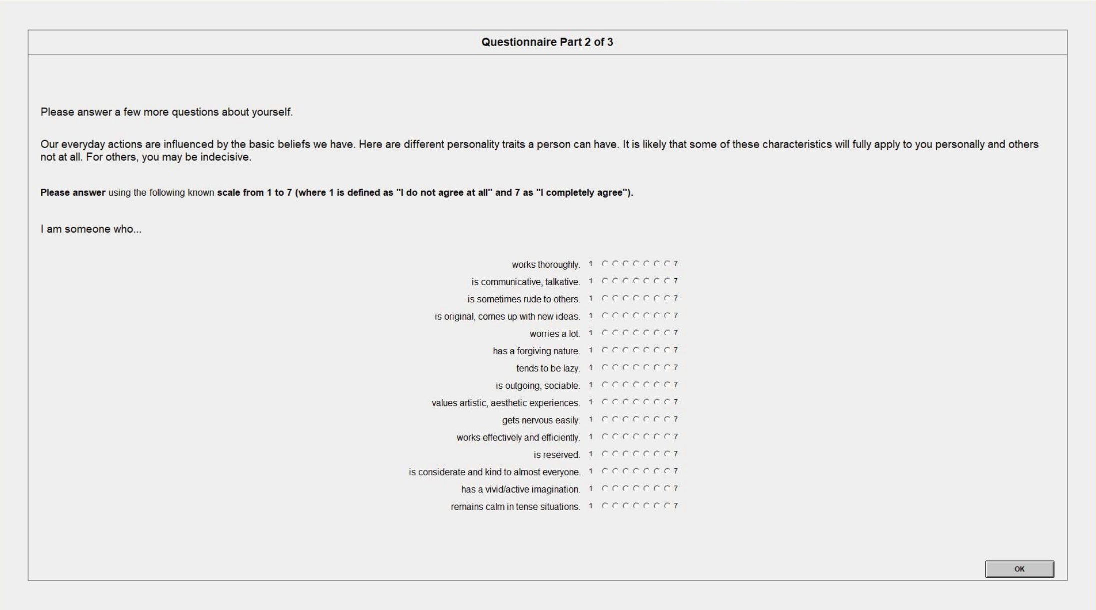
Please answer the questions regarding the performance rating scale C

Please answer using the following **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I have the feeling that

the performance rating scale is favorably desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale is fairly desgined. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows appropriate evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows kind evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows generous evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow good-natured evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow fair evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow outstanding evaluations. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale allows disadvantageous evaluation of employees. 1 ○ ○ ○ ○ ○ ○ ○ 7
the performance rating scale does not allow bonuses to be assigned generously. 1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.32: Questions on Big Five Personality Traits



Questionnaire Part 2 of 3

Please answer a few more questions about yourself.

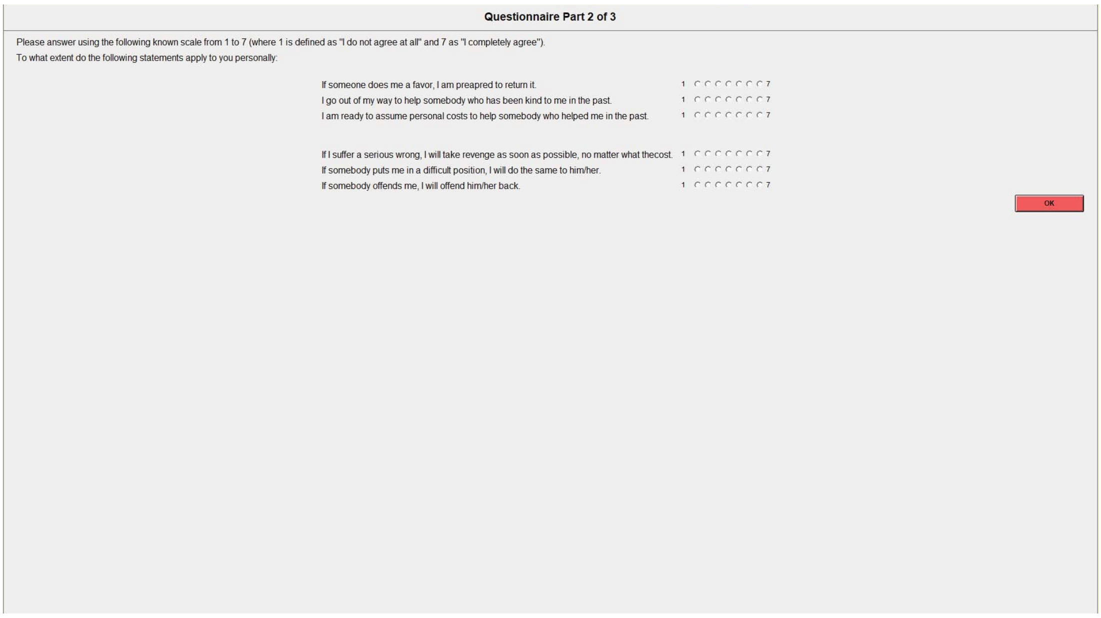
Our everyday actions are influenced by the basic beliefs we have. Here are different personality traits a person can have. It is likely that some of these characteristics will fully apply to you personally and others not at all. For others, you may be indecisive.

Please answer using the following known **scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").**

I am someone who...

works thoroughly. 1 ○ ○ ○ ○ ○ ○ ○ 7
is communicative, talkative. 1 ○ ○ ○ ○ ○ ○ ○ 7
is sometimes rude to others. 1 ○ ○ ○ ○ ○ ○ ○ 7
is original, comes up with new ideas. 1 ○ ○ ○ ○ ○ ○ ○ 7
worries a lot. 1 ○ ○ ○ ○ ○ ○ ○ 7
has a forgiving nature. 1 ○ ○ ○ ○ ○ ○ ○ 7
tends to be lazy. 1 ○ ○ ○ ○ ○ ○ ○ 7
is outgoing, sociable. 1 ○ ○ ○ ○ ○ ○ ○ 7
values artistic, aesthetic experiences. 1 ○ ○ ○ ○ ○ ○ ○ 7
gets nervous easily. 1 ○ ○ ○ ○ ○ ○ ○ 7
works effectively and efficiently. 1 ○ ○ ○ ○ ○ ○ ○ 7
is reserved. 1 ○ ○ ○ ○ ○ ○ ○ 7
is considerate and kind to almost everyone. 1 ○ ○ ○ ○ ○ ○ ○ 7
has a vivid/active imagination. 1 ○ ○ ○ ○ ○ ○ ○ 7
remains calm in tense situations. 1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.33: Questions on Positive and Negative Reciprocity



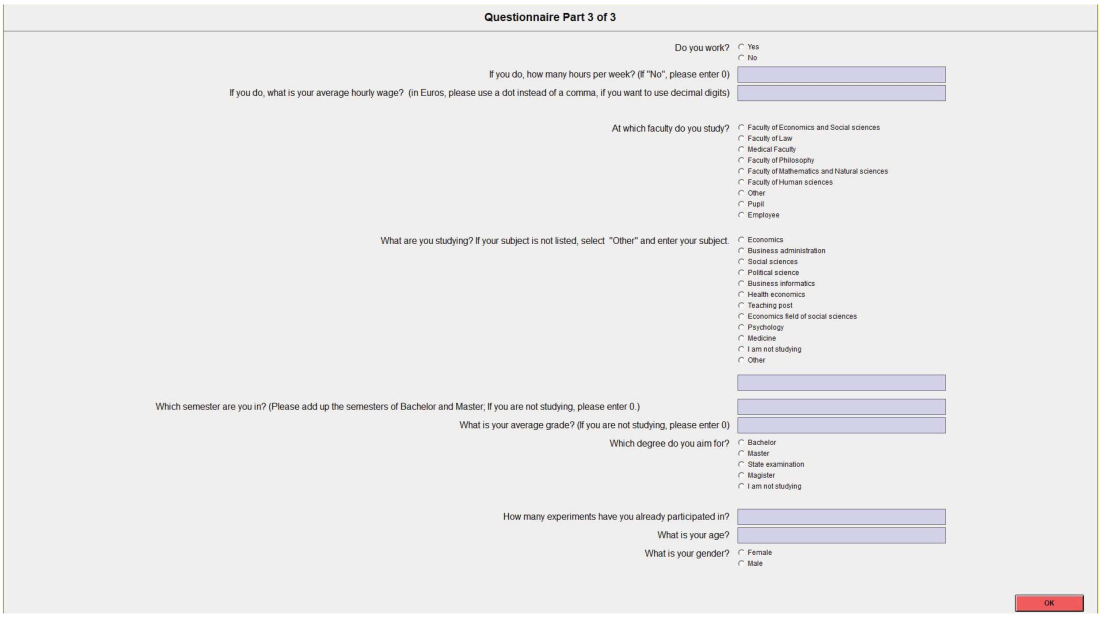
Questionnaire Part 2 of 3

Please answer using the following known scale from 1 to 7 (where 1 is defined as "I do not agree at all" and 7 as "I completely agree").
To what extent do the following statements apply to you personally:

If someone does me a favor, I am preapred to return it. 1 ○ ○ ○ ○ ○ ○ ○ 7
I go out of my way to help somebody who has been kind to me in the past. 1 ○ ○ ○ ○ ○ ○ ○ 7
I am ready to assume personal costs to help somebody who helped me in the past. 1 ○ ○ ○ ○ ○ ○ ○ 7

If I suffer a serious wrong, I will take revenge as soon as possible, no matter what thecost. 1 ○ ○ ○ ○ ○ ○ ○ 7
If somebody puts me in a difficult position, I will do the same to him/her. 1 ○ ○ ○ ○ ○ ○ ○ 7
If somebody offends me, I will offend him/her back. 1 ○ ○ ○ ○ ○ ○ ○ 7

Figure 3.34: Demographic Questions



Questionnaire Part 3 of 3

Do you work? ○ Yes ○ No

If you do, how many hours per week? (If "No", please enter 0)

If you do, what is your average hourly wage? (in Euros, please use a dot instead of a comma, if you want to use decimal digits)

At which faculty do you study?
○ Faculty of Economics and Social sciences
○ Faculty of Law
○ Medical Faculty
○ Faculty of Philosophy
○ Faculty of Mathematics and Natural sciences
○ Faculty of Human sciences
○ Other
○ Pupil
○ Employee

What are you studying? If your subject is not listed, select "Other" and enter your subject.
○ Economics
○ Business administration
○ Social sciences
○ Political science
○ Business informatics
○ Health economics
○ Teaching post
○ Economics field of social sciences
○ Psychology
○ Medicine
○ I am not studying
○ Other

Which semester are you in? (Please add up the semesters of Bachelor and Master; If you are not studying, please enter 0.)

What is your average grade? (If you are not studying, please enter 0)

Which degree do you aim for?
○ Bachelor
○ Master
○ State examination
○ Magister
○ I am not studying

How many experiments have you already participated in?

What is your age?

What is your gender? ○ Female ○ Male

Chapter 4

How to Induce Sustainable Customer Buying: The Roles of Sustainability Messages and Price Discounts

Large amounts of food are wasted globally, which contributes to the climate crisis and decreases retailers' profits. To reduce food waste at retailers, we analyze how sustainability messages and price discounts can increase sales of earlier expiring items. Research on food waste reduction has emphasized price discounts. We analyze an alternative or supplemental approach for food waste reduction that uses sustainability messages to incentivize purchases of earlier expiring items. We conducted an online experiment where subjects chose between earlier expiring and longer lasting items. As non-monetary incentive to buy earlier expiring items, we show a sustainability message. As monetary incentive, we offer different price discounts. We find that 1) displaying a sustainability message induces more subjects to buy earlier expiring items; 2) the higher the price discounts, the more subjects buy earlier expiring items; 3) some subjects do not change their behavior, or crowd out when receiving price discounts; 4) a sustainability message induces a) more subjects to buy earlier expiring items independent of whether they are discounted or not, b) less subjects to buy earlier expiring items only when they are discounted, and c) more subjects who crowd out when receiving a marginal discount to switch back to buying earlier expiring items when receiving higher discounts. Thus, retailers can incentivize purchases of earlier expiring items not only by price discounts, but also by sustainability messages. Understanding how different customer types respond to either incentive can help retailers to offer each incentive only to customers who most likely respond with buying expiring items.

4.1 Introduction

About one third of the world's food production and about 890 billion Euros are wasted globally every year (Shukla et al. 2019). A considerable portion of food waste occurs in grocery retailing, when perished products are discarded due to spoilage. This is partly driven by customer expectations to find fresh products in fully stocked stores (Noleppa and Cartsburg 2015). If there are multiple items of a product with different expiration dates on the shelf, many customers have strong preferences to buy longer lasting items (Broekmeulen and Van Donselaar 2009, Broekmeulen and Bakx 2010). Estimates how often customers buy longer lasting items as opposed to earlier expiring items of perishable products vary between product categories and range from 10 to 66% (Bastiaansen 2019). This buying behavior can cause food waste of earlier expiring items in grocery stores. Incentivizing customers to purchase earlier expiring items and thus more sustainable buying behavior would reduce such waste.

Customers care about (food) waste and value ethical behavior (Trudel and Cotte 2009, Commission 2017). However, when it comes to actual purchase decisions, their actions are not necessarily consistent with their attitudes (Auger and Devinney 2007, Commission 2008, Devinney et al. 2010).

We analyze how more sustainable buying behavior can be induced and study a non-monetary and a monetary mechanism to incentivize people to buy earlier expiring items. As non-monetary mechanism, we analyze sustainability messages. If customers are conscious about food waste, but tend to buy longer lasting instead of earlier expiring items, it might help to make them more aware of the consequences of their buying decisions. We provide a message where we state that buying earlier expiring items can save them from being discarded at the retailer. This might increase customers' utility from buying earlier expiring items and increase sales of these items. As monetary mechanism, we analyze price discounts. By considering that earlier expiring items will not last as long as longer lasting ones, customers might be concerned that the items spoil before consumption. To compensate for the expected higher monetary loss, retailers can offer earlier expiring items at a price discount to induce purchases of the respective items.

It is not clear whether non-monetary mechanisms will be successful to incentivize people to buy earlier expiring items and reduce food waste, and how they interact with monetary mechanisms. Since sustainability messages do not have any monetary effect, there is no monetary incentive for customers to buy earlier expiring instead of longer lasting items. Price discounts have a monetary effect, but it is not clear in advance whether any price discount would suffice, or whether a certain threshold has to be reached for a discount to become effective. Moreover, the two mechanisms could cancel each other out when individuals show a crowding out effect, as has been observed in studies related to ours (Frey and Oberholzer-Gee 1997, Gneezy and Rustichini 2000, Bénabou and Tirole 2006). Accordingly, the interaction effect of using price discounts with sustainability messages are not clear-cut.

We designed an experiment where subjects made buying decisions for five identical perishable products. For each product, subjects had to choose between an earlier expiring and a longer lasting item. There was one decision round for each product, that is, five rounds in total. In the first decision round, both items of the first product were sold at the same price. In consecutive decision rounds, the earlier expiring item was discounted. We used a Baseline treatment without a sustainability message and a Sustainability Message treatment where a sustainability message was shown.

We find that more subjects buy the earlier expiring item in the Sustainability Message treatment than in the Baseline treatment. Not surprisingly, the higher the price discount, the more subjects buy the earlier expiring item. Consequently, a retailer could consider showing sustainability messages and offering earlier expiring items at a price discount to induce customers to purchase these items. However, this would be costly, because some customers would buy earlier expiring items without price discounts and hence receive unnecessary price discounts. Others would switch from buying the earlier expiring item to buying the longer lasting item when a price discount is introduced. A retailer who understands how different customers respond to sustainability messages and price discounts can leverage this information with customized approaches.

We identify four types of customers. Type AE (Always Expiring) customers always buy earlier expiring items, regardless of whether they are discounted or not. Type NE (Never Expiring) customers never buy earlier expiring items, whether they are discounted or not. Type PS (Price Sensitive) customers buy longer lasting items without price discounts, but switch to earlier expiring items when price discounts are sufficiently high. Type CO (Crowding Out) customers buy earlier expiring items without price discounts, but switch to longer lasting items when a marginal price discount is offered. Some of these Type CO customers switch back to buying earlier expiring items when price discounts are sufficiently high (crowding in). In our experiments, the majority of subjects can be classified as one of these types.

When comparing buying decisions of the Sustainability Message treatment with those of the Baseline treatment, we find that the share of Type AE customers is larger and the share of Type PS customers is smaller when a sustainability message is shown. Moreover, we observe more crowding in when a sustainability message is shown.

Our results suggest that companies who know the types of their customers can efficiently incentivize customers to purchase earlier expiring items by a targeted promotion policy: Provide sustainability messages to all customers and price discounts to price sensitive customers only.

The contribution of our work is to improve the understanding of the customer buying process for perishable products. We analyze how sustainability messages and price discounts incentivize customers to buy earlier expiring items to reduce food waste. We develop a Behavioral Model that describes individual and aggregate buying behavior in response to sustainability messages and price discounts. We design an experiment to test the potential effects of both measures on individual and aggregate buying decisions. Four types of buying behavior can be distinguished and we can assign most subjects to one of them. We identify for which of these customer types sustainability messages and price discounts are effective and develop managerial recommendations.

4.2 Behavioral Model

In this section, we develop a Behavioral Model to describe the potential effect of providing price discounts, sustainability messages, or both on customer buying decisions.

4.2.1 Setting

We consider an expected utility maximizing customer who must choose between two items of a product, one with an earlier expiration date than the other. For notational convenience, we refer to the item with the earlier expiration date as the "expiring item" and to the item with the longer expiration date as the "lasting item". At the purchase time, customers do not know the specific consumption date, but they know the distribution of the consumption dates. If a purchased item expires before the consumption date it must be re-purchased.

4.2.2 Monetary Preferences

If both items have the same price, an expected profit maximizing customer prefers the lasting item, because it is less likely to expire before consumption than the expiring item. This decision is optimal for the customer but results in higher expected food waste at the retailer.

Price discounts of the expiring item can induce purchases of the expiring item. They reduce the expected monetary costs associated with choosing the expiring item as opposed to the lasting item and compensate customers for the increased risk of having to re-purchase the product.

Research showed that price discounts are a powerful tool to induce purchases of expiring items. For example, Smith and Agrawal (2017) and Chua et al. (2017) model markdown pricing decisions in a retail setting. Smith and Agrawal (2017) analyze how prices can be decreased over time to reduce leftover inventory. Chua et al. (2017) demonstrate that price discounts can redirect customers from lasting to expiring products to avoid leftover inventory that would otherwise be wasted.

The above suggests that discounting expiring items increases their purchases. However, it is not clear how price discounts compare to and interact with sustainability messages in affecting buying decisions. We address this issue in our research.

4.2.3 Food Saving Preferences

Research in behavioral economics and behavioral management suggests that sustainability messages might also incentivize purchases of expiring items. They aim at increasing the non-monetary value of expiring items and thereby compensating customers for the increased spoilage risk.

Studies find that individuals make decisions considering not only their own outcome but also the well-being and costs incurred to others: For example, people have concerns for fairness and reciprocity and behave altruistically (see for an overview Fehr and Schmidt 1999, 2003). These social preferences explain cooperation, charitable donations, and the voluntary provision of public goods (Fehr and Fischbacher 2002, 2003) and are also relevant for inventory decisions (see for example Bolton et al. 2012, Becker-Peth et al. 2013, Papier and Thonemann 2021). Customers might care about food waste at the retailer and its consequences for the environment. Buying an expiring item as opposed to a lasting item reduces expected food waste at the retailer. Compared to an equally priced lasting item, it increases customers' expected costs, but can trigger non-monetary customer utility. If this utility exceeds the disutility of higher expected costs, individuals might still choose the expiring item without price discounts.

Sustainability messages emphasize saving food at the retailer and might increase the customers' utility from food saving preferences. This suggests that sustainability messages increase purchases of expiring items. Research has not analyzed how addressing non-monetary food waste saving preferences affect customer buying behavior. We contribute to filling this gap with our research. We note that purchasing expiring items increases the probability that the respective items expire at home. However, while retailers must discard products after the expiration date, customers can still eat most of them since they are still "safe and wholesome" (Food Safety and Inspection Service U.S. Department of Agriculture 2019). In this case, purchasing expiring items can contribute to an overall reduction of food waste. If items are not edible after the expiration date, food waste might be shifted from the retailer to the customer. However, individuals can prevent items from perishing by refrigerating them or adjusting their meal plans, which mitigates the potential of food waste at home, and contributes to an overall reduction of food waste. In any case, the potential waste shift needs to be carefully considered when interpreting our results.

4.2.4 Crowding out

The literature on the interaction of intrinsic and extrinsic motivation finds that monetary incentives can trump intrinsic motivation (see for example Frey and Oberholzer-Gee 1997, Bénabou and Tirole 2003). Frey and Oberholzer-Gee (1997) conclude that in the presence of crowding out price incentives are less effective than standard theory suggests. Gneezy and Rustichini (2000) find that the effectiveness of incentives depends on the size of the incentives. They observe that small incentives crowd out pro-social behavior but the effect vanishes if incentives are sufficiently high. Potential reasons for crowding out in grocery shopping settings could be that individuals interpret price discounts as a signal that choosing lasting items is the social norm (Gneezy et al. 2011) or that expiring items bear further risks that must be compensated (for a related argument see Frey and Oberholzer-Gee 1997).

The effectiveness of price discounts might depend on individuals' preferences for avoiding food waste and the crowding out effect. Price discounts might be less effective for customers who care about food waste at the retailer than for customers who only care about money. If the effect observed by Gneezy and Rustichini (2000) is also present in our setting, then the negative effect of price discounts should only occur for marginal price discounts.

4.2.5 Model

Consider a customer who evaluates a lasting item L and an expiring item E of a given product. The customer's utility consists of the utility from consuming the product, which is reduced by the disutility from paying the purchase price and, if the item expires before consumption, having to pay the price again. The utility of the lasting item L is

$$U_L(r) = U_C - U_R r - U_R r I_T(T^C > T_L^E). \tag{4.1}$$

$U_C \geq 0$ denotes the utility from consuming the product, $U_R \geq 0$ the (dis)utility from paying the purchase price, and $r > 0$ the regular purchase price. The indicator function $I_T()$ equals one if the consumption date T^C is after the expiration date of the lasting item T_L^E. If $I_T()$ is equal to one, the product must be re-purchased at the regular price r. The utility of the expiring item E is

$$U_E(r,\alpha,M) = U_C - U_R r(1-\alpha) - U_R r I_T(T^C > T_E^E) + U_S + U_M I_M - U_O I_O(\alpha > 0). \tag{4.2}$$

The first three terms are similar to those of the lasting item, except that the expiring item may be priced at a discount $0 \leq \alpha < 1$, and T_E^E denotes the expiration date of the expiring item. $U_S \geq 0$ is the utility gained from potentially saving the expiring item from being discarded at the retailer. $U_M \geq 0$ represents additional utility from potentially saving the expiring item when the retailer displays a sustainability message M – in which case the variable $I_M() = 1$ – as it increases the customers' focus on their sustainability attitude. $U_O \geq 0$ is the utility loss customers might experience if a discount is given ($\alpha > 0$), which only applies if the indicator function $I_O()$ equals one.

Customers prefer the expiring item if the expected utility of choosing it is greater than the expected utility of choosing the lasting item. The expected utility difference $\triangle_{EL}EU$ between choosing the expiring and the lasting item is

$$\triangle_{EL}EU = EU_E - EU_L = U_0 + U_M I_M + U_R r\alpha - U_O I_O, \tag{4.3}$$

$$where\ U_0 = U_S - U_R r(P_E - P_L)$$

denotes the expected utility difference if neither a discount nor a sustainability message is provided. We denote the probability that the lasting item expires before the consumption date by P_L and the probability that the expiring item expires before the consumption date by P_E, with $P_L < P_E$. The probability difference $(P_E - P_L) > 0$ resembles the increased spoilage risk of the expiring item. U_0 is positive if an individual's utility for saving food U_S compensates their expected costs associated with choosing the expiring item $U_R r(P_E - P_L)$.

4.2.6 Hypotheses Development

We develop hypotheses based on the expected utility difference $\triangle_{EL}EU$ (Equation (4.3)) of the previous subsection. The higher the expected utility difference $\triangle_{EL}EU$, the higher is the probability that customers choose the expiring item and hence the higher the share of customers choosing the expiring item. We compare customers who receive a sustainability message to those who do not. When receiving a sustainability message ($I_M = 1$), customers put more attention to their sustainability attitude and receive additional expected utility from potentially saving the expiring item. This additional expected utility is captured by U_M in Equation (4.3), i.e. $\triangle_{I_M \in \{0,1\}} \triangle_{EL}EU = U_M \geq 0$, and we hypothesize:

Hypothesis 1: *For a given price discount, the share of individuals buying the expiring item is larger with than without a sustainability message.*

Price discounts provide a monetary compensation for the disutility from purchasing an expiring item as opposed to a lasting one. A price discount reduces the disutility from purchasing an item (U_R) by the price discount (α).

The resulting marginal utility is greater or equal to zero, i.e. $\frac{\partial \triangle_{EL}EU}{\partial \alpha} = U_R r \geq 0$, and we hypothesize:

Hypothesis 2: *a) Without and b) with a sustainability message and for strictly positive price discounts ($\alpha > 0$), the share of individuals buying the expiring item increases in the price discount.*

If an expiring item is initially not discounted ($\alpha = I_O = 0$) and a marginal price discount is introduced, we obtain a marginal change in utility of $\lim_{\alpha \to 0} \frac{\partial \triangle_{EL}EU}{\partial \alpha} = U_R r - U_O$. The price discount increases the utility of the discounted expiring item compared to the regularly-priced expiring item by $U_R r \alpha$ and the crowding out effect reduces it by U_O. Depending on the values of U_R, r, α, and U_O, the expected utility of the expiring item and the predicted share of individuals buying it can increase or decrease when a price discount is introduced. Thus, we cannot state a hypothesis on the effect of introducing a price discount on the share of individuals buying the expiring item. A decrease or increase of the share of subjects buying the expiring item from no discount to a marginal discount would indicate that overall a crowding out effect outweighs or succumbs to a price discount effect.

To obtain an indication whether a crowding out effect exists in our setting, we analyze whether individuals who buy the expiring item at the regular price switch to buying the lasting item when they are offered a marginal price discount. A decrease in the share of these individuals would indicate a stronger crowding out effect, but an increase would not reject the existence of a crowding out effect.

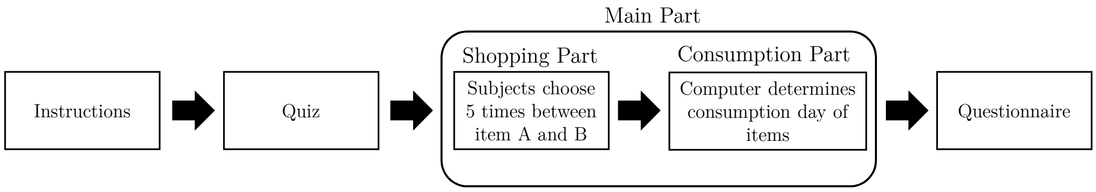
4.3 Experimental Design

Overview We designed an experiment that simulates a typical situation that customers face when buying perishable products. We present the experiment design, protocol, and subject pool in the following subsections.

Before running the main experiment, we ran the same experiment with a smaller number of subjects as a pilot. Based on the pilot, we determined the sample size for our main experiment using a power analysis. Please see Appendix 4.B.3 for the results of the pilot experiment, which are consistent with the results of our main experiment.

Experimental Details We implemented a grocery shopping setting that comprised two successive parts: Shopping and Consumption. In the Shopping part, subjects chose between an expiring item ”A” and a lasting item ”B” for five perishable products. The Consumption part spanned a hypothetical period of 10 days. The expiring item expired after 8 days, the lasting item expired after 9 days. Subjects did not know when an item would be consumed in the Consumption part. However, they learned the consumption pattern for each product, which was the same for all products and followed a uniform distribution, that is, on

Figure 4.1: Experimental Procedure


Main Part
Shopping Part
Consumption Part
Instructions
Quiz
Subjects choose 5 times between item A and B
Computer determines consumption day of items
Questionnaire

each of the 10 days of the Consumption part, it was equally likely that an item was consumed. Figure 4.1 illustrates the sequence of the different parts of our experiment. Please refer to the Appendix 4.C for screenshots of the experiment.

The five decisions differed only in the discount given for the expiring item: In the first decision, the prices of the expiring and lasting item were both 10 experimental currency units (ECUs). In the second to fifth decision, the expiring item was sold at a discount of 1%, 5%, 15%, and 30%, respectively. We did not randomize the decision order to analyze how subjects react to the introduction and increase of price discounts and to control for individual characteristics.

Subjects had a budget of 100 ECUs at the beginning of the experiment. Each purchase reduced the budget by the price paid for an item. If the consumption day was after the expiration day of an item, the computer automatically bought the product again and the corresponding cost of 10 ECUs was deducted from the budget. After the experiment, subjects received the remaining budget as payout in addition to a show-up fee of $2.50. 11 ECUs were converted into $1.

Treatment Variation We ran two treatments, a Baseline treatment and a Sustainability Message treatment that differed only in the message that was shown in the Sustainability Message treatment, but not in the Baseline treatment. In the Sustainability Message treatment, we displayed the following sustainability message in the instructions, and on the decision screen: "If you choose the item with the shorter expiration date, you can potentially avoid that it is discarded and reduce the amount of food waste at the retailer. This is because a perishable item can no longer be sold when it reaches its expiration date at the retailer." In the quiz, we asked a true/false question about the content of the message checking whether subjects read and understood the sustainability message.

Experimental Protocol and Subject Pool The experiment was conducted online with Qualtrics and a self-developed JavaScript. Upon accessing the experiment, subjects were randomly assigned to either the Baseline or Sustainability Message treatment, and this assignment was kept throughout the experiment. After reading the instructions subjects had to pass a quiz to enter the decision phase. The quiz comprised five identical questions in both treatments. In the Sustainability Message treatment, we asked an additional true/false question on the content of the sustainability message. To ensure that subjects had understood the instructions, subjects could only participate if they had answered all questions correctly within two attempts. After the decision rounds, subjects answered a questionnaire about their picking behavior, their shopping and sustainability attitudes, and their risk preference. At the conclusion of the experiment we asked subjects about their age, gender, level of education, employment status, and annual income from all sources before taxes.

This study and the pilot are preregistered under AEARCTR-0008303.

We recruited subjects on Amazon's Mechanical Turk (MTurk) online labor market (Buhrmester et al. 2011, Lee et al. 2018, Aguinis et al. 2021) using the service of CloudResearch.com (Litman et al. 2017) to manage our Human Intelligence Tasks (HITs). We invited only residents of the United States that had completed at least 100 HITs with an approval rate of at least 95% to ensure that subjects were familiar with MTurk and understood English instructions.

Selective attrition is not a concern in our study. 580 subjects started our experiment. 40 subjects dropped-out, and 150 subjects failed the quiz resulting in 390 subjects who finished the experiment. There are no statistically significant differences in the drop-out rates ($\chi^2(1) = 0.358$, Pr $= 0.550$) or the quiz failure rates ($\chi^2(1) = 1.157$, Pr $= 0.282$) between the Sustainability Message treatment and the Baseline treatment. We exclude eight subjects where the JavaScript was disabled and hence we cannot be sure whether our experiment ran properly. Moreover, we exclude nine subjects who finished the experiment but opened the experiment link multiple times and, for example, saw both treatments. The resulting sample size is 373 subjects with 189 subjects in the Baseline treatment, and 184 subjects in the Sustainability Message treatment.

See Table 4.7 in the Appendix 4.A for detailed sample demographics. Earnings including a \$2.50 participation fee ranged between \$3.60 and \$7.51 depending on subjects' item choice and the randomly determined consumption days. The median experiment duration was 11.6 minutes converting to a median hourly wage of \$34.29, which are considerable earnings on MTurk.

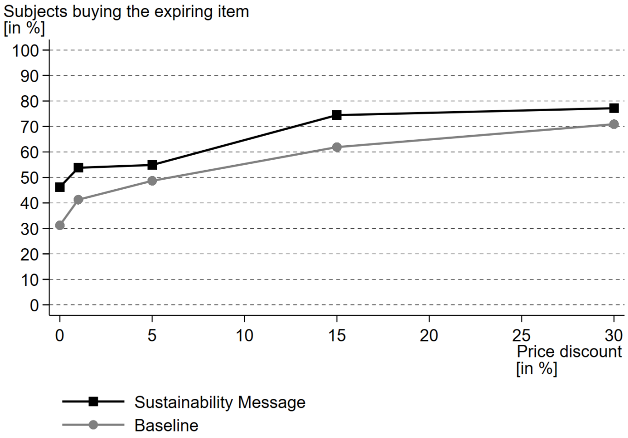
4.4 Results

We analyze whether more subjects buy the expiring item if they receive a sustainability message (Hypothesis 1) and whether more subjects buy the expiring item with increasing price discounts (Hypothesis 2) in Section 4.4.1. We examine whether a crowding out effect exists in Section 4.4.2. In Section 4.4.3, we identify four customer types that differ in their buying behavior and analyze the effect of a sustainability message and price discounts across types. In Section 4.4.4, we discuss individual differences between the four customer types.

4.4.1 Test of Hypotheses 1 and 2

A central question of our study is whether sustainability messages affect customer buying behavior. We start by comparing the share of subjects buying the expiring item in the Baseline treatment and the Sustainability Message treatment (Figure 4.2). We observe that more subjects purchase the expiring item in the Sustainability Message treatment than in the Baseline treatment at all price discounts,

Figure 4.2: Effects of Sustainability Messages and Price Discounts on Buying Behavior



which indicates support for Hypothesis 1. We also observe that more subjects buy the expiring item as the price discounts increase, which indicates support for Hypothesis 2.

We test Hypotheses 1 and 2 formally using a Linear Probability Model. We introduce a dummy as dependent variable that equals one if the expiring item was chosen and zero otherwise.

As independent variables, we include a treatment dummy for the sustainability message and an integer variable for the price discounts. We cluster standard errors on the subject level. Table 4.1 shows the results for models without and with controls in columns (1) and (2), respectively. The control variables are individuals' demographic background such as education, income, age or gender, their risk preference, and their shopping and sustainability attitude (see Appendix 4.C for the full questionnaire). The following analyses rely on the model without controls, but the results are similar for the model with controls.

The estimated average treatment effect is that the probability of individuals buying the expiring item is 10.5% higher with than without a sustainability message. The coefficient is significantly different from zero ($p < .01$, two-sided), providing support for Hypothesis 1. Moreover, the coefficient of the price discount variable is positive and significantly different from zero ($p < .01$, two-sided), which indicates

Table 4.1: Effects of Sustainability Messages and Price Discounts on Buying Behavior

Dependent Variable Y =1 if Expiring Item Chosen; =0 Otherwise	Overall		Baseline		Sustainability Message	
	(1)	(2)	(3)	(4)	(5)	(6)
Sustainability Message	0.1051***	0.1089***				
	(.03)	(.03)				
Price Discount	0.0110***	0.0110***	0.0100***	0.0100***	0.0088***	0.0088***
	(.00)	(.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.3960***	0.6048***	0.4295***	0.3929*	0.5385***	0.5039**
	(.02)	(.14)	(.03)	(.21)	(.03)	(.20)
Observations	1865	1865	756	756	736	736
Subjects	373	373	189	189	184	184
Controls	No	Yes	No	Yes	No	Yes

Notes: Pooled ordinary least squares regressions on individual buying behavior are performed.
* p < 0.10, ** p < 0.05, *** p < 0.01, two-sided. Robust standard errors are in parentheses, clustered at the individual level.
We obtain similar results when running Logit or Probit regressions.

support for Hypothesis 2. To test the effect of strictly positive price discounts without (Hypohesis 2 a) and with a sustainability message (Hypohesis 2 b), we run the same regression specification separately for the Baseline and Sustainability Message treatment and exclude the first buying decisions without a discount. Table 4.1 shows the results for the Baseline treatment without and with controls in columns (3) and (4), and for the Sustainability Message treatment in columns (5) and (6). Across all regression models, the coefficient of the price discount variable is positive and significantly different from zero ($p < .01$, two-sided), which provides support for Hypotheses 2.

4.4.2 Analysis of Crowding out Effect

When introducing a 1% price discount, there are two potential effects that can result in opposing buying behavior. A crowding out effect can result in fewer customers buying the expiring item when a price discount is introduced, whereas the price discount effect can result in more customers buying the expiring item. An interesting question is thus which effect is stronger or whether these effects offset each other. In our data, we observe an increase in the share of people buying the expiring item as we increase the discount from zero to 1%. This indicates that the price discount effect is stronger than the crowding out effect on the aggregate

Table 4.2: Reaction to 1% Price Discount of Subjects Buying the Expiring Item at the Regular Price

	Buying Behavior	
	Expiring Item	Lasting Item
Baseline	63%	37%***
Sustainability Message	68%	32%***

Note: *** $p < 0.01$, two-sided Fisher's exact tests.

level.

To investigate the existence of a crowding out effect, we analyze whether subjects buying the expiring item at the regular price switch to buying the lasting item when we introduce a marginal price discount for the expiring item (see Table 4.2). If a substantial crowding out effect existed in our setting, that is, a crowding out effect that is greater than the effect size of a 1% price discount, it should be detectable in this subset.

We find that 37% and 32% of the subjects buying the expiring item at the regular price in the Baseline and Sustainability Message treatment, respectively crowd out and switch to buying the lasting item when we introduce a price discount of 1%. These shares are significantly different from 0% in both treatments (Fisher's exact tests, $p = .000$, two-sided) and not significantly different from each other (Fisher's exact tests, $p = .592$, two-sided).

Our analyses indicate that a crowding out effect exists. We do not observe the effect in the aggregate data shown in Figure 4.2, because the crowding out effect that we observe in the subset of subjects buying the expiring item at the regular price is smaller than the price discount effect for subjects buying the lasting item at a regular price.

4.4.3 Effects of Price Discounts and Sustainability Messages Across Customer Types

In this section, we analyze the effects of price discounts and sustainability messages on individual buying behavior. The results of the aggregate data indicate that retailers can increase the sales of expiring items with price discounts and sustainability messages. However, some subjects buying the expiring item at the regular price switch to buying the lasting item when price discounts are introduced. Thus, the aggregate effect of price discounts and sustainability messages consists of opposing subset effects, and we next analyze the effects of price discounts and sustainability messages on individual buying behavior.

Definition of Customer Types

Based on the Behavioral Model in Section 4.2.5, we identify four different customer types that differ in their buying behavior. We have summarized criteria when a customer type buys an expiring or lasting item based on their expected utility in Table 4.3 where we distinguish between the expected utility gained without and with price discounts.

Type AE (Always Expiring) customers always buy the expiring item. These customers gain higher expected utility from buying the expiring item than from buying the lasting item without and with price discounts. As can be observed from Table 4.3, these customers value the potential of saving the expiring item (U_S) more than they dislike its spoilage risk ($U_R r(P_E - P_L)$) in the absence of price discounts. With price discounts, they value the potential of saving the expiring item and the price effect of discounts more than the spoilage risk and crowding out effect of price discounts.

In contrast, Type NE (Never Expiring) customers never buy the expiring item. These customers do not gain higher expected utility from buying the expiring item

Table 4.3: Definition of Customer Types

Customer Type	Buying Behavior	Criteria	
		Without Price Discount	With Price Discount
Always Expiring	Expiring item	$U_S > U_R r(P_E - P_L)$	$U_S + U_R r\ \alpha > U_R r(P_E - P_L) + U_O$
Never Expiring	Lasting item	$U_S < U_R r(P_E - P_L)$	$U_S + U_R r\ \alpha < U_R r(P_E - P_L) + U_O$
Price Sensitive	Lasting item Expiring item	$U_S < U_R r(P_E - P_L)$	If $0 < \alpha' \leq \alpha^{PS}$: $U_S + U_R r\ \alpha' < U_R r\ (P_E - P_L) + U_O$ If $\alpha'' > \alpha^{PS}$: $U_S + U_R r\ \alpha'' > U_R r\ (P_E - P_L) + U_O$
Crowding Out	Lasting item Expiring item	$U_S > U_R r(P_E - P_L)$	If $0 < \alpha^* \leq \alpha^{CO}$: $U_S + U_R r\ \alpha^* < U_R r\ (P_E - P_L) + U_O$ If $\alpha^{**} > \alpha^{CO}$: $U_S + U_R r\ \alpha^{**} > U_R r\ (P_E - P_L) + U_O$

than from buying the lasting item without and with price discounts. Without price discounts, these customers value the potential of saving the expiring item less than they dislike its spoilage risk. With price discounts, they value the potential of saving the expiring item and the price effect of discounts less than they dislike the spoilage risk of the expiring item and the crowding out effect of price discounts.

Type PS (Price Sensitive) customers react to whether a price discount is below or above their price discount threshold (α^{PS}). They buy the lasting item for price discounts $\alpha^{'} \leq \alpha^{PS}$ but switch to buying the expiring item for price discounts $\alpha^{''} > \alpha^{PS}$. These customers gain lower expected utility from buying the expiring item than from buying the lasting item for price discounts $\alpha^{'} \leq \alpha^{PS}$. For price discounts $\alpha^{'} \leq \alpha^{PS}$, these customers value the potential of saving the expiring item and the price effect of discounts less than they dislike the spoilage risk and the crowding out effect of price discounts. However, for price discounts $\alpha^{''} > \alpha^{PS}$, it is vice versa.

Type CO (Crowding Out) buy the expiring item without price discounts but switch to buying the lasting item when a marginal discount is below their crowding out threshold, i.e., $0 < \alpha^* \leq \alpha^{CO}$. They switch back to buying the expiring item when price discounts $\alpha^{**} > \alpha^{CO}$ are offered. These customers gain higher expected utility from buying the expiring item than from buying the lasting item without price discounts. For price discounts $\alpha^* \leq \alpha^{CO}$ they gain lower expected utility from buying the expiring item than from buying the lasting item. For price discounts $\alpha^{**} > \alpha^{CO}$ it is vice versa.

Table 4.4: Shares of Subjects Across Customer Types

Customer Type	Baseline	Sustainability Message	Difference	p-value
AE	7.9%	23.9%	16.0%	0.000
NE	8.5%	5.4%	-3.0%	0.311
PS	43.9%	34.2%	-9.7%	0.057
CO	7.4%	10.3%	2.9%	0.365
Other	32.3%	26.1%	-6.2%	0.211

Notes: AE, Always Expiring; NE, Never Expiring; PS, Price Sensitive; CO, Crowding Out. We report p-values of two-sided Fisher's exact tests.

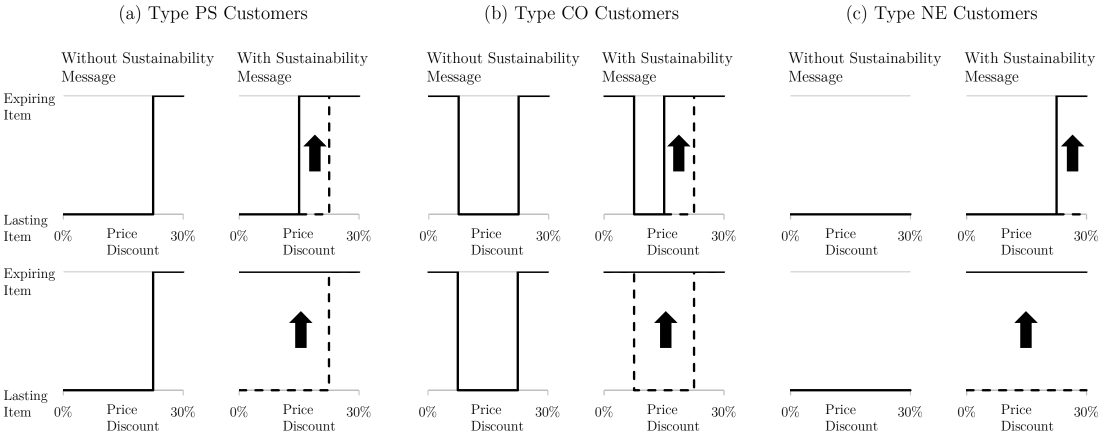
These customers value the potential of saving the expiring item more than the disutility of its spoilage risk without price discounts. For price discounts $\alpha^* \leq \alpha^{CO}$ the potential of saving the expiring item and the price effect of discounts are smaller than the crowding out effect of price discounts and the disutility of the spoilage risk. Moreover, for price discounts $\alpha^{**} > \alpha^{CO}$ it is vice versa.

We analyze individual buying behavior in both treatments and cluster subjects into the four customer types defined above. We can classify 68% and 74% of subjects in the Baseline and Sustainability Message treatment, respectively using these four customer types. In the Baseline treatment, we find that about 8% of the subjects are Type AE, 9% Type NE, 44% Type PS and 7% Type CO as can be seen in column (1) of Table 4.4.

Effect of Sustainability Messages on Buying Behavior Across Customers Types

We now analyze whether the buying behavior of each customer type changes when sustainability messages are shown. The Behavioral Model predicts that the expected utility of buying an expiring item increases by $U_M \geq 0$ when sustainability messages are shown. Figure 4.3 schematically depicts the potential change in buying behavior when sustainability messages are shown. Small effect sizes of U_M are visualized in the top row whereas large effect sizes can be seen in the bottom row. By definition, the buying behavior of Type AE customers cannot change towards buying the expiring item and hence does not differ with or without

Figure 4.3: Schematic Effect of Sustainability Messages on Buying Behavior Across Customer Types



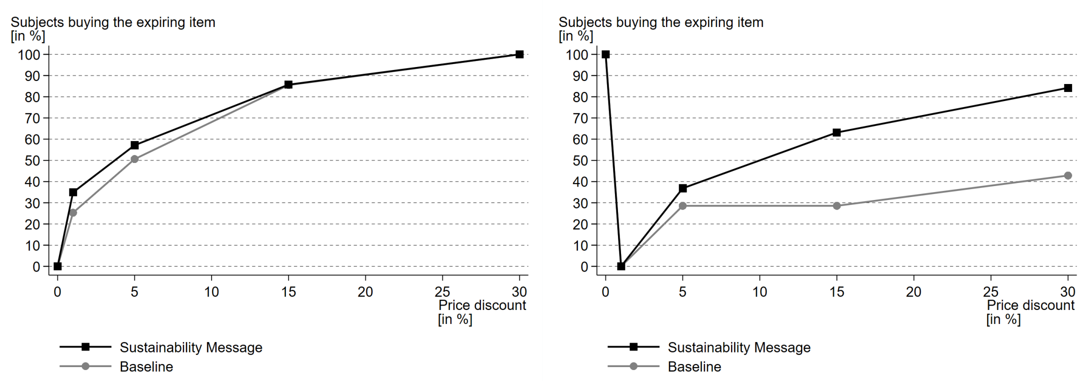
sustainability messages.

As Figure 4.3 shows, the Behavioral Model predicts that for a given effect size U_M, Type PS, CO, and NE customers buy the expiring item at a smaller discount with than without sustainability messages. Consequently, we expect that the share of Type PS and CO subjects who switch (back) to buying the expiring item is larger in the Sustainability Message treatment than in the Baseline treatment (see the top row of Figures 4.3 (a) & (b)). By definition, we cannot observe any differences in the percentage of subjects buying the expiring item if customers are of type AE and NE in the Sustainability Message treatment.

We compare the predictions of the Behavioral Model to the observations for Type PS and CO subjects in our experiment. Figure 4.4 visualizes the buying behavior of these subjects in the Sustainability Message treatment and in the Baseline treatment in our experiment.

Figure 4 (a) shows the buying behavior of Type PS subjects for the Sustainability Message and Baseline treatment. Compared to the Baseline treatment, we observe similar and not significantly different shares of subjects buying the expiring item at all discounts (Fisher's exact test, two-sided, $p > 0.1$). We thus do not find that the share of Type PS subjects buying the expiring item is larger in the Sustainability Message treatment than in the Baseline treatment for any given

Figure 4.4: Effect of Sustainability Messages on Buying Behavior of Types PS and CO Subjects



(a) Buying Behavior of Type PS Subjects **(b)** Buying Behavior of Type CO Subjects

discount.

Figure 4 (b) shows the buying behavior of Type CO subjects for the Sustainability Message and Baseline treatment. Compared to the Baseline treatment, we observe larger shares of subjects buying the expiring item at price discounts of 5%, 15% and 30% in the Sustainability Message treatment, which is in line with the predictions of the Behavioral Model. The differences for the discounts of 15% and 30% are significant (Fisher's exact test, two-sided, $p = .080$ and $p = .024$ for 15% and 30%, respectively). We thus find that the share of Type CO subjects who switch to buying the expiring item is larger in the Sustainability Message treatment than in the Baseline treatment for discounts $> 5\%$.

We next analyze whether the share of subjects within each customer type differs between the Sustainability Message and the Baseline treatment. The bottom row of Figures 4.3 (a) & (b) shows that if the effect size of U_M is sufficiently large, customers showing a buying behavior of Type PS or CO in the Baseline treatment might show a buying behavior of Type AE in the Sustainability Message treatment. Moreover, Figure 4.3 (c) shows that dependent on the effect size of U_M, customers of Type NE in the Baseline treatment might show a buying behavior of Type PS or AE in the Sustainability Message treatment. Thus, we expect a larger share of Type AE subjects and a smaller share of Type NE and CO subjects in the

Sustainability Message treatment than in the Baseline treatment. Compared to the Baseline treatment, some PS Type subjects might show an AE Type buying behavior and some NE Type subjects might show a PS Type buying behavior in the Sustainability Message treatment. Accordingly, we cannot predict whether the share of Type PS subjects is larger or smaller in the Sustainability Message treatment than in the Baseline treatment.

Column (3) of Table 4.4 shows the differences in the share of subjects between treatments across the four customer types. Column (4) of Table 4.4 reports the *p*-values of two-sided Fisher's exact tests testing whether these shares differ significantly between treatments. As predicted by the Behavioral Model, we see a significantly larger share of Type AE subjects in the Sustainability Message treatment than in the Baseline treatment. Moreover, we see a significantly smaller share of Type PS subjects in the Sustainability Message treatment than in the Baseline treatment. One explanation could be that more Type PS subjects show a Type AE behavior than Type NE subjects show a Type PS behavior.

To summarize, we observe four customer types that constitute the aggregate buying behavior as predicted by the Behavioral Model: Type AE who always buys the expiring item. Type NE who never buys the expiring item. Type PS who switches from buying the lasting to the expiring item when price discounts are sufficiently high. Type CO who crowds out and switches from buying the expiring to the lasting item when price discounts are introduced. This suggests that only Type PS customers should receive price discounts as these individuals are the only ones where a positive effect on sales of expiring items can be expected.

Moreover, we find that the aggregate positive effect of sustainability messages on sales of expiring items is driven by 1) different behavior within customer types, and 2) different shares of subjects across types. A significantly larger share of Type CO subjects switches back to buying the expiring item in the Sustainability Message than in the Baseline treatment (crowding in). Additionally, the share of Type AE subjects is significantly larger in the Sustainability Message treatment

Table 4.5: Mean of Individual Characteristics Across Customer Types

Individual Characteristics	Customer Type				
	AE	NE	PS	CO	Overall
Income	5.39	5.58	*4.95*	5.91	5.23
Education	4.44	5.00	4.58	5.00	4.64
Employed	0.86	1.00	0.92	0.97	0.92
Risk preference	4.47	6.54	4.67	7.52	5.17
Age	41.78	38.04	38.89	36.06	39.10
Female	0.56	0.46	0.47	0.39	0.48
I avoid wasting food.	6.20	5.81	6.17	5.76	6.09
Sustainability is important to me.	5.81	5.85	5.82	5.58	5.79
Animal welfare is important to me.	5.90	5.69	5.78	5.61	5.78
I can afford to buy sustainable products if I want to.	5.27	5.62	*4.86*	5.18	5.06
I pay attention to the price when I shop.	6.34	5.92	*6.58*	6.36	6.43
Number of Subjects	59	26	146	33	264

Notes: AE, Always Expiring; NE, Never Expiring; PS, Price Sensitive; CO, Crowding Out.

than in the Baseline while the share of Type PS subjects is significantly smaller than in the Baseline. Thus, a retailer might consider providing sustainability messages to all customer types as a positive effect on sales of expiring items can be expected.

4.4.4 Individual Differences Between Customer Types

Retailers would benefit from being able to distinguish the four customer types and hence anticipate their reaction to price discounts and sustainability messages. This distinction would allow them to use price discounts and sustainability messages where they are most effective.

At the end of the experiment, we asked subjects standard demographic questions and elicited their risk preferences. We also asked subjects about their shopping and sustainability attitude by asking whether they agree to the following statements on a 7-point scale: ”I avoid wasting food.”, ”Sustainability is important to me.”, ”Animal welfare is important to me.”, ”I can afford to buy sustainable products if I want to.”, and ”I pay attention to the price when I shop.”.

Table 4.5 summarizes the mean for each characteristic and customer type. The last row contains the number of subjects of each type in both, Baseline and Sustainability Message Treatment. For detailed numbers per treatment, we refer to Table 4.8 and Table 4.9 in Appendix 4.B.1.

A retailer would profit from targeting price discounts only at customers who are likely to switch from buying lasting to buying expiring items with price discounts, that is, the Type PS customers. To identify these customers, the retailer could analyze their characteristics. We observe from Table 4.5 that Type PS subjects reported the lowest income (mean of 4.95). They also gave the lowest rating for the question ”I can afford to buy sustainable products if I want to.”, which means that they cannot afford sustainable products as easily as subjects of other types. Moreover, they gave the highest rating for the question ”I pay attention to the price when I shop.”, which is consistent with being price sensitive.

We compare the answers of Type PS subjects to those of other types with two-sided Wilcoxon rank-sum tests in Table 4.6 (see Tables 4.10 and 4.11 in Appendix 4.B.2 for an overview per treatment). If we compare all Type PS subjects to the remaining ones (i.e., non PS), we find significant differences for all of them (see last column $\overline{PS}$). Looking into the comparison in more detail, we observe significant income differences between Type PS and CO subjects, but not between Type PS and AE or NE subjects. We observe significant differences for all types compared to PS on the question about being able to afford sustainable products. Regarding the question whether subjects pay attention to price, there is only a significant difference between Type PS and NE subjects. These findings suggest

Table 4.6: Tests Comparing Individual Characteristics of Type PS Subjects With Those of Other Subjects

Individual Characteristics	Customer Type			
	AE	NE	CO	$\overline{\text{PS}}$
Income	0.3819	0.1789	0.0495	0.0560
I can afford to buy sustainable products if I want to.	0.0369	0.0131	0.0802	0.0026
I pay attention to the price when I shop.	0.1420	0.0046	0.2081	0.0116
Number of Subjects	59	26	33	118

Notes: AE, Always Expiring; NE, Never Expiring; CO, Crowding Out; $\overline{\text{PS}}$, Not Price Sensitive, i.e. all subjects except Type Price Sensitive. We report *p*-values of two-sided Wilcoxon rank-sum tests.
We compare the distribution of a variable between Type Price Sensitive subjects and the respective other subject type.

that a retailer who is aware of these characteristics might predict customers' reactions to price discounts and sustainability messages.

4.5 Conclusion

We analyze how sustainability messages affect the behavior of customers when buying perishable products and how this effect interacts with price discounts. It is not clear whether sustainability messages affect sales of expiring items. Individuals report to care about sustainability but it has not been studied if sustainability aspects incentivize sales of expiring items. Moreover, the literature suggests a positive effect of price discounts but findings on the effect of combining non-monetary with monetary incentives are ambiguous. Some studies find a crowding out effect while others report a null or a positive effect when both incentives are present. Thus, the effect of combining sustainability messages with price discounts on sales of expiring items is not clear.

We ran an online experiment where subjects choose between an expiring and a lasting item for perishable products. In our experiments, we find that sustainability messages and price discounts can induce customers to buy expiring items and thus contribute to food waste reduction. The combination of both incentives increases aggregate sales of expiring items. We also see a crowding out effect for some subjects. However, the negative effect is not observable on the aggregate level since it is out-weighted by the positive effect of price discounts on sales of expiring items. We find that it can be beneficial for retailers to consider the characteristics of a customer, since both incentives do not work for every customer in the same way.

The findings are not only interesting from an academic perspective, but also relevant for retail managers. An important insight is the overall positive effect of sustainability messages, which shows that there is a clear benefit from promoting sustainability in retailing and for retailers to show sustainability-related messages in stores. When seeing a sustainability message more subjects buy the expiring item irrespective of whether it is discounted or not and less subjects buy the expiring item only when it is discounted. Additionally, more subjects who crowd out when they receive a marginal price discount switch back to buying the expiring item when they receive higher price discounts (crowding in).

Moreover, we identify four customer types that respond differently to price discounts. One customer type reacts positively to price discounts while the other customer types respond negatively to it or not at all. A retailer could consider offering price discounts only to customers where they have a positive effect and not to those where price discounts do not or negatively affect demand, as these would incur unnecessary costs to the retailer.

One of the main limitations of our study is that we tested subjects' behavior in an online experiment without real incentive to make sustainable decisions and where food waste was not directly visible. Interestingly though, we still observe significant effects of a sustainability message, which means that this is likely a lower bound to what would be observable in practice, where these effects might be even more pronounced. A field experiment would provide further insights and could also be used to test whether the effects differ by product characteristics.

4.6 Acknowdlegements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1–390838866. This study is preregistered under AEARCTR-0008303. Please see Appendix 4.C for the experimental instructions and screenshots.

Appendix of Chapter 2

4.A Sample Demographics

Table 4.7: Sample Demographics

Demographics	Percentage (N=373)
Age[1]	38.03 (10.45)
Female	46.92
Highest level of education	
Less than High school degree	0.27
High school graduate	6.43
Vocational/technical school	4.83
Some college	14.75
Bachelor's degree	58.18
Master's degree	13.40
Doctoral degree	1.61
Advanced professional degree (JD, MD, MBA, etc.)	0.54
Employment status	
Working (paid employee)	79.62
Working (self-employed)	14.48
Not working	5.36
Other	0.54
Annual income from all sources before taxes	
$10,000 or less	6.70
$10,001 to $20,000	6.97
$20,001 to $30,000	12.60
$30,001 to $40,000	10.72
$40,001 to $50,000	16.35
$50,001 to $60,000	17.16
$60,001 to $70,000	8.31
$70,001 to $80,000	7.24
Over $80,000	13.94

Note: [1] Mean in years (standard deviation)

4.B Further Analyses

4.B.1 Analysis of Individual Characteristics Across Customer Types

Table 4.8: Mean of Individual Characteristics Across Customer Types in the Baseline Treatment

Individual Characteristics			Customer Type		
	AE	NE	PS	CO	Overall
Income	5.53	5.00	*4.67*	6.29	4.99
Education	4.33	4.94	4.53	5.14	4.63
Employed	1.00	1.00	*0.94*	1.00	0.96
Risk preference	5.47	6.69	*4.59*	7.93	5.32
Age	39.00	39.25	*40.52*	34.57	39.53
Female	0.40	0.44	*0.49*	0.36	0.46
I avoid wasting food.	6.40	5.94	6.31	5.64	6.20
Sustainability is important to me.	5.40	5.88	5.76	5.50	5.70
Animal welfare is important to me.	5.33	5.94	5.80	5.29	5.70
I can afford to buy sustainable products if I want to.	5.33	5.75	*4.83*	5.07	5.03
I pay attention to the price when I shop.	6.53	5.94	*6.59*	6.21	6.46
Number of Subjects	15	16	83	14	128

Notes: AE, Always Expiring; NE, Never Expiring; PS, Price Sensitive; CO, Crowding Out.

Table 4.9: Mean of Individual Characteristics Across Customer Types in the Sustainability Message Treatment

Individual Characteristics			Customer Type		
	AE	NE	PS	CO	Overall
Income	5.34	6.50	*5.30*	5.63	5.45
Education	4.48	5.10	4.65	4.89	4.66
Employed	0.82	1.00	0.90	0.95	0.89
Risk preference	4.14	6.30	4.78	7.21	5.02
Age	42.73	36.10	36.75	37.16	38.69
Female	0.61	0.50	0.44	0.42	0.50
I avoid wasting food.	6.14	5.60	5.98	5.84	5.99
Sustainability is important to me.	5.95	5.80	5.89	5.63	5.87
Animal welfare is important to me.	6.09	5.30	5.76	5.84	5.85
I can afford to buy sustainable products if I want to.	5.25	5.40	*4.89*	5.26	5.10
I pay attention to the price when I shop.	6.27	5.90	*6.56*	6.47	6.40
Number of Subjects	44	10	63	19	136

Notes: AE, Always Expiring; NE, Never Expiring; PS, Price Sensitive; CO, Crowding Out.

4.B.2 Comparing Individual Characteristics of Type PS Subjects With Those of Other Subjects

Table 4.10: Tests Comparing Individual Characteristics of Type PS Subjects With Those of Other Subjects in the Baseline Treatment

Individual Characteristics	Customer Type			
	AE	NE	CO	$\overline{\text{PS}}$
Income	0.2247	0.3492	0.0255	0.0280
I can afford to buy sustainable products if I want to.	0.1492	0.0263	0.4743	0.0237
I pay attention to the price when I shop.	0.5179	0.0159	0.2751	0.0320
Employed	0.8551	0.8119	0.9010	0.2195
Risk preference	0.3604	0.0061	0.0000	0.0001
Age	0.7048	0.9268	0.0259	0.1876
Female	0.6996	0.8900	0.5125	0.4054
Number of Subjects	15	16	14	45

Notes: AE, Always Expiring; NE, Never Expiring; CO, Crowding Out; $\overline{\text{PS}}$, Not Price Sensitive, i.e. all subjects except Type Price Sensitive.
We report p-values of two-sided Wilcoxon rank-sum tests.
We compare the distribution of a variable between Type Price Sensitive subjects and the respective other subject type.

Table 4.11: Tests Comparing Individual Characteristics of Type PS Subjects With Those of Other Subjects in the Sustainability Message Treatment

Individual Characteristics	Customer Type			
	AE	NE	CO	$\overline{\text{PS}}$
Income	0.8778	0.1511	0.6944	0.6592
I can afford to buy sustainable products if I want to.	0.1323	0.2977	0.0834	0.0430
I pay attention to the price when I shop.	0.2365	0.1276	0.5100	0.1325
Number of Subjects	44	10	19	73

Notes: AE, Always Expiring; NE, Never Expiring; CO, Crowding Out; $\overline{\text{PS}}$, Not Price Sensitive, i.e. all subjects except Type Price Sensitive.
We report p-values of two-sided Wilcoxon rank-sum tests.
We compare the distribution of a variable between Type Price Sensitive subjects and the respective other subject type.

4.B.3 Results of Pilot Experiment

Figure 4.5: Effects of Sustainability Messages and Price Discounts on Buying Behavior in the Pilot

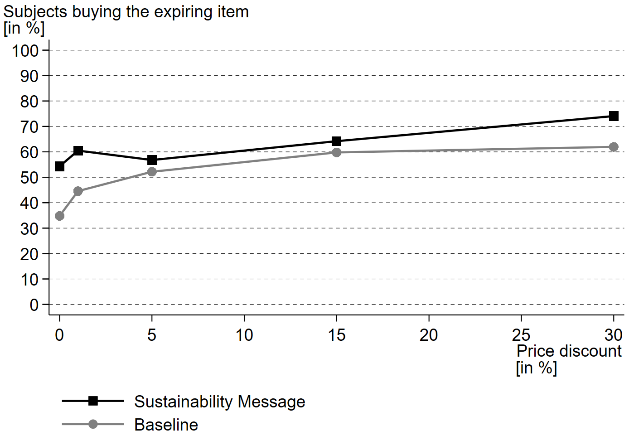
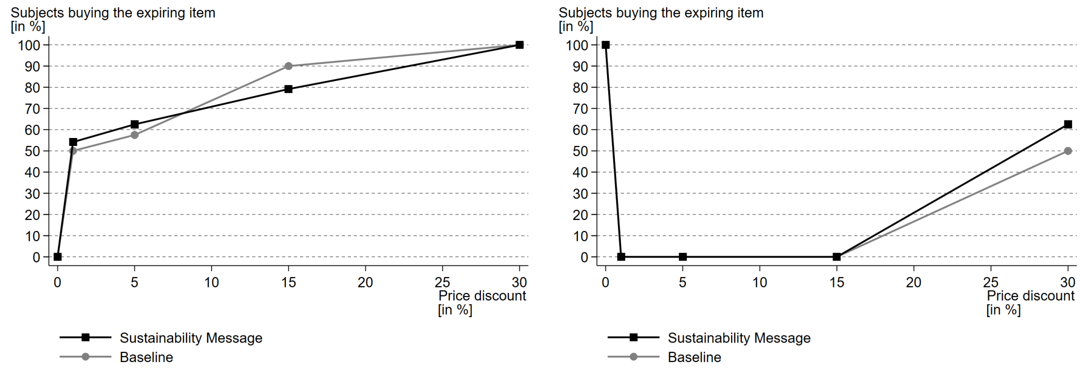


Table 4.12: Effects of Sustainability Messages and Price Discounts on Buying Behavior in the Pilot

Dependent Variable Y =1 if Expiring Item Chosen; =0 Otherwise	Overall		Baseline		Sustainability Message	
	(1)	(2)	(3)	(4)	(5)	(6)
Sustainability Message	0.1132**	0.0796*				
	(.05)	(.05)				
Price Discount	0.0068**	0.0068**	0.0055**	0.0055**	0.0054**	0.0054**
	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
Constant	0.4375***	-0.0173	0.4761***	0.3022	0.5698***	-0.3780
	(.03)	(.27)	(.05)	(.38)	(.05)	(.41)
Observations	865	865	368	368	324	324
Subjects	173	173	92	92	81	81
Controls	No	Yes	No	Yes	No	Yes

Notes: Pooled ordinary least squares regressions on individual buying behavior are performed.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, two-sided. Robust standard errors are in parentheses, clustered at the individual level.
We obtain similar results when running Logit or Probit regressions.

Figure 4.6: Effect of Sustainability Messages on Buying Behavior of Types PS and CO Subjects in the Pilot



(a) Buying Behavior of Type PS Subjects

(b) Buying Behavior of Type CO Subjects

Table 4.13: Shares of Subjects Across Customer Types in the Pilot

Customer Type	Baseline	Sustainability Message	Difference	*p*-value
AE	7.6%	29.6%	22.0%	0.000
NE	8.7%	3.7%	-5.0%	0.222
PS	43.5%	29.6%	-13.8%	0.082
CO	10.9%	9.9%	-1.0%	1.000
Other	29.3%	27.2%	-2.2%	0.866

Notes: AE, Always Expiring; NE, Never Expiring; PS, Price Sensitive; CO, Crowding Out.
We report *p*-values of two-sided Fisher's exact tests.

4.C Screenshots of the Experiment

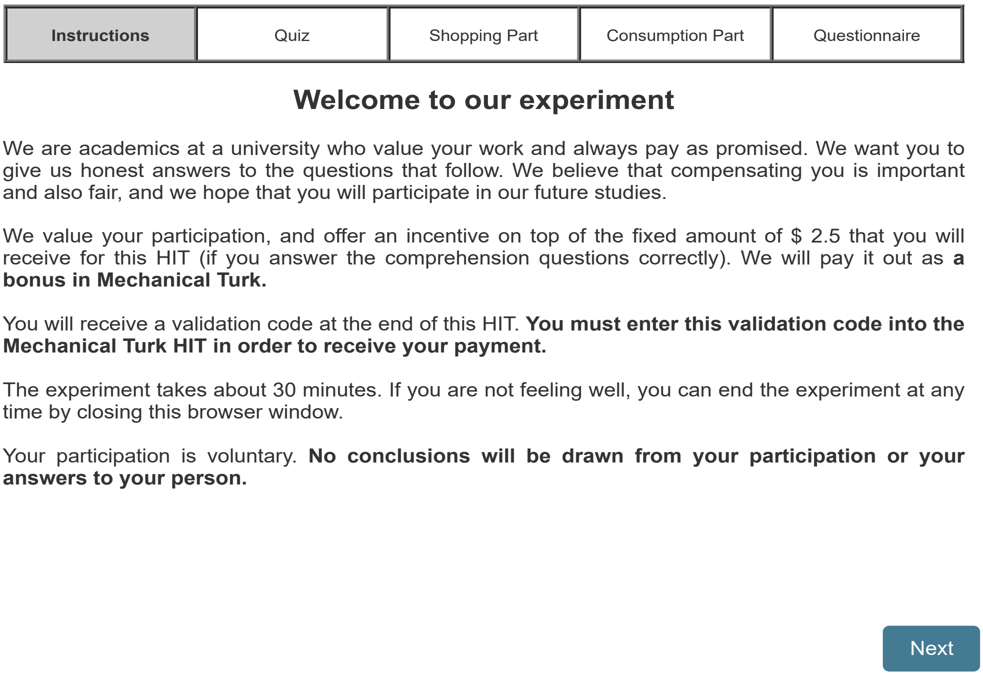
In the following, we present screenshots of our experiment. The framed sustainability message shown in the instructions (Figure 4.9) and on the decision screen (Figures 4.13 and 4.14) as well as the sixth quiz question (Figure 4.12) were only displayed in the Sustainability Message but not in the Baseline treatment. These were the only differences between the Sustainability Message and Baseline treatment. The correct answers to the questions of the quiz are selected on the respective screenshots (Figure 4.10-4.12).

We show screenshots of the first and fifth decision of the main part, when the expiring item A was offered at no discount and a discount of 30%. In the decision rounds 2, 3, and 4 item A was offered at discounts of 1%, 5%, and 15%.

We show a randomly generated example of the validation screen (Figure 4.8) and decisions (Figure 4.15).

4.C.1 Welcome Screen and Instructions

Figure 4.7: Welcome Screen



Welcome to our experiment

We are academics at a university who value your work and always pay as promised. We want you to give us honest answers to the questions that follow. We believe that compensating you is important and also fair, and we hope that you will participate in our future studies.

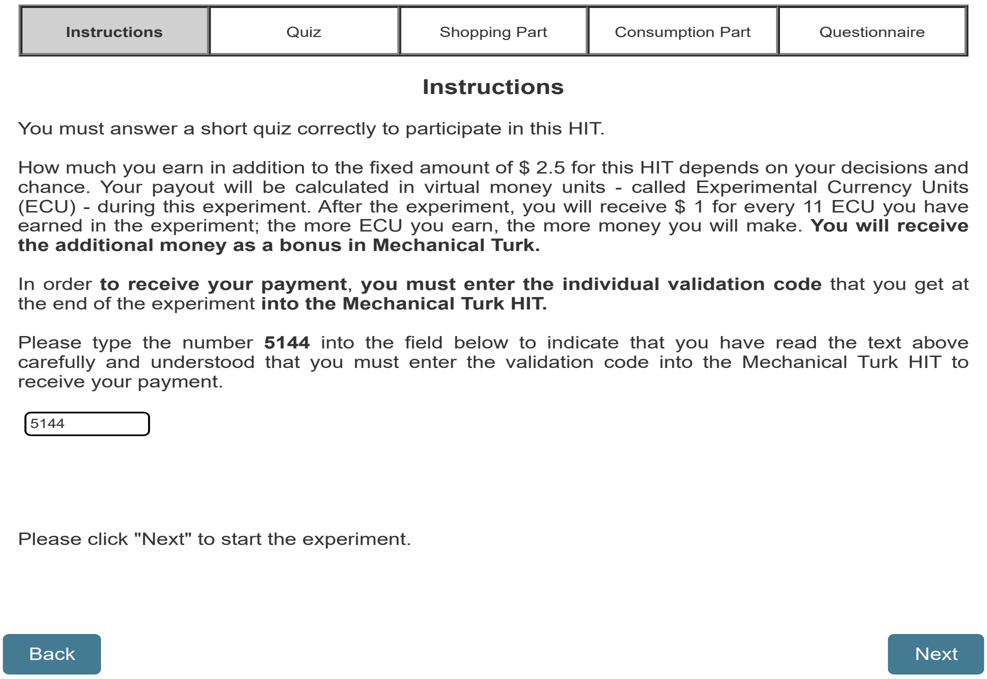
We value your participation, and offer an incentive on top of the fixed amount of $ 2.5 that you will receive for this HIT (if you answer the comprehension questions correctly). We will pay it out as **a bonus in Mechanical Turk.**

You will receive a validation code at the end of this HIT. **You must enter this validation code into the Mechanical Turk HIT in order to receive your payment.**

The experiment takes about 30 minutes. If you are not feeling well, you can end the experiment at any time by closing this browser window.

Your participation is voluntary. **No conclusions will be drawn from your participation or your answers to your person.**

Figure 4.8: Instructions I



Instructions

You must answer a short quiz correctly to participate in this HIT.

How much you earn in addition to the fixed amount of $ 2.5 for this HIT depends on your decisions and chance. Your payout will be calculated in virtual money units - called Experimental Currency Units (ECU) - during this experiment. After the experiment, you will receive $ 1 for every 11 ECU you have earned in the experiment; the more ECU you earn, the more money you will make. **You will receive the additional money as a bonus in Mechanical Turk.**

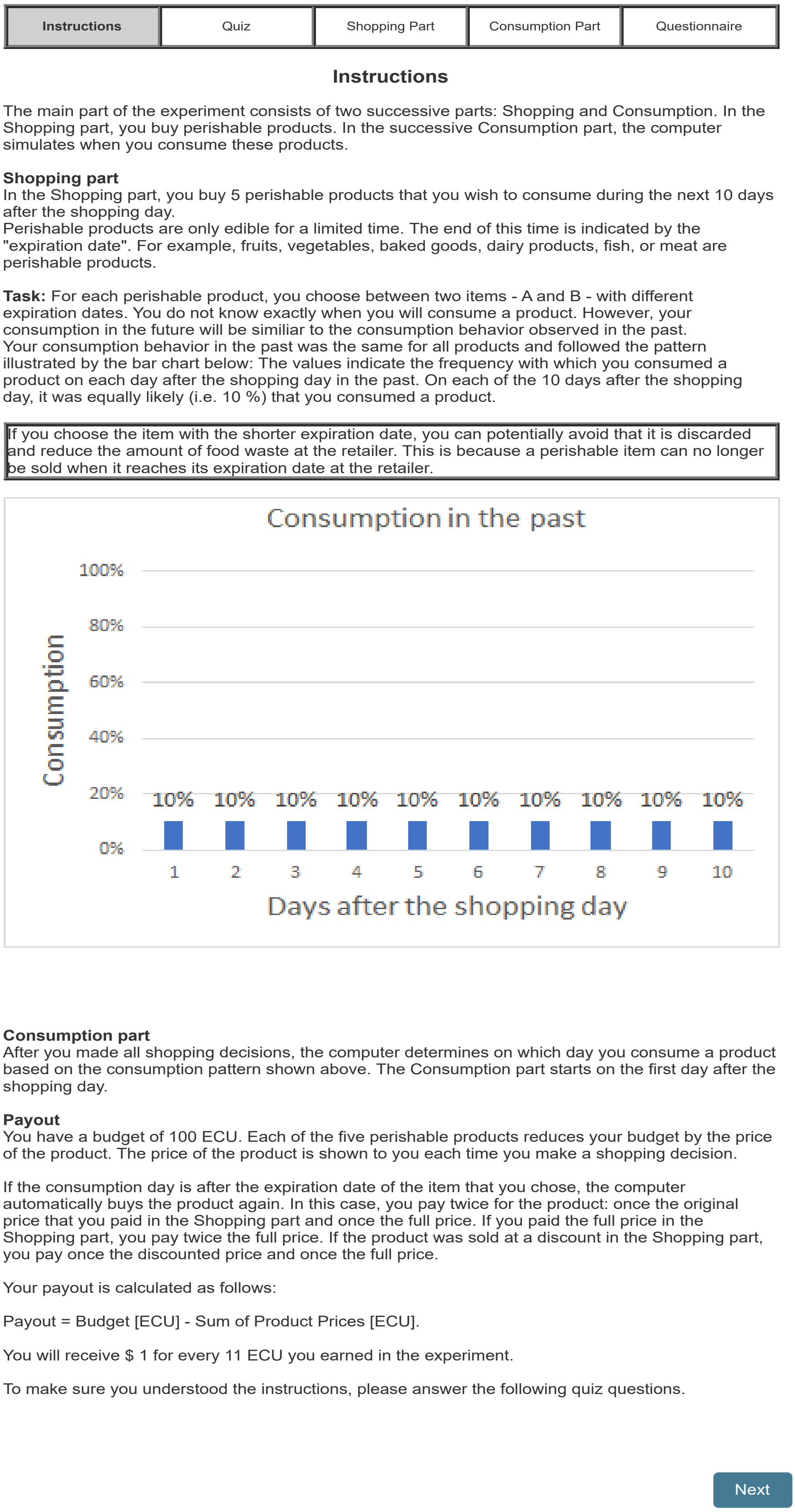
In order **to receive your payment**, **you must enter the individual validation code** that you get at the end of the experiment **into the Mechanical Turk HIT.**

Please type the number **5144** into the field below to indicate that you have read the text above carefully and understood that you must enter the validation code into the Mechanical Turk HIT to receive your payment.

5144

Please click "Next" to start the experiment.

Figure 4.9: Instructions II



Instructions

The main part of the experiment consists of two successive parts: Shopping and Consumption. In the Shopping part, you buy perishable products. In the successive Consumption part, the computer simulates when you consume these products.

Shopping part
In the Shopping part, you buy 5 perishable products that you wish to consume during the next 10 days after the shopping day.
Perishable products are only edible for a limited time. The end of this time is indicated by the "expiration date". For example, fruits, vegetables, baked goods, dairy products, fish, or meat are perishable products.

Task: For each perishable product, you choose between two items - A and B - with different expiration dates. You do not know exactly when you will consume a product. However, your consumption in the future will be similiar to the consumption behavior observed in the past.
Your consumption behavior in the past was the same for all products and followed the pattern illustrated by the bar chart below: The values indicate the frequency with which you consumed a product on each day after the shopping day in the past. On each of the 10 days after the shopping day, it was equally likely (i.e. 10 %) that you consumed a product.

If you choose the item with the shorter expiration date, you can potentially avoid that it is discarded and reduce the amount of food waste at the retailer. This is because a perishable item can no longer be sold when it reaches its expiration date at the retailer.

Consumption part
After you made all shopping decisions, the computer determines on which day you consume a product based on the consumption pattern shown above. The Consumption part starts on the first day after the shopping day.

Payout
You have a budget of 100 ECU. Each of the five perishable products reduces your budget by the price of the product. The price of the product is shown to you each time you make a shopping decision.

If the consumption day is after the expiration date of the item that you chose, the computer automatically buys the product again. In this case, you pay twice for the product: once the original price that you paid in the Shopping part and once the full price. If you paid the full price in the Shopping part, you pay twice the full price. If the product was sold at a discount in the Shopping part, you pay once the discounted price and once the full price.

Your payout is calculated as follows:

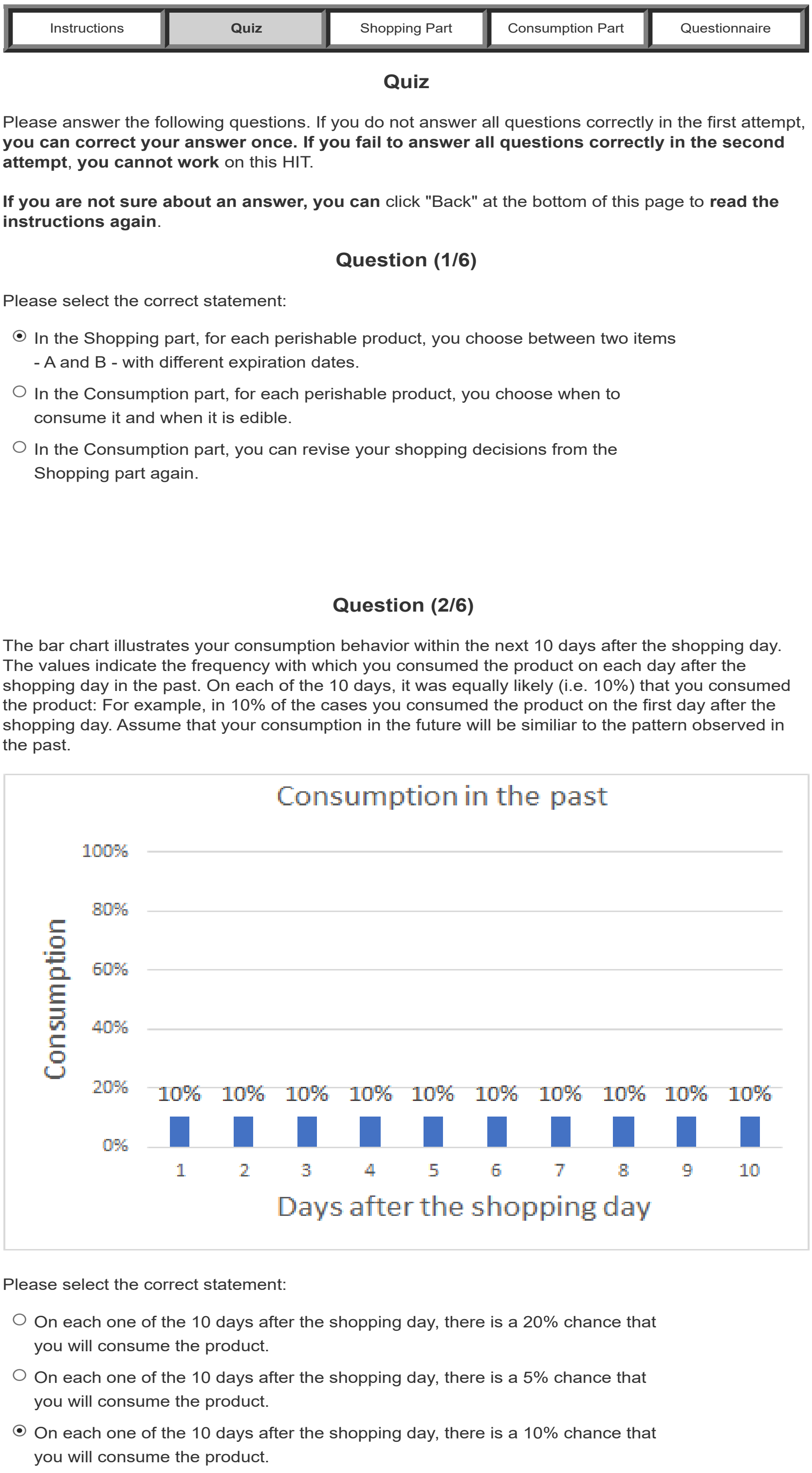
Payout = Budget [ECU] - Sum of Product Prices [ECU].

You will receive $ 1 for every 11 ECU you earned in the experiment.

To make sure you understood the instructions, please answer the following quiz questions.

4.C.2 Quiz

Figure 4.10: Quiz Questions I



Quiz

Please answer the following questions. If you do not answer all questions correctly in the first attempt, **you can correct your answer once. If you fail to answer all questions correctly in the second attempt**, **you cannot work** on this HIT.

If you are not sure about an answer, you can click "Back" at the bottom of this page to **read the instructions again**.

Question (1/6)

Please select the correct statement:

● In the Shopping part, for each perishable product, you choose between two items - A and B - with different expiration dates.

○ In the Consumption part, for each perishable product, you choose when to consume it and when it is edible.

○ In the Consumption part, you can revise your shopping decisions from the Shopping part again.

Question (2/6)

The bar chart illustrates your consumption behavior within the next 10 days after the shopping day. The values indicate the frequency with which you consumed the product on each day after the shopping day in the past. On each of the 10 days, it was equally likely (i.e. 10%) that you consumed the product: For example, in 10% of the cases you consumed the product on the first day after the shopping day. Assume that your consumption in the future will be similiar to the pattern observed in the past.

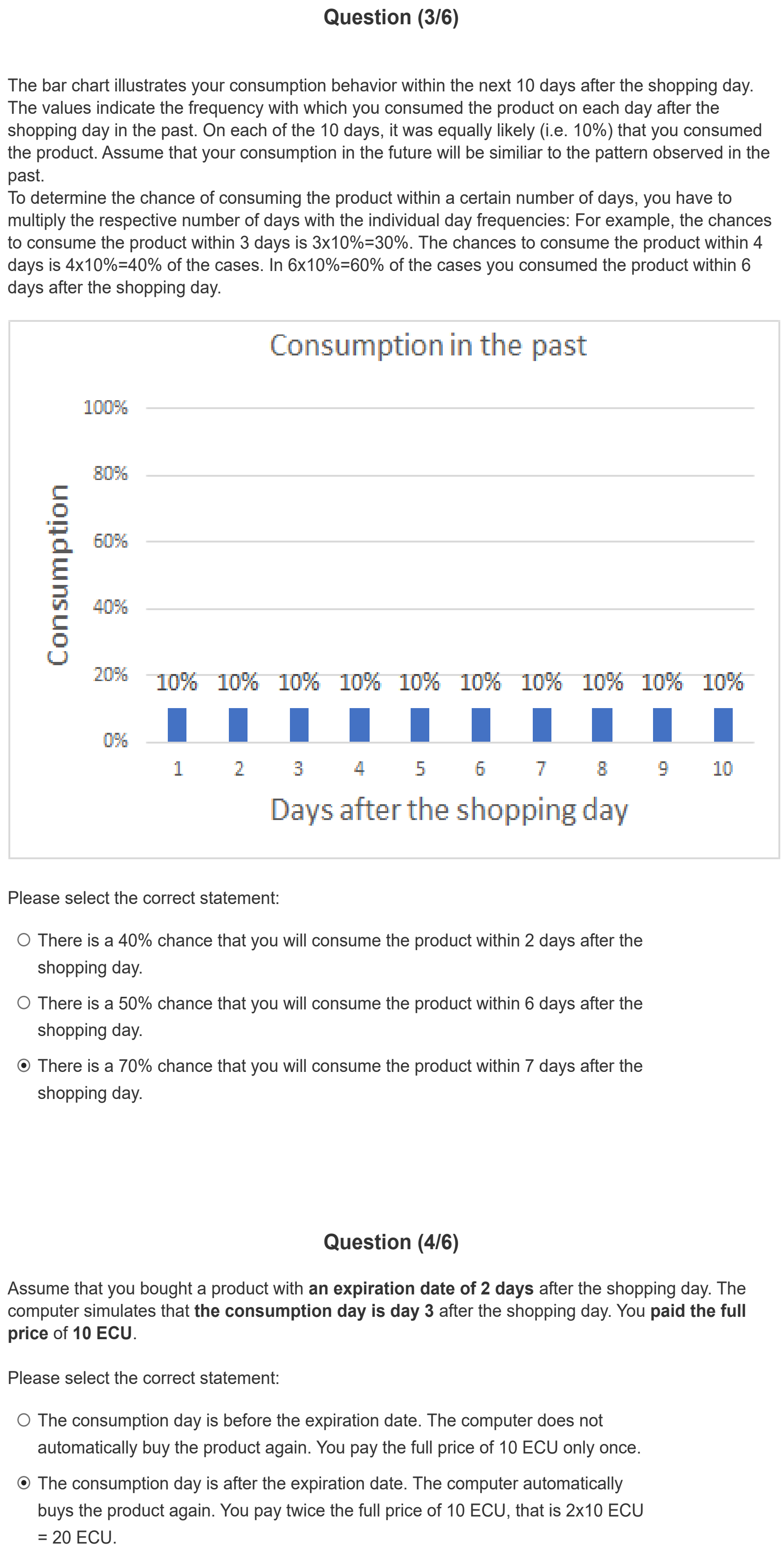
Please select the correct statement:

○ On each one of the 10 days after the shopping day, there is a 20% chance that you will consume the product.

○ On each one of the 10 days after the shopping day, there is a 5% chance that you will consume the product.

● On each one of the 10 days after the shopping day, there is a 10% chance that you will consume the product.

Figure 4.11: Quiz Questions II



Question (3/6)

The bar chart illustrates your consumption behavior within the next 10 days after the shopping day. The values indicate the frequency with which you consumed the product on each day after the shopping day in the past. On each of the 10 days, it was equally likely (i.e. 10%) that you consumed the product. Assume that your consumption in the future will be similiar to the pattern observed in the past.

To determine the chance of consuming the product within a certain number of days, you have to multiply the respective number of days with the individual day frequencies: For example, the chances to consume the product within 3 days is 3x10%=30%. The chances to consume the product within 4 days is 4x10%=40% of the cases. In 6x10%=60% of the cases you consumed the product within 6 days after the shopping day.

Please select the correct statement:

○ There is a 40% chance that you will consume the product within 2 days after the shopping day.

○ There is a 50% chance that you will consume the product within 6 days after the shopping day.

● There is a 70% chance that you will consume the product within 7 days after the shopping day.

Question (4/6)

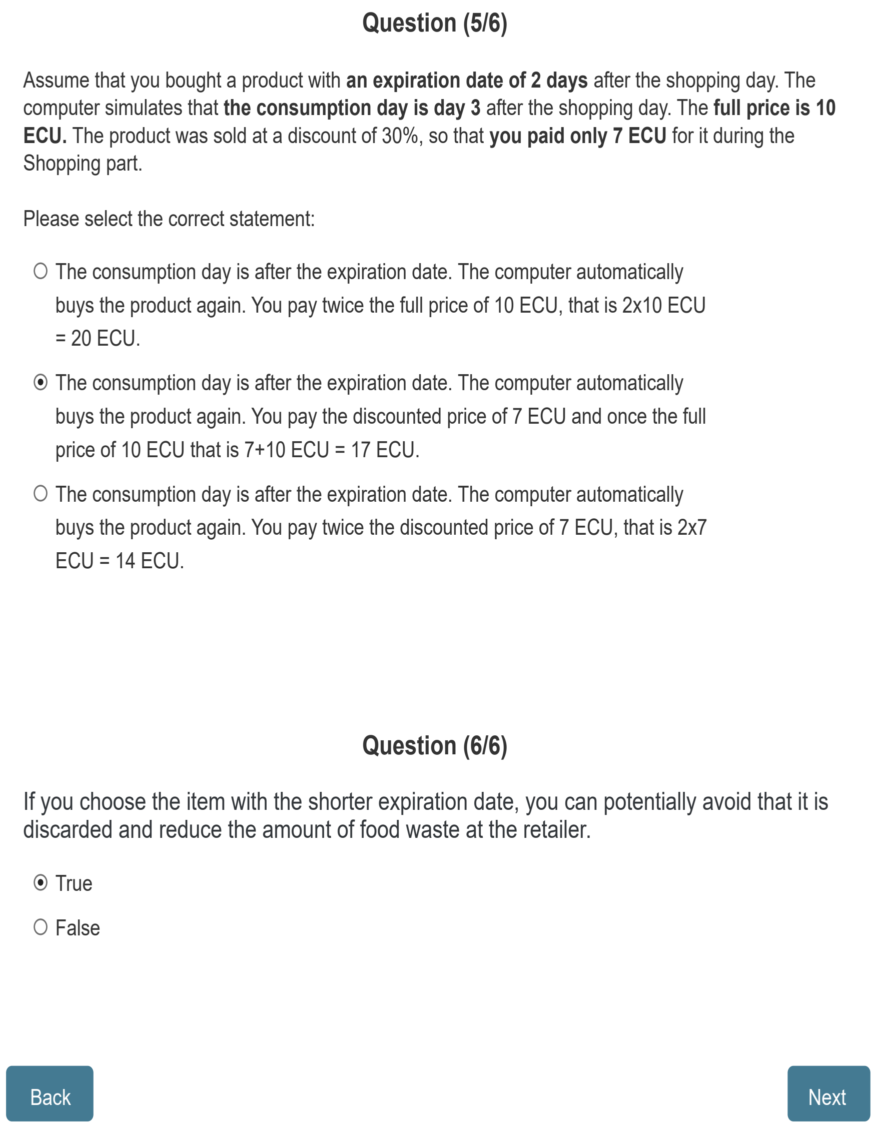
Assume that you bought a product with **an expiration date of 2 days** after the shopping day. The computer simulates that **the consumption day is day 3** after the shopping day. You **paid the full price** of **10 ECU**.

Please select the correct statement:

○ The consumption day is before the expiration date. The computer does not automatically buy the product again. You pay the full price of 10 ECU only once.

● The consumption day is after the expiration date. The computer automatically buys the product again. You pay twice the full price of 10 ECU, that is 2x10 ECU = 20 ECU.

Figure 4.12: Quiz Questions III



Question (5/6)

Assume that you bought a product with **an expiration date of 2 days** after the shopping day. The computer simulates that **the consumption day is day 3** after the shopping day. The **full price is 10 ECU.** The product was sold at a discount of 30%, so that **you paid only 7 ECU** for it during the Shopping part.

Please select the correct statement:

○ The consumption day is after the expiration date. The computer automatically buys the product again. You pay twice the full price of 10 ECU, that is 2x10 ECU = 20 ECU.

● The consumption day is after the expiration date. The computer automatically buys the product again. You pay the discounted price of 7 ECU and once the full price of 10 ECU that is 7+10 ECU = 17 ECU.

○ The consumption day is after the expiration date. The computer automatically buys the product again. You pay twice the discounted price of 7 ECU, that is 2x7 ECU = 14 ECU.

Question (6/6)

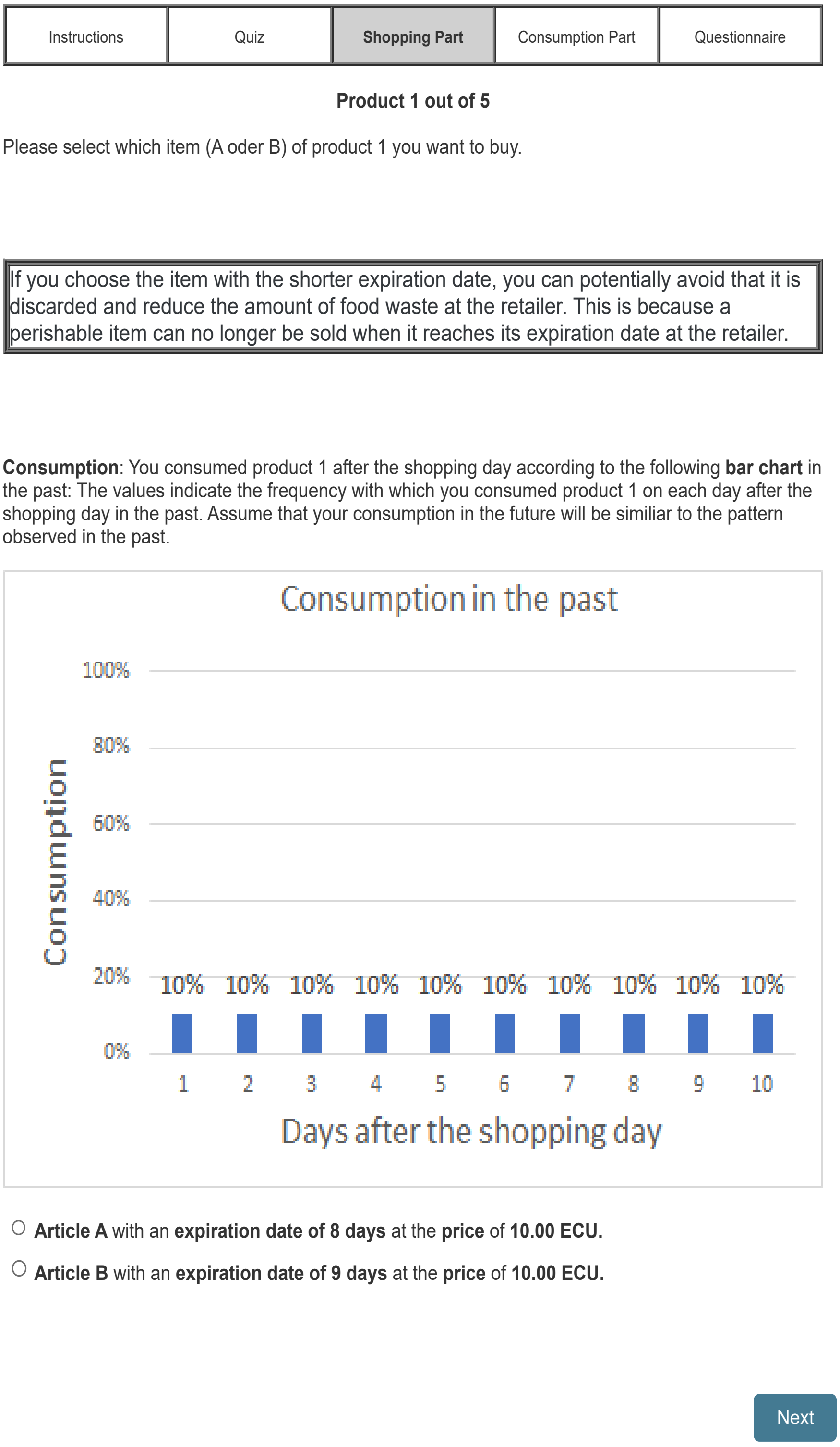
If you choose the item with the shorter expiration date, you can potentially avoid that it is discarded and reduce the amount of food waste at the retailer.

● True

○ False

4.C.3 Main Part

Figure 4.13: Shopping Part I: Decision Screen Without Discount



Product 1 out of 5

Please select which item (A oder B) of product 1 you want to buy.

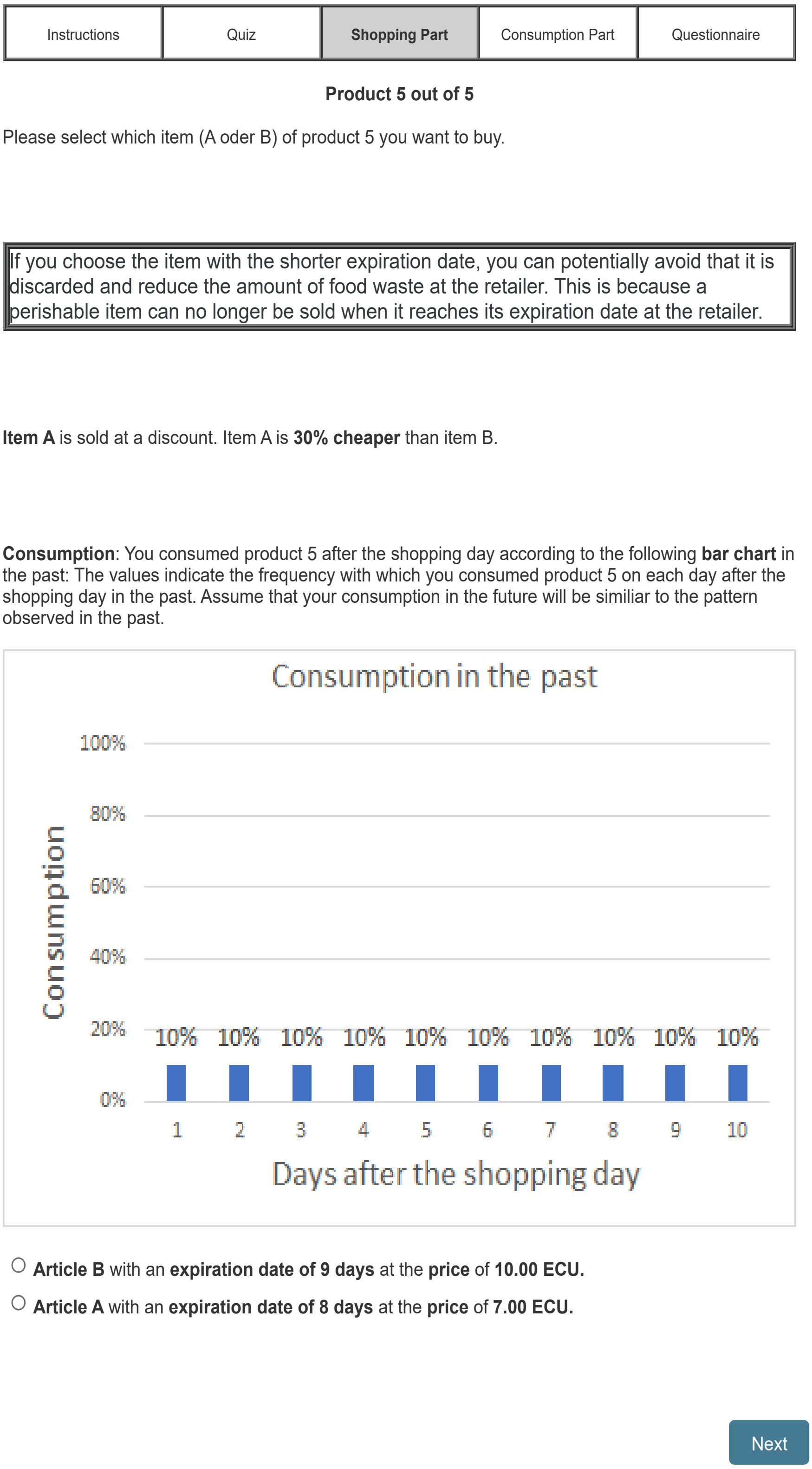
If you choose the item with the shorter expiration date, you can potentially avoid that it is discarded and reduce the amount of food waste at the retailer. This is because a perishable item can no longer be sold when it reaches its expiration date at the retailer.

Consumption: You consumed product 1 after the shopping day according to the following **bar chart** in the past: The values indicate the frequency with which you consumed product 1 on each day after the shopping day in the past. Assume that your consumption in the future will be similiar to the pattern observed in the past.

○ **Article A** with an **expiration date of 8 days** at the **price** of **10.00 ECU.**

○ **Article B** with an **expiration date of 9 days** at the **price** of **10.00 ECU.**

Figure 4.14: Shopping Part II: Decision Screen With Discount of 30%



Product 5 out of 5

Please select which item (A oder B) of product 5 you want to buy.

If you choose the item with the shorter expiration date, you can potentially avoid that it is discarded and reduce the amount of food waste at the retailer. This is because a perishable item can no longer be sold when it reaches its expiration date at the retailer.

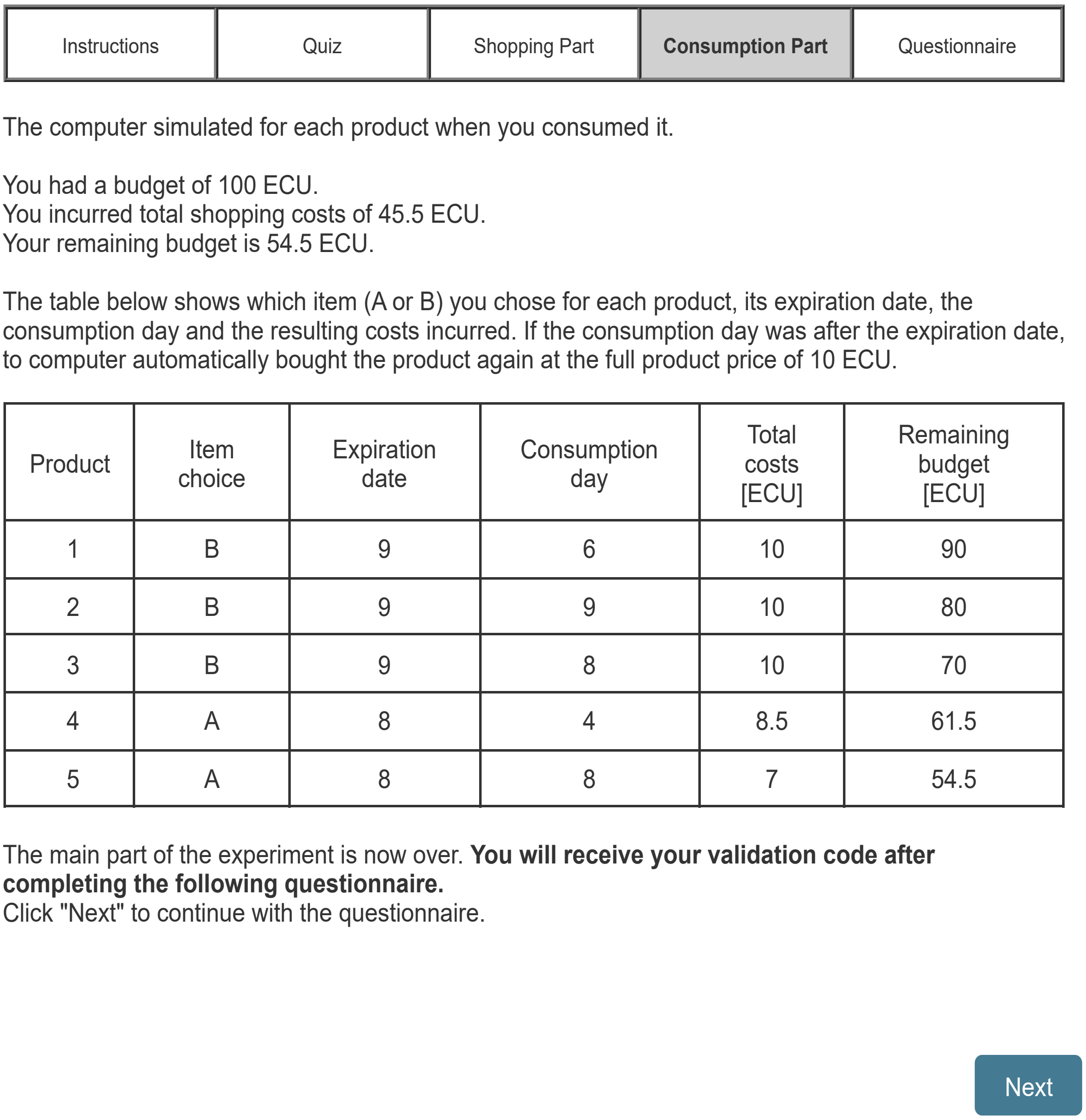
Item A is sold at a discount. Item A is **30% cheaper** than item B.

Consumption: You consumed product 5 after the shopping day according to the following **bar chart** in the past: The values indicate the frequency with which you consumed product 5 on each day after the shopping day in the past. Assume that your consumption in the future will be similiar to the pattern observed in the past.

○ **Article B** with an **expiration date of 9 days** at the **price** of **10.00 ECU.**

○ **Article A** with an **expiration date of 8 days** at the **price** of **7.00 ECU.**

Figure 4.15: Consumption Part



The computer simulated for each product when you consumed it.

You had a budget of 100 ECU.
You incurred total shopping costs of 45.5 ECU.
Your remaining budget is 54.5 ECU.

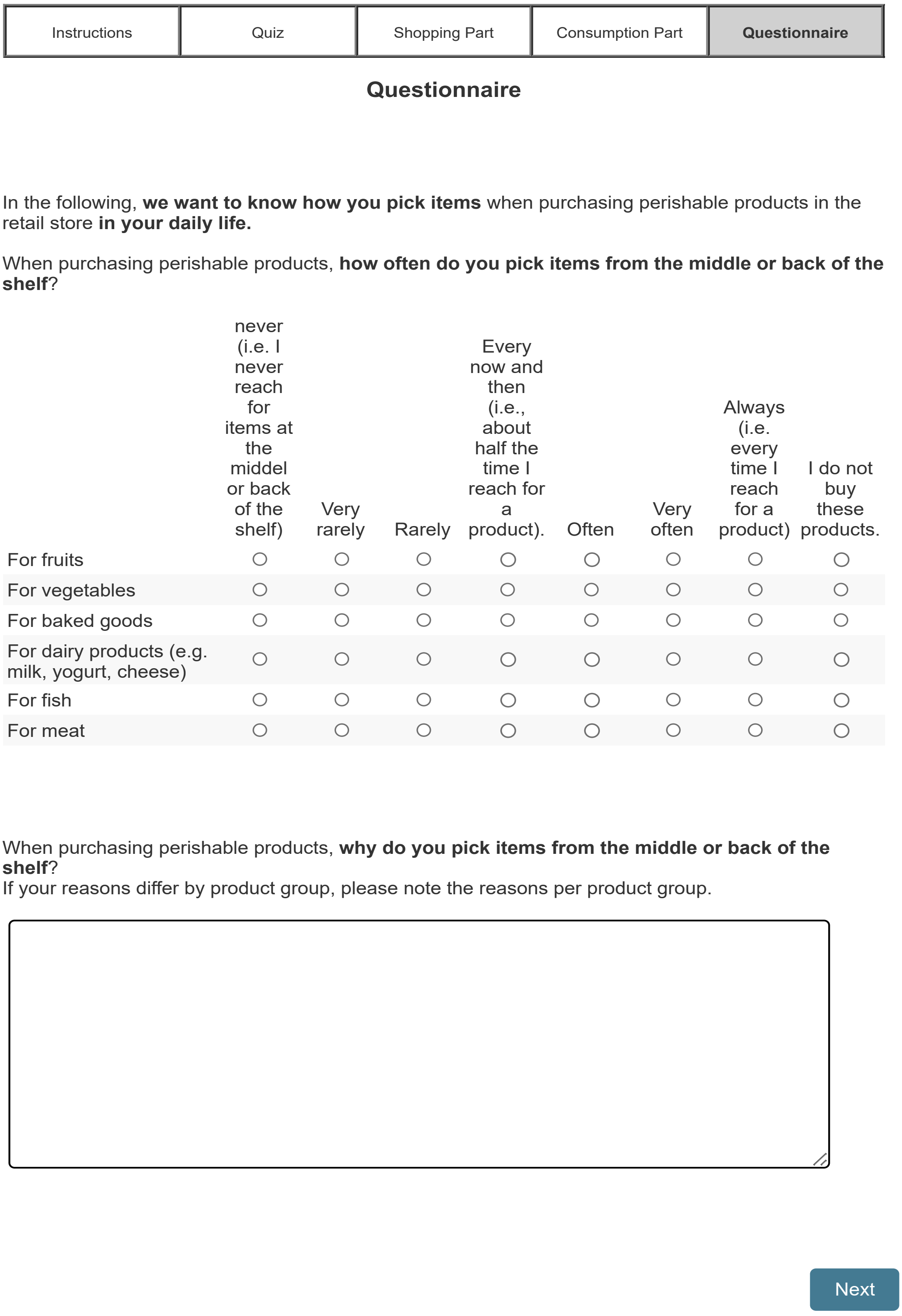
The table below shows which item (A or B) you chose for each product, its expiration date, the consumption day and the resulting costs incurred. If the consumption day was after the expiration date, to computer automatically bought the product again at the full product price of 10 ECU.

Product	Item choice	Expiration date	Consumption day	Total costs [ECU]	Remaining budget [ECU]
1	B	9	6	10	90
2	B	9	9	10	80
3	B	9	8	10	70
4	A	8	4	8.5	61.5
5	A	8	8	7	54.5

The main part of the experiment is now over. **You will receive your validation code after completing the following questionnaire.**
Click "Next" to continue with the questionnaire.

4.C.4 Questionnaire

Figure 4.16: Questions on Picking Behavior I



Questionnaire

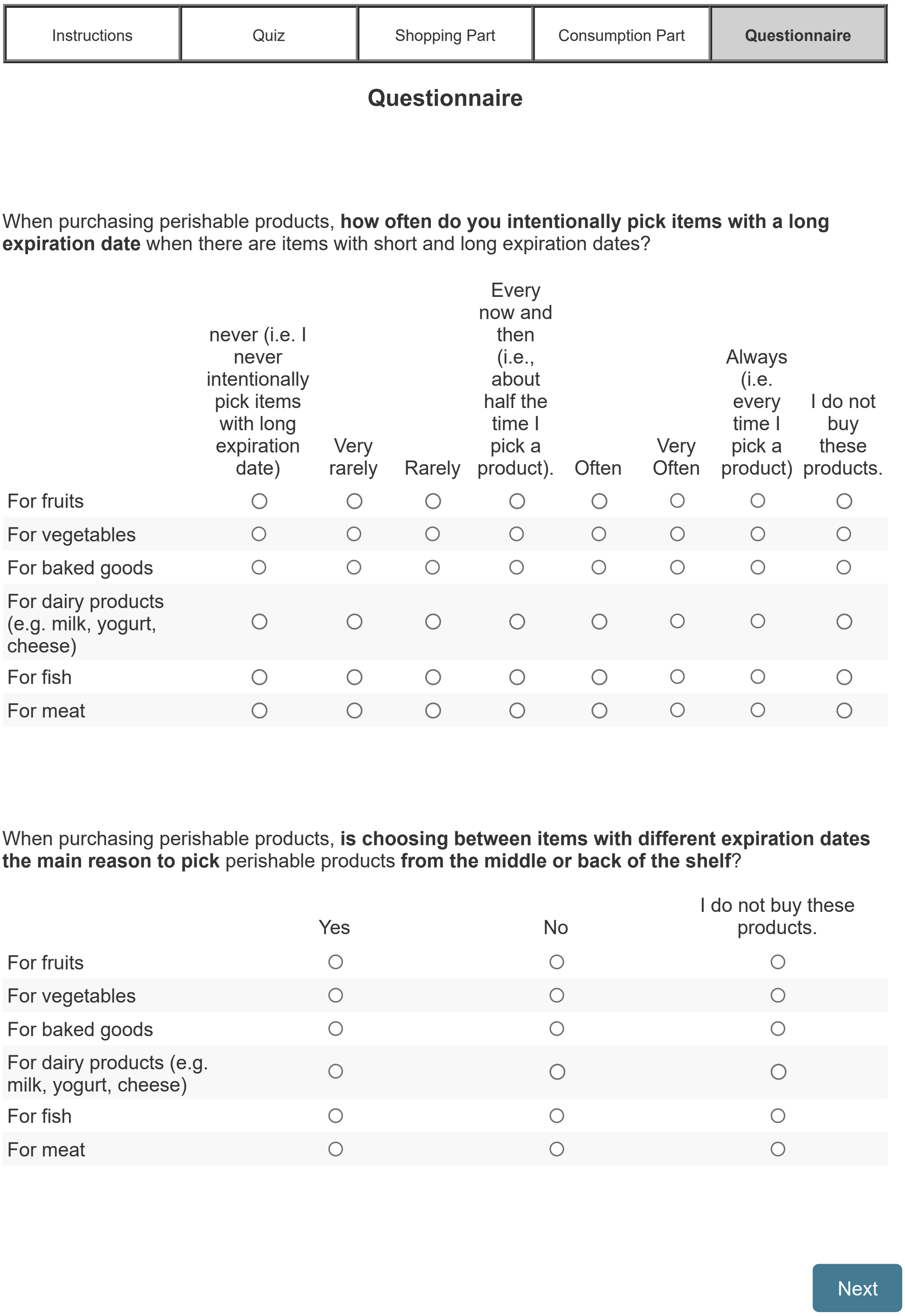
In the following, **we want to know how you pick items** when purchasing perishable products in the retail store **in your daily life.**

When purchasing perishable products, **how often do you pick items from the middle or back of the shelf**?

	never (i.e. I never reach for items at the middel or back of the shelf)	Very rarely	Rarely	Every now and then (i.e., about half the time I reach for a product).	Often	Very often	Always (i.e. every time I reach for a product)	I do not buy these products.
For fruits	○	○	○	○	○	○	○	○
For vegetables	○	○	○	○	○	○	○	○
For baked goods	○	○	○	○	○	○	○	○
For dairy products (e.g. milk, yogurt, cheese)	○	○	○	○	○	○	○	○
For fish	○	○	○	○	○	○	○	○
For meat	○	○	○	○	○	○	○	○

When purchasing perishable products, **why do you pick items from the middle or back of the shelf**?
If your reasons differ by product group, please note the reasons per product group.

Figure 4.17: Questions on Picking Behavior II



Questionnaire

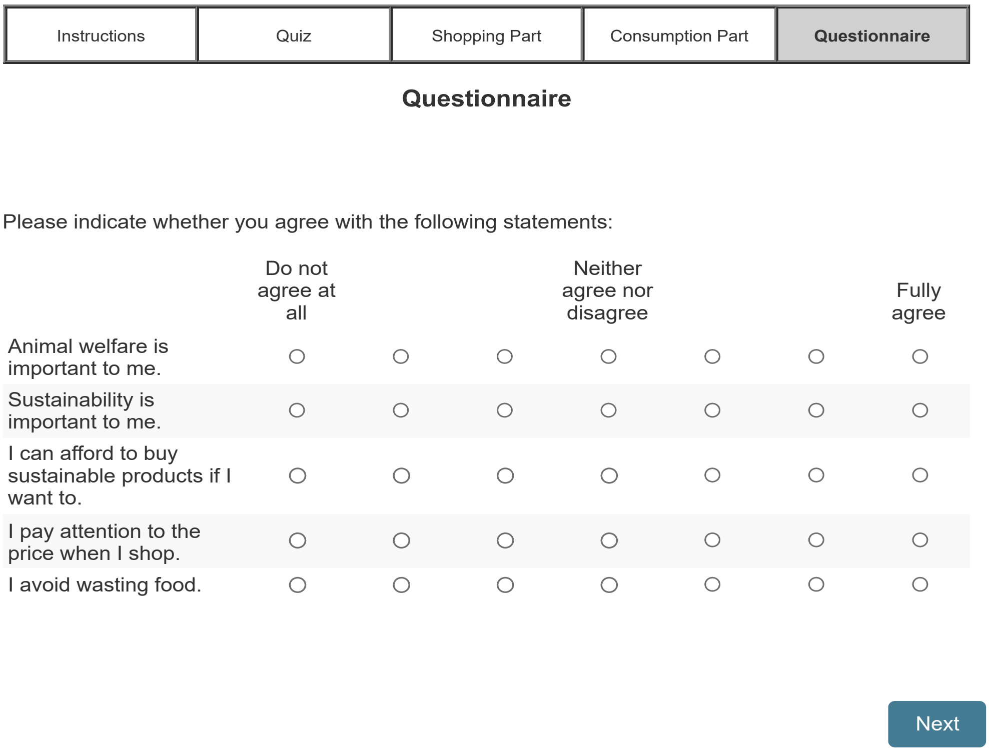
When purchasing perishable products, **how often do you intentionally pick items with a long expiration date** when there are items with short and long expiration dates?

	never (i.e. I never intentionally pick items with long expiration date)	Very rarely	Rarely	Every now and then (i.e., about half the time I pick a product).	Often	Very Often	Always (i.e. every time I pick a product)	I do not buy these products.
For fruits	○	○	○	○	○	○	○	○
For vegetables	○	○	○	○	○	○	○	○
For baked goods	○	○	○	○	○	○	○	○
For dairy products (e.g. milk, yogurt, cheese)	○	○	○	○	○	○	○	○
For fish	○	○	○	○	○	○	○	○
For meat	○	○	○	○	○	○	○	○

When purchasing perishable products, **is choosing between items with different expiration dates the main reason to pick** perishable products **from the middle or back of the shelf**?

	Yes	No	I do not buy these products.
For fruits	○	○	○
For vegetables	○	○	○
For baked goods	○	○	○
For dairy products (e.g. milk, yogurt, cheese)	○	○	○
For fish	○	○	○
For meat	○	○	○

Figure 4.18: Questions on Shopping and Sustainability Attitudes

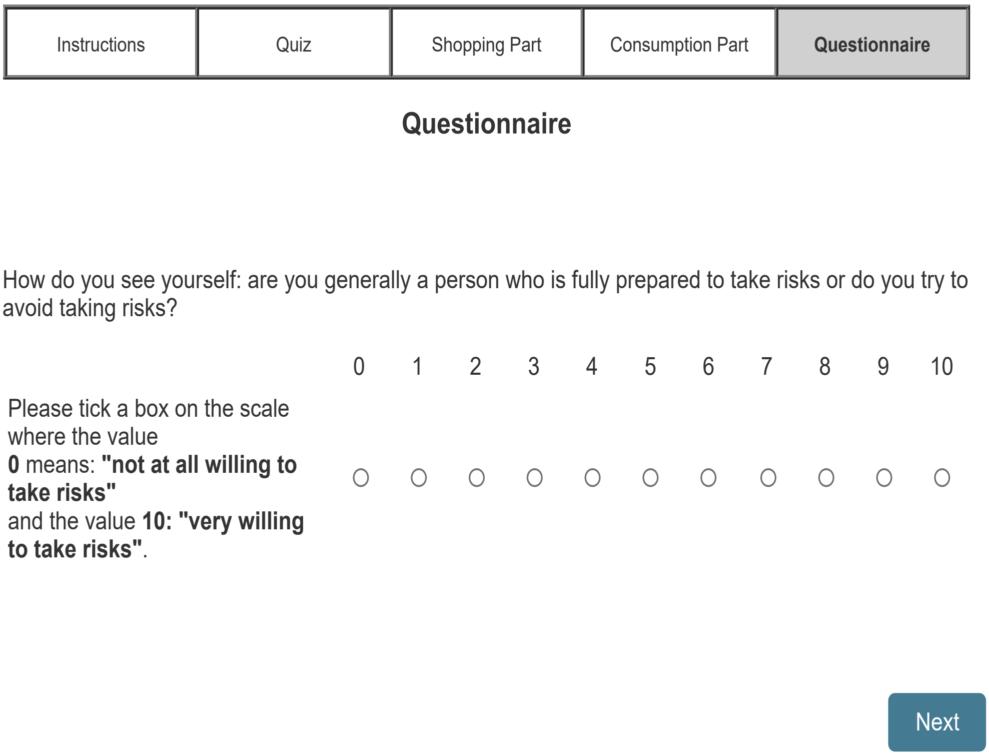


Questionnaire

Please indicate whether you agree with the following statements:

	Do not agree at all			Neither agree nor disagree			Fully agree
Animal welfare is important to me.	○	○	○	○	○	○	○
Sustainability is important to me.	○	○	○	○	○	○	○
I can afford to buy sustainable products if I want to.	○	○	○	○	○	○	○
I pay attention to the price when I shop.	○	○	○	○	○	○	○
I avoid wasting food.	○	○	○	○	○	○	○

Figure 4.19: Question on Risk Attitude

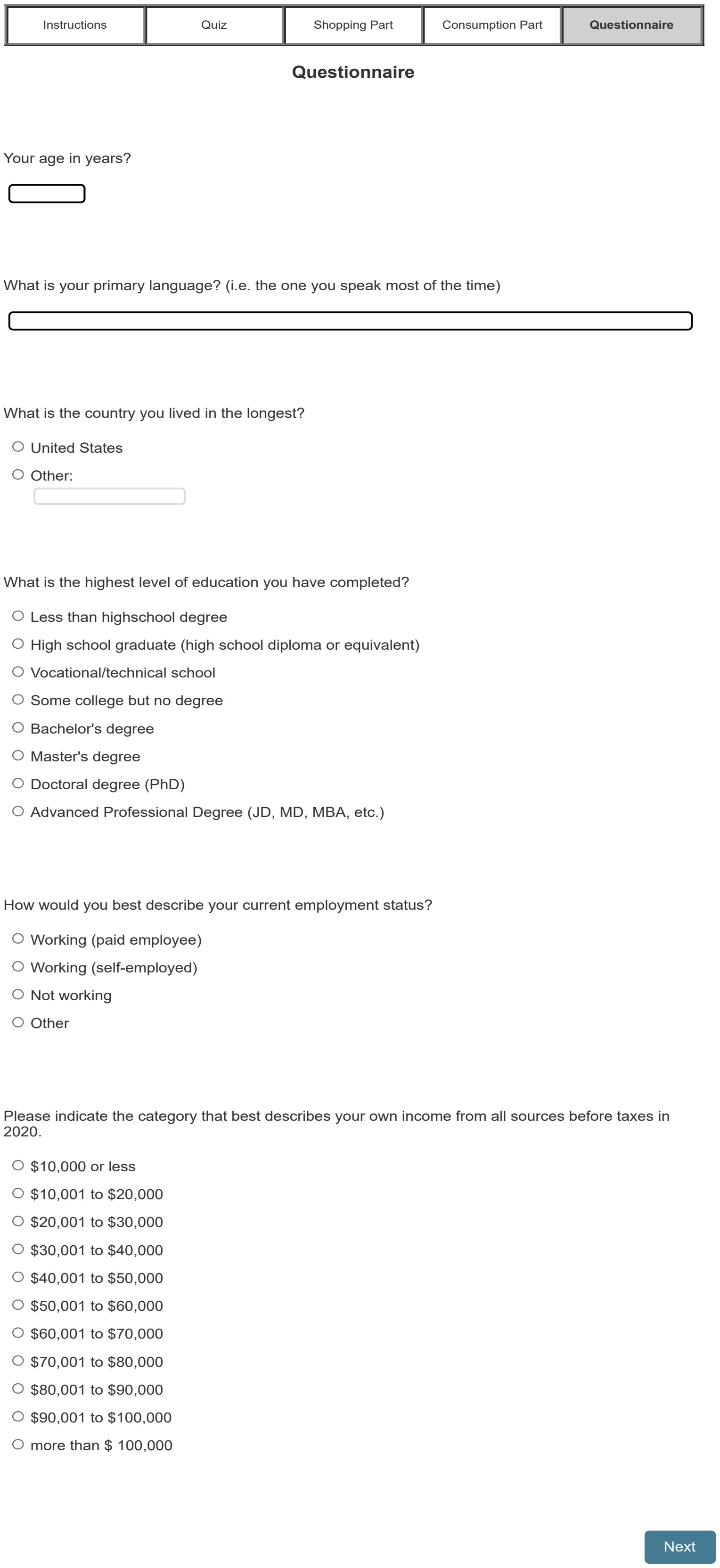


Questionnaire

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

	0	1	2	3	4	5	6	7	8	9	10
Please tick a box on the scale where the value **0** means: **"not at all willing to take risks"** and the value **10: "very willing to take risks"**.	○	○	○	○	○	○	○	○	○	○	○

Figure 4.20: Demographic Questions I



Questionnaire

Your age in years?

What is your primary language? (i.e. the one you speak most of the time)

What is the country you lived in the longest?

○ United States

○ Other:

What is the highest level of education you have completed?

○ Less than highschool degree

○ High school graduate (high school diploma or equivalent)

○ Vocational/technical school

○ Some college but no degree

○ Bachelor's degree

○ Master's degree

○ Doctoral degree (PhD)

○ Advanced Professional Degree (JD, MD, MBA, etc.)

How would you best describe your current employment status?

○ Working (paid employee)

○ Working (self-employed)

○ Not working

○ Other

Please indicate the category that best describes your own income from all sources before taxes in 2020.

○ $10,000 or less

○ $10,001 to $20,000

○ $20,001 to $30,000

○ $30,001 to $40,000

○ $40,001 to $50,000

○ $50,001 to $60,000

○ $60,001 to $70,000

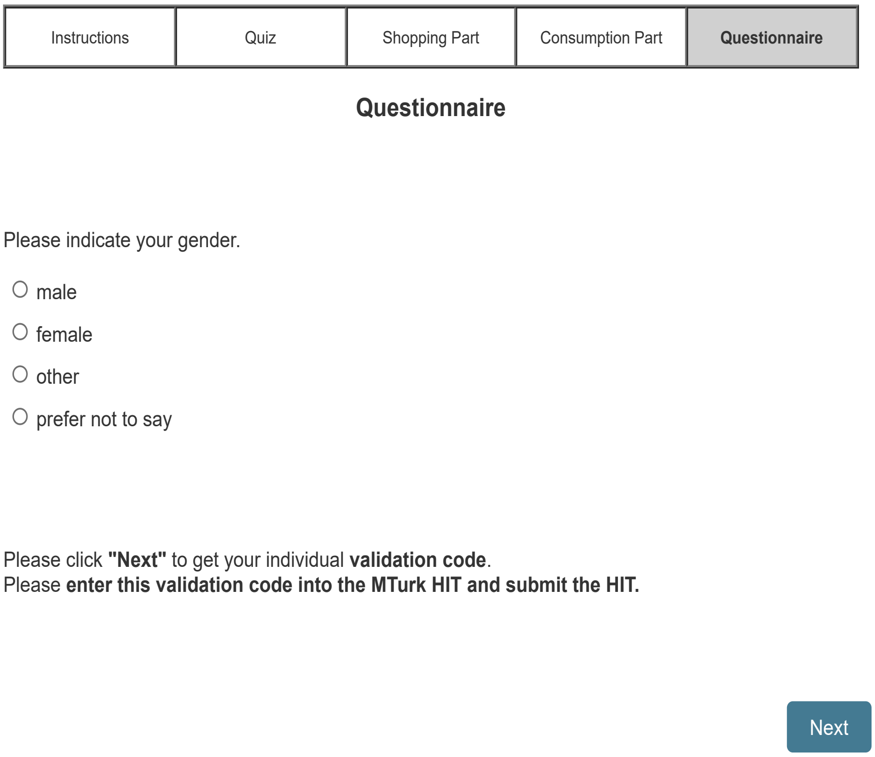
○ $70,001 to $80,000

○ $80,001 to $90,000

○ $90,001 to $100,000

○ more than $ 100,000

Figure 4.21: Demographic Questions II



Questionnaire

Please indicate your gender.

○ male

○ female

○ other

○ prefer not to say

Please click **"Next"** to get your individual **validation code**.
Please **enter this validation code into the MTurk HIT and submit the HIT.**

Chapter 5

Conclusion

This dissertation aims at improving the understanding how firms can design non-monetary and monetary incentives to affect behavior of employees and customers. In three research projects, we investigate how the incentive design of firms affects employee performance in short-term and long-term employer-employee relations (Chapters 2 and 3) as well as customer buying behavior (Chapter 4). In all three research projects, we designed and ran experiments to identify the causal effect of different incentive mechanisms on human behavior. For the research in Chapter 2, we developed a novel real effort task to test the performance effect of unused low rating categories. For the research in Chapter 4, we developed a behavioral model to describe how individual and aggregate buying behavior changes in response to non-monetary and monetary incentives.

Our research shows that firms can influence employee and customer behavior with their incentive design choice. We find that both the actual provision and the perception of non-monetary and monetary incentives can stimulate individual behavior that benefits the outcomes of firms. The (perceived) combination of both incentive types also induces behavior that increases the output of firms in our settings. Moreover, we observe that individuals react heterogeneously to either incentive type and the combination of both. We also see that the performance effect of incentive schemes depends on whether or not individuals are repeatedly exposed to them and can react dynamically. As a result, when designing incentive schemes firms may consider the composition of their employee and customer groups and whether or not these groups are repeatedly exposed to them.

In this chapter, we first summarize the key results of the three dissertation projects and then discuss directions for future research.

5.1 Key Results

In **Chapter 2 and Chapter 3**, we analyze whether firms should employ unused low rating categories in rating scales of performance appraisals.

In **Chapter 2**, we test the performance effect of an unused low category in rating scales in short-term employer-employee relations. We find that individuals who received the lowest possible rating category worked significantly more when they did not learn that the additional low rating category was unused. We do, however, not observe that an unused lower category in rating scales raised average performance in short-term employer-employee relations – independent of whether individuals learned or not that the respective category was unused. Furthermore, individuals perceived rating scales with an unused low rating category as unkind when they did not know that the respective category was unused. Moreover, they perceived rating scales with an unused low rating category as more kind when they knew that the respective category was unused. Overall, it seems that individuals do not only focus on their monetary incentives and personal performance rating but also consider the design and kindness of rating scales in short-term employer-employee relations. Low performing individuals seem to focus more on their monetary incentives and personal performance rating than the other individuals.

In **Chapter 3**, we analyze the performance effect of an unused low category in rating scales in long-term employer-employee relations, when employees receive multiple ratings and can react dynamically. We find that showing an unused low category in rating scales raised total performance by 20.92% when individuals did not learn that the respective category was unused. In line with the results from our investigation of short-term relations (Chapter 2), performance was not higher in the first period. However, over time individuals worked increasingly more when being evaluated on a scale with an unused low rating category. In line with the results of Chapter 2, we observe a stronger response of low performers as the performance effect was driven by these individuals. As in Chapter 2, the presence of an unused low rating category affected the perceived kindness of rating scales. However, this was only the case when individuals knew all rating scales and thereby had a reference point for the evaluation of a rating scale. Moreover, performance was not significantly higher when individuals learned that the low rating category in the rating scale was unused. Overall, the results suggest that it may be valuable for firms to employ low rating categories in rating scales in long-term employer-employee relations even when these categories are never used. Individuals' monetary incentives and personal performance ratings seem to be more important than the kindness of rating scales when individuals receive ratings repeatedly over time. Our analysis also suggests that (perceived) higher incentives induced by an unused low rating category drive performance results rather than reciprocal reactions to higher personal performance ratings. This mechanism seems to be especially strong for low performing individuals as they showed a stronger performance increase than the other individuals, which is in line with the results of the investigation of short-term employer-employee relations (Chapter 2).

In **Chapter 4**, we analyze how sustainability messages, price discounts, and the combination of both induce customers to buy earlier expiring items of perishable products. We find that displaying sustainability messages and offering price discounts induced purchases of earlier expiring items. As predicted by our behavioral model, we observe different customer types that showed heterogeneous reactions to price discounts: Some individuals did not change their behavior while others switched to buying earlier expiring items or crowded out when receiving price discounts. Moreover, we find that the positive effect on purchases of earlier expiring items in the presence of sustainability messages was driven by different behavior within customer types and different shares of individuals across these types. Overall, our results suggest that retailers can utilize sustainability messages and price discounts to increase purchases of earlier expiring items and thereby reduce food waste at grocery stores. Additionally, retailers who know how sustainability messages and price discounts affect the purchase behavior of different customer types can leverage this information with customized promotions to increase purchases of earlier expiring items.

5.2 Critical Review and Future Research

In all research projects that constitute this dissertation, we used controlled experimental conditions to test the effect of specific incentive mechanisms on human behavior: For the research in Chapters 2 and 3, we ran two controlled online field experiments and a controlled laboratory experiment, respectively to test the effect of unused low rating categories in performance appraisals on employee performance. For the research in Chapter 4, we conducted a controlled online experiment to analyze the effect of sustainability messages and price discounts on customer buying behavior. While this approach allows to isolate and clearly identify the influence of the incentive mechanisms of interest, it excludes by design the influences of other potential aspects that are most likely also present in firm settings. It is thus an interesting area of future research to complement our findings by testing the effect of dummy rating categories in performance appraisals as well as sustainability messages and price discounts on human behavior in other settings such as firms or other organizations. This enhances the external validity of our results and the understanding which other influences are in place in other environments and how these interact with the ones present in our settings.

In **Chapters 2 and 3**, we analyze the effect of one dummy category on employee performance. We choose a binary design – dummy category yes / no – to identify the influence of a dummy category that is not dependent on the number of dummy categories. However, firms also employ rating scales that contain more than one dummy category. Hence, it would be valuable to test whether or not the number of dummy categories also affect employee performance.

Another interesting field of research is whether our results hold across different types of tasks. Firms use performance appraisals for all types of tasks but we tested the performance effect of dummy categories in tedious tasks that demand low cognitive effort. Using this type of task allows to assume that intrinsic motivation does not significantly affect performance in our setting. However, it is not clear whether the same effect of feedback on performance would be observed for cognitively-demanding tasks. Accordingly, we encourage future research to test the effect of dummy categories on employee performance with experiments where individuals work on cognitively-demanding tasks.

In **Chapter 4**, we ran an online experiment that mimics the inventory decision customers face when purchasing perishable products. The online environment of the experiment might create a situation similar to online shopping. When shopping in grocery stores, however, customers decision to purchase earlier expiring items probably also depends on situational factors such as the position of a product within a store and on the shelf, or whether or not other customers are present. Moreover, to identify a general decision pattern independent of specific product types, we intentionally did not specify for which type of perishable products individuals made purchase decisions. However, customers might decide differently between buying earlier expiring or longer lasting items of for example meat or vegetables and thus dependent on the type of perishable products. By design, we also did not analyze how sustainability messages and price discounts affect customer buying behavior over long time frames or across product categories. There might be a spill-over effect that could be relevant for demand planning and inventory management in grocery stores. As a result, we welcome future research to analyze how the aforementioned aspects affect customer buying behavior and how they interact with the general decision pattern we observed in our research.

Finally, we want to direct future research to two experimental design challenges that we find applicable and relevant for all experimental research in (behavioral) management science and (behavioral) economics. In Chapters 2 and 3, individuals were not informed about all possible rating scales when they worked on the task since we believe this is most representative for the situation in firms. Thereby potential reciprocal reactions to the design of rating scales might be less pronounced since individuals had no reference point to judge whether a given scale is more or less kind. This is one of numerous examples for situations in which empirical researchers face trade-offs between implementing experimental settings that approximate institutional settings and implementing settings that allow to test all theoretically possible mechanisms. We leave the discussion about how this trade-off decision should be optimally made to future research. Moreover, in Chapter 4, individuals' decision did not affect actual food waste at a retailer. We chose this experimental design since we did not find a proper replicate of food waste or saving at a retailer for our experimental setting. To circumvent the risk of potentially testing another mechanism, we decided that individuals' decision had no effect on actual food waste at a retailer. In our experience, this is a common challenge researchers face when designing laboratory or online experiments to test the influence of non-monetary mechanisms. We invite future research to discuss on how to decide whether or not to include potentially inaccurate mechanisms in such settings.

References

Adams, J. S. (1963). Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, 67(5):422–436.

Aguinis, H., Villamor, I., and Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, 47(4):823–837.

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

Akerlof, G. A. and Yellen, J. L. (1988). Fairness and unemployment. *American Economic Review*, 78(2):44–49.

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

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–55.

Auger, P. and Devinney, T. M. (2007). Do what consumers say matter? the misalignment of preferences with unconstrained ethical intentions. *Journal of Business Ethics*, 76(4):361–383.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7):435–452.

Bandiera, O., Barankay, I., and Rasul, I. (2011). Field experiments with firms. *The Journal of Economic Perspectives*, 25(3):63–82.

Barankay, I. (2011). Rankings and social tournaments: Evidence from a crowd-sourcing experiment. *Working Paper, University of Pennsylvania, Philadelphia.*

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. *Working Paper, Wharton School, University of Pennsylvania, Philadelphia.*

Bastiaansen, G. (2019). Implementation of expiration date visibility at jumbo supermarkten. Technical report, Eindhoven University of Technology, Eindhoven.

Becker-Peth, M., Katok, E., and Thonemann, U. W. (2013). Designing buyback contracts for irrational but predictable newsvendors. *Management Science*, 59(8):1800–1816.

Bellemare, C. and Sebald, A. (2019). Self-confidence and reactions to subjective performance evaluations. *IZA Discussion Paper No. 12215*.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3):489–520.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Berger, J., Harbring, C., and Sliwka, D. (2013). Performance appraisals and the impact of forced distribution–an experimental investigation. *Management Science*, 59(1):54–68.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *Accounting Review*, 86(5):1549–1575.

Bolton, G. E. and Ockenfels, A. (2012). Behavioral economic engineering. *Journal of Economic Psychology*, 33(3):665–676.

Bolton, G. E., Ockenfels, A., and Thonemann, U. W. (2012). Managers and students as newsvendors. *Management Science*, 58(12):2225–2233.

Bowles, S. and Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Bretz, R. D. J., Milkovich, G. T., and Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, 18(2):321–352.

Broekmeulen, R. A. C. M. and Bakx, C. H. M. (2010). *In-store replenishment procedures for perishable inventory in a retail environment with handling costs and storage constraints*. BETA publicatie : working papers. Technische Universiteit Eindhoven.

Broekmeulen, R. A. C. M. and Van Donselaar, K. H. (2009). A heuristic to manage perishable inventory with batch ordering, positive lead-times, and time-varying demand. *Computers & Operations Research*, 36(11):3013–3018.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data. *Perspectives on Psychological Science*, 6(1):3–5.

Chandler, J., Mueller, P., and Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1):112–130.

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

Chua, G. A., Mokhlesi, R., and Sainathan, A. (2017). Optimal discounting and replenishment policies for perishable products. *International Journal of Production Economics*, 186:8–20.

Chung, D. J. and Narayandas, D. (2017). Incentives versus reciprocity: Insights from a field experiment. *Journal of Marketing Research*, 54(4):511–524.

Clarkson, M. E. (1995). A stakeholder framework for analyzing and evaluating corporate social performance. *Academy of Management Review*, 20(1):92–117.

Commission, E. (2008). Special eurobarometer 295. attitudes of european citizens toward the environment. Brussels, Belgium.

Commission, E. (2017). Special eurobarometer 468. attitudes of european citizens toward the environment. Brussels, Belgium.

Condly, S. J., Clark, R. E., and Stolovitch, H. D. (2003). The effects of incentives on workplace performance: A meta-analytic review of research studies. *Performance improvement quarterly*, 16(3):46–63.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Devinney, T. M., Auger, P., and Eckhardt, G. M. (2010). *The myth of the ethical consumer*. Cambridge University Press.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Ellingsen, T. and Johannesson, M. (2007). Paying respect. *Journal of Economic Perspectives*, 21(4):135–149.

Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness-intentions matter. *Games and Economic Behavior*, 62(1):287–303.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Fehr, E. and Fischbacher, U. (2002). Why social preferences matter – the impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, 112(478):C1–C33.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E., Gachter, S., and Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 65(4):833.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459.

Fehr, E., Klein, A., and Schmidt, K. M. (2007). Fairness and contract design. *Econometrica*, 75(1):121–154.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–140.

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

Fehr, E. and Schmidt, K. M. (2003). Theories of fairness and reciprocity – evidence and economic applications. In Dewatripont, M., Hansen, L. P., and Turnovsky, S. J., editors, *Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs*, volume I, pages 208–257. Cambridge University Press, Cambridge.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Food Safety and Inspection Service U.S. Department of Agriculture (2019). Food product dating. `https://www.fsis.usda.gov/food-safety/safe-food-handling-and-preparation/food-safety-basics/food-product-dating`. Accessed: 2022-10-11.

Frederiksen, A., Lange, F., and Kriechel, B. (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior and Organization*, 134:408–429.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, 87(4):746–755.

Garlick, R. (2022). Incentive marketplace estimate research study. Technical report, Incentive Federation.

Gill, D., Kissova, Z., Lee, J., and Prowse, V. L. (2019). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*, 65(2):494–507.

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

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht*, volume 63, pages 79–93. Göttingen, Germany.

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

Hoffmann, C. and Thommes, K. (2020). Can digital feedback increase employee performance and energy efficiency in firms? evidence from a field experiment. *Journal of Economic Behavior and Organization*, 180:49–65.

Holland, K. (2006). Performance reviews: Many need improvement. `http://www.nytimes.com/2006/09/10/business/yourmoney/10mgmt.html`. Accessed: 2022-10-11.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91.

Horton, J. J. (2010). Employer expectations, peer effects and productivity: Evidence from a series of field experiments. *arXiv preprint arXiv:1008.2437*.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory:

Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.

Infratest Sozialforschung, T. N. S. (2012). SOEP 2005 - Erhebungsinstrumente 2005 (Welle 22) des Sozio-oekonomischen Panels. *SOEP Survey Papers*, (103: Series A.Berlin: DIW/SOEP).

Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk Marketplace. *XRDS*, 17(2):16–21.

Köszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature*, 52(4):1075–1118.

Kube, S., Maréchal, M. A., and Puppe, C. (2012). The currency of reciprocity - gift-exchange in the workplace. *American Economic Review*, 102(4):1644–1662.

Landy, F. J. and Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87(1):72–107.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Lee, Y. S., Seo, Y. W., and Siemsen, E. (2018). Running behavioral operations experiments using Amazon's Mechanical Turk. *Production and Operations Management*, 27(5):973–989.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.

Levitt, S. D. and Neckermann, S. (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy*, 30(4):639–657.

List, J. and Rasul, I. (2011). Field experiments in labor economics. volume 4A, chapter 02, pages 103–228. Elsevier, 1 edition.

Litman, L., Robinson, J., and Abberbock, T. (2017). Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2):433–442.

Mellström, C. and Johannesson, M. (2008). Crowding out in blood donation: was titmuss right? *Journal of the European Economic Association*, 6(4):845–863.

Noleppa, S. and Cartsburg, M. (2015). *Das grosse Wegschmeissen: vom Acker bis zum Verbraucher: Ausmaß und Umwelteffekte der Lebensmittelverschwendung in Deutschland*. WWF Deutschland.

Ockenfels, A., Sliwka, D., and Werner, P. (2015). Bonus payments and reference point violations. *Management Science*, 61(7):1496–1513.

Oswald, A. J., Proto, E., and Sgroi, D. (2015). Happiness and productivity. *Journal of Labor Economics*, 33(4):789–822.

Paolacci, G., Chandler, J., and Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, 5(5):411–419.

Papier, F. and Thonemann, U. W. (2021). Effect of social preferences on sales and operations planning. *Operations Research*, 69(5):1368–1395.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72(6):407–418.

Prendergast, C. J. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302.

Robinson, J., Rosenzweig, C., Moss, A. J., and Litman, L. (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PLoS ONE*, 14(12):1–29.

Schmidt, K. M. (2017). Contributions of Oliver Hart and Bengt Holmström to contract theory. *The Scandinavian Journal of Economics*, 119(3):489–511.

Sebald, A. and Walzl, M. (2014). Subjective performance evaluations and reciprocity in principal-agent relations. *Scandinavian Journal of Economics*, 116(2):570–590.

Shukla, P., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Pörtner, H., Roberts, D., Zhai, P., Slade, R., Connors, S., Van Diemen, R., et al. (2019). Ipcc, 2019: Climate change and land: an ipcc special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.

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

Smith, S. A. and Agrawal, N. (2017). Optimal markdown pricing and inventory allocation for retail chains with inventory dependent demand. *Manufacturing & Service Operations Management*, 19(2):290–304.

Titmuss, R. M. (1970). The gift relationship: From human blood to social policy. london: Allen and unwin.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

Trudel, R. and Cotte, J. (2009). Does it pay to be good? *MIT Sloan Management Review*, 50(2):61.

U.S. Census Bureau (2015). 2015 management and organizational practices survey. `https://www.census.gov/data/tables/2015/econ/mops/2015-survey-release.html`. Accessed: 2022-10-11.

Villeval, M. C. (2020). Performance feedback and peer effects. In Zimmermann, K. F., editor, *Handbook of Labor, Human Resources and Population Economics*, pages 1–38. Springer International Publishing, Cham.

Vogt, T. (2021). On rating scales in performance appraisals: Performance effects of an unused low rating category in short-term interactions. *Working Paper, University of Cologne, Germany.*

White, K., Habib, R., and Hardisty, D. J. (2019). How to shift consumer behaviors to be more sustainable: A literature review and guiding framework. *Journal of Marketing*, 83(3):22–49.