Technical University of Munich
TUM School of Management
Professorship of Managerial Economics

Three Essays on the Economics of Managerial Decision-Making

Magnus Emanuel Strobel

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Sebastian Goerg

Prüfer der Dissertation:
1. Prof. Dr. Michael Kurschilgen
2. Prof. Dr. Hanna Hottenrott

Die Dissertation wurde am 26.02.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Management am 15.03.2021 angenommen.

Abstract

This dissertation consists of three studies in which I contribute to the scientific literature on managerial decision-making.[1] Each study focuses on an important situation that managerial decision-makers regularly face. In the first study, I shed light on the selection of new employees by answering the question of how incentives to fake affect the predictive power of personality assessments. In the second study, I focus on the negotiation of wages by answering the question if employers wage-discriminate against applicants based on their signaled prosociality. Finally, in the third study, I analyze the enforcement of rule compliance by answering the question of whether an increase in monitoring reduces sabotage in contests.

[1]Please note that all studies in my dissertation are based on research that I jointly conducted with my co-authors. Hence, throughout this dissertation, whenever I refer to my research, "I" actually stands for "my co-authors and me". Of course, all remaining errors are mine.

Zusammenfassung

Diese Dissertation besteht aus drei Studien, mit denen ich einen Beitrag zur wissenschaftlichen Literatur über die Entscheidungsfindung von Managern leiste. Jede dieser Studien befasst sich mit einer wichtigen Situation, mit der Entscheidungsträger im Management regelmäßig konfrontiert werden. In der ersten Studie beschäftige ich mich mit der Auswahl neuer Mitarbeiter, indem ich die Frage beantworte, wie Anreize zur Täuschung die Vorhersagekraft von Methoden zur Persönlichkeitsbewertung beeinflussen. In der zweiten Studie befasse ich mich mit Gehaltsverhandlungen, indem ich die Frage beantworte, ob Arbeitgeber Bewerber aufgrund ihrer signalisierten Prosozialität in Bezug auf Gehälter diskriminieren. In der dritten Studie analysiere ich die Durchsetzung der Einhaltung von Regeln, indem ich die Frage beantworte, ob eine Erhöhung der Überwachung zu einer Reduktion von Sabotage in Wettbewerben führt.

Acknowledgement

Throughout my PhD studies, I have received a great deal of support and assistance, for which I am very grateful.

First and foremost, I would like to express my deepest gratitude to my scientific supervisor and friend, Michael Kurschilgen, who successfully guided me throughout all phases of this dissertation. Michael, thank you for believing in my potential, the countless hours you invested in our research projects, and the many learning opportunities that I was allowed to experience on this journey. You are a true role model, not just scientifically but also on a personal level.

I am very grateful to Hanna Hottenrott for being my second supervisor and mentor. Hanna, thank you for your in-depth scientific guidance, encouragement, and general support along the way. I still remember with joy how you surprised me with a little refreshment package after I taught my first lecture in the main auditorium. It was a great pleasure to enjoy your amazing support on this journey.

Additionally, I sincerely thank my great co-authors, Thomas Daske, Julian Hackinger, Julia Holzmann, and Michael Kurschilgen, for their valuable contributions to our common research projects.

I would also like to thank all my colleagues and friends with whom I had the pleasure to interact with. Thank you for your support on this journey and all the many memorable moments during our time together at TUM.

My sincere thanks go to all our student research assistants for their valuable support. In particular, I would like to thank Julian Schwierzy and Christoph Kecht for their dedicated help on my research projects.

I deeply thank Jeffrey Hancock (Stanford University), Drazen Prelec and Danica Mijovic-Prelec (Massachusetts Institute of Technology), and Colin Camerer (California Institute of Technology) for giving me the great opportunity to conduct research stays at their respective universities and thereby highly enriching my academic and personal development.

I am forever thankful to all my friends who supported me at different stages during my PhD. A special thanks goes to Peter Schmidt and Lars Hornbaker for their valuable help and contributions.

Most importantly, I truly thank Olivia Köhler for her unconditional love and continuous support. Without you, none of this would have been possible. Words cannot describe how thankful I am to have you by my side. It has been a wonderful journey so far – and I am fully convinced that with you by my side, the best is yet to come.

I am deeply thankful to Benedikt and Elias Strobel. Thank you for being the best possible brothers by always pushing me above and beyond my limits and gently picking me up when I fall. Lastly, I would like to express my deepest gratitude to my parents Claudia and Bruno Strobel. Thank you for your unconditional love, for encouraging and supporting me in following my dreams, and for making me who I am.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Research questions

In this dissertation, I study the economics of managerial decision-making. According to Bazerman and Moore (2009), managerial decision-making is the process of defining the problem, identifying and weighing the criteria, generating alternatives, rating each alternative on each criterion, and computing the optimal decision. While a plethora of information sources exist that provide practical guidance on how to make optimal decisions, the complexity of human behavior often causes allegedly optimal actions to not result in the desired outcome. For instance, fines issued for late pickups in a kindergarten may increase rather than decrease the number of parents who come late to pick up their children (Gneezy and Rustichini, 2000). Such adverse effects do not only result from fines but can also result from rewards. For instance, when schools are rewarded on student test scores, the result may be that teachers cheat more to improve their pupils' test results (Jacob and Levitt, 2003). To better understand how to account for the "human factor" in optimal decision-making, I study three important situations that managerial decision-makers regularly face, namely the selection of new employees, the negotiation of wages, and the enforcement of rule compliance.

First, I study a situation in which managerial decision-makers try to identify the most suitable candidate among all applicants for an advertised job vacancy. To evaluate their suitability, managerial decision-makers typically use self-reported personality tests to assess the personality traits of job applicants. This approach proves to be highly predictive in the absence of incentives to fake (see Ozer and Benet-Martínez, 2006, for a review on the predictiveness of the Big Five on many different outcomes). However, in the presence of incentives to fake, this approach fails to predict the suitability of job applicants (see Viswesvaran and Ones, 1999, for a meta-analysis on the fakability of personality inventories). For instance, when applicants know which personality traits are desired by the employer, they exaggerate these personality traits in a self-serving way, making the results of the personality test useless. To better understand how personality traits can be validly assessed in the presence of incentives to fake, I conducted my first study, which is presented in Chapter 2. In this study, I examine the research question of how incentives to fake affect the predictive power of personality assessments.

Second, I shed light on a situation in which managerial decision-makers try to make applicants optimal wage offers. From the perspective of rational employers, an optimal wage offer corresponds to the reservation wage of the applicant. While underpaying leads to the rejection of the offer, overpaying is equally inefficient as the applicant would have also accepted the offer for a lower wage. To estimate applicants' reservation wages, employers can use different signals. For instance, prosociality is commonly associated with a preference for meaning over money. This is because applicants who engaged in prosocial activities, such as volunteering, deliberately missed out on monetary compensation. Hence, applicants who signal prosociality should have lower reservation wages which should translate into lower wage offers from rational employers. To investigate whether this is the case, I conducted my second study, which is presented in Chapter 3. In this study, I examine the research question of whether employers wage-discriminate against applicants based on their signaled prosociality.

Third, I research a situation in which managerial decision-makers try to prevent sabotage among employees. Sabotage among employees can be caused by contests, such as promotions for higher positions. Contests are a widespread way to incentivize employees to exert productive effort. However, in contests, contestants have two strategies to be successful, namely by exerting productive effort or by sabotaging others (Lazear, 1989). To prevent sabotage among employees, managerial decision-makers typically resort to deterrence, such as an increase in monitoring. To analyze the effectiveness of this measure, I conducted my third study, which is presented in Chapter 4. In this study, I examine the research question of whether an increase in monitoring reduces sabotage.

1.2 Data and methodology

To answer my research questions, I conducted three separate studies using different kinds of data and methodologies.

In Chapter 2, I used data from an online experiment to study the question of how incentives to fake affect the predictive power of personality assessments. More precisely, I study assessments of job applicants' cooperativeness. In the online experiment, I first elicited subjects' cooperativeness before they performed additional personality tests. As a treatment manipulation, I varied the incentives to fake being cooperative between subjects. Based on the data from this experiment, I compare the predictive power of two different approaches for assessing subjects' cooperativeness. First, I study the established approach of assessments based on personality (Big Five) scores obtained from self-reported personality tests. Second, I study the novel approach of assessments based on linguistic (LIWC) scores obtained from written self-descriptions. To make the cooperativeness predictions, I used state-of-the-art machine learning techniques.

In Chapter 3, I used data from a field and laboratory experiment to study the question of whether employers wage-discriminate against applicants based on their signaled prosociality. In the field experiment, which was a correspondence study, I sent out résumés with questionnaires to human resources (HR) managers in Germany. On the résumés, I varied the prosociality of the fictitious applicants by including different work experiences. Each HR manager received one of four possible résumés. In the attached questionnaire, HR managers were asked to state a hypothetical wage offer for the applicant and to estimate the applicant's reservation wage. In the laboratory experiment, subjects played an ultimatum game in which the proposers received a signal on the prosociality of their matched responder before they decided on their offer. The prosociality signal was the responder's donation amount to a charity in a preceding dictator game. Methodology-wise, I used regression analyses and non-parametric tests to analyze the data from the experiments.

In Chapter 4, I used publicly available data from professional soccer to study the question of whether an increase in monitoring reduces sabotage. To gather the data, I used automated web-scraping. Since sabotage is typically forbidden, it is difficult to observe it in the field. Therefore, sports contests, in which rule-breaking can be used as a proxy for sabotage, provide a suitable alternative (Chowdhury and Gürtler, 2015). In this study, I introduce a novel proxy for sabotage in sports, namely substitutions that are due to an in-match injury. In contrast to existing proxies for sabotage, this measure offers the advantage that it allows to directly capture the deterrence effect. To analyze my research question, I took advantage of a quasi-natural experiment caused by the introduction of the Video Assistant Referee (VAR) in the 1. Bundesliga but not in the 2. Bundesliga. Methodology-wise, I used this rule change to study the effect of an increase in monitoring with a difference-in-differences approach.

1.3 Related literature

Each study in this dissertation relates to different strands of the scientific literature. In particular, they are embedded in the existing literature as follows.

The study in Chapter 2 relates to the scientific literature on personality assessments. Personality is defined as "[...] the consistent set of traits, attitudes, emotions, and behaviors that people have." (Boyd and Pennebaker, 2017, p. 63). While there are many different theories that aim to provide a taxonomy for personality, the most widely accepted theory is the so-called trait approach. It assumes "[...] that our cognitions, emotions, and behaviors are determined by a number of consistent and relatively stable traits." (Matz et al., 2016, p. 36). The most influential and widely accepted trait theory is the Big Five model (Goldberg, 1990). It posits that the personality traits of Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N) fully subsume people's personality. The most common way to assess personality is by means of self-reported personality tests. In these tests, test takers indicate on a Likert scale how much a statement applies to them, which can then be aggregated into personality scores. While self-reported personality tests offer the advantage of easy applicability, they come along with the disadvantage that they "[...] reflect only one aspect of personality – people's explicit theories of what they think they are like." (Boyd and Pennebaker, 2017, p. 63). To overcome this shortcoming, an emergent strand of the literature proposes that personality can also be assessed based on linguistic features (Boyd and Pennebaker, 2017; Pennebaker et al., 2003; Pennebaker and King, 1999). The basic idea behind this approach is that language reflects personality since it is embedded in one's linguistic style. Furthermore, "[l]anguage use is relatively reliable overtime, internally consistent, and differs considerably between people." (Boyd and Pennebaker, 2017, p. 63). The ascent of language as a predictor for personality is strongly linked to the advances in text-analysis programs. While several text-analysis programs that allow obtaining linguistic features exist (see Pennebaker et al., 2003, for an overview of psychological word count approaches), the most prominent and widely used is Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015). LIWC is a text-analysis tool that analyzes the emotional, cognitive, and structural components of language. It compares each word in a given text with an internal library and returns a score for each linguistic category. Each score is calculated as the share of words that fall into the respective category over all words in the text (Tausczik and Pennebaker, 2010). The literature shows that LIWC scores exhibit test-retest reliability, external validity, and internal consistency (Pennebaker and King, 1999). However, a problem with LIWC is that it ignores context, irony, sarcasm, and idioms (Tausczik and Pennebaker, 2010). In the absence of faking, both approaches are predictive for a variety of life outcomes (see Ozer and Benet-Martínez, 2006, for a literature review on the correlations between personality and consequential outcomes; see Pennebaker et al. (2003) for a summary of the literature that links natural word use to personality). In the presence of faking, personality scores obtained from self-reported personality tests lose their predictive power (Viswesvaran and Ones, 1999). With respect to linguistic scores, the scientific literature shows that there are some differences between liars and truth-tellers in terms of their linguistic expression (Hauch et al., 2015). However, it is still an open question whether linguistic scores retain their predictive power in the presence of faking – a research gap that I address with my first study.

The study in Chapter 3 relates to the scientific literature on the meaning of work and labor market outcomes. The meaning of work literature extends the classical models on labor supply, which view work as an exchange of time and effort for money. It integrates the idea that workers do not only care about money but also care about other aspects of their work, such as its meaningfulness and the mission of their company (see Rosso et al., 2010, for an overview of the meaning of work literature). The integration of meaning as a source of utility into models of labor supply has profound implications for incentive theory (Cassar and Meier, 2018). For instance, the related literature documents a positive effect of meaning (Kosfeld et al., 2017; Chandler and Kapelner, 2013), missions (Carpenter and Gong, 2016; Gerhards, 2015), and charitable donations (Cassar, 2019; Tonin and Vlassopoulos, 2015; Imas, 2014) on workers' exerted effort. However, when being used instrumentally, prosocial incentives in the form of charitable donations backfire by leading to lower effort

(Cassar and Meier, 2017). Furthermore, the literature documents a negative effect of employers' social responsibility (Burbano, 2016) and the perceived meaning of the task (Ariely et al., 2008) on reservation wages. This negative relationship also translates into lower wages paid by socially responsible firms (Nyborg and Zhang, 2013) and principals who make charitable donations (Cassar, 2019).[1] With respect to the prosociality of the worker, the literature on labor market outcomes shows that volunteering has a positive effect on being invited for a job interview (Piopiunik et al., 2020; Baert and Vujić, 2018). Furthermore, past or current volunteering activity increases current and future wages (see Table 1 in Baert and Vujić, 2018, for an overview of the literature on the returns to volunteering in the labour market). Most studies that investigate the effect of volunteering on wages rely on survey data (Cozzi et al., 2017; Sauer, 2015; Hackl et al., 2007; Prouteau and Wolff, 2006) which makes it difficult to identify causal relationships. There is only little causal evidence on this topic (see Day and Devlin, 1998, 1997, for studies that use IV modeling to identify causality) – a lack of evidence that I address with my second study.

The study in Chapter 4 relates to the scientific literature on (sabotage in) contests (see Connelly et al., 2014, for a literature review on contest theory). A contest is defined as "[...] a situation in which individuals expend irretrievable resources to win valuable prize(s)." (Chowdhury and Gürtler, 2015, p. 135). The literature on tournament theory respectively contests originates from the seminal work of Lazear and Rosen (1981). The basic idea of tournament theory at that time was "[...] that firms induce effort from employees by effectively pooling some portion of wages from all the employees at one rank into the wages at the next highest rank, giving each the opportunity to win promotion to that rank." (Connelly et al., 2014, p. 18). The advantage of relative performance evaluations over absolute performance evaluations is that they are robust to common shocks (Nalebuff and Stiglitz, 1983), and workers' relative positions are typically less costly to observe (Lazear and Rosen, 1981). However, as first pointed out by Lazear (1989), a major problem of relative performance evaluations is that contestants can not only be successful by exerting productive effort but also by exerting destructive effort, i. e., sabotage (see Chowdhury and Gürtler, 2015, for a literature review on sabotage in contests). Sabotage is defined as "[...] any (costly) actions that one worker takes that adversely affect output of another." (Lazear, 1989, p. 563). While sabotage might prove beneficial for the saboteur, it adversely affects overall welfare. Reasons for this are the oppor-

[1] The result of Cassar (2019) only holds for non-motivated principals.

tunity costs of sabotage for the saboteurs and the decrease in productive output of the sabotaged contestants (Chowdhury and Gürtler, 2015). Furthermore, the expectation of being sabotaged has a discouragement effect (Gürtler and Münster, 2013, 2010), and sabotage may even lead to adverse selection into the contest (Münster, 2007). To address the problem of sabotage, there are two main policy approaches, namely reducing the benefits of sabotage or increasing its costs (Chowdhury and Gürtler, 2015). Policies that reduce the benefits of sabotage include reducing the spread between the winning and losing prizes (Lazear, 1989) and increasing the number of contestants (Konrad, 2000). Regarding policies that increase the costs of sabotage, the deterrence hypothesis posits that crime decreases in the certainty or the severity of punishment (Becker, 1968). Empirical evidence from the laboratory shows that revealing the identity of saboteurs (Harbring et al., 2007) and constant pairings (Yavas and Vandegrift, 2010) reduce sabotage among contestants. Empirical evidence from the field, showing a negative relationship between sabotage and its punishment, comes from Balafoutas et al. (2012). In addition, there is mixed empirical evidence from the field on the effect of monitoring on sabotage (Allen, 2016; Heckelman and Yates, 2003; Levitt, 2002; McCormick and Tollison, 1984). While McCormick and Tollison (1984) found that an increase in monitoring reduces sabotage, Heckelman and Yates (2003) found no such effect, and Allen (2016) even found that it may increase sabotage. Hence, the empirical evidence from the field is rather inconclusive – an issue that I try to resolve by providing new evidence with my third study.

1.4 Results and contributions

The results of my dissertation help to better understand managerial decision-making. More precisely, I contribute to the scientific literature as follows.

In Chapter 2, I find that in the absence of incentives to fake, machine learning classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions, whereas classifiers based on linguistic scores succeed. In the presence of incentives to fake, both approaches fail to make significantly better than chance predictions.

The scientific contribution of this study is twofold. First, it extends the existing literature on personality assessments by providing experimental evidence on the predictive power of personality and linguistic scores in the presence of incentives to fake. This topic, with respect to personality scores, has previously been addressed by the scientific literature (see Viswesvaran and Ones, 1999, for a meta-analysis on the fakability of personality inventories), however a lack of empirical evidence exists with respect to linguistic scores. Furthermore, it breaks new ground by using machine learning classifiers to study the predictive power of personality and linguistic scores. Second, this study adds to the existing literature on personality assessments by providing new insights on the predictive power of personality and linguistic scores in the absence of incentives to fake.

In practice, the insights of this study are the first step towards tools that will allow managerial decision-makers to validly assess personality traits based on linguistic features, even in the presence of faking. The ascent of voice assistants, like Alexa and Google Assistant, drastically increases the availability of linguistic data. This data creates new opportunities for these companies to analyze their customers' personality traits by using their linguistic expressions. Knowledge about their customers' personality traits, in turn, allows companies to deduce their preferences and thus also their willingness to pay (WTP) for certain products. For instance, extroverts are likely to have a higher WTP for a fancy garment than introverts. If companies engage in price discrimination based on their customers' personality traits, extraverts will have an incentive to fake when interacting with these voice assistants. However, these attempts will be pointless if future studies manage to prove that linguistic features serve as a robust personality predictor in the presence of faking.

In Chapter 3, I do not find empirical evidence to support the hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers. In the field experiment, the hypothetical wage offers by HR managers are not affected by the prosociality of the fictitious applicants, signaled by the work experiences on their résumés. Furthermore, I find that estimated reservation wages by HR managers are not affected by the prosociality of the fictitious applicants. Likewise, in the laboratory experiment, proposers' offers in the ultimatum game are not affected by responders' donation amounts in the strategic dictator game. In line with this finding, I also do not find that responders' reservation wages, measured by their minimum acceptance thresholds in the ultimatum game, are affected by their prosociality. Overall, these results provide empirical evidence that signaling prosociality does not backfire financially by leading to lower wage offers.

The scientific contribution of this study is threefold. First, it contributes to the literature on labor market outcomes by providing causal evidence on the effect of applicants' signaled prosociality on their wage offers. Although there exists a plethora of studies on this topic (Cozzi et al., 2017; Sauer, 2015; Hackl et al., 2007; Prouteau and Wolff, 2006), causal evidence is rather sparse (Day and Devlin, 1998, 1997). Second, this study contributes to the meaning of work literature by investigating the link between applicants' prosociality and their reservation wages. While reservation wages decrease in the prosociality and meaningfulness of the work (Burbano, 2016; Ariely et al., 2008), it is not clear whether the prosociality of workers moderates this effect. Third, this study contributes methodologically to the field of correspondence studies by adding a novel treatment manipulation to signal prosociality, namely the prosociality of previous work experiences. This approach allows to better control for other skills and competencies that are associated with prosociality which also affect wages.

In practice, the findings of this study have implications for both employers and applicants. The finding that reservation wages are not linked to applicants' prosociality helps employers when deciding on their wage offers. The insight that signaled prosociality does not affect wage offers suggests to applicants that employers do not use this signal to their detriment.

In Chapter 4, I do not find empirical evidence to support the deterrence hypothesis. In contrast to its prediction, I do not find empirical evidence that an increase in monitoring reduces sabotage. More precisely, the introduction of the VAR does not reduce the probability that a substitution that is due to an in-match injury takes place during a match. Furthermore, it also does not reduce the number of substitutions that are due to an in-match injury. This holds for both home and away teams.

The scientific contribution of this study is threefold. First, it contributes to the empirical literature on sabotage in contests by providing further empirical evidence on the deterrence hypothesis. While the theoretical prediction of the deterrence hypothesis is unambiguous, the empirical literature is rather inconclusive (Allen, 2016; Heckelman and Yates, 2003; Levitt, 2002; McCormick and Tollison, 1984). Second, it contributes to the empirical literature on the home bias by studying how an increase in monitoring, and therefore a reduction in the wiggle room of referees, affects sabotage of home and away teams. Shedding light on this question provides a more nuanced picture of the overall effect of an increase in monitoring by also analyzing the sub-effects. Third, it contributes methodologically by adding a novel proxy for sabotage in sports, namely substitutions that are due to an in-match injury. This novel proxy overcomes the shortcomings of existing proxies (e. g., cards in soccer (Deutscher and Schneemann, 2017; Bartling et al., 2015; del Corral et al., 2009; Garicano and Palacios-Huerta, 2005), time penalties in ice hockey (Allen, 2016; Heckelman and Yates, 2003; Levitt, 2002), and shido in judo (Balafoutas et al., 2012)), by not being affected by changes in the detection probability and errors and favoritism of referees.

In practice, the insights from this study show managerial decision-makers that deterrence, implemented by an increase in monitoring, does not necessarily lead to a reduction of sabotage. Instead, the findings suggest that the effectiveness of this policy intervention might be contingent on the affected individuals' perceptions of the situation in which they are involved.

1.5 Dissertation outline and summary

This dissertation is structured as follows. In Chapter 2, I examine the first research question by presenting the study *"How do incentives to fake affect the predictive power of personality assessments? An experimental study."* In Chapter 3, I answer the second research question with the study *"Do employers wage-discriminate against applicants based on their signaled prosociality? Empirical evidence from the field and laboratory."* In Chapter 4, I provide answers to the third research question by presenting the study *"Does an increase in monitoring reduce sabotage? Empirical evidence from professional soccer."* Finally, I conclude this dissertation in Chapter 5 by summarizing my results and outlining avenues for future research. References for all three studies are provided in Chapter 6.

Table 1.1: Overview of the research studies.

	Chapter 2	Chapter 3	Chapter 4
Title	• How do incentives to fake affect the predictive power of personality assessments? An experimental study	• Do employers wage-discriminate against applicants based on their signaled prosociality? Empirical evidence from the field and laboratory	• Does an increase in monitoring reduce sabotage? Empirical evidence from professional soccer
Research question	• How do incentives to fake affect the predictive power of personality assessments?	• Do employers wage-discriminate against applicants based on their signaled prosociality?	• Does an increase in monitoring reduce sabotage?
Data and methods	• Online experiment • Linguistic Inquiry and Word Count (LIWC) • Machine learning	• Field and laboratory experiment • Correspondence study • Non-parametric tests • Regression analyses	• Quasi-natural experiment • Ordinary least squares difference-in-differences estimation
Related literature	• Personality assessments • Faking in personality assessments	• Work meaning as a source of utility • Labor market outcomes	• Sabotage in contests • Deterrence hypothesis
Results	• In the absence of incentives to fake, linguistic scores allow to make valid cooperativeness predictions • In the presence of incentives to fake, neither linguistic nor personality scores allow to make valid cooperativeness predictions	• Employers do not wage-discriminate against applicants based on their signaled prosociality • Applicants' reservation wages do not depend on their prosociality	• An increase in monitoring does not reduce sabotage • A reduction in the wiggle room for referees does not affect the sabotage effort of home and away teams
Scientific contributions	• Studies the predictive power of linguistic (LIWC) and personality (Big Five) scores in the presence of incentives to fake • Provides further evidence on the predictive power of linguistic (LIWC) and personality (Big Five) scores in the absence of incentives to fake	• Provides causal evidence on the effect of applicants' signaled prosociality on their wage offers • Studies the relationship between applicants' prosociality and their reservation wages • Adds previous work experience as a novel treatment manipulation to signal prosociality in correspondence studies	• Provides further empirical evidence on the deterrence hypothesis • Studies how a reduction in the wiggle room of referees affects sabotage of home and away teams • Studies sabotage with a novel proxy, namely substitutions that are due to an in-match injury
Co-authors	• Michael Kurschilgen	• Julia Holzmann	• Thomas Daske • Julian Hackinger
Contributions	• Developing the research question and design - both authors contributed equally • Programming and conducting the experiment • Programming the machine learning part • Conducting the data analysis • Writing the paper	• Developing the research question and design - both authors contributed equally • Deriving the hypothesis • Programming the experiments • Conducting the experiments - both authors contributed equally • Conducting the data analysis - I made the main contribution • Writing the paper	• Developing the research question • Developing the research design - all authors contributed equally • Crawling the data - Julian Hackinger and I contributed equally • Conducting the data analysis - I made the main contribution • Writing the paper

2. How do incentives to fake affect the predictive power of personality assessments? An experimental study

Abstract

Personality traits are frequently assessed in situations where respondents have incentives to fake. We study how incentives to fake affect the predictive power of personality assessments. Using experimental data, we compare the predictive power of two different approaches to predict the cooperativeness of job applicants: predictions by machine learning classifiers based on personality (Big Five) scores obtained from self-reported personality tests and linguistic (LIWC) scores obtained from written self-descriptions. For each approach, we trained different classifiers with data from subjects without incentives to fake their cooperativeness and evaluated the classifiers' predictions with data from subjects with incentives to fake. Our results show that in the absence of incentives to fake, machine learning classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions, whereas classifiers based on linguistic scores succeed. In the presence of incentives to fake, both approaches fail to make significantly better than chance predictions.

Keywords: Big Five; Faking; Machine learning; LIWC; Personality

JEL Codes: C38; C71; C93

Authors: Michael Kurschilgen and Magnus Strobel

2.1 Introduction

Assessments of personality traits are ubiquitous. For instance, they are applied in hiring processes to assess job applicants' suitability for vacant positions. According to a survey among 344 Society for Human Resource Management members, about 22 percent use personality tests to evaluate job candidates (SHRM, 2014).

Many situations in which personality traits are assessed are plagued with faking. Faking is defined as "[...] dishonest impression management or intentional distortion of responses to interview questions or misrepresentation in order to create a good impression." (Levashina and Campion, 2006, p. 301). Often, there are incentives to fake. For instance, if undetected, job applicants who are applying for a position that involves a lot of collaborative work increase their chances of being hired if they exaggerate required personality traits such as their agreeableness.

This raises the question of how incentives to fake affect the predictive power of personality assessments. We shed light on this question by following the example of hiring processes. In hiring processes, among other criteria (e. g., wage demand, skills, etc.), applicants are typically selected based on their personality traits. In our study, we focus on selections of applicants based on the personality trait of cooperativeness. We chose cooperativeness because it constitutes a highly sought-after personality trait in today's job market. This is due to the collaborative nature of many jobs that require the ability to work in teams. According to a study, nearly 75% of employers rate teamwork and collaboration as "very important" (Queens University of Charlotte, 2021).

During hiring processes, companies use different approaches to obtain data for the assessment of personality traits. In this study, we compare the predictive power of cooperativeness assessments based on data from the following two approaches. First, we study cooperativeness assessments based on data from an established approach. That is, we study how incentives to fake affect the predictive power of cooperativeness assessments based on personality (Big Five) scores obtained from self-reported personality tests. Second, we study cooperativeness assessments based on data from a rather novel approach. That is, we study how incentives to fake affect the predictive power of cooperativeness assessments based on linguistic (LIWC) scores obtained from written self-descriptions (e. g., cover letters).

To compare the two approaches in the absence and presence of incentives to fake, we conducted an online experiment consisting of seven stages. In stage one, subjects played a one-shot public goods game (henceforth, referred to as the first public goods game). Subjects' contributions in this public goods game served as our measure for their true cooperativeness (labels).[1] In stage two, we introduced a treatment manipulation by varying whether subjects have (*Treatment* group) or do not have (*Control* group) salient incentives to fake their cooperativeness. In particular, subjects in the *Treatment* group were told that a committee would evaluate their cooperativeness based on the responses in the subsequent stages, with the 40% most cooperative being eligible for a bonus. In stage three, subjects completed a written self-description of 3,000 characters in which they were asked to describe themselves by elaborating on their skills, hobbies, experiences, dreams, and hopes. From these written self-descriptions, we used Linguistic Inquiry and Word Count (LIWC) to obtain the linguistic scores for our machine learning part (features).[2] In stage four, subjects performed a 10-item personality test of the Big-Five dimensions. From these self-reported personality tests, we obtained the personality scores for our machine learning part (features). In stage five, subjects played another public goods game (henceforth, referred to as the second public goods game). This second public goods game served as a manipulation check. In stage six, we elicited subjects' beliefs about their position in the cooperativeness ranking before they completed a socio-demographic questionnaire in stage seven.

In two separate steps, we compared the predictive power of personality scores vs. linguistic scores in the absence (*Control* group) and in the presence (*Treatment* group) of incentives to fake. In the first step, we used supervised learning to train different machine learning classifiers with data from the *Control* group. As labels, we predicted whether subjects' true cooperativeness is above the median. As features, we used personality (Big Five) scores and linguistic (LIWC) scores, which we obtained from the self-reported personality test and written self-description. To find the best hyperparameter set and parameters, we used nested stratified k-fold cross-validation. To evaluate the performance of our models, we used Matthews correlation coefficient. After selecting the best models, we trained each classifier with the best hyperparameter set on the *Control* group's entire data. In the second step, based on the data from the *Treatment* group, we compared the predictions of these classifiers with predictions of a dummy classifier.

[1] Labels are the outputs of a machine learning model.

[2] Features are the inputs of a machine learning model.

Our results show that in the absence of incentives to fake, machine learning classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions, whereas classifiers based on linguistic scores succeed. In the presence of incentives to fake, both approaches fail to make significantly better than chance predictions.

The scientific contribution of this study is twofold. First, it extends the existing literature on personality assessments by providing experimental evidence on the predictive power of personality and linguistic scores in the presence of incentives to fake. In particular, it breaks new ground by using machine learning classifiers to study the predictive power of personality and linguistic scores. Regarding personality scores, there exists already ample scientific evidence which documents that personality scores are prone to faking. The meta-study of Viswesvaran and Ones (1999) shows that all Big Five factors are equally fakable. In our study, we provide evidence that incentives to fake being cooperative distort assessments of the personality trait of Neuroticism. Regarding linguistic scores, the literature documents that "[...] relative to truth-tellers, liars experienced greater cognitive load, expressed more negative emotions, distanced themselves more from events, expressed fewer sensory-perceptual words, and referred less often to cognitive processes." (Hauch et al., 2015, p. 307). We add to these findings by showing that incentives to fake being cooperative distort the LIWC category of *Drives*. Second, this study adds to the existing literature on personality assessments by providing new insights on the predictive power of personality and linguistic scores in the absence of incentives to fake. So far, the scientific literature documents that the personality traits of *Extraversion* (Koole et al., 2016) and *Agreeableness* (Koole et al., 2016; Kagel and McGee, 2014; Volk et al., 2011) predict cooperation. This study adds the findings that the personality traits of *Openness*, *Conscientiousness*, and *Agreeableness* are positively and significantly correlated with cooperativeness. In addition, it shows that the LIWC categories of *3rd pers plural*, *Common Adverbs*, *Anxiety*, *Health*, *Drives*, and *Religion* are positively and significantly correlated with cooperation. Furthermore, the LIWC categories of *Sadness*, *Future focus*, and *Periods* are negatively and significantly correlated with cooperation.

The remainder of this paper is structured as follows. In Section 2.2, we derive the hypotheses of our study. In Section 2.3, we outline the design and procedure of our experiment. In Section 2.4, we describe the methodology of our machine learning approach. We present our results in Section 2.5 before we conclude by discussing them in Section 2.6.

2.2 Hypotheses

The use of personality scores obtained from self-reported personality tests as a personality predictor is based on the idea that personality (e. g., people's habitual patterns of thought, feeling, and action) can be subsumed by a certain set of personality traits. Personality is defined as "[...] the consistent set of traits, attitudes, emotions, and behaviors that people have." (Boyd and Pennebaker, 2017, p. 63). While many theories try to provide a taxonomy for people's personalities, the most widely accepted theory is the so-called trait approach. It assumes "[...] that our cognitions, emotions, and behaviors are determined by a number of consistent and relatively stable traits." (Matz et al., 2016, p. 36). The most influential and widely accepted trait theory is the Big Five model (Goldberg, 1990). It posits that personality can be captured along five different traits. The five overarching traits – which are also often referred to by their acronym "*OCEAN*" – are *Openness* (O), *Conscientiousness* (C), *Extroversion* (E), *Agreeableness* (A), and *Neuroticism* (N). Each of these traits consists of different facets which can be measured by different items. For instance, *Openness* includes the facets of *fantasy*, *aesthetics*, *feelings*, *actions*, *ideas*, and *values* which can be captured by the items "I have a vivid imagination" and "I have difficulty understanding abstract ideas" (Matz et al., 2016). The Big Five personality traits are typically assessed through self-reported personality tests. In these tests, test-takers indicate on a five-point Likert scale how much each item applies to them. Finally, based on the responses to the associated items, a score can be calculated for each personality trait, which allows comparing test-takers.

In the absence of incentives to fake, personality scores are predictive for a variety of life outcomes, including individual outcomes, interpersonal outcomes, and social institutional outcomes (see Ozer and Benet-Martínez, 2006, for a literature review on the correlation between personality and consequential outcomes). For instance, Ozer and Benet-Martínez (2006) lists correlations between the Big Five personality traits and physical health (Caspi et al., 2005), Axis I disorders (Trull and Sher, 1994), self-concept and identity (Clancy and Dollinger, 1993), peer and family relationships (Jensen-Campbell et al., 2002), romantic relationships (Robins et al., 2002), occupational choice and performance (Barrick et al., 2003; Larson et al., 2002), political attitudes and values (van Hiel et al., 2004; Heaven and Bucci, 2001; Saucier, 2000), volunteerism and community involvement (Carlo et al., 2005), and criminality (Walton and Roberts, 2004). Furthermore, the Big Five personality traits are also correlated with intelligence (Ackerman and Heggestad, 1997), economic preferences (Becker et al., 2012), behavior (Paunonen and Ashton, 2001), consumer behavior (Kassarjian, 2018), and job performance (Ziegler et al., 2014; Neal et al., 2012; Judge et al., 1999).

In the presence of incentives to fake, personality scores lose their predictive power. A summary of the scientific literature on faking in personality tests shows strong evidence for response-distortion (e. g., faking) (Morgeson et al., 2007). Furthermore, the meta-study by Viswesvaran and Ones (1999) shows that all Big Five factors are equally fakable. Their results show that effect sizes are lower for fake good (i. e., making a favorable impression) than for fake bad treatments (i. e., making an unfavorable impression). The reason that self-reported personality tests are prone to faking is their straightforward phrasing, which makes it quite obvious which traits are assessed by the specific items. For instance, the statement "I see myself as critical, quarrelsome" appears to provide an answer to the question of how well applicants integrate into teams. Indeed, many studies corroborate this hunch by showing that the associated personality trait of *Agreeableness* predicts cooperation (Koole et al., 2016; Kagel and McGee, 2014; Volk et al., 2011). Hence, job applicants who want to be perceived as a team player will simply respond with "Disagree strongly" to this statement. To attenuate the problem of faking, many approaches have been proposed. For instance, they include limiting the response time (Holden et al., 2001), a social desirability correction to approximate individuals' honest scores (Ellingson et al., 1999), instructional warnings (Dwight and Donovan, 2003), and forced-choice formats (Heggestad et al., 2006; Jackson et al., 2000). While all these approaches may provide some remedy to the problem of faking, none of them is able to resolve it fully (Hogan et al., 2007). Therefore, we conjecture that in the presence of incentives to fake, assessments based on personality (Big Five) scores obtained from self-reported personality tests fail to make better than chance cooperativeness predictions.

The use of linguistic scores obtained from written self-descriptions as a personality predictor is based on the idea that language reflects personality since it is embedded in one's linguistic style. Furthermore, "[l]anguage use is relatively reliable overtime, internally consistent, and differs considerably between people." (Boyd and

Pennebaker, 2017, p. 63). The ascent of language as a predictor for personality is strongly linked to the advances in text-analysis programs. While there exist several text-analysis programs that allow obtaining linguistic features (see Pennebaker et al., 2003, for an overview of psychological word count approaches), the gold standard to obtain linguistic scores from linguistic samples is Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015). LIWC is a text analysis tool that analyzes the emotional, cognitive, and structural components of language. It compares each word of a given text with an internal library that contains words from different linguistic categories. For instance, these categories include social processes (e. g., "Family: dad, daughter, aunt"), cognitive processes (e. g., "Causation: therefore, reason"), and perceptual processes (e. g., "Feel: feel, sleek"). Each time a word matches a word in one of the categories, it increases the count of this category by one (Tausczik and Pennebaker, 2010). As an output, LIWC calculates a score for each category, which gives the percentage share of words in this category over all words in the text. The related literature shows that LIWC scores exhibit test-retest reliability, external validity, and internal consistency (Pennebaker and King, 1999). However, one of the disadvantages of the LIWC approach is that it ignores context, i. e., it falsely codes irony, sarcasm, and idioms (Tausczik and Pennebaker, 2010).

In the absence of incentives to fake, linguistic scores are predictive for personality traits. Within the literature that uses language to predict personality, there are two approaches that are predominately followed (Boyd and Pennebaker, 2017). The first, and most prominent, approach is to connect language to the results of self-reported personality tests (Schwartz et al., 2013; Yarkoni, 2010; Hirsh and Peterson, 2009; Mehl et al., 2006; Pennebaker and King, 1999). For instance, Pennebaker and King (1999) find modest but reliable correlations between the Five-Factor personality dimension and language use. In line with this result, Hirsh and Peterson (2009) find that, in self-narratives, all Big Five personality traits are strongly and significantly associated with word use patterns theoretically appropriate to the trait. Additionally, language in modern-day forms of communication, such as Facebook status updates (Schwartz et al., 2013) and texts written by bloggers (Yarkoni, 2010), also prove to be predictive for Big Five personality traits. Although this approach proves to be quite successful, it is problematic since it treats the results of the self-reported personality tests as "ground-truth", which, however, are prone to several limitations. The second approach is to directly connect language to personality processes instead of the results of self-reported personality tests. This approach has proven to be successful in linking language to academic success (Pennebaker et al., 2014; Robinson et al., 2013), longevity (Penzel et al., 2017; Pressman and Cohen, 2007), suicidal tendencies (Wiltsey Stirman and Pennebaker, 2001), standing in social hierarchies (Kacewicz et al., 2014), emotional upheavals (Pennebaker and Lay, 2002; Stone and Pennebaker, 2002), relationship satisfaction (Slatcher et al., 2008), cooperation (Rand et al., 2015), age (Pennebaker and Stone, 2003), and gender (Newman et al., 2008). Furthermore, the relationship between language and personality holds in different contexts, like self-narratives Hirsh and Peterson (2009), daily diaries, writing assignments, essays, and journal abstracts (Pennebaker and King, 1999), one-hour life history interviews (Fast and Funder, 2008), bereavement stories (Baddeley and Singer, 2008), and recordings of day-to-day speech (Mehl et al., 2006).

In the presence of incentives to fake, linguistic scores might retain their predictive power by being robust to faking. A reason for their robustness as a personality predictor could be that, while it is relatively easy for people to change what they say, it is more difficult for them to change how they say it. That is, while some linguistic scores change when lying, others remain unchanged (Hauch et al., 2015; Hancock et al., 2007; Bond and Lee, 2005; Zhou et al., 2004; Newman et al., 2003). The literature shows that "[...] relative to truth-tellers, liars experienced greater cognitive load, expressed more negative emotions, distanced themselves more from events, expressed fewer sensory-perceptual words, and referred less often to cognitive processes." (Hauch et al., 2015, p. 307). While these linguistic characteristics are affected by lying, their findings imply that all other linguistic characteristics do not significantly change. Therefore, we conjecture that in the presence of incentives to fake, assessments based on linguistic (LIWC) scores obtained from written self-descriptions allow making better than chance cooperativeness predictions.

2.3 Experiment

2.3.1 Design and procedure

To answer the question of how incentives to fake affect the predictive power of personality assessments, we conducted the following online experiment as illustrated in Figure 2.1 (see Section 2.7.5 for screenshots of the experiment).[3]

[3]Please note that we omitted some stages in Figure 2.1 that we did not use for our data analysis.

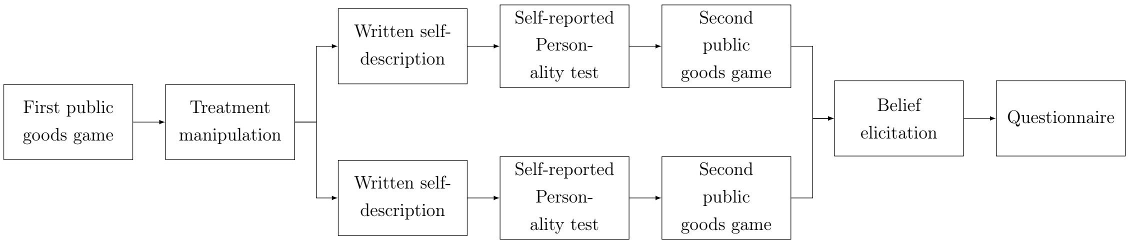


Figure 2.1: Design and procedure of the online experiment.

In stage one, we elicited subjects' true cooperativeness with a one-shot public goods game. In the game, subjects were endowed with 20 points which they could either retain or invest into a common project.[4] The payoff from the public goods game was as follows:

$$\pi_i = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j \tag{2.1}$$

where g_i is the subject's contribution and g_j are the other players' contributions to the project, and 0.4 constitutes the marginal payoff of contributing to the public good. Each group consisted of four players. The socially optimal decision is full contribution by all players in the group. However, for rational money-maximizing individuals, the unique Nash equilibrium is complete free-riding by all subjects. Hence, the setup resembles a teamwork situation in which individual team members have an incentive to free-ride by contributing less than the other team members. Subjects' contributions (g_i) to the public good serve as our discrete measure for their true cooperativeness.

In stage two, subjects were shown the treatment manipulation. To this end, subjects were randomly assigned to the *Control* or *Treatment* group. The probability of being assigned to either the *Control* or the *Treatment* group was 75% and 25%, respectively. Subjects in both groups were shown the following information:

On the following pages, we ask you to complete 3 additional personality tests.

Additionally, subjects in the *Treatment* group were shown the following treatment information:

[4]The €/point exchange rate was 0.1.

Based on these 3 personality tests, a committee will decide if you belong to the 40% most cooperative participants. If you belong to the 40% most cooperative participants, you will receive a bonus of €10.[5]

Hence, in contrast to subjects in the *Control* group, subjects in the *Treatment* group had a salient monetary incentive to fake being cooperative in the subsequent stages.

In stage three, subjects performed a written self-description of 3,000 characters. We set this threshold such that it resembles the length of a typical one-page cover letter. In this task, subjects were asked to describe themselves, by, for instance, elaborating on their skills, hobbies, experiences, dreams, and hopes.

In stage four, subjects performed a self-reported 10-item Big Five personality test (Gosling et al., 2003). This personality test is on a 7-point Likert scale and contains two items per personality trait.

In stage five, subjects performed the conditional cooperation test by Fischbacher et al. (2001). Subjects first made an unconditional contribution decision, similar to the one described above. Next, subjects made a second type of contribution decision, i. e., they filled out a contribution table. In the contribution table, subjects had to indicate their contribution amount, given the 21 average contribution amounts (rounded to integers) of the three other group members. For three out of the four group members, the unconditional contribution decision was payoff relevant, and for the other remaining group member, the contribution table determined the payoff. The payoff relevant contribution decision was determined randomly.[6]

In stage six, subjects had to state their belief about their position in the cooperativeness ranking. As belief accuracy is significantly higher when beliefs are incentivized (Gächter and Renner, 2010), we paid subjects a bonus of €5 if their stated position in the cooperativeness raking (out of 100 participants) matched their assigned position.

[5]The ranking was made based on subjects' cooperativeness scores. Each subjects' cooperativeness score was the average of three scores, one for each additional personality test. For the written self-description, a student research assistant read all texts and then assigned a score from zero to one for each subject's perceived cooperativeness. For the Big Five personality test, it was the *Agreeableness* score, scaled from zero to one. For the second public goods game, it was the average contribution decision of the unconditional and conditional contribution decisions, scaled from zero to one.

[6]Please note that we did an ex post matching of groups, since a real-time matching was not practicably feasible due to the online setting.

Lastly, in stage seven, subjects completed a socio-demographic questionnaire in which we also asked subjects about their native language.

2.3.2 Procedural details and descriptive statistics

We programmed the experiment in oTree (Chen et al., 2016) and conducted it online with students from the subject pool of experimenTUM, the experimental laboratory at the Technical University of Munich. Between June 30 and July 9, 2020, we conducted 17 online sessions with a total of 400 subjects. In our data analysis, we excluded all subjects (10) who did not comply with our requirements for the written self-descriptions (e. g., by using copy & paste). Furthermore, since we analyzed the linguistic characteristics of subjects' written self-descriptions, we excluded all subjects (94) who indicated in the socio-demographic questionnaire that their mother tongue is not German. Thereby we guarantee that our results are not driven by a lack of oracy of non-native speakers. This left us with a total of 296 subjects for our analysis (217 subjects in the *Control* group and 79 subjects in the *Treatment* group).

2.4 Machine learning approach

We tackled our research question by comparing the predictive power of personality scores and linguistic scores in the absence and presence of incentives to fake. More precisely, we predicted whether subjects' true cooperativeness is above the median of the *Control* group or not. To make these predictions, we used machine learning. In particular, we used supervised learning with data from the *Control* group to train different machine learning classifiers for the prediction task. Supervised learning is a type of machine learning that uses labeled training data (labels and features) to train machine learning models. In our approach, we used the following labels, features, and machine learning classifiers.[7]

As labels, we used subjects' contribution decision (g_i) in the first public goods game. Since we did not introduce any treatment differences at this point, it served as our measure for the subjects' true cooperativeness. We converted subjects' true cooperativeness to a binary label. Based on a median split of subjects' contribution

[7]We used the Python programming language to implement this approach.

decisions, we assigned them a label of 1 if they contributed more than the median of the *Control* group and 0 otherwise.

As features, we used data from the following two assessment tasks for our machine learning models. First, we used personality scores obtained from the self-reported personality tests. More concretely, we used the raw responses to the 10-item personality test of the Big-Five dimensions of Gosling et al. (2003). Second, we used linguistic scores obtained from the written self-descriptions. In particular, we used LIWC (Internal German Dictionary 2015) to obtain 97 linguistic scores from the written self-descriptions (Pennebaker et al., 2015).[8]

In machine learning, feature selection is primarily used to remove non-informative or redundant features. This helps to reduce computation time, improve prediction performance, and gain a better understanding of the data. Methods for feature selection can be divided into two categories, unsupervised and supervised methods, where the latter can be further divided into wrapper, filter, and embedded methods (Chandrashekar and Sahin, 2014). In our machine learning approach, we used a filter method to identify and select the most informative and relevant personality and linguistic scores. Filter methods evaluate the relevance of features by considering their relationship with the labels and only retaining those for the training of the model which pass some criterion. Based on data from the *Control* group, we conducted a correlation analysis between subjects' true cooperativeness and their personality and linguistic scores. We only retained those personality and linguistic scores for which the Pearson correlation coefficient is significant at the 10% level (see Table 2.1 for the retained categories).

As machine learning models, we used several classifiers for binary classification from the scikit-learn machine learning library (Pedregosa et al., 2011). In particular, we used the following classifiers: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and Multi-layer Perceptron Neural Network. Furthermore, to have a benchmark for our predictions, we used the following dummy classifiers: Dummy Minority (always predicts the minority class), Dummy Majority (always predicts the majority class), Dummy Stratified (predicts based on the training set's class distribution), and Dummy Uniform (predicts uniformly at random). To draw an analogy to hiring processes, each of these dummy classifiers represents a certain type of human resources (HR) manager. Dummy Minority represents HR managers who naively think all of the applicants are above the

[8]Please find an overview of all 97 LIWC categories in Section 2.7.1.

median in terms of their cooperativeness. Dummy Majority represents HR managers who naively think all of the applicants are below the median in terms of their cooperativeness. Dummy Uniform represents HR managers who decide randomly by tossing a fair coin whether current applicants are above or below the median in terms of their cooperativeness. Lastly, Dummy Stratified represents HR managers who decide based on their experience on the distribution of past applicants' cooperativeness whether current applicants are above the median in terms of their cooperativeness or not. Since Dummy Stratified best represents sophisticated HR managers, we used it as a benchmark in our analysis to test if the other classifiers perform significantly better than chance.

To improve the performance of our machine learning models, we used *nested stratified k-fold cross-validation* to find the best hyperparameter set. More concretely, the meaning of the terms is as follows.

The term *k-fold cross-validation* describes a resampling procedure that allows training and testing/evaluating machine learning models on small data samples. Machine learning models are typically evaluated using a test dataset, which consists of samples that were held out from the training process. The performance of the model is then determined by the model's ability to classify unseen samples correctly. In k-fold cross-validation, the data is split into k equal-sized non-overlapping folds. Each of these k folds is then used exactly once as a test/validate set to evaluate the model, while the remaining k-1 folds are used as a training set. Hence, k-fold cross-validation solves the trade-off between increasing the train vs. increasing the test set. While a larger train set increases the chances of building a model that generalizes well to unseen data, a larger test set increases the chances that the testing error converges to the generalization error.

The term *stratified* signifies that each fold in the cross-validation preserves the percentage of samples of both classes. This ensures that each fold is a good representative of the entire data set. Hence, this approach solves the problem that "unlucky splits", in which one class is under- or overrepresented in the training or testing set, lead to biased results.

The term *nested* describes an approach in which there is one k-fold cross-validation loop (inner loop) within another (outer loop). It addresses the problem of overfitting, which arises with a single cross-validation loop when hyperparameters are also optimized. Overfitting arises in a single cross-validation loop since the same data is used to tune model parameters and evaluate model performance. To overcome this problem, nested cross-validation splits the data in train, validate, and test folds so that evaluating the model performance (on the test folds) can be done separately from tuning the model parameters (on the train and validate folds).

To evaluate and select the best models, we used Matthews correlation coefficient. The advantage of MCC over other measures like the F1 or Accuracy score is that it only produces a high score "[...] if the prediction obtained good results in all of the four confusion matrix categories (true positives [TP], false negatives [FN], true negatives [TN], and false positives [FP]), proportionally both to the size of positive elements and the size of negative elements in the dataset." (Chicco and Jurman, 2020, p. 1). In essence, MCC is a correlation coefficient between the observed and predicted binary classifications:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.2}$$

It ranges from +1 (perfect prediction) to −1 (worst prediction), with 0 being equivalent to a random prediction.

To train our machine learning models, we used the following approach, which is illustrated in Figure 2.2.[9]

[9]Please note that we have two data sets, namely the personality scores from the self-reported personality tests and the linguistic scores from the written self-descriptions. Therefore, the following steps are all performed twice, once with each data set.

Figure 2.2: Machine learning approach: Nested stratified 5-fold cross-validation.

In the outer loop, the data set was divided into five equal-sized non-overlapping folds, which preserved the percentage of samples of both classes. Each of these five folds was then used exactly once as a test set to evaluate the model, while the remaining four folds were used as a training set (see 1 to 5 of the outer loop of Figure 2.2).

In the inner loop, we used *GridSearchCV* from the *scikit-learn* library to identify and return the best hyperparameter set and parameter values for all our classifiers.

It automatically conducts an exhaustive cross-validated grid-search over specified parameter grids. For each classifier, we specified the parameter grid as follows.[10] For hyperparameters that take non-numeric inputs, we included all predefined choices of the hyperparameter. For hyperparameters that take numeric inputs, we specified three different values (the default value and reasonable below-default/above-default values equidistantly from the default).[11] Based on the specified grid, *GridSearchCV* iterated over all possible hyperparameter combinations (x) and returned the best model for each classifier and training set of the outer loop. In particular, it divided the training set from the outer loop into five equal-sized non-overlapping folds, which preserved the percentage of samples of both classes. Each of these five folds was then used exactly once as a validation set to evaluate the model, while the remaining four folds were used as a training set. Using a *Pipeline* from the *imbalanced-learn* library, we preprocessed the data in each iteration of the inner loop as follows. First, the features were standardized by subtracting the mean and scaling to unit variance. Second, the minority class was randomly oversampled to increase the number of observations in our data set. Next, *GridSearchCV* identified each classifier's best model by calculating the average error of each possible hyperparameter combination on the inner loop (see 5.x.1 to 5.x.5 of the inner loop of Figure 2.2).

After we obtained the best hyperparameter set for each classifier from the inner loop's cross-validation, we selected the hyperparameter set that best generalizes to out-of-sample data. That is, we evaluated their predictions with the test data from the outer loop and chose the hyperparameter set with the best generalization error (evaluated with MCC). Lastly, for each classifier and identified hyperparameter set, we tuned the parameters by training each classifier's model with the entire data from the *Control* group. With the trained models, we then made predictions based on the *Treatment* group's features and evaluated these predictions.

2.5 Results

In the following, we present the results of our data analysis. First, we show the results of the first public goods game from which we derived subjects' cooperativeness (Section 2.5.1). Next, we present our feature selection (Section 2.5.2) before we

[10]See Section 2.7.2 for the entire grid.

[11]Please note that in rare cases, we deviated from this rule by specifying other values.

analyze the predictions in the absence (Section 2.5.3) and presence (Section 2.5.4) of incentives to fake.

2.5.1 True cooperativeness

In our machine learning approach, we predicted whether subjects' true cooperativeness is above the median or not based on their personality and linguistic scores. As a measure for the subjects' true cooperativeness, we used their contribution in the first public goods game.

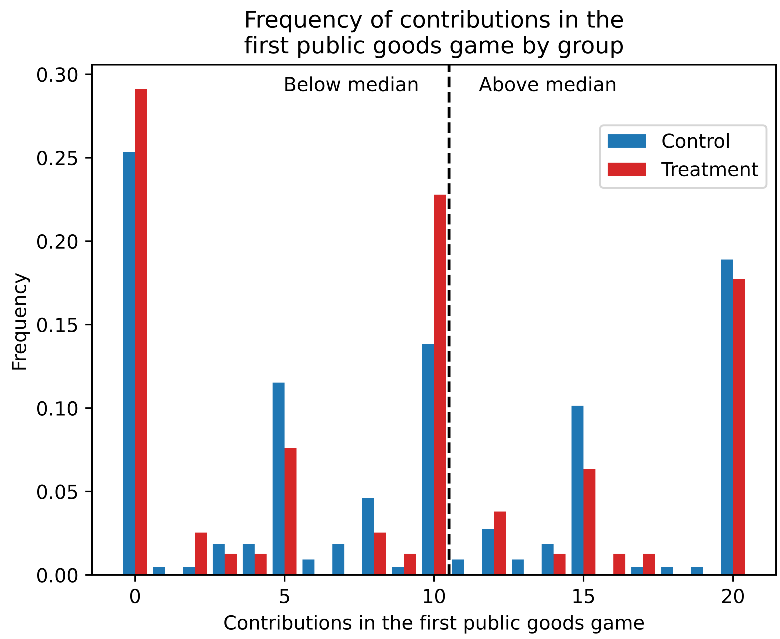


Figure 2.3: Frequency of contributions in the first public goods game by group.

Notes: In the first public goods game, subjects had to decide how much of their endowment (20 points) they want to contribute to a public good.

Figure 2.3 shows the frequency of subjects' contributions in our first public goods game by their treatment group. Since we did not introduce a treatment difference when subjects played the first public goods game, we do not expect any significant differences in terms of the contributions between the two groups. A two-sided Mann-Whitney U test between the *Control* (mean=9.06) and *Treatment* (mean=8.66) group confirms this assumption (p=0.695). This finding is also corroborated by a two-sided Kolmogorov-Smirnov test (p=0.992) between the two distributions. In general, both distributions show five frequent contribution amounts, namely 0, 5, 10, 15, and 20 points, with complete free-riding being the most frequent choice. The share of free-riders in our experiment is in line with the findings of Fischbacher et al. (2001) who found that about 30% can be classified as purely selfish. Regarding the median split for the label of our machine learning approach, we find that out of the 217 subjects in the *Control* group, 80 (≈37%) contributed more than the median, and 137 (≈63%) contributed as much as or less than the median. In the *Treatment* group, out of the 79 subjects, 25 (≈32%) contributed more than the median, and 54 (≈68%) contributed as much as or less than the median (of the *Control* group).

2.5.2 Feature selection

To identify the most predictive features for our binary classification task, we first analyzed, for subjects in the *Control* group, the correlation between their true cooperativeness and their personality and linguistic scores.

Table 2.1: Correlations between subjects' true cooperativeness and their personality scores.

Personality trait/Item	**r**	**p**
Openness	**0.185**	**0.006**
Open to new experiences, complex	0.141	0.039
Conventional, uncreative	-0.149	0.028
Conscientiousness	**0.117**	**0.085**
Dependable, self-disciplined	0.160	0.018
Disorganized, careless	-0.051	0.455
Extraversion	**0.038**	**0.581**
Extraverted, enthusiastic	0.016	0.814
Reserved, quiet	-0.050	0.460
Agreeableness	**0.195**	**0.004**
Critical, quarrelsome	-0.112	0.101
Sympathetic, warm	0.218	0.001
Neuroticism	**0.019**	**0.783**
Anxious, easily upset	-0.031	0.647
Calm, emotionally stable	0.002	0.973

Notes: The analysis only includes subjects from the *Control* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). r gives Pearson correlation coefficients. p gives the respective p-values. Bold text indicates personality traits. In the two rows below each personality trait are the associated items.

For the *Control* group, Table 2.1 shows the Pearson correlation coefficients (r) and respective p-values for the correlations between subjects' true cooperativeness and their personality scores. On the trait-level, the results show that *Openness*, *Conscientiousness*, and *Agreeableness* are all positively and significantly correlated with subjects' true cooperativeness at the 10% level. Hence, we corroborate the results from the literature on personality traits and cooperation, which find a positive association between *Agreeableness* and cooperation (Koole et al., 2016; Kagel and McGee, 2014; Volk et al., 2011). On the item-level, we find that the statements "I see myself as ..." (1) "Open to new experiences, complex", (2) "Conventional, uncreative", (3) "Dependable, self-disciplined", and (4) "Sympathetic, warm" are significantly correlated with subjects' true cooperativeness at the 10% level. To solely use informative and relevant data for the training of our machine learning models, we only selected personality scores (i. e., item scores) as features that are significant at the 10% level.

Table 2.2: Correlations between subjects' true cooperativeness and their linguistic scores.

LIWC category	**Label**	**Examples**	**r**	**p**
3rd pers plural	they	sie, deren, ihrem	0.126	0.063
Common Adverbs	adverb	außerdem, dabei, gar	0.115	0.090
Anxiety	anx	ängstlich, besorgt	0.137	0.043
Sadness	sad	schluchzen, träne, trauer	-0.146	0.032
Health	health	erkältet, klinik, medikament	0.154	0.023
Drives	drives	freund, erfolg, gemobbt	0.117	0.085
Future focus	focusfuture	bald, später, wird	-0.127	0.063
Religion	relig	fromm, kirche	0.113	0.097
Periods	Period	-	-0.120	0.078

Notes: The analysis only includes subjects from the *Control* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions. r gives Pearson correlation coefficients. p gives the respective p-values.

For the *Control* group, Table 2.2 shows the Pearson correlation coefficients (r) and respective p-values for the correlations between subjects' true cooperativeness and their linguistic scores. The table only includes correlations that are significant at the 10% level. As can be seen, nine LIWC categories are associated with subjects' true cooperativeness. While *Sadness*, *Future focus*, and *Periods* are negatively correlated with subjects' true cooperativeness, *3rd pers plural*, *Common Adverbs*, *Anxiety*, *Health*, *Drives*, and *Religion* show positive correlations. Since the table only includes results that are significant at the 10% level, it shows all the linguistic scores that we selected as features for the training of our machine learning models.

2.5.3 Predictions in the absence of incentives to fake

Following the machine learning approach, as described in Section 2.4, we trained different machine learning models using the selected personality and linguistic scores from the *Control* group as features.

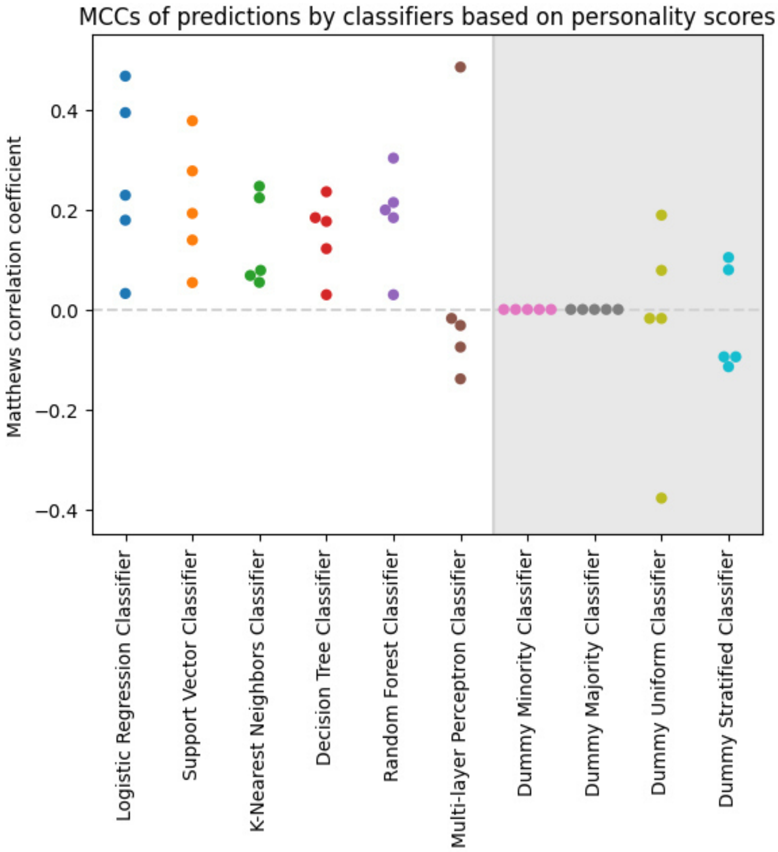


Figure 2.4: Swarmplots showing Matthews correlation coefficients for predictions on subjects' true cooperativeness based on personality scores in the training process.

Notes: The analysis only includes subjects from the *Control* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). The predictions were made by classifiers in the outer loop of a nested stratified 5-fold cross-validation.

Figure 2.4 shows the distribution of the MCCs for predictions by classifiers based on personality scores in the outer loop of the training process. As can be seen, except for predictions based on the Multi-layer Perceptron Classifier, the MCCs are all greater than zero.[12] For four out of six classifiers (Logistic Regression Classifier, Support Vector Classifier, Decision Tree Classifier, and Random Forest Classifier), the median MCCs are close to 0.2, indicating a weak positive relationship between the predicted cooperativeness for subjects and their true cooperativeness.

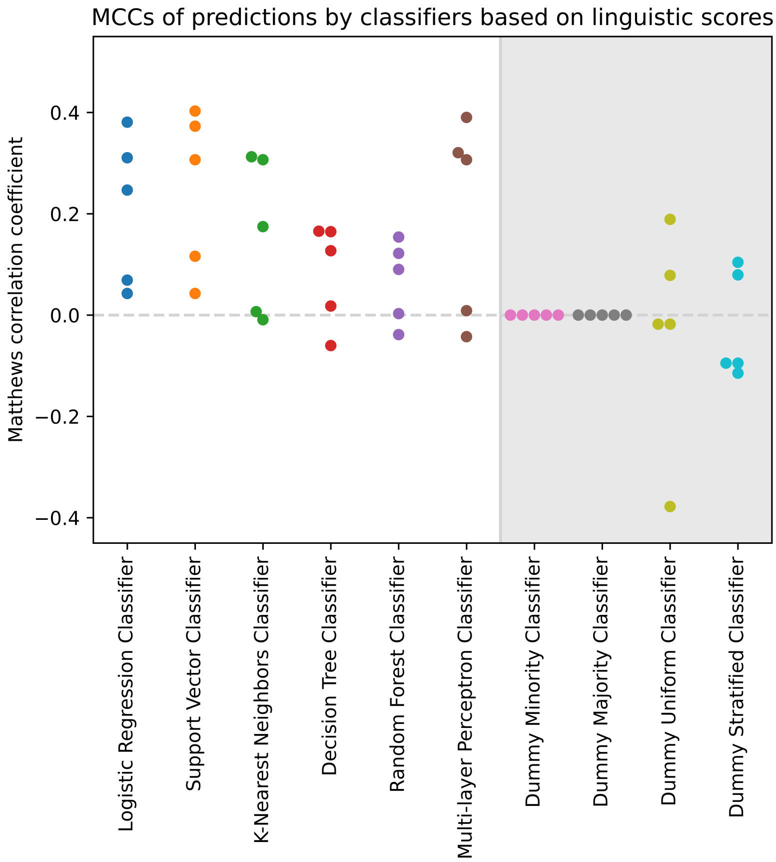


Figure 2.5: Swarmplots showing Matthews correlation coefficients for predictions on subjects' true cooperativeness based on linguistic scores in the training process.

Notes: The analysis only includes subjects from the *Control* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions. The predictions were made by classifiers in the outer loop of a nested stratified 5-fold cross-validation.

Figure 2.5 shows the distribution of the MCCs for predictions by classifiers based on linguistic scores in the outer loop of the training process. As can be seen, most MCCs are greater than zero, with median MCCs ranging from 0.09 (negligible relationship) to 0.31 (moderate positive relationship).[13]

[12]Ignoring the MCCs of the predictions by the dummy classifiers.
[13]Ignoring the MCCs of the predictions by the dummy classifiers.

To test whether our classifiers' predictions are significantly better than chance, we conducted pairwise McNemar's tests to check if there are differences between the predictions of the Dummy Stratified Classifier (i. e., random predictions) and the other classifiers (Dietterich, 1998; McNemar, 1947). As discussed in Section 2.4, we chose the Dummy Stratified Classifier as a benchmark since it best represents sophisticated HR managers. For each classifier, we tested the predictions of the model with the best hyperparameter set based on the MCCs of the outer loop of the training process. We used the predictions of this model, since it has the same hyperparameter set as the final model which we subsequently used for the prediction on the *Treatment* data.

Table 2.3: Results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness based on personality/linguistic scores and those of the Dummy Stratified Classifier.

Scores	**Classifier**	**p**	**True/ True**	**True/ False**	**False/ True**	**False/ False**
Personality	Logistic Regression	0.115	19	14	6	5
Personality	Support Vector	0.383	17	13	8	6
Personality	K-Nearest Neighbors	0.238	15	12	6	10
Personality	Decision Tree	0.167	15	13	6	10
Personality	Random Forest	0.481	18	11	7	7
Personality	Multi-layer Perceptron	0.115	19	14	6	5
Linguistic	Logistic Regression	0.049	17	13	4	9
Linguistic	Support Vector	0.030	17	14	4	8
Linguistic	K-Nearest Neighbors	0.108	13	17	8	6
Linguistic	Decision Tree	0.690	10	14	11	8
Linguistic	Random Forest	0.286	13	14	8	9
Linguistic	Multi-layer Perceptron	0.064	14	17	7	6

Notes: Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions. Results show the p-values of McNemar's tests and the associated contingency tables. True/-False gives a correct/wrong prediction, where the first term represents the classifier based on personality/linguistic scores, and the second term represents the Dummy Stratified Classifier.

Table 2.3 shows the results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness based on their personality and linguistic scores and those of the Dummy Stratified Classifier. The results show that no classifier based on personality scores makes significantly better than chance predictions.

Result 1. *In the absence of incentives to fake, no classifier based on personality scores makes significantly better than chance predictions.*

Hence, we cannot meaningfully interpret their predictions in the next part when we use these classifiers to make predictions on the data of the *Treatment* group. However, the McNemar's tests which compare predictions by the Logistic Regression Classifier and Multi-layer Perceptron Classifier with the Dummy Stratified Classifier, achieve a p-value of 11.5% and are therefore almost significant at the 10% level. Thus, albeit with a grain of salt, we can take a look at the predictions of these two classifiers to gain an insight into the predictive power of personality scores for predictions on cooperativeness in the presence of incentives to fake.

Regarding predictions based on linguistic scores, three classifiers (Logistic Regression, Support Vector, and Multi-layer Perceptron) make significantly better than chance predictions.

Result 2. *In the absence of incentives to fake, three classifiers (Logistic Regression Classifier, Support Vector Classifier, and Multi-layer Perceptron Classifier) make significantly better than chance predictions.*

Therefore, we can meaningfully interpret their predictions in the next part when we use these classifiers to make predictions on the data of the *Treatment* group. Overall, we conclude that, with our data, only linguistic scores achieve significantly better than chance predictions in the absence of incentives to fake.

The large variance in the distribution of the MCCs in Figure 2.4 and Figure 2.5 shows the strong dependence of the classifiers' performance on the train-test split of the outer loop. To reduce the impact of the train-test split on the performance of the classifiers, we retrained each classifier's final model with the entire data of the *Control* group before making predictions based on the data from the *Treatment* group.

2.5.4 Predictions in the presence of incentives to fake

In this section, we study the predictive power of cooperativeness predictions when subjects have incentives to fake being cooperative. To this end, we used the personality and linguistic scores from the *Treatment* group as features to predict whether subjects' true cooperativeness is above the median or not.

However, before studying the effect of incentives to fake, we first need to verify that our treatment manipulation was effective. To check whether this is the case, we conducted the following two tests. First, we studied the within-dimension of subjects' cooperativeness. That is, we tested if there is a significant difference between subjects' contributions in the first and second public goods game in the *Treatment* group but not in the *Control* group. Second, we studied the between-dimension of subjects' cooperativeness. More precisely, we tested if there is a significant difference between the contributions in the second public goods game between subjects from the *Control* and *Treatment* group (see Section 2.7.3 for additional information on these dimensions).

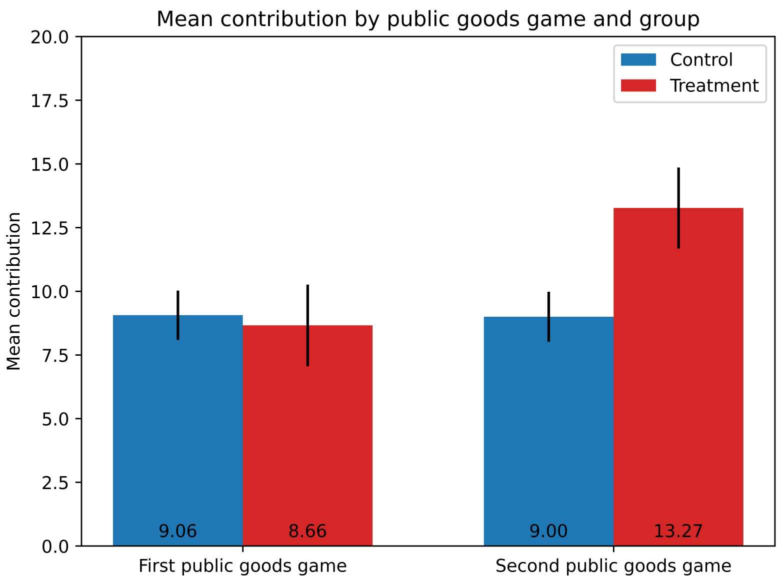


Figure 2.6: Mean contributions in the first and second public goods game by group.

Notes: Mean contributions in the first and second public goods game of the *Control* (blue) and *Treatment* (red) group. In the first and second public goods game, subjects had to decide how much of their endowment (20 points) they want to contribute to a public good. Before playing the second public goods game, subjects in the *Treatment* group were incentivized to fake being cooperative. The error bars indicate 95% confidence intervals.

Figure 2.6 shows the mean contributions in the first and second public goods game of the *Control* (blue) and *Treatment* (red) group. A pairwise comparison shows that there is no significant difference between the contributions of subjects in the *Control* group between the first (mean=9.06) and second (mean=9.00) public goods game (two-sided Wilcoxon signed-rank test: p=0.773). However, there is a significant difference between the contributions of subjects in the *Treatment* group between the first (8.66) and second (13.27) public goods game (two-sided Wilcoxon signed-rank test: $p<0.001$). Furthermore, there is also a significant difference between the contributions of subjects in the *Control* (mean=9.00) and *Treatment* (mean=13.27) group in the second public goods game (two-sided Mann-Whitney U test: $p<0.001$). Based on these findings, we conclude that our treatment manipulation was highly effective. That is, subjects in the *Treatment* group tried to fake being more cooperative, whereas this was not the case for subjects in the *Control* group.

Next, we study the effect of incentives to fake on personality scores obtained from self-reported personality tests.

Table 2.4: Two-sided Mann–Whitney U tests between subjects' personality scores from the *Control* and *Treatment* group.

Personality trait/Item	**Mean score (*Control*)**	**Mean score (*Treatment*)**	**p**
Openness	5.46	5.54	0.579
Open to new experiences, complex	5.97	6.11	0.288
Conventional, uncreative	3.05	3.03	0.935
Conscientiousness	5.74	5.89	0.397
Dependable, self-disciplined	5.76	5.86	0.648
Disorganized, careless	2.29	2.08	0.245
Extraversion	4.77	4.74	0.654
Extraverted, enthusiastic	4.98	4.97	0.604
Reserved, quiet	3.43	3.49	0.747
Agreeableness	5.16	5.37	0.214
Critical, quarrelsome	3.34	3.08	0.189
Sympathetic, warm	5.66	5.81	0.753
Neuroticism	4.99	5.37	0.017
Anxious, easily upset	2.99	2.68	0.052
Calm, emotionally stable	4.97	5.42	0.025

Notes: Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). p gives the respective p-values.

Table 2.4 shows the results of two-sided Mann–Whitney U tests between subjects' personality scores from the *Control* and *Treatment* group. As can be seen, only the scores of the personality trait of *Neuroticism* and its two associated items differ significantly between the *Control* and *Treatment* group (two-sided Mann-Whitney U test: $p=0.017$). However, since we did not use these items as features to train our machine learning classifiers, this does not pose a problem. In contrast to what we expected, subjects in the *Treatment* group did not fake to be more agreeable than those in the *Control* group. Hence, overall, our selected personality scores prove to be quite robust to incentives to fake. Thus, they could potentially be a good predictor for subjects' true cooperativeness in the presence of incentives to fake.

Result 3. *Our selected personality scores prove to be quite robust to incentives to fake.*

Next, we study the effect of incentives to fake on our selected linguistic scores obtained from written self-descriptions.

Table 2.5: Two-sided Mann–Whitney U tests between subjects' linguistic scores from the *Control* and *Treatment* group.

LIWC category	**Label**	**Mean score (*Control*)**	**Mean score (*Treatment*)**	**p**
3rd pers plural	they	0.31	0.42	0.317
Common Adverbs	adverb	6.36	6.06	0.219
Anxiety	anx	0.20	0.15	0.239
Sadness	sad	0.70	0.66	0.597
Health	health	0.96	0.92	0.982
Drives	drives	9.08	10.34	< 0.001
Future focus	focusfuture	0.85	0.86	0.982
Religion	relig	0.15	0.13	0.354
Periods	Period	5.88	5.53	0.008

Notes: Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions. p gives the respective p-values.

Table 2.5 shows the results of two-sided Mann–Whitney U tests between subjects' linguistic scores from the *Control* and *Treatment* group. As can be seen, there are only two linguistic scores that differ significantly between the two groups, which are "Drives", which includes examples such as "freund, erfolg, gemobbt", and *Periods.*[14] Hence, overall, our selected linguistic scores prove to be quite robust to incentives to fake. Thus, they should be a good predictor for subjects' true cooperativeness in the presence of incentives to fake.

Result 4. *Our selected linguistic scores prove to be quite robust to incentives to fake.*

[14]The English translation of these examples is: "friend, success, harassed."

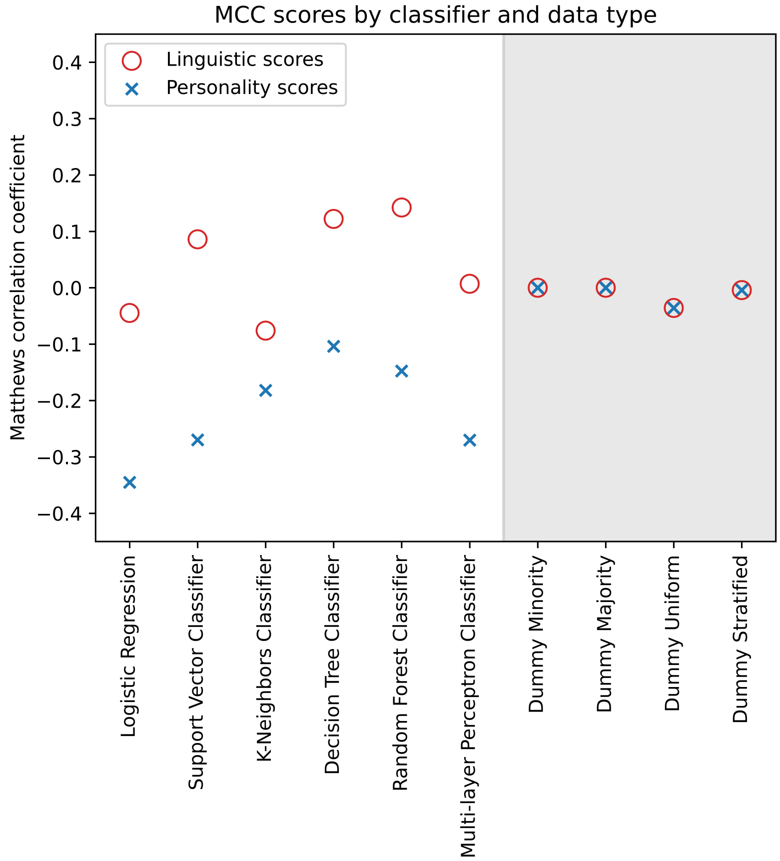


Figure 2.7: Matthews correlation coefficients for predictions on subjects' true cooperativeness based on their personality/linguistic scores.

Notes: The analysis only includes subjects from the *Treatment* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions.

Figure 2.7 shows, for subjects in the *Treatment* group, the MCCs for predictions on subjects' true cooperativeness based on their personality and linguistic scores by different classifiers.[15] As can be seen, the MCCs of each classifier are higher when the prediction was based on linguistic scores than when it was based on personality scores. This observation suggests that predictions by classifiers based on linguistic scores are superior to predictions by classifiers based on personality scores in the presence of incentives to fake. Furthermore, the results show that the MCCs of three classifiers (Support Vector, Decision Tree, and Random Forest) based on linguistic

[15]See Section 2.7.4 for a robustness check with the evaluation scores for Precision, Recall, Accuracy, and F1.

scores are higher than those of the dummy classifiers. In contrast, all MCCs of predictions by classifiers based on personality scores are lower than those of the dummy classifiers. To test whether predictions by classifiers based on personality and linguistic scores are significantly different than chance predictions (i. e., better than the Dummy Stratified Classifier), we conducted pairwise McNemar's tests.

Table 2.6: Results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness based on their personality/linguistic scores and those of the Dummy Stratified Classifier.

Scores	**Classifier**	**p**	**True/ True**	**True/ False**	**False/ True**	**False/ False**
Personality	Logistic Regression	0.006	15	9	26	29
Personality	Support Vector	0.041	17	11	24	27
Personality	K-Nearest Neighbors	0.200	22	11	19	27
Personality	Decision Tree	0.243	19	14	22	24
Personality	Random Forest	0.362	23	12	18	26
Personality	Multi-layer Perceptron	0.024	17	10	24	28
Linguistic	Logistic Regression	0.302	14	19	27	19
Linguistic	Support Vector	0.636	23	22	18	16
Linguistic	K-Nearest Neighbors	0.749	20	18	21	20
Linguistic	Decision Tree	0.451	22	25	19	13
Linguistic	Random Forest	0.222	24	26	17	12
Linguistic	Multi-layer Perceptron	0.644	18	19	23	19

Notes: The analysis only includes subjects from the *Treatment* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions. Results show the p-values of McNemar's tests and the associated contingency tables. True/False gives a correct/wrong prediction, where the first term represents the classifier based on personality/linguistic scores, and the second term represents the Dummy Stratified Classifier.

Table 2.6 shows the results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness based on their personality and linguistic scores and those of the Dummy Stratified Classifier. The results show that, for classifiers based on personality scores, the McNemar's tests for three classifiers (Logistic Regression, Support Vector, and Multi-layer Perceptron) achieve significant p-values. However, in all these cases, the MCCs are below those of the Dummy Stratified Classifier. Therefore, we conclude that in the presence of incentives to fake, assessments based on personality (Big Five) scores obtained from self-reported personality tests fail to make better than chance cooperativeness predictions.

Result 5. *In the presence of incentives to fake, classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions.*

However, this finding raises the question of whether the significant inverse relationship between personality scores and subjects' true cooperativeness might be useful as a reversed predictor. If that is the case, then personality scores, even though they are prone to faking, might still be useful when it comes to predicting subjects' true personality traits. This question ought to be addressed by future studies in this field.

The results of Table 2.6 show that, for classifiers based on linguistic scores, the predictions are not significantly better than those of a Dummy Stratified Classifier. Therefore, we conclude that in the presence of incentives to fake, assessments based on linguistic (LIWC) scores obtained from written self-descriptions fail to make better than chance cooperativeness predictions.

Result 6. *In the presence of incentives to fake, classifiers that make predictions on subjects' cooperativeness based on linguistic scores fail to make significantly better than chance predictions.*

2.6 Conclusion

In this study, we analyze how incentives to fake affect the predictive power of personality assessments. We provide answers to this question by following the example of hiring processes, which are typically poised with faking. More precisely, we study a situation in which companies try to predict applicants' true cooperativeness based on data from tasks that are common during an application process, namely self-reported personality tests and written self-descriptions.

To gather data that mimic this situation, we conducted an experiment in which we elicited subjects' true cooperativeness through a first public goods game. Next, we introduced a treatment manipulation by incentivizing subjects in the *Treatment* group, but not in the *Control* group, to fake being cooperative. Subsequently, subjects performed a written self-description, a self-reported personality test, and a second public goods game.

To shed light on our research question, we studied the predictive power of predictions on subjects' true cooperativeness based on data from the self-reported personality test and written self-description in the presence of incentives to fake. To make these predictions, we used machine learning classifiers which we trained with supervised learning on data from the *Control* group. As labels, we used the binary outcome of whether subjects' true cooperativeness is above the median of the *Control* group or not. As features, we used personality (Big Five) and linguistic (LIWC) scores, which we obtained from the self-reported personality test and written self-description. With the trained classifiers, we then predicted, for subjects in the *Treatment* group, whether their true cooperativeness is above the median of the *Control* group or not. To study the predictive power of these predictions, we used Matthews correlation coefficient, and as a robustness check, other evaluation metrics for binary classification.

Our results show that in the absence of incentives to fake, machine learning classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions, whereas classifiers based on linguistic scores succeed. In the presence of incentives to fake, both approaches fail to make significantly better than chance predictions.

The limited ability of our classifiers to make better than chance predictions could be due to shortcomings in our methodological approach, which can be mitigated in future studies as follows. First, written self-descriptions are more prone to deception than spoken self-descriptions. This is because written responses allow for sufficient time to reflect, whereas spoken responses require immediate action, which yields spontaneous, undisguised responses. Hence, assessing applicants' personality traits based on spoken self-descriptions might yield better results. Second, the scientific literature shows that linguistic characteristics are dependent on age and gender (Newman et al., 2008; Pennebaker et al., 2003). Hence, using linguistic features that are tailored to demographic characteristics might prove to yield better results. Third, in terms of its grammar, syntax, and other characteristics, the German language differs considerably from other languages. Due to the peculiarities of the German language, assessing personality traits from linguistic characteristics might work better in other languages. Fourth, extending our approach to predict other personality traits that are more strongly correlated with linguistic features, for example, extroversion (Mairesse et al., 2007), might considerably improve the results. Fifth, given that many machine learning models reach their full potential only with large data sets, the results drawn from our very small sample size might improve considerably by using more data.

These shortcomings provide an ideal starting point for future research. Overall, our findings underline the promising nature of language as a predictor for personality traits, even when incentives to fake are present. Hence, following this approach might prove to be a fruitful avenue for future research in this field.

2.7 Appendix

2.7.1 LIWC categories

Table 2.7: LIWC categories
Source: Modified version of Table 1 by Meier et al. (2015).

Category	**Label**	**Examples**
Word Count	WC	-
Summary Variables		
Analytic Thinking	Analytic	-
Clout	Clout	-
Authentic	Authentic	-
Emotional tone	Tone	-
Words/sentence	WPS	-
Words > 6 letters	Sixltr	-
Dictionary words	Dic	-
Linguistic Dimensions		
Total function words	funct	es, zu, nicht, sehr
Total pronouns	pronoun	ich, sie, man
Personal pronouns	ppron	ich, sie, ihm
1st pers singular	i	ich, mir mein
1st pers plural	we	wir, uns, unsere
2nd person	you_total	du, dein, dich, sie, ihr, euch
2nd pers singular	you_sing	du, dein, dich
2nd pers plural	you_plur	euch, euer, ihr,
2nd pers formal	you_formal	sie, ihr, ihnen
3rd person	other	sie, ihr, ihm, deren, ihrem
3rd pers singular	shehe	sie, ihr, ihm
3rd pers plural	they	sie, deren, ihrem
Impersonal pronouns	ipron	man, all, manche
Articles	article	ein, der, die, nen
Prepositions	prep	ab, auf, danach
Auxiliary verbs	auxverb	bin, habt, geht's
Common Adverbs	adverb	außerdem, dabei, gar
Conjunctions	conj	anstatt, auch, und
Negations	negate	kein, nein, nichts
Other Grammar		
Common verbs	verb	abreist, besuchen, esse
Common adjectives	adj	lange, frei, schön
Comparisons	compare	ähnlich, älter, wichtiger

Interrogatives	interrog	inwiefern, wann, warum
Numbers	number	acht, eins, halb
Quantifiers	quant	viel, wenig, ziemlich
Psychological Processes		
Affective processes	affect	glücklich, weinen
Positive emotion	posemo	glücklich, liebe, schön
Negative emotion	negemo	beleidigt, bösartig, heulen
Anxiety	anx	ängstlich, besorgt
Anger	anger	hass, sauer, zorn
Sadness	sad	schluchzen, träne, trauer
Social processes	social	gesellig, kumpel, reden
Family	family	papa, tochter, tante
Friends	friend	bro, kumpel
Female references	female	frau, mädchen, weiblich
Male references	male	bruder, mann, onkel
Cognitive processes	cogproc	denken, weil, wissen
Insight	insight	denken, realisieren
Causation	cause	deswegen, grund
Discrepancy	discrep	sollte, wollte
Tentative	tentat	eventuell, vielleicht
Certainty	certain	immer, sicher
Differentiation	differ	aber, sonst
Perceptual processes	percept	fühle, höre, schauen
See	see	angeschaut, sehe, sicht
Hear	hear	höre, klang, zuhören
Feel	feel	fühle, fühlt, glatt
Biological processes	bio	essen, blut, schmerz
Body	body	arm, kopf, muskel
Health	health	erkältet, klinik, medikament
Sexual	sexual	geil, heiß, nackt
Ingestion	ingest	hunger, mahlzeit, pizza
Drives	drives	freund, erfolg, gemobbt
Affiliation	affiliation	allianz, freund, sozial
Achievement	achieve	besser, erfolg, sieg
Power	power	gemobbt, herrscher
Reward	reward	jubel, medaille
Risk	risk	gefahr, kritisch
Time orientations		
Past focus	focuspast	früher, gestern, war
Present focus	focuspresent	aktuell, bin, heute
Future focus	focusfuture	bald, später, wird
Relativity	relative	gegend, region, plötzlich
Motion	motion	ankunft, auto, gehen
Space	space	unten, über, klein
Time	time	ab, bisher, dauerhaft
Personal concerns		
Work	work	beruf, job, hochschule
Leisure	leisure	aktivität, kino, reise
Home	home	sofa, wohnzimmer
Money	money	rechnung, schuld, teuer
Religion	relig	fromm, kirche

Death	death	begräbnis, tod
Informal language	informal	aufm, lol, cool
Swear words	swear	depp, drecksack, motherfucker
Netspeak	netspeak	likes, lol, ok
Assent	assent	gell, genau, ja
Nonfluencies	nonflu	äh, oh, hm
Fillers	filler	naja, wasweißich, sozusagen
-	AllPunc	-
-	Period	-
-	Comma	-
-	Colon	-
-	SemiC	-
-	QMark	-
-	Exclam	-
-	Dash	-
-	Quote	-
-	Apostro	-
-	Parenth	-
-	OtherP	-

Table 2.7 shows all the LIWC categories of the German dictionary (version LIWC2015) that we used for our study.

2.7.2 Hyperparameter grid

The following code shows the hyperparameter grid, which we specified for the *GridSearchCV* from the *scikit-learn* library.

```
classifier_and_hyperparameter = {
    'LogisticRegression': (LogisticRegression(),
            {
            'logisticregression__penalty': ['l1', 'l2', 'elasticnet', 'none'],
            'logisticregression__dual': [True, False],
            'logisticregression__tol': list(10.0 ** np.arange(-5, -2)),
            'logisticregression__C': list(10.0 ** np.arange(-1, 2)),
            'logisticregression__fit_intercept': [True, False],
            'logisticregression__intercept_scaling': list(10.0 ** np.arange(-1, 2)),
            'logisticregression__class_weight': ['balanced', None],
            'logisticregression__random_state': [random_seed],
            'logisticregression__solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
            'logisticregression__max_iter': list(10.0 ** np.arange(1, 4)),
            'logisticregression__multi_class': ['auto'],
            'logisticregression__verbose': [0],
            'logisticregression__warm_start': [True, False],
            'logisticregression__n_jobs': [-1],
            'logisticregression__l1_ratio': [float(x) for x in np.linspace(start=0, stop=1, num=3)]
            }),
    'SupportVectorClassifier': (SVC(),
            {
            'svc__C': list(10.0 ** np.arange(-1, 2)),
            'svc__kernel': ['linear', 'poly', 'rbf', 'sigmoid', 'precomputed'],
            'svc__degree': [2, 3, 4],
            'svc__gamma': ['scale', 'auto'],
            'svc__coef0': [float(x) for x in np.linspace(start=-1, stop=1, num=3)],
            'svc__shrinking': [True, False],
```

```
            'svc__probability': [True, False],
            'svc__tol': list(10.0 ** np.arange(-4, -1)),
            'svc__cache_size': [float(200)],
            'svc__class_weight': [None, 'balanced'],
            'svc__verbose': [False],
            'svc__max_iter': [-1],
            'svc__decision_function_shape': ['ovo', 'ovr'],
            'svc__break_ties': [False],
            'svc__random_state': [random_seed]
            }),
    'KNeighborsClassifier': (KNeighborsClassifier(),
            {
            'kneighborsclassifier__n_neighbors': [int(x) for x in np.linspace(start=1, stop=10, num=10)],
            'kneighborsclassifier__weights': ['uniform', 'distance'],
            'kneighborsclassifier__algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
            'kneighborsclassifier__leaf_size': [10, 20, 30],
            'kneighborsclassifier__p': [2],
            'kneighborsclassifier__metric': ['minkowski'],
            'kneighborsclassifier__metric_params': [None],
            'kneighborsclassifier__n_jobs': [-1]
            }),
    'DecisionTreeClassifier': (DecisionTreeClassifier(),
            {
            'decisiontreeclassifier__criterion': ['gini', 'entropy'],
            'decisiontreeclassifier__splitter': ['best', 'random'],
            'decisiontreeclassifier__max_depth': [None],
            'decisiontreeclassifier__min_samples_split': [2],
            'decisiontreeclassifier__min_samples_leaf': [1],
            'decisiontreeclassifier__min_weight_fraction_leaf': [float(0)],
            'decisiontreeclassifier__max_features': ['auto', 'sqrt', 'log2', None],
            'decisiontreeclassifier__random_state': [random_seed],
            'decisiontreeclassifier__max_leaf_nodes': [None],
            'decisiontreeclassifier__min_impurity_decrease': [float(0)],
            'decisiontreeclassifier__min_impurity_split': [float(0)],
            'decisiontreeclassifier__class_weight': ['balanced', None],
            'decisiontreeclassifier__ccp_alpha': [float(0)]
            }),
    'RandomForestClassifier': (RandomForestClassifier(),
            {
            'randomforestclassifier__n_estimators': [int(x) for x in np.linspace(start=10, stop=100, num=10)],
            'randomforestclassifier__criterion': ['gini', 'entropy'],
            'randomforestclassifier__max_depth': [None],
            'randomforestclassifier__min_samples_split': [2],
            'randomforestclassifier__min_samples_leaf': [1],
            'randomforestclassifier__min_weight_fraction_leaf': [float(0)],
            'randomforestclassifier__max_features': ['auto', 'sqrt', 'log2'],
            'randomforestclassifier__max_leaf_nodes': [None],
            'randomforestclassifier__min_impurity_decrease': [float(0)],
            'randomforestclassifier__bootstrap': [True, False],
            'randomforestclassifier__oob_score': [True, False],
            'randomforestclassifier__n_jobs': [-1],
            'randomforestclassifier__random_state': [random_seed],
            'randomforestclassifier__verbose': [0],
            'randomforestclassifier__warm_start': [True, False],
            'randomforestclassifier__class_weight': ['balanced', None],
            'randomforestclassifier__ccp_alpha': [float(0)],
            'randomforestclassifier__max_samples': [None]
            }),
    'MLPClassifier': (MLPClassifier(),
            {
            'mlpclassifier__hidden_layer_sizes': [(5,), (10,), (15,), (20,),
                                                  (5, 5), (10, 10), (15, 15), (20, 20)],
            'mlpclassifier__activation': ['identity', 'logistic', 'tanh', 'relu'],
            'mlpclassifier__solver': ['lbfgs', 'sgd', 'adam'],
            'mlpclassifier__alpha': list(10.0 ** np.arange(-5, -2)),
            'mlpclassifier__batch_size': ['auto'],
            'mlpclassifier__learning_rate': ['constant', 'invscaling', 'adaptive'],
            'mlpclassifier__learning_rate_init': list(10.0 ** np.arange(-4, -1)),
            'mlpclassifier__power_t': [0.5],
            'mlpclassifier__max_iter': [200],
            'mlpclassifier__shuffle': [True, False],
```

```
            'mlpclassifier__random_state': [random_seed],
            'mlpclassifier__tol': list(10.0 ** np.arange(-5, -2)),
            'mlpclassifier__verbose': [False],
            'mlpclassifier__warm_start': [True, False],
            'mlpclassifier__momentum': [0.9],
            'mlpclassifier__nesterovs_momentum': [True],
            'mlpclassifier__early_stopping': [False],
            'mlpclassifier__validation_fraction': [0.1],
            'mlpclassifier__beta_1': [0.9],
            'mlpclassifier__beta_2': [0.999],
            'mlpclassifier__epsilon': list(10.0 ** np.arange(-8, -7)),
            'mlpclassifier__n_iter_no_change': [10],
            'mlpclassifier__max_fun': [15000]
            }),
    'DummyMinority': (DummyClassifier(),
            {
            'dummyclassifier__strategy': ['constant'],
            'dummyclassifier__random_state': [random_seed],
            'dummyclassifier__constant': [1]
            }),
    'DummyMajority': (DummyClassifier(),
            {
            'dummyclassifier__strategy': ['most_frequent'],
            'dummyclassifier__random_state': [random_seed]
            }),
    'DummyUniform': (DummyClassifier(),
            {
            'dummyclassifier__strategy': ['uniform'],
            'dummyclassifier__random_state': [random_seed]
            }),
    'DummyStratified': (DummyClassifier(),
            {
            'dummyclassifier__strategy': ['stratified'],
            'dummyclassifier__random_state': [random_seed]
            })
    }
```

2.7.3 Manipulation check

In this section, we provide more information on our manipulation check by presenting the results of the between and within dimension on a more granular level.

2.7.3.1 Within dimension

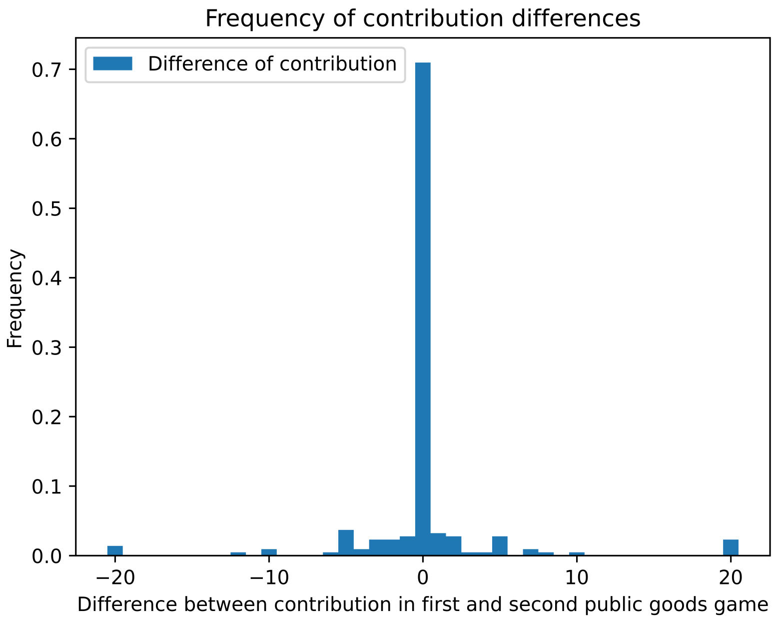


Figure 2.8: Frequency of differences between the contribution in the first and second public goods game for subjects in the *Control* group.

Notes: The analysis only includes subjects from the *Control* group. In the first and second public goods game, subjects had to decide how much of their endowment (20 points) they want to contribute to a public good.

Figure 2.8 shows the within-dimension of subjects' contributions in the *Control* group on a more fine-grained level. In particular, it shows the difference between subjects' contribution in the first and second public goods game on the x-axis and the respective frequency of observations on the y-axis. The figure shows that a large majority (about 71%) are internally consistent by contributing the same amount in both public goods games. Hence, we conclude that subjects did not exhibit boredom or wealth effects that might affect our results.

2.7.3.2 Between dimension

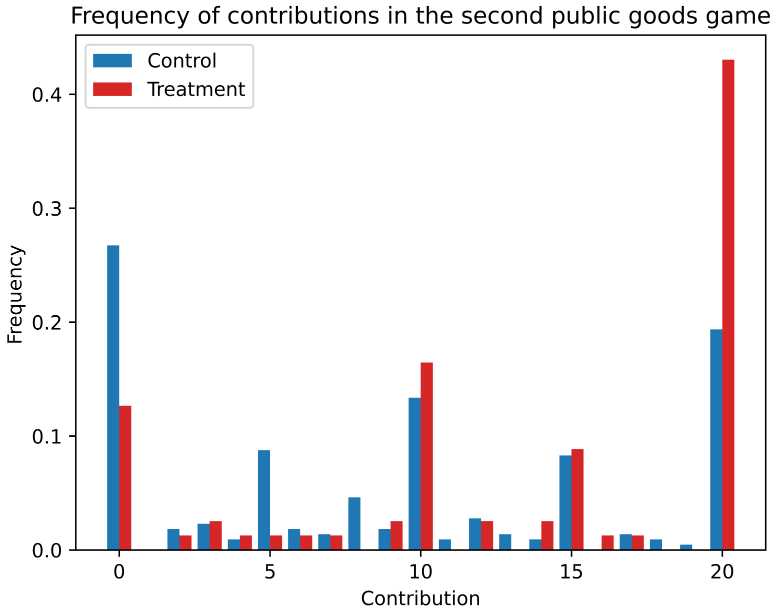


Figure 2.9: Frequency of contributions in the second public goods game: *Control* (blue) vs. *Treatment* (red).

Notes: In the second public goods game, subjects had to decide how much of their endowment (20 points) they want to contribute to a public good. Before playing the second public goods game, subjects in the *Treatment* group were incentivized to fake being cooperative.

Table 2.9 shows the results of the between-dimension of subjects' contributions on a more fine-grained level. In particular, it presents the contributions in the second public goods game for subjects in the *Control* (blue) and *Treatment* (red) group. The results of a two-sided Kolmogorov-Smirnov test shows that the cumulative distribution functions of the two groups differ significantly ($p<0.001$). Therefore, we conclude that our treatment manipulation was highly effective.

2.7.4 Robustness check

As a robustness check, we present in the following an analysis of our prediction results with different evaluation metrics. In line with our example of hiring processes, we chose these evaluation metrics to be best applicable for companies in different situations, depending on the characteristics of the labor market. The two key criteria for the characteristics of the labor market are the strength of employees' rights and the unemployment rate. If employees are very well protected, and it is difficult for employers to fire them, then employers want to minimize false positives. That is, employers want to avoid hiring an applicant whom they believe is very cooperative, but in fact, is not. Contrary, if there is a shortage of qualified applicants in the labor market, employers want to minimize false negatives. That is, employers do not want cooperative applicants to slip through their fingers. To account for these special situations, we use the following evaluation metrics.

Precision is best suited if false positives are very costly, i. e., if employees' rights are very well protected. It is defined as the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Recall is best suited if false negatives are very costly, i. e., if there is a shortage of cooperative applicants on the job market. It is defined as the ratio of correctly predicted positive observations to all observations in the actual class.

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

Accuracy is best suited if false positives and false negatives have similar costs, i. e., if it is equally important that employees' rights are very well protected and that there is a shortage of cooperative applicants on the job market. It is defined as the ratio of correctly predicted observations to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

F1 is best suited if false positives and false negatives have similar costs and classes are imbalanced, i. e., if there are more non-cooperative than cooperative applicants, or vice versa. It is defined as the weighted average of Precision and Recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2} \times (FP + FN)} \tag{2.6}$$

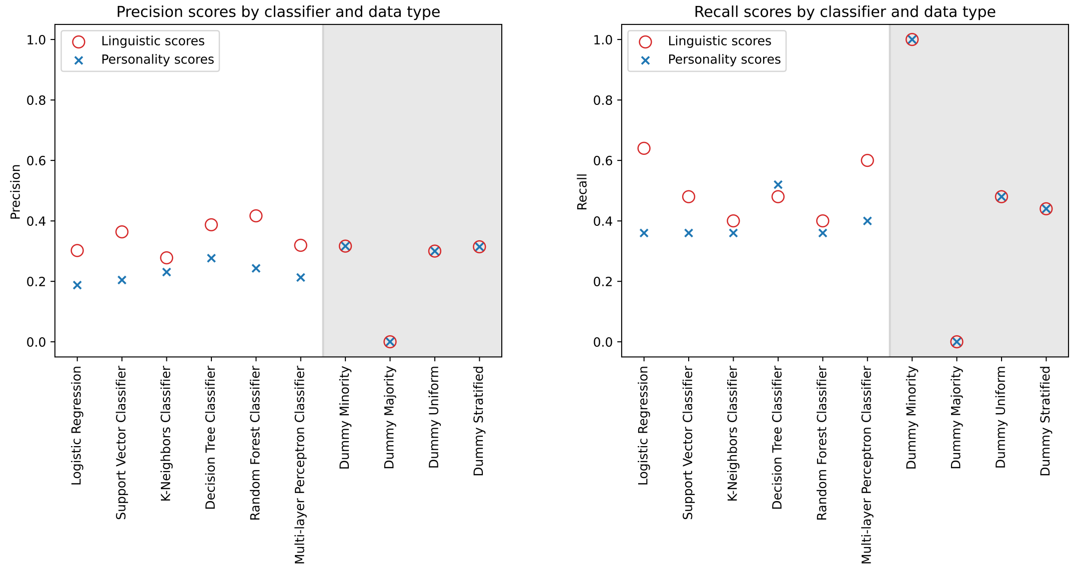


Figure 2.10: Precision (left) and Recall (right) scores for predictions on subjects' true cooperativeness based on their personality/linguistic scores (*Treatment* group).

Notes: The analysis only includes subjects from the *Treatment* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions.

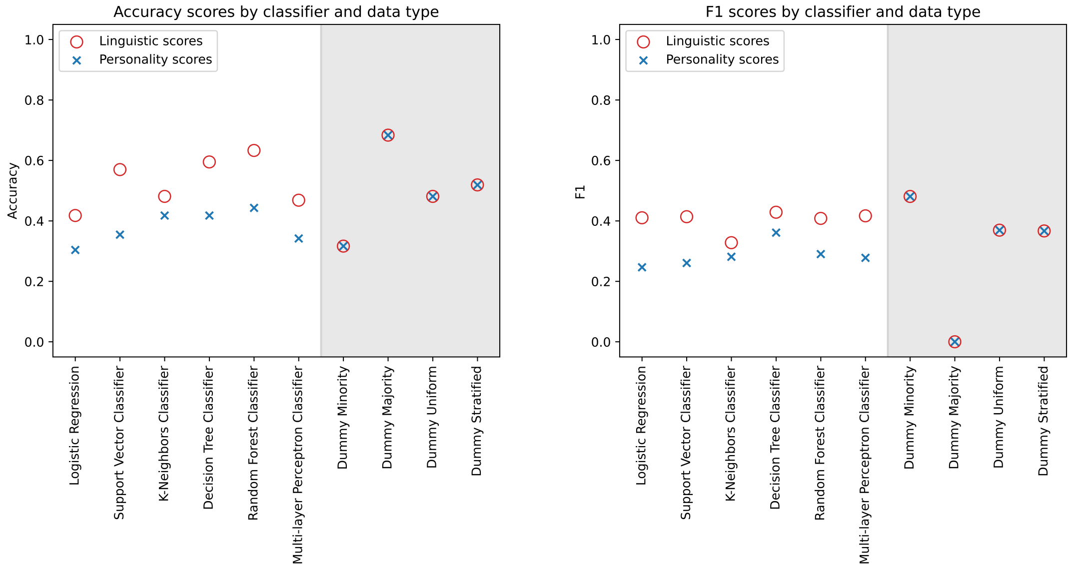
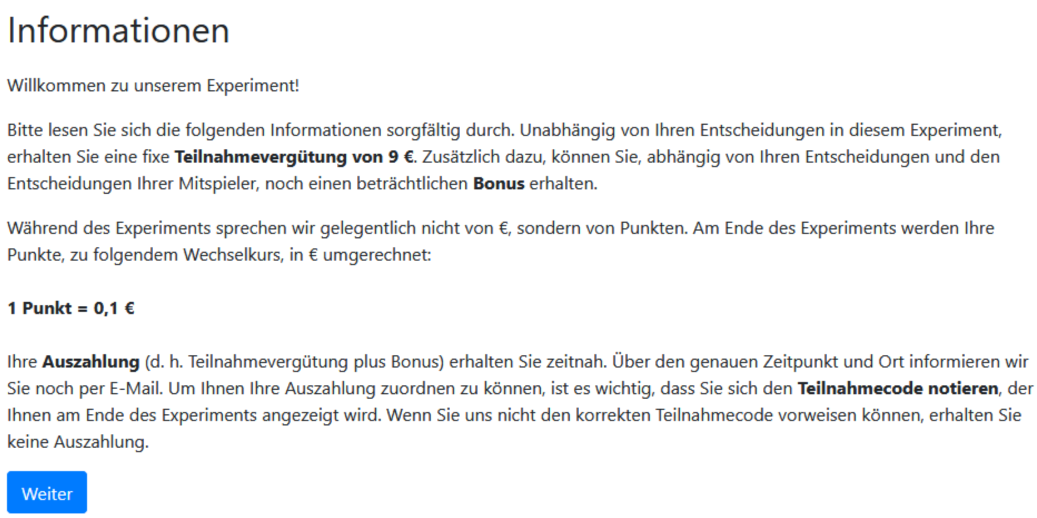


Figure 2.11: Accuracy (left) and F1 (right) scores for predictions on subjects' true cooperativeness based on their personality/linguistic scores (*Treatment* group).

Notes: The analysis only includes subjects from the *Treatment* group. Subjects' true cooperativeness is given by their contribution ([0, 20]) in a one-shot public goods game. Personality scores are given by the 7-point Likert scale responses to the items of the Big Five personality test by Gosling et al. (2003). Linguistic scores are given by the scores for the LIWC categories obtained from subjects' written self-descriptions.

Table 2.10 and 2.11 present the results of our predictions based on the data from the *Treatment* group with the additional evaluation metrics. As can be seen, classifiers that make predictions on subjects' cooperativeness based on linguistic scores appear to perform consistently better than those which use personality scores, independently of the evaluation metric. That is, independent of the characteristics of the labor market, linguistic scores appear to be the better features. However, since these predictions are not significantly better than chance, we cannot stress this visual observation. Overall, these results are in line with our main findings that are based on MCC as the evaluation metric.

2.7.5 Screenshots experiment



Informationen

Willkommen zu unserem Experiment!

Bitte lesen Sie sich die folgenden Informationen sorgfältig durch. Unabhängig von Ihren Entscheidungen in diesem Experiment, erhalten Sie eine fixe **Teilnahmevergütung von 9 €**. Zusätzlich dazu, können Sie, abhängig von Ihren Entscheidungen und den Entscheidungen Ihrer Mitspieler, noch einen beträchtlichen **Bonus** erhalten.

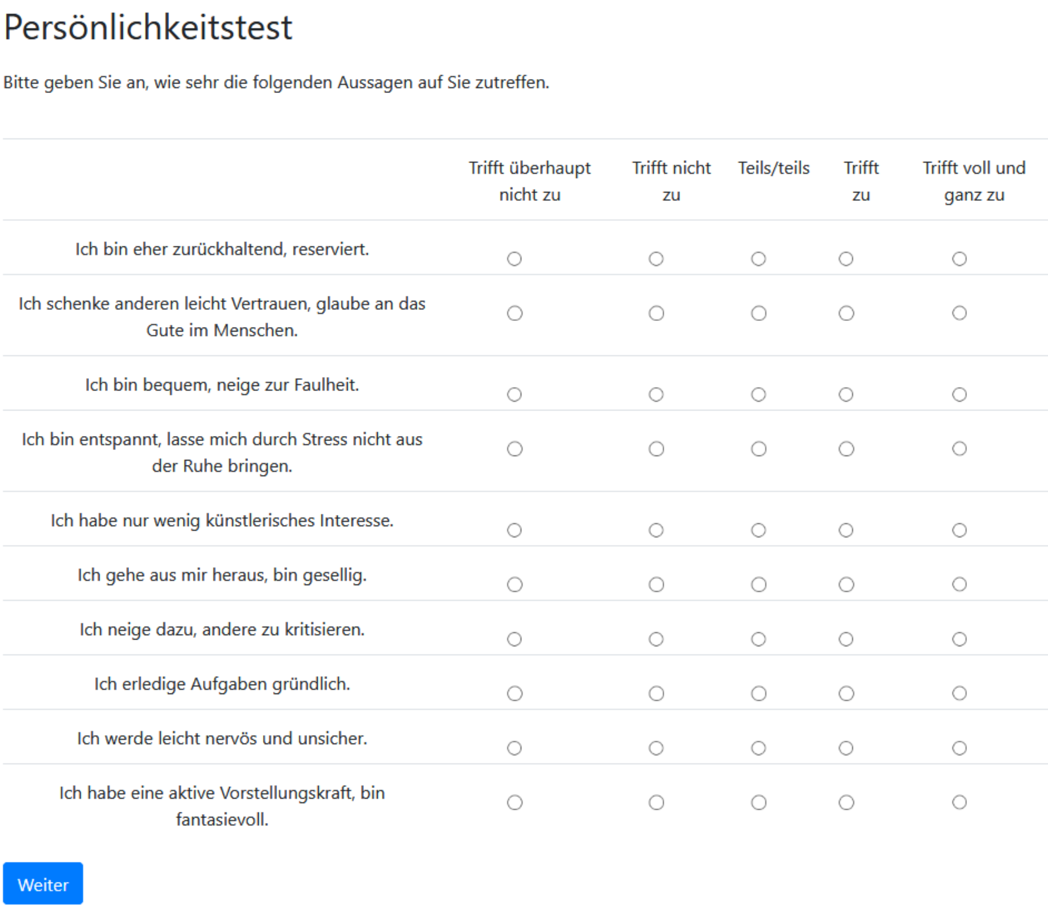
Während des Experiments sprechen wir gelegentlich nicht von €, sondern von Punkten. Am Ende des Experiments werden Ihre Punkte, zu folgendem Wechselkurs, in € umgerechnet:

1 Punkt = 0,1 €

Ihre **Auszahlung** (d. h. Teilnahmevergütung plus Bonus) erhalten Sie zeitnah. Über den genauen Zeitpunkt und Ort informieren wir Sie noch per E-Mail. Um Ihnen Ihre Auszahlung zuordnen zu können, ist es wichtig, dass Sie sich den **Teilnahmecode notieren**, der Ihnen am Ende des Experiments angezeigt wird. Wenn Sie uns nicht den korrekten Teilnahmecode vorweisen können, erhalten Sie keine Auszahlung.

Weiter

Figure 2.12: Screenshot of the online experiment: page 1.



Persönlichkeitstest

Bitte geben Sie an, wie sehr die folgenden Aussagen auf Sie zutreffen.

	Trifft überhaupt nicht zu	Trifft nicht zu	Teils/teils	Trifft zu	Trifft voll und ganz zu
Ich bin eher zurückhaltend, reserviert.	○	○	○	○	○
Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	○	○	○	○	○
Ich bin bequem, neige zur Faulheit.	○	○	○	○	○
Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	○	○	○	○	○
Ich habe nur wenig künstlerisches Interesse.	○	○	○	○	○
Ich gehe aus mir heraus, bin gesellig.	○	○	○	○	○
Ich neige dazu, andere zu kritisieren.	○	○	○	○	○
Ich erledige Aufgaben gründlich.	○	○	○	○	○
Ich werde leicht nervös und unsicher.	○	○	○	○	○
Ich habe eine aktive Vorstellungskraft, bin fantasievoll.	○	○	○	○	○

Weiter

Figure 2.13: Screenshot of the online experiment: page 2.

Instruktionen

Im Folgenden finden Sie die Instruktionen für die nächste Aufgabe.

Zu Beginn werden alle Teilnehmer zufällig in **4er Gruppen** aufgeteilt. Jedes Gruppenmitglied erhält initial **20 Punkte**. Ihre Aufgabe besteht darin zu entscheiden, wie Sie Ihre Punkte aufteilen möchten. Sie können die Punkte dabei in ein **gemeinsames Projekt** beitragen oder alternativ **für sich behalten**. Die Konsequenzen dieser Entscheidung werden Ihnen weiter unten noch ausführlicher erläutert.

Durch das Bewegen eines Schiebers entscheiden Sie, wie Sie Ihre Punkte aufteilen möchten. Bitte überlegen Sie sich Ihre Aufteilung genau. Sobald Sie auf die nächste Seite gewechselt sind, können Sie Ihre Entscheidung nicht mehr revidieren.

Sobald alle Gruppenmitglieder ihre Entscheidungen getroffen haben, wird berechnet, wie viele Punkte jedes Gruppenmitglied zum gemeinsamen Projekt beigetragen hat. Davon hängt ihr Gesamteinkommen ab, welches Sie in dieser Aufgabe erhalten. Ihr Gesamteinkommen (in Punkten) berechnet sich dabei wie folgt:

Gesamteinkommen = **Einkommen aus dem gemeinsamen Projekt** + **Einbehaltene Punkte**

Wobei ihr Einkommen aus dem gemeinsamen Projekt sich wie folgt berechnet:

Einkommen aus dem gemeinsamen Projekt = Summe aller Beiträge in das gemeinsame Projekt x 0,4

Beispiele:
Jedes Gruppenmitglied hat einen Betrag von 20 Punkten.

- Wenn jedes Gruppenmitglied (inklusive Ihnen) 0 Punkte zum Projekt beiträgt, verdienen Sie 20 Punkte (Einkommen aus dem gemeinsamen Projekt: 0 Punkte (0 x 0,4 = 0) & Einbehaltene Punkte: 20 Punkte).
- Wenn jedes Gruppenmitglied (inklusive Ihnen) 20 Punkte zum Projekt beiträgt, verdienen Sie 32 Punkte (Einkommen aus dem gemeinsamen Projekt: 32 Punkte (80 x 0,4 = 32) & Einbehaltene Punkte: 0 Punkte).
- Wenn Sie 0 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 60 Punkte zum Projekt beitragen, verdienen Sie 44 Punkte (Einkommen aus dem gemeinsamen Projekt: 24 Punkte (60 x 0,4 = 24) & Einbehaltene Punkte: 20 Punkte).
- Wenn Sie 20 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 0 Punkte zum Projekt

Figure 2.14: Screenshot of the online experiment: page 3 (part 1).

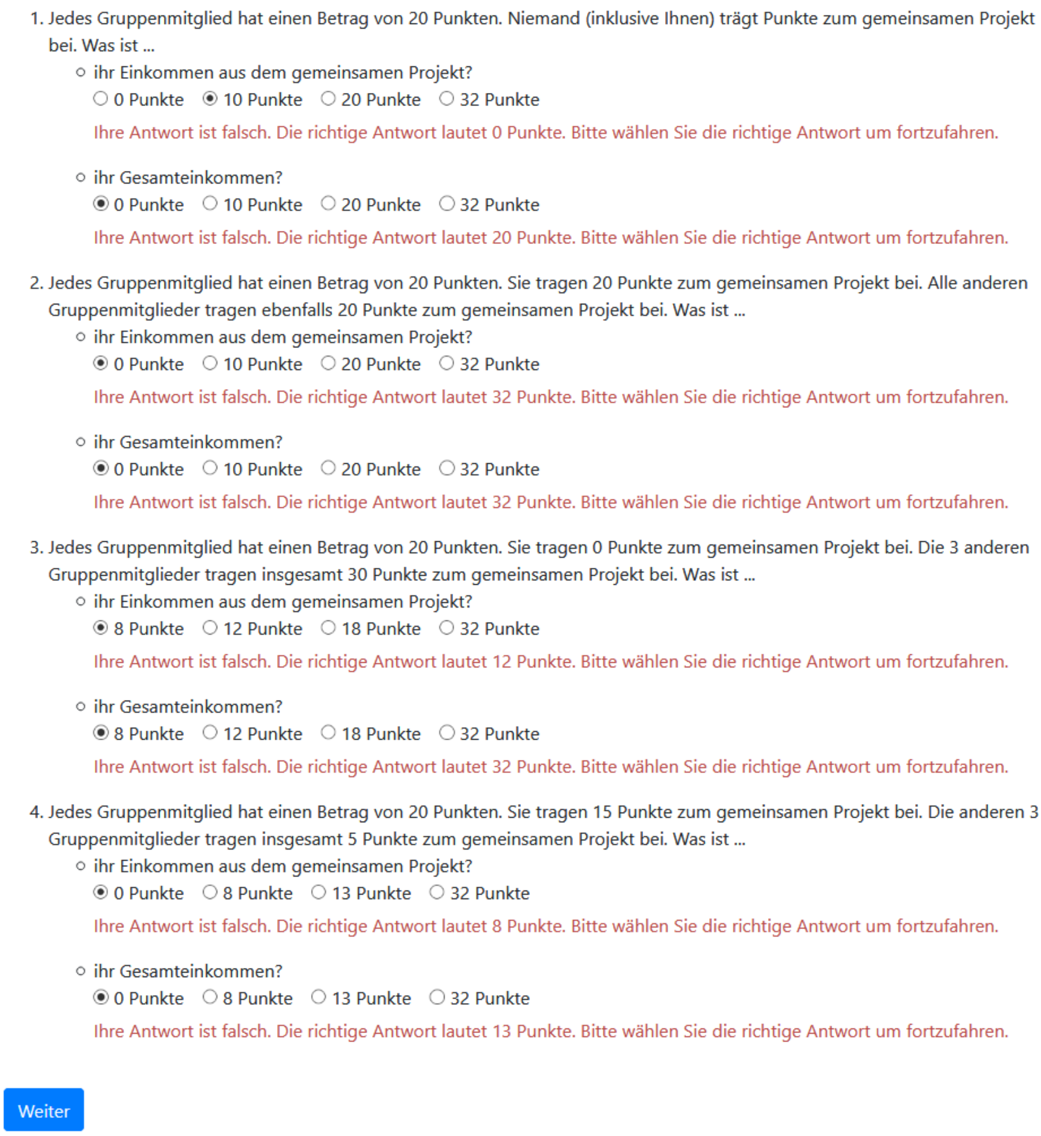
- Wenn Sie 20 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 0 Punkte zum Projekt beitragen, verdienen Sie 8 Punkte (Einkommen aus dem gemeinsamen Projekt: 8 Punkte (20 x 0,4 = 8) & Einbehaltene Punkte: 0 Punkte).

Kontrollfragen

1. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Niemand (inklusive Ihnen) trägt Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?
     ○ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte
   - ihr Gesamteinkommen?
     ○ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte
2. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 20 Punkte zum gemeinsamen Projekt bei. Alle anderen Gruppenmitglieder tragen ebenfalls 20 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?
     ○ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte
   - ihr Gesamteinkommen?
     ○ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte
3. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 0 Punkte zum gemeinsamen Projekt bei. Die 3 anderen Gruppenmitglieder tragen insgesamt 30 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?
     ○ 8 Punkte ○ 12 Punkte ○ 18 Punkte ○ 32 Punkte
   - ihr Gesamteinkommen?
     ○ 8 Punkte ○ 12 Punkte ○ 18 Punkte ○ 32 Punkte
4. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 15 Punkte zum gemeinsamen Projekt bei. Die anderen 3 Gruppenmitglieder tragen insgesamt 5 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?
     ○ 0 Punkte ○ 8 Punkte ○ 13 Punkte ○ 32 Punkte
   - ihr Gesamteinkommen?
     ○ 0 Punkte ○ 8 Punkte ○ 13 Punkte ○ 32 Punkte

Weiter

Figure 2.15: Screenshot of the online experiment: page 3 (part 2).



1. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Niemand (inklusive Ihnen) trägt Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?

     ○ 0 Punkte ◉ 10 Punkte ○ 20 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 0 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
   - ihr Gesamteinkommen?

     ◉ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 20 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
2. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 20 Punkte zum gemeinsamen Projekt bei. Alle anderen Gruppenmitglieder tragen ebenfalls 20 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?

     ◉ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 32 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
   - ihr Gesamteinkommen?

     ◉ 0 Punkte ○ 10 Punkte ○ 20 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 32 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
3. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 0 Punkte zum gemeinsamen Projekt bei. Die 3 anderen Gruppenmitglieder tragen insgesamt 30 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?

     ◉ 8 Punkte ○ 12 Punkte ○ 18 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 12 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
   - ihr Gesamteinkommen?

     ◉ 8 Punkte ○ 12 Punkte ○ 18 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 32 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
4. Jedes Gruppenmitglied hat einen Betrag von 20 Punkten. Sie tragen 15 Punkte zum gemeinsamen Projekt bei. Die anderen 3 Gruppenmitglieder tragen insgesamt 5 Punkte zum gemeinsamen Projekt bei. Was ist ...
   - ihr Einkommen aus dem gemeinsamen Projekt?

     ◉ 0 Punkte ○ 8 Punkte ○ 13 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 8 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.
   - ihr Gesamteinkommen?

     ◉ 0 Punkte ○ 8 Punkte ○ 13 Punkte ○ 32 Punkte

     Ihre Antwort ist falsch. Die richtige Antwort lautet 13 Punkte. Bitte wählen Sie die richtige Antwort um fortzufahren.

Weiter

Figure 2.16: Screenshot of the online experiment: page 3 (part 2 with the error message).

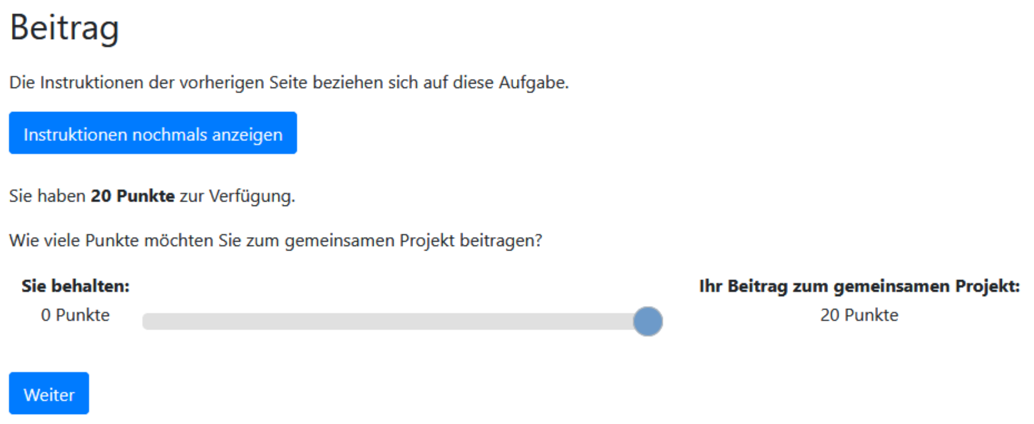


Figure 2.17: Screenshot of the online experiment: page 4.

Beitrag

Die Instruktionen der vorherigen Seite beziehen sich auf diese Aufgabe.

Instruktionen nochmals anzeigen

Zu Beginn werden alle Teilnehmer zufällig in **4er Gruppen** aufgeteilt. Jedes Gruppenmitglied erhält initial **20 Punkte**. Ihre Aufgabe besteht darin zu entscheiden, wie Sie Ihre Punkte aufteilen möchten. Sie können die Punkte dabei in ein **gemeinsames Projekt** beitragen oder alternativ **für sich behalten**. Die Konsequenzen dieser Entscheidung werden Ihnen weiter unten noch ausführlicher erläutert.

Durch das Bewegen eines Schiebers entscheiden Sie, wie Sie Ihre Punkte aufteilen möchten. Bitte überlegen Sie sich Ihre Aufteilung genau. Sobald Sie auf die nächste Seite gewechselt sind, können Sie Ihre Entscheidung nicht mehr revidieren.

Sobald alle Gruppenmitglieder ihre Entscheidungen getroffen haben, wird berechnet, wie viele Punkte jedes Gruppenmitglied zum gemeinsamen Projekt beigetragen hat. Davon hängt ihr Gesamteinkommen ab, welches Sie in dieser Aufgabe erhalten. Ihr Gesamteinkommen (in Punkten) berechnet sich dabei wie folgt:

Gesamteinkommen = **Einkommen aus dem gemeinsamen Projekt** + **Einbehaltene Punkte**

Wobei ihr Einkommen aus dem gemeinsamen Projekt sich wie folgt berechnet:

Einkommen aus dem gemeinsamen Projekt = Summe aller Beiträge in das gemeinsame Projekt x 0,4

Beispiele:
Jedes Gruppenmitglied hat einen Betrag von 20 Punkten.

- Wenn jedes Gruppenmitglied (inklusive Ihnen) 0 Punkte zum Projekt beiträgt, verdienen Sie 20 Punkte (Einkommen aus dem gemeinsamen Projekt: 0 Punkte (0 x 0,4 = 0) & Einbehaltene Punkte: 20 Punkte).
- Wenn jedes Gruppenmitglied (inklusive Ihnen) 20 Punkte zum Projekt beiträgt, verdienen Sie 32 Punkte (Einkommen aus dem gemeinsamen Projekt: 32 Punkte (80 x 0,4 = 32) & Einbehaltene Punkte: 0 Punkte).
- Wenn Sie 0 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 60 Punkte zum Projekt beitragen, verdienen Sie 44 Punkte (Einkommen aus dem gemeinsamen Projekt: 24 Punkte (60 x 0,4 = 24) & Einbehaltene Punkte: 20 Punkte).

Figure 2.18: Screenshot of the online experiment: page 4 (collapsed - part 1).

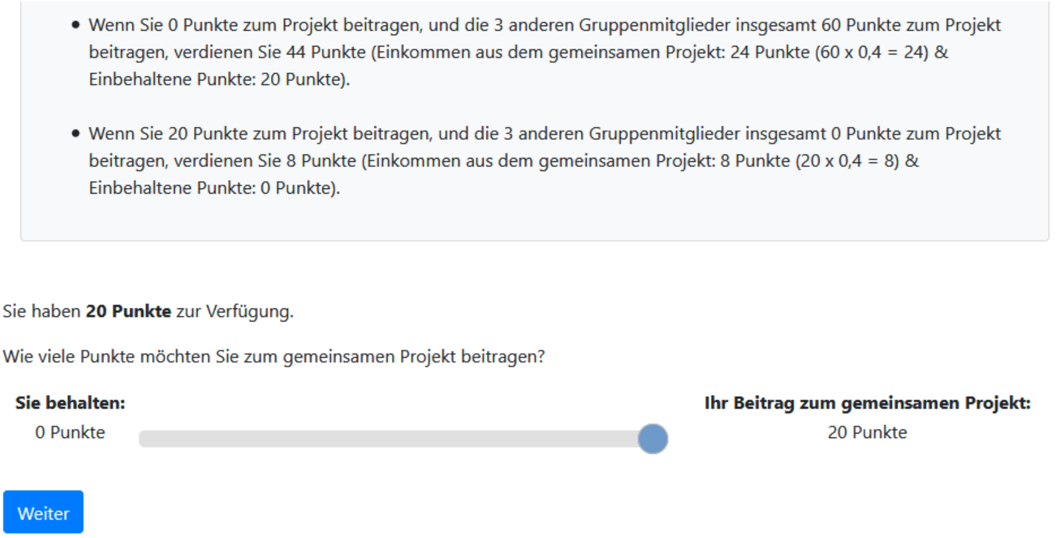


Figure 2.19: Screenshot of the online experiment: page 4 (collapsed - part 2).

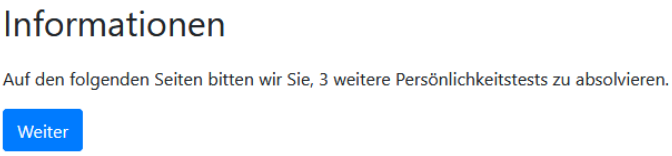


Figure 2.20: Screenshot of the online experiment: page 5 (control group).

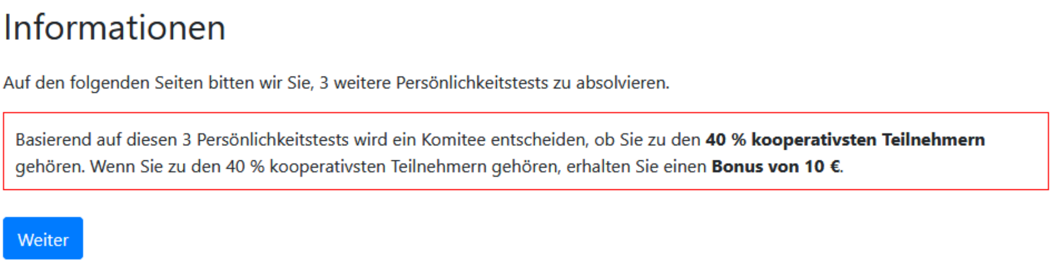
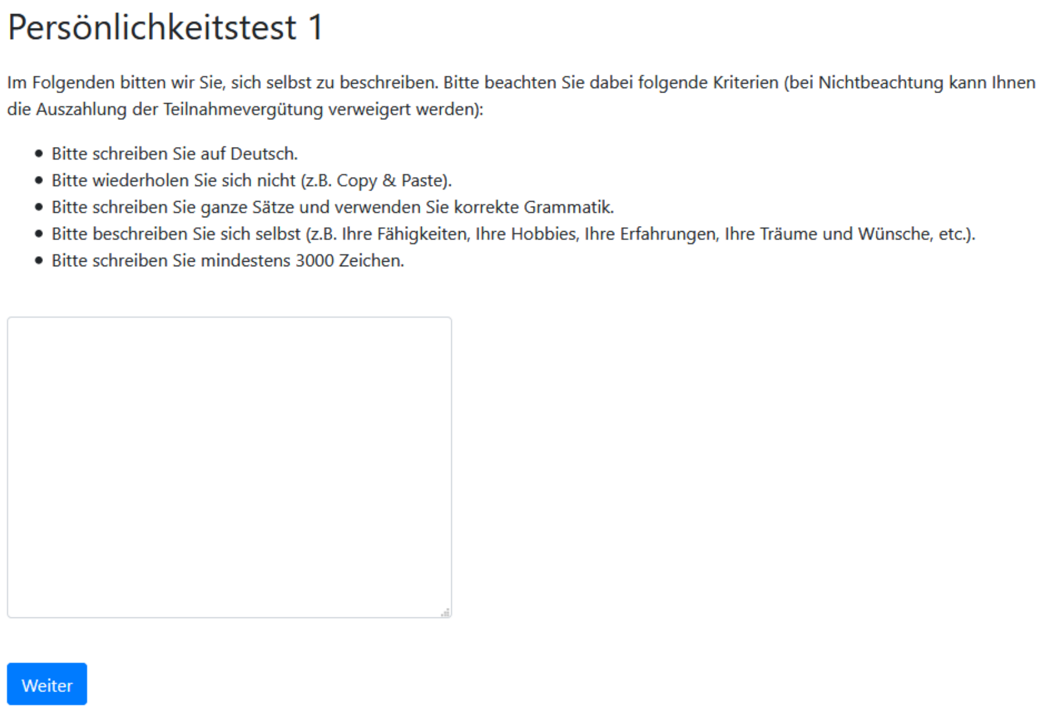


Figure 2.21: Screenshot of the online experiment: page 5 (treatment group).



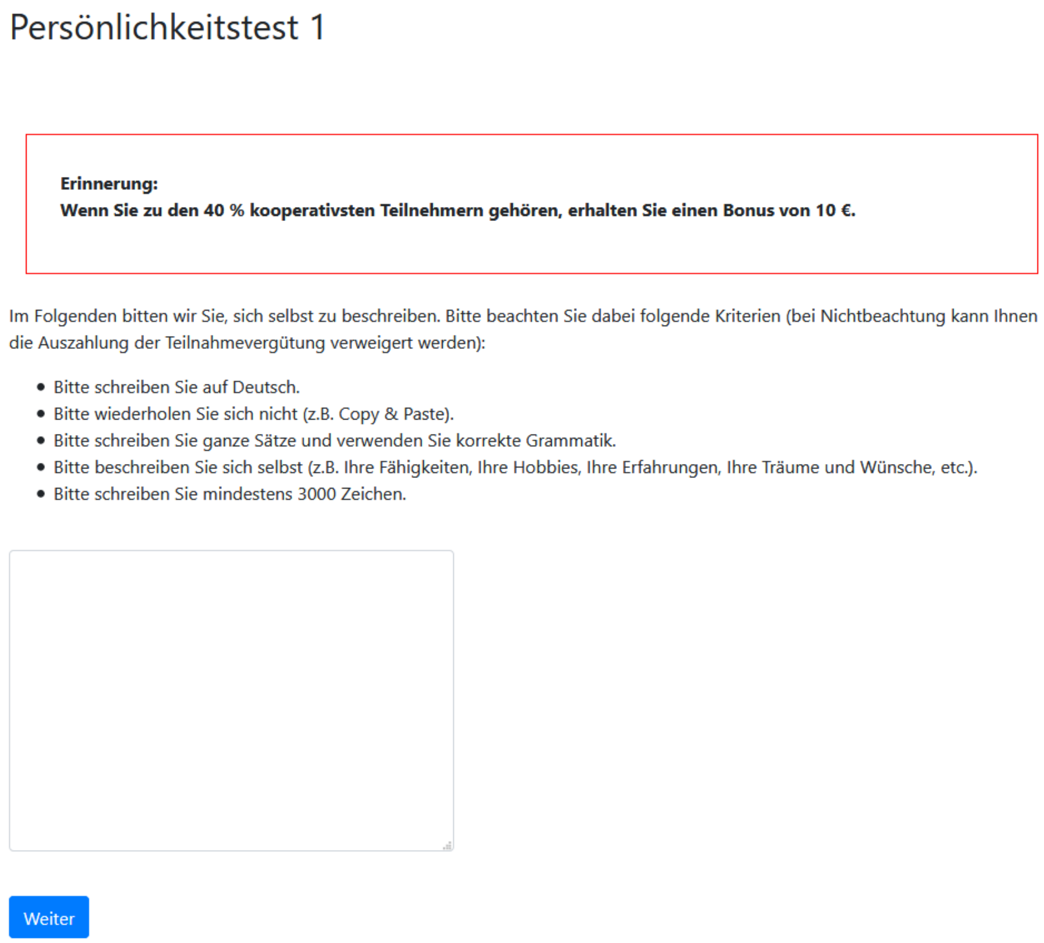
Persönlichkeitstest 1

Im Folgenden bitten wir Sie, sich selbst zu beschreiben. Bitte beachten Sie dabei folgende Kriterien (bei Nichtbeachtung kann Ihnen die Auszahlung der Teilnahmevergütung verweigert werden):

- Bitte schreiben Sie auf Deutsch.
- Bitte wiederholen Sie sich nicht (z.B. Copy & Paste).
- Bitte schreiben Sie ganze Sätze und verwenden Sie korrekte Grammatik.
- Bitte beschreiben Sie sich selbst (z.B. Ihre Fähigkeiten, Ihre Hobbies, Ihre Erfahrungen, Ihre Träume und Wünsche, etc.).
- Bitte schreiben Sie mindestens 3000 Zeichen.

Weiter

Figure 2.22: Screenshot of the online experiment: page 6 (control group).



Persönlichkeitstest 1

Erinnerung:
Wenn Sie zu den 40 % kooperativsten Teilnehmern gehören, erhalten Sie einen Bonus von 10 €.

Im Folgenden bitten wir Sie, sich selbst zu beschreiben. Bitte beachten Sie dabei folgende Kriterien (bei Nichtbeachtung kann Ihnen die Auszahlung der Teilnahmevergütung verweigert werden):

- Bitte schreiben Sie auf Deutsch.
- Bitte wiederholen Sie sich nicht (z.B. Copy & Paste).
- Bitte schreiben Sie ganze Sätze und verwenden Sie korrekte Grammatik.
- Bitte beschreiben Sie sich selbst (z.B. Ihre Fähigkeiten, Ihre Hobbies, Ihre Erfahrungen, Ihre Träume und Wünsche, etc.).
- Bitte schreiben Sie mindestens 3000 Zeichen.

Weiter

Figure 2.23: Screenshot of the online experiment: page 6 (treatment group).

Persönlichkeitstest 2

Im Folgenden finden Sie eine Reihe von Persönlichkeitseigenschaften, die mehr oder weniger stark auf Sie zutreffen. Bitte markieren Sie für jede Aussage, inwieweit sie auf Sie zutrifft oder nicht. Sie sollen diese Einstufung jeweils für Paare von Eigenschaften vornehmen, auch wenn möglicherweise die eine Eigenschaft stärker zutrifft als die andere.

Ich sehe mich selbst als ...

	Trifft überhaupt nicht zu	Trifft größtenteils nicht zu	Trifft eher nicht zu	Weder zutreffend noch unzutreffend	Trifft eher zu	Trifft größtenteils zu	Trifft voll und ganz zu
... extravertiert, begeistert.	○	○	○	○	○	○	○
... kritisch, streitsüchtig.	○	○	○	○	○	○	○
... zuverlässig, selbstdiszipliniert.	○	○	○	○	○	○	○
... ängstlich, leicht aus der Fassung zu bringen.	○	○	○	○	○	○	○
... offen für neue Erfahrungen, vielschichtig.	○	○	○	○	○	○	○
... zurückhaltend, still.	○	○	○	○	○	○	○
... verständnisvoll, warmherzig.	○	○	○	○	○	○	○
... unorganisiert, achtlos.	○	○	○	○	○	○	○
... gelassen, emotional stabil.	○	○	○	○	○	○	○

Figure 2.24: Screenshot of the online experiment: page 7 (control group - part 1).

... gelassen, emotional stabil.	○	○	○	○	○	○	○
... konventionell, unkreativ.	○	○	○	○	○	○	○

Weiter

Figure 2.25: Screenshot of the online experiment: page 7 (control group - part 2).

Persönlichkeitstest 2

Erinnerung:
Wenn Sie zu den 40 % kooperativsten Teilnehmern gehören, erhalten Sie einen Bonus von 10 €.

Im Folgenden finden Sie eine Reihe von Persönlichkeitseigenschaften, die mehr oder weniger stark auf Sie zutreffen. Bitte markieren Sie für jede Aussage, inwieweit sie auf Sie zutrifft oder nicht. Sie sollen diese Einstufung jeweils für Paare von Eigenschaften vornehmen, auch wenn möglicherweise die eine Eigenschaft stärker zutrifft als die andere.

Ich sehe mich selbst als ...

	Trifft überhaupt nicht zu	Trifft größtenteils nicht zu	Trifft eher nicht zu	Weder zutreffend noch unzutreffend	Trifft eher zu	Trifft größtenteils zu	Trifft voll und ganz zu
... extravertiert, begeistert.	○	○	○	○	○	○	○
... kritisch, streitsüchtig.	○	○	○	○	○	○	○
... zuverlässig, selbstdiszipliniert.	○	○	○	○	○	○	○
... ängstlich, leicht aus der Fassung zu bringen.	○	○	○	○	○	○	○
... offen für neue Erfahrungen, vielschichtig.	○	○	○	○	○	○	○
... zurückhaltend, still.	○	○	○	○	○	○	○

Figure 2.26: Screenshot of the online experiment: page 7 (treatment group - part 1).

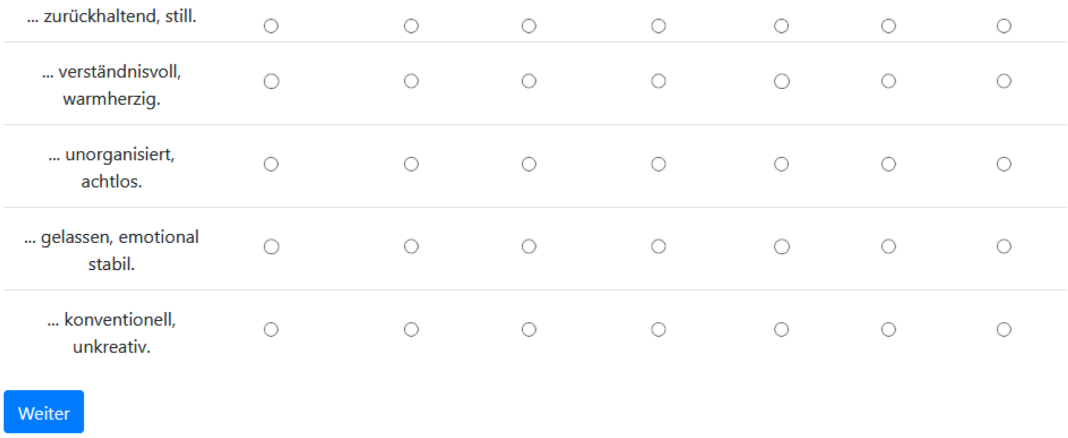


Figure 2.27: Screenshot of the online experiment: page 7 (treatment group - part 2).

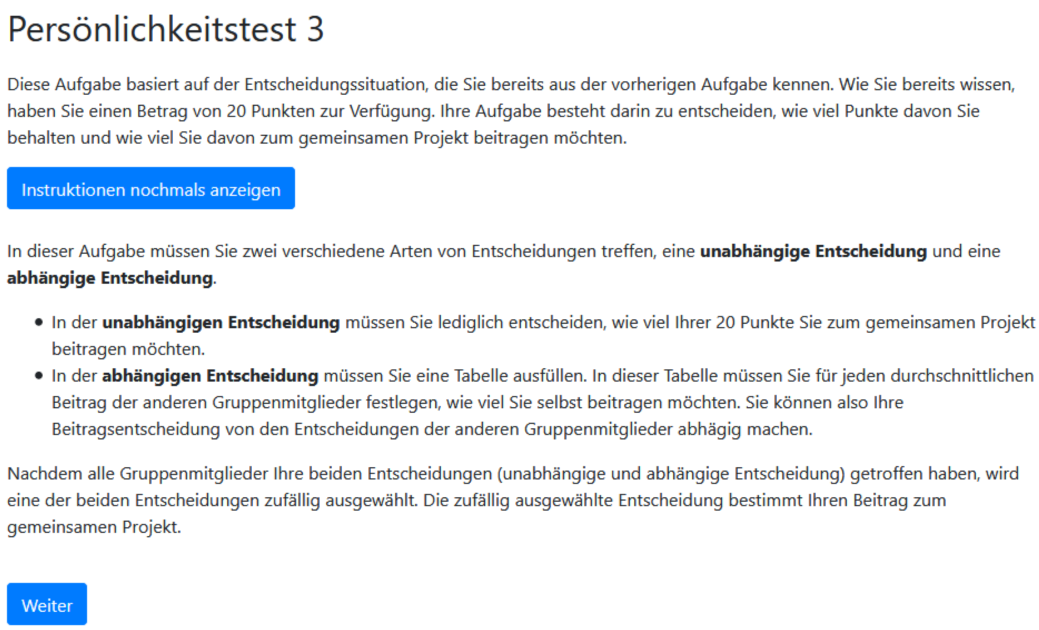
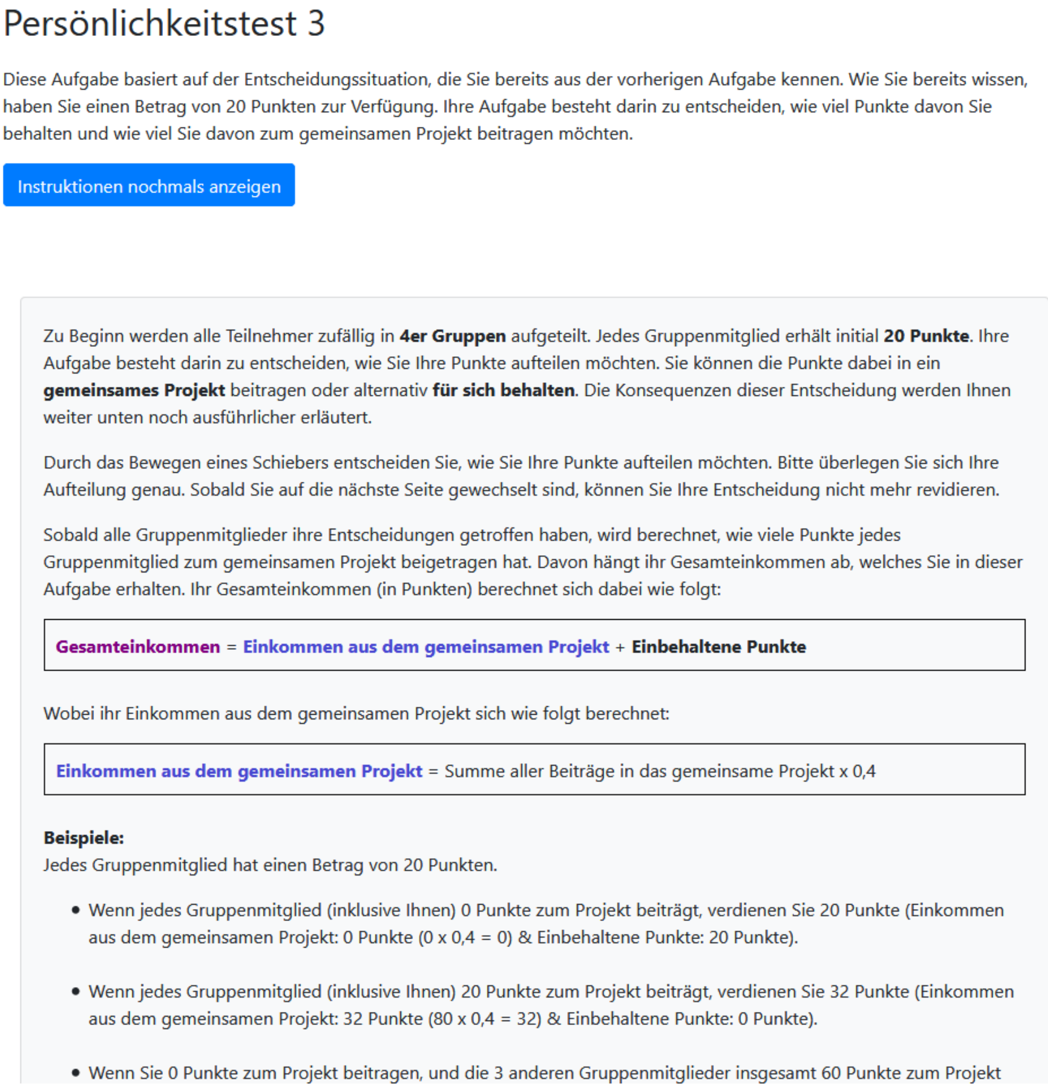


Figure 2.28: Screenshot of the online experiment: page 8 (control group).



Persönlichkeitstest 3

Diese Aufgabe basiert auf der Entscheidungssituation, die Sie bereits aus der vorherigen Aufgabe kennen. Wie Sie bereits wissen, haben Sie einen Betrag von 20 Punkten zur Verfügung. Ihre Aufgabe besteht darin zu entscheiden, wie viel Punkte davon Sie behalten und wie viel Sie davon zum gemeinsamen Projekt beitragen möchten.

Instruktionen nochmals anzeigen

Zu Beginn werden alle Teilnehmer zufällig in **4er Gruppen** aufgeteilt. Jedes Gruppenmitglied erhält initial **20 Punkte**. Ihre Aufgabe besteht darin zu entscheiden, wie Sie Ihre Punkte aufteilen möchten. Sie können die Punkte dabei in ein **gemeinsames Projekt** beitragen oder alternativ **für sich behalten**. Die Konsequenzen dieser Entscheidung werden Ihnen weiter unten noch ausführlicher erläutert.

Durch das Bewegen eines Schiebers entscheiden Sie, wie Sie Ihre Punkte aufteilen möchten. Bitte überlegen Sie sich Ihre Aufteilung genau. Sobald Sie auf die nächste Seite gewechselt sind, können Sie Ihre Entscheidung nicht mehr revidieren.

Sobald alle Gruppenmitglieder ihre Entscheidungen getroffen haben, wird berechnet, wie viele Punkte jedes Gruppenmitglied zum gemeinsamen Projekt beigetragen hat. Davon hängt ihr Gesamteinkommen ab, welches Sie in dieser Aufgabe erhalten. Ihr Gesamteinkommen (in Punkten) berechnet sich dabei wie folgt:

Gesamteinkommen = **Einkommen aus dem gemeinsamen Projekt** + **Einbehaltene Punkte**

Wobei ihr Einkommen aus dem gemeinsamen Projekt sich wie folgt berechnet:

Einkommen aus dem gemeinsamen Projekt = Summe aller Beiträge in das gemeinsame Projekt x 0,4

Beispiele:
Jedes Gruppenmitglied hat einen Betrag von 20 Punkten.

- Wenn jedes Gruppenmitglied (inklusive Ihnen) 0 Punkte zum Projekt beiträgt, verdienen Sie 20 Punkte (Einkommen aus dem gemeinsamen Projekt: 0 Punkte (0 x 0,4 = 0) & Einbehaltene Punkte: 20 Punkte).
- Wenn jedes Gruppenmitglied (inklusive Ihnen) 20 Punkte zum Projekt beiträgt, verdienen Sie 32 Punkte (Einkommen aus dem gemeinsamen Projekt: 32 Punkte (80 x 0,4 = 32) & Einbehaltene Punkte: 0 Punkte).
- Wenn Sie 0 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 60 Punkte zum Projekt

Figure 2.29: Screenshot of the online experiment: page 8 (control group - collapsed, part 1).

- Wenn Sie 0 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 60 Punkte zum Projekt beitragen, verdienen Sie 44 Punkte (Einkommen aus dem gemeinsamen Projekt: 24 Punkte (60 x 0,4 = 24) & Einbehaltene Punkte: 20 Punkte).

- Wenn Sie 20 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 0 Punkte zum Projekt beitragen, verdienen Sie 8 Punkte (Einkommen aus dem gemeinsamen Projekt: 8 Punkte (20 x 0,4 = 8) & Einbehaltene Punkte: 0 Punkte).

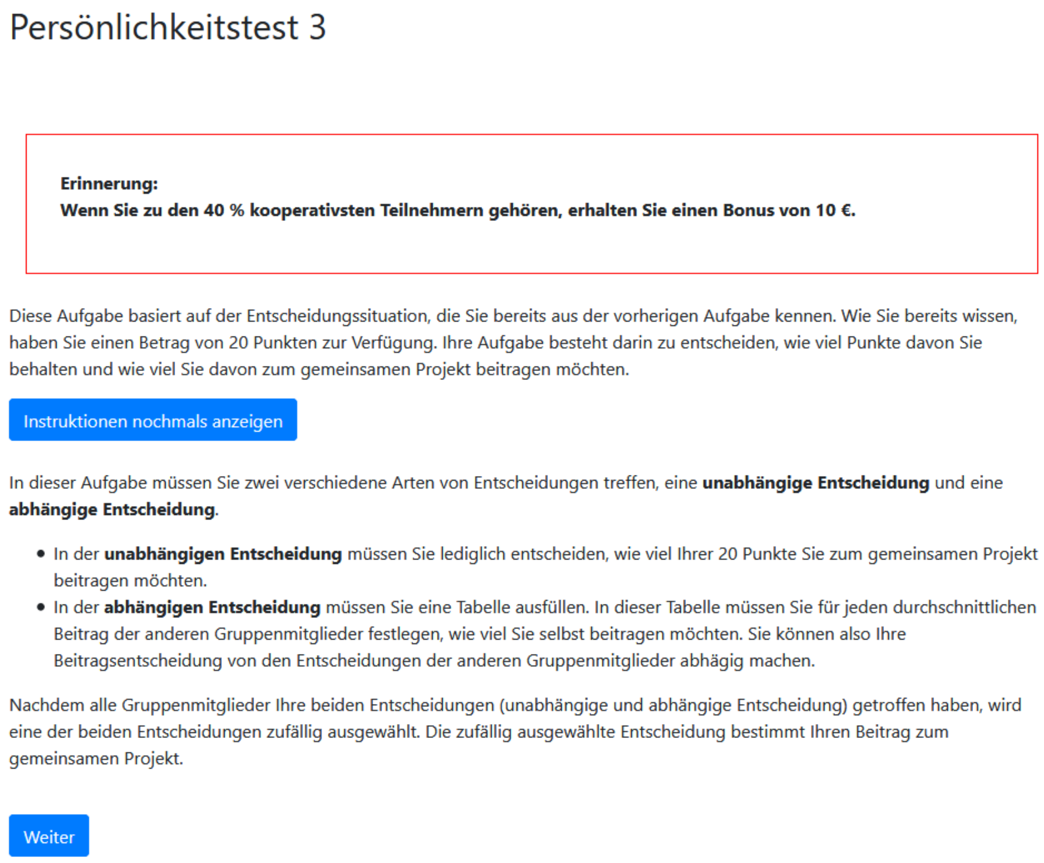
In dieser Aufgabe müssen Sie zwei verschiedene Arten von Entscheidungen treffen, eine **unabhängige Entscheidung** und eine **abhängige Entscheidung**.

- In der **unabhängigen Entscheidung** müssen Sie lediglich entscheiden, wie viel Ihrer 20 Punkte Sie zum gemeinsamen Projekt beitragen möchten.
- In der **abhängigen Entscheidung** müssen Sie eine Tabelle ausfüllen. In dieser Tabelle müssen Sie für jeden durchschnittlichen Beitrag der anderen Gruppenmitglieder festlegen, wie viel Sie selbst beitragen möchten. Sie können also Ihre Beitragsentscheidung von den Entscheidungen der anderen Gruppenmitglieder abhägig machen.

Nachdem alle Gruppenmitglieder Ihre beiden Entscheidungen (unabhängige und abhängige Entscheidung) getroffen haben, wird eine der beiden Entscheidungen zufällig ausgewählt. Die zufällig ausgewählte Entscheidung bestimmt Ihren Beitrag zum gemeinsamen Projekt.

Weiter

Figure 2.30: Screenshot of the online experiment: page 8 (control group - collapsed, part 2).



Persönlichkeitstest 3

Erinnerung:
Wenn Sie zu den 40 % kooperativsten Teilnehmern gehören, erhalten Sie einen Bonus von 10 €.

Diese Aufgabe basiert auf der Entscheidungssituation, die Sie bereits aus der vorherigen Aufgabe kennen. Wie Sie bereits wissen, haben Sie einen Betrag von 20 Punkten zur Verfügung. Ihre Aufgabe besteht darin zu entscheiden, wie viel Punkte davon Sie behalten und wie viel Sie davon zum gemeinsamen Projekt beitragen möchten.

Instruktionen nochmals anzeigen

In dieser Aufgabe müssen Sie zwei verschiedene Arten von Entscheidungen treffen, eine **unabhängige Entscheidung** und eine **abhängige Entscheidung**.

- In der **unabhängigen Entscheidung** müssen Sie lediglich entscheiden, wie viel Ihrer 20 Punkte Sie zum gemeinsamen Projekt beitragen möchten.
- In der **abhängigen Entscheidung** müssen Sie eine Tabelle ausfüllen. In dieser Tabelle müssen Sie für jeden durchschnittlichen Beitrag der anderen Gruppenmitglieder festlegen, wie viel Sie selbst beitragen möchten. Sie können also Ihre Beitragsentscheidung von den Entscheidungen der anderen Gruppenmitglieder abhägig machen.

Nachdem alle Gruppenmitglieder Ihre beiden Entscheidungen (unabhängige und abhängige Entscheidung) getroffen haben, wird eine der beiden Entscheidungen zufällig ausgewählt. Die zufällig ausgewählte Entscheidung bestimmt Ihren Beitrag zum gemeinsamen Projekt.

Weiter

Figure 2.31: Screenshot of the online experiment: page 8 (treatment group).

Persönlichkeitstest 3

Erinnerung:
Wenn Sie zu den 40 % kooperativsten Teilnehmern gehören, erhalten Sie einen Bonus von 10 €.

Diese Aufgabe basiert auf der Entscheidungssituation, die Sie bereits aus der vorherigen Aufgabe kennen. Wie Sie bereits wissen, haben Sie einen Betrag von 20 Punkten zur Verfügung. Ihre Aufgabe besteht darin zu entscheiden, wie viel Punkte davon Sie behalten und wie viel Sie davon zum gemeinsamen Projekt beitragen möchten.

Instruktionen nochmals anzeigen

Zu Beginn werden alle Teilnehmer zufällig in **4er Gruppen** aufgeteilt. Jedes Gruppenmitglied erhält initial **20 Punkte**. Ihre Aufgabe besteht darin zu entscheiden, wie Sie Ihre Punkte aufteilen möchten. Sie können die Punkte dabei in ein **gemeinsames Projekt** beitragen oder alternativ **für sich behalten**. Die Konsequenzen dieser Entscheidung werden Ihnen weiter unten noch ausführlicher erläutert.

Durch das Bewegen eines Schiebers entscheiden Sie, wie Sie Ihre Punkte aufteilen möchten. Bitte überlegen Sie sich Ihre Aufteilung genau. Sobald Sie auf die nächste Seite gewechselt sind, können Sie Ihre Entscheidung nicht mehr revidieren.

Sobald alle Gruppenmitglieder ihre Entscheidungen getroffen haben, wird berechnet, wie viele Punkte jedes Gruppenmitglied zum gemeinsamen Projekt beigetragen hat. Davon hängt ihr Gesamteinkommen ab, welches Sie in dieser Aufgabe erhalten. Ihr Gesamteinkommen (in Punkten) berechnet sich dabei wie folgt:

Gesamteinkommen = **Einkommen aus dem gemeinsamen Projekt** + **Einbehaltene Punkte**

Wobei ihr Einkommen aus dem gemeinsamen Projekt sich wie folgt berechnet:

Einkommen aus dem gemeinsamen Projekt = Summe aller Beiträge in das gemeinsame Projekt x 0,4

Beispiele:

Figure 2.32: Screenshot of the online experiment: page 8 (treatment group - collapsed, part 1).

Beispiele:
Jedes Gruppenmitglied hat einen Betrag von 20 Punkten.

- Wenn jedes Gruppenmitglied (inklusive Ihnen) 0 Punkte zum Projekt beiträgt, verdienen Sie 20 Punkte (Einkommen aus dem gemeinsamen Projekt: 0 Punkte (0 x 0,4 = 0) & Einbehaltene Punkte: 20 Punkte).

- Wenn jedes Gruppenmitglied (inklusive Ihnen) 20 Punkte zum Projekt beiträgt, verdienen Sie 32 Punkte (Einkommen aus dem gemeinsamen Projekt: 32 Punkte (80 x 0,4 = 32) & Einbehaltene Punkte: 0 Punkte).

- Wenn Sie 0 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 60 Punkte zum Projekt beitragen, verdienen Sie 44 Punkte (Einkommen aus dem gemeinsamen Projekt: 24 Punkte (60 x 0,4 = 24) & Einbehaltene Punkte: 20 Punkte).

- Wenn Sie 20 Punkte zum Projekt beitragen, und die 3 anderen Gruppenmitglieder insgesamt 0 Punkte zum Projekt beitragen, verdienen Sie 8 Punkte (Einkommen aus dem gemeinsamen Projekt: 8 Punkte (20 x 0,4 = 8) & Einbehaltene Punkte: 0 Punkte).

In dieser Aufgabe müssen Sie zwei verschiedene Arten von Entscheidungen treffen, eine **unabhängige Entscheidung** und eine **abhängige Entscheidung**.

- In der **unabhängigen Entscheidung** müssen Sie lediglich entscheiden, wie viel Ihrer 20 Punkte Sie zum gemeinsamen Projekt beitragen möchten.
- In der **abhängigen Entscheidung** müssen Sie eine Tabelle ausfüllen. In dieser Tabelle müssen Sie für jeden durchschnittlichen Beitrag der anderen Gruppenmitglieder festlegen, wie viel Sie selbst beitragen möchten. Sie können also Ihre Beitragsentscheidung von den Entscheidungen der anderen Gruppenmitglieder abhägig machen.

Nachdem alle Gruppenmitglieder Ihre beiden Entscheidungen (unabhängige und abhängige Entscheidung) getroffen haben, wird eine der beiden Entscheidungen zufällig ausgewählt. Die zufällig ausgewählte Entscheidung bestimmt Ihren Beitrag zum gemeinsamen Projekt.

Weiter

Figure 2.33: Screenshot of the online experiment: page 8 (treatment group - collapsed, part 2).

Persönlichkeitstest 3

Bitte treffen Sie nun Ihre **unabhängige Entscheidung**. Sie haben **20 Punkte** zur Verfügung.

Wie viele Punkte möchten Sie zum gemeinsamen Projekt beitragen?

Sie behalten:	**Ihr Beitrag zum gemeinsamen Projekt:**
0 Punkte	20 Punkte

Weiter

Figure 2.34: Screenshot of the online experiment: page 9 (control group).

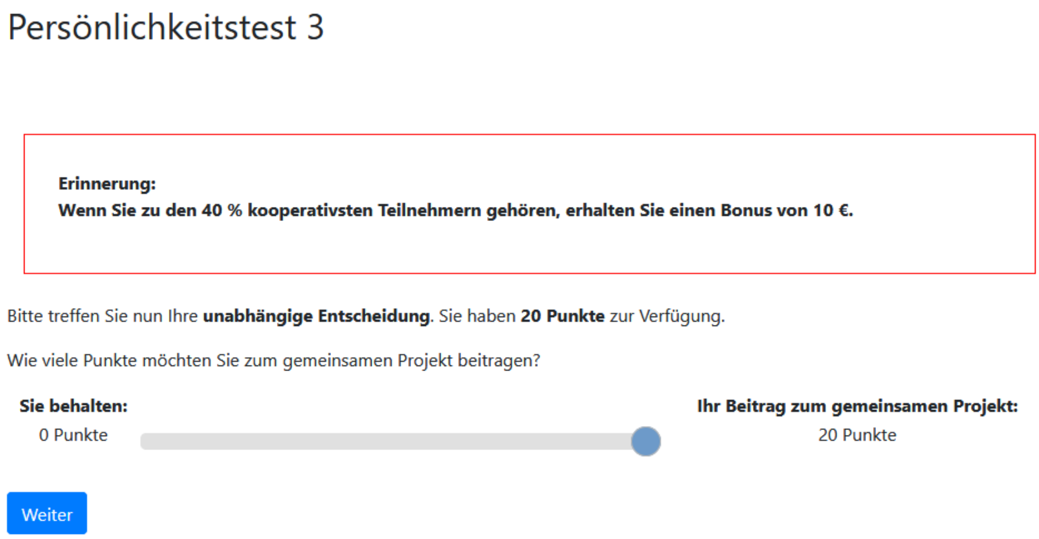


Figure 2.35: Screenshot of the online experiment: page 9 (treatment group).

Persönlichkeitstest 3

Bitte treffen Sie nun Ihre **abhängige Entscheidung**. Wir bitten Sie dabei für jeden durchschnittlichen Beitrag (auf ganze Zahlen gerundet) Ihrer Gruppenmitglieder in das gemeinsame Projekt festzulegen, wie viel Sie abhägig davon selbst beitragen möchten. Für jede Entscheidung haben Sie **20 Punkte** zur Verfügung.

Wenn meine Gruppenmitglieder durchschnittlich ... beitragen	**Dann behalte ich ...**	**Dann trage ich ... bei**
0 Punkte	0 Punkte	20 Punkte
1 Punkt	0 Punkte	20 Punkte
2 Punkte	0 Punkte	20 Punkte
3 Punkte	0 Punkte	20 Punkte
4 Punkte	0 Punkte	20 Punkte
5 Punkte	0 Punkte	20 Punkte
6 Punkte	0 Punkte	20 Punkte
7 Punkte	0 Punkte	20 Punkte
8 Punkte	0 Punkte	20 Punkte
9 Punkte	0 Punkte	20 Punkte
10 Punkte	0 Punkte	20 Punkte

Figure 2.36: Screenshot of the online experiment: page 10 (control group - part 1).

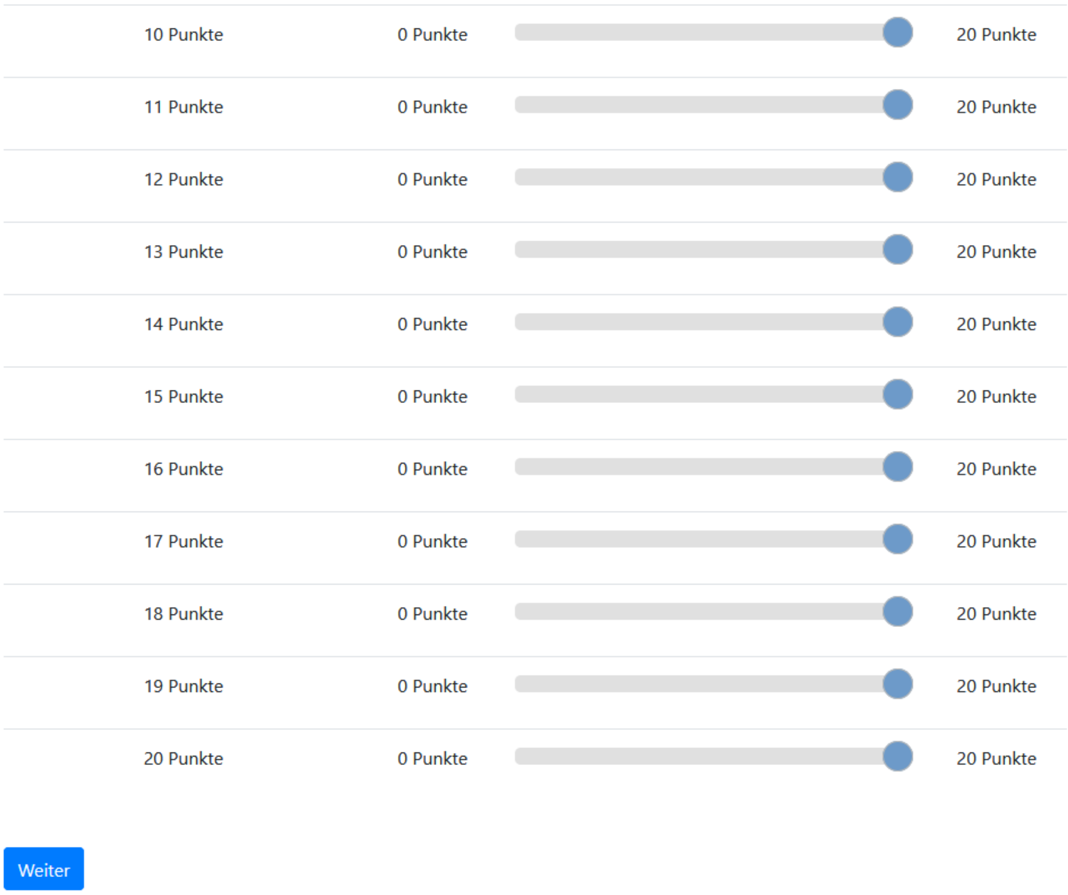
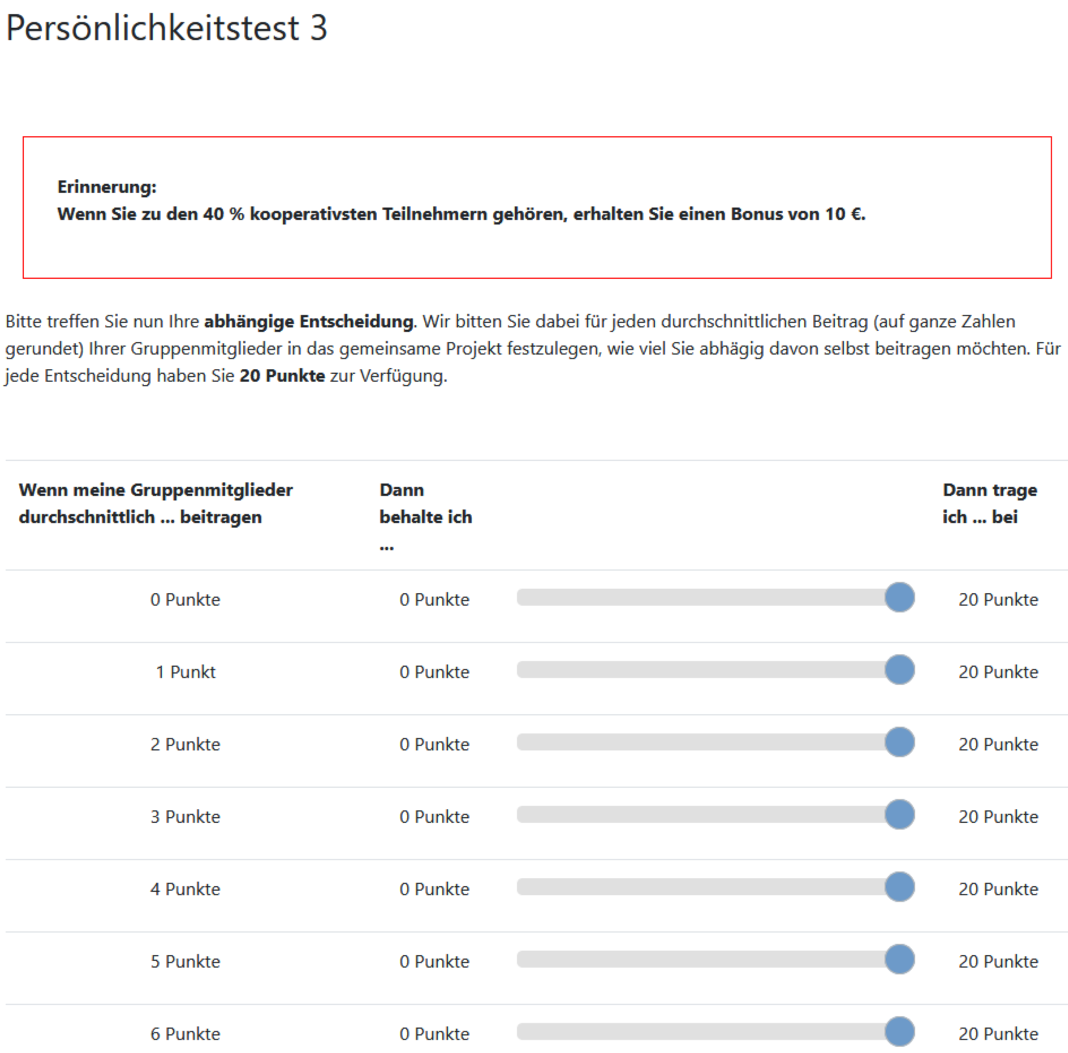


Figure 2.37: Screenshot of the online experiment: page 10 (control group - part 2).



Persönlichkeitstest 3

Erinnerung:
Wenn Sie zu den 40 % kooperativsten Teilnehmern gehören, erhalten Sie einen Bonus von 10 €.

Bitte treffen Sie nun Ihre **abhängige Entscheidung**. Wir bitten Sie dabei für jeden durchschnittlichen Beitrag (auf ganze Zahlen gerundet) Ihrer Gruppenmitglieder in das gemeinsame Projekt festzulegen, wie viel Sie abhägig davon selbst beitragen möchten. Für jede Entscheidung haben Sie **20 Punkte** zur Verfügung.

Wenn meine Gruppenmitglieder durchschnittlich ... beitragen	**Dann behalte ich ...**	**Dann trage ich ... bei**
0 Punkte	0 Punkte	20 Punkte
1 Punkt	0 Punkte	20 Punkte
2 Punkte	0 Punkte	20 Punkte
3 Punkte	0 Punkte	20 Punkte
4 Punkte	0 Punkte	20 Punkte
5 Punkte	0 Punkte	20 Punkte
6 Punkte	0 Punkte	20 Punkte

Figure 2.38: Screenshot of the online experiment: page 10 (treatment group - part 1).

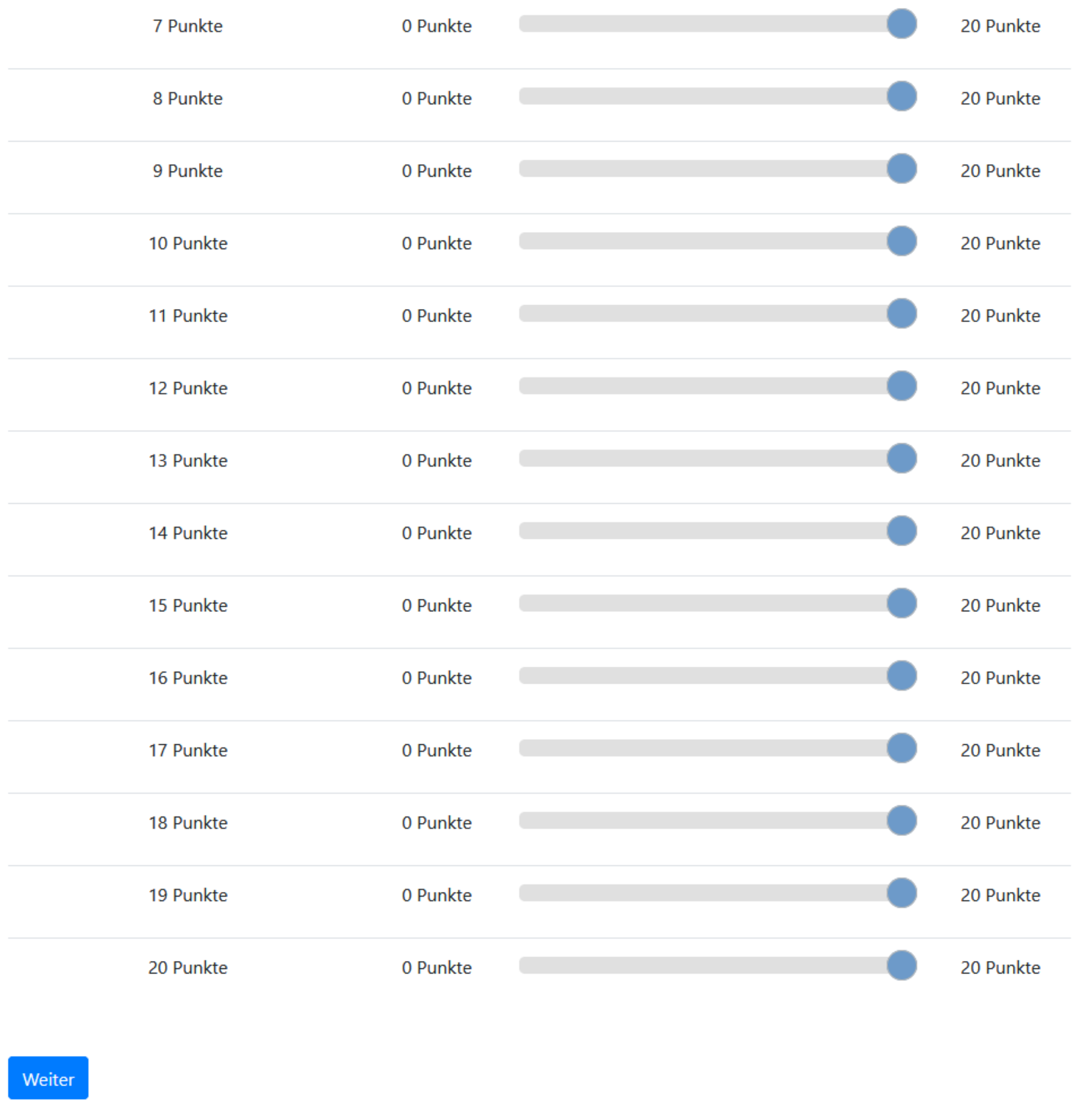


Figure 2.39: Screenshot of the online experiment: page 10 (treatment group - part 2).

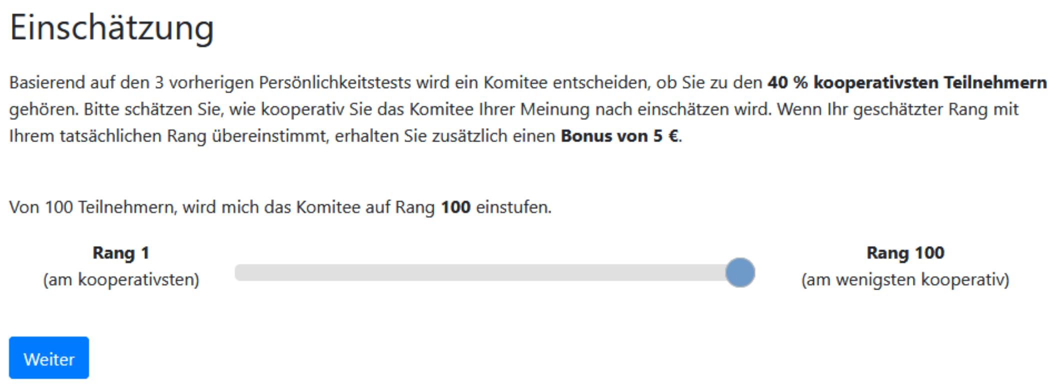


Figure 2.40: Screenshot of the online experiment: page 11.

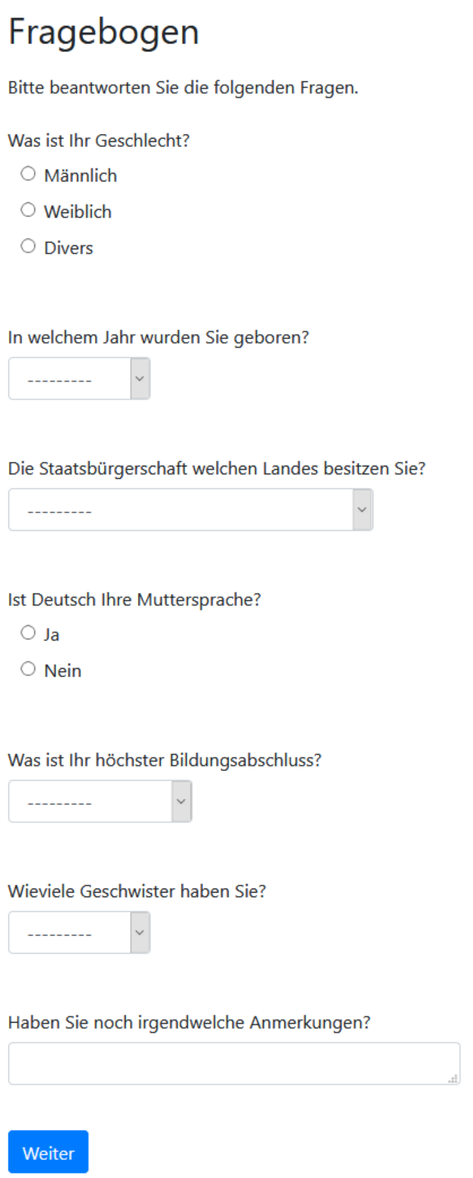


Figure 2.41: Screenshot of the online experiment: page 12.

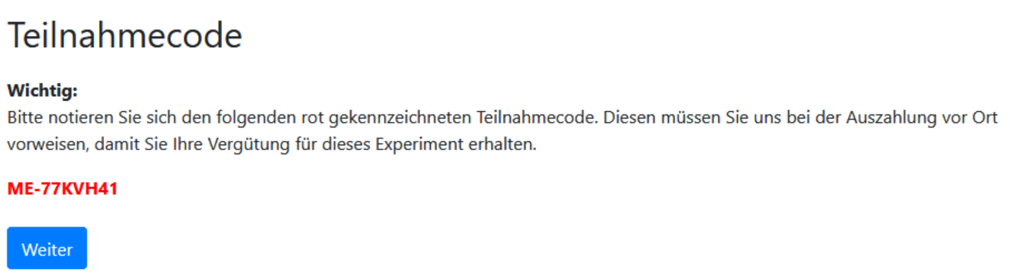
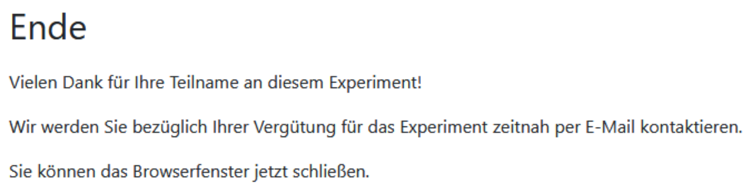


Figure 2.42: Screenshot of the online experiment: page 13.


Ende

Vielen Dank für Ihre Teilname an diesem Experiment!

Wir werden Sie bezüglich Ihrer Vergütung für das Experiment zeitnah per E-Mail kontaktieren.

Sie können das Browserfenster jetzt schließen.

Figure 2.43: Screenshot of the online experiment: page 14.

3. Do employers wage-discriminate against applicants based on their signaled prosociality? Empirical evidence from the field and laboratory

Abstract

Theory predicts that workers' reservation wages decrease in their prosociality. Yet, it is common that job applicants signal their prosociality to future employers (by their choice of internships, volunteering, etc.). We study if employers wage-discriminate against applicants based on their signaled prosociality. To answer this question, we conducted a field experiment with human resources (HR) managers and a laboratory experiment. Our results show that signals on applicants' prosociality do not affect their wage offers. Furthermore, we do not find empirical evidence that applicants' reservation wages depend on their prosociality.

Keywords: Discrimination; Donation; Prosociality; Résumé; Wage

JEL Codes: J39; D64

Authors: Julia Holzmann and Magnus Strobel

3.1 Introduction

A generation of meaning-seeking talent is currently entering the job market. In 2018, 30.3% of all young Americans aged between 16 to 29 engaged in volunteering (CIRCLE, 2018). This popularity is reflected on applicants' résumés, which often exhibit a significant amount of volunteering, social commitment, and engagement in charitable activities. This raises the question of how signals on applicants' prosociality affect their wage offers.

Theory predicts that workers' reservation wages decrease in their prosociality. We derive this prediction from a utility function by Kesternich et al. (2020). In their model, workers derive utility both from their wage and the meaning of their work. It follows from the model that ceteris paribus, the higher the marginal utility that workers derive from the meaning of their work (i. e., the higher their prosociality), the lower the reservation wages that they require to obtain their reservation utility (i. e., the utility for which they accept the job offer). Hence, this theoretical prediction suggests that employers can infer applicants' reservation wages from their true prosociality. If applicants' signaled prosociality is a good proxy for their true prosociality, rational employers should wage-discriminate against applicants based on their signaled prosociality. Thus, we conjecture that the higher the signaled prosociality of applicants, the lower their wage offers.

In practice, employers can use different signals of the applicants to assess their prosociality. For instance, they can use the work experiences on their résumés. Prosocial work is commonly associated with lower wages. Hence, applicants who engaged in prosocial work deliberately missed out on monetary compensation. This reveals their preference for meaning over money (i. e., their prosociality) which in turn serves as a signal on their reservation wages.

In this study, we test the hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers. To this end, we conducted both a field and laboratory experiment.

The field experiment consisted of two parts, a pre-study in the laboratory and the main experiment in the field. In the main experiment, we sent out résumés of fictitious job applicants, together with a questionnaire, to human resources (HR) managers in Germany. Each company received one of four possible résumés, which was randomly selected. The résumés differed only with respect to the applicants' prosociality and their gender. The applicants' prosociality was signaled by their previous work experiences. The résumés signaling high prosociality included internships at Tesla Inc. and Innogy SE, whereas the résumés signaling low prosociality included internships at Porsche AG and RWE AG. The applicants' gender was signaled by their names on the résumés. In the accompanying questionnaire, we asked the HR managers to make, based on the résumé, a hypothetical wage offer to their applicant. Furthermore, we asked them to estimate the reservation wage of the fictitious applicant. In the pre-study, we identified the companies that served as our treatment manipulation in the main experiment. In particular, out of several companies, we selected companies that vary in the perceived social commitment of their interns but are as equal as possible in their competitiveness to be accepted and the importance certain skills to work at these companies. This selection ensures that any observed treatment differences in the main experiment are driven by the difference in the prosociality that is attributed to the applicants and not by differences in their attributed skills.

In the laboratory experiment, subjects played a non-strategic dictator game, a strategic dictator game, and an ultimatum game. In both dictator games, subjects had to decide how much of their endowment they want to keep for themselves and how much they want to donate to a charity of their choice. Between the non-strategic and the strategic dictator game, subjects were informed that they would also play an ultimatum game in which the roles of proposers and responders would be randomly assigned. Additionally, subjects were informed that, prior to making their respective offer in the ultimatum game, proposers would be shown their matched responder's donation amount in the strategic dictator game. We interpret subjects' donation amounts in the non-strategic dictator game as their true prosociality and subjects' donation amounts in the strategic dictator game as their signaled prosociality. Furthermore, we interpret proposers' offers as their wage offers and responders' minimum acceptance thresholds as their reservation wages.

We do not find empirical evidence to support the hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers. In the field experiment, the hypothetical wage offers by HR managers are not affected by the prosociality of the fictitious applicants, signaled by the work experiences on their résumés. Furthermore, we find that estimated reservation wages by HR managers are not affected by the prosociality of the fictitious applicants. Likewise, in the laboratory experiment, proposers' offers in the ultimatum game are not affected by responders' donation amounts in the strategic dictator game. In line with this finding, we also do not find that responders' reservation wages, measured by their minimum acceptance thresholds in the ultimatum game, are affected by their prosociality. Overall, our results provide empirical evidence that signaling prosociality does not backfire financially by leading to lower wage offers.

The scientific contribution of this study is threefold. First, it contributes to the literature on labor market outcomes by providing causal evidence on the effect of applicants' signaled prosociality on their wage offers. The existing literature documents a positive relationship between workers' prosociality and their wages (Cozzi et al., 2017; Sauer, 2015; Hackl et al., 2007; Prouteau and Wolff, 2006). Studies in this field show that wage premiums for volunteering do not depend on the duration of the volunteering (Eberl and Krug, 2020), are higher for males than for females (Day and Devlin, 1997), and are higher in the public than the private sector (Prouteau and Wolff, 2006). However, only a few studies provide causal evidence for the relationship between prosociality and wages (Day and Devlin, 1998, 1997). This study provides additional causal evidence from a correspondence study and the controlled environment of an experimental laboratory. Second, this study contributes to the meaning of work literature by investigating the link between applicants' prosociality and their reservation wages. While several studies document that reservation wages decrease in the prosociality and meaningfulness of the work (Burbano, 2016; Ariely et al., 2008), there is a lack of empirical evidence on the effect of workers' prosociality on their reservation wages. Shedding light on this link helps to better understand the underlying mechanism behind employers' wage offer decisions. Third, this study contributes methodologically to the field of correspondence studies by adding a novel treatment manipulation to signal prosociality. So far, prosociality has mostly been signaled by volunteering (Piopiunik et al., 2020; Wallrodt and Thieme, 2020). In contrast, this study introduces the prosociality of previous work experiences as a signal on the applicants' prosociality. By including internships that only differ with respect to their prosociality but are similar in most other regards, this approach is better able to identify the isolated effect of prosociality than existing studies, which typically either include a volunteering experience or not.

The remainder of this study is structured as follows. In Section 3.2, we provide a theoretical framework and derive our hypothesis. In Section 3.3, we outline the design and procedure of our experiments before presenting the results in Section 3.4. In Section 3.5, we conclude by discussing our findings and experimental shortcomings. In Section 3.6, we present additional information on our experiments.

3.2 Theoretical framework

In this section, we derive a theoretical prediction for the effect of applicants' signaled prosociality on their wage offers. Our prediction is based on a utility function by Kesternich et al. (2020).[1] In their model, the utility U of a worker is given by:

$$U(w, \theta, x, e) = w + \theta m(x, e) - c(e) \tag{3.1}$$

where w denotes the worker's wage, θ denotes how important work meaning is for the worker, $m(x, e)$ denotes how much meaning is generated depending on the job's characteristics x and the worker's effort e, and $c(e)$ denotes the worker's effort costs.[2]

In the utility function above, prosociality is captured by θ, which gives the worker's marginal return on effort in working for a prosocial employer. This reflects the real-world observation that the marginal return on effort in working for a prosocial employer increases with the prosociality of the worker. For instance, ceteris paribus, a prosocial doctor derives more effort-related utility from working for Doctors Without Borders than an antisocial doctor (see Cassar and Meier, 2018, for a general discussion of this idea).

Kesternich et al. (2020) assume a situation in which an employer offers a job to a worker at the wage w. Based on this wage offer, the worker decides to accept or reject the job offer. When rejecting the job offer, the worker realizes the outside option utility $\bar{U}$. When accepting the job offer, the worker decides how much effort e to exert. The worker's optimal effort $e^\star(w, \theta, x)$ is then implicitly defined by the first order condition $\frac{\partial U(w,\theta,x,e)}{\partial e} = \theta \frac{\partial m(x,e)}{\partial e} - \frac{\partial c(e)}{\partial e} = 0$. A worker decides to accept the job offer if and only if w is sufficiently large so that $U(w, \theta, x, e^\star(w, \theta, x)) \geq \bar{U}$. Hence, a worker's reservation wage $w^\star(\theta, x)$ is implicitly defined by the following indifference condition.

$$w + \theta m(x, e^\star(w, \theta, x)) - c(e^\star(w, \theta, x)) = \bar{U} \tag{3.2}$$

[1]Their utility function is a reduced form of the more general utility function by Cassar and Meier (2018).

[2]We assume a convex cost function $c(e)$ (i. e., $\frac{\partial c(e)}{\partial e} > 0$ and $\frac{\partial^2 c(e)}{\partial e^2} > 0$) and that $m(x, e)$ is concave in the job meaning x (i. e., $\frac{\partial m(x,e)}{\partial x} > 0$ and $\frac{\partial^2 m(x,e)}{\partial x^2} < 0$) and the worker's effort e (i. e., $\frac{\partial m(x,e)}{\partial e} > 0$ and $\frac{\partial^2 m(x,e)}{\partial e^2} < 0$). Furthermore, it holds that: $\lim_{e \to 0} \frac{\partial m(x,e)}{\partial e} = \infty$, $\lim_{e \to \infty} \frac{\partial m(x,e)}{\partial e} = 0$, $\lim_{e \to 0} \frac{\partial c(e)}{\partial e} = 0$, and $\lim_{e \to \infty} \frac{\partial c(e)}{\partial e} = \infty$.

By implicitly differentiating the indifference condition 3.2, we obtain the relationship between the reservation wage and the marginal return on effort in working for a prosocial employer.[3]

$$\frac{\partial w^\star(\theta,x)}{\partial\theta} = -m(x,e^\star(w,\theta,x)) \tag{3.3}$$

Since $m(x,e^\star(w,\theta,x)) > 0$, Equation 3.3 shows that workers' reservation wages decrease with their prosociality. Hence, if applicants' signaled prosociality is a good proxy for their true prosociality, then it should hold that applicants with higher signals of prosociality have lower reservation wages. We conjecture that rational employers take this relationship into account when deciding on their wage offers and therefore discriminate against workers based on their signaled prosociality.

Hypothesis. *The higher the signaled prosociality of applicants, the lower their wage offers.*

3.3 Experimental design and procedure

To test whether employers wage-discriminate against applicants based on their signaled prosociality, we conducted a field and laboratory experiment.

3.3.1 Field experiment

Our field experiment was divided into two parts, a pre-study in the laboratory and a main experiment in the field.

[3]We derive this relationship as follows:

$\frac{\partial w^\star(\theta,x)}{\partial\theta} + \frac{\partial\theta m(x,e^\star(w,\theta,x))}{\partial\theta} - \frac{\partial c(e^\star(w,\theta,x))}{\partial\theta} = \frac{\partial\bar{U}}{\partial\theta} \overset{!}{=} 0$

$\Leftrightarrow \frac{\partial w^\star(\theta,x)}{\partial\theta} + \frac{\partial\theta m(x,e^\star(w,\theta,x))}{\partial\theta} - \frac{\partial c(e^\star(w,\theta,x))}{\partial\theta} = 0$

$\Leftrightarrow \frac{\partial w^\star(\theta,x)}{\partial\theta} + m(x,e^\star(w,\theta,x)) + \theta\frac{\partial m(x,e^\star(w,\theta,x))}{\partial\theta} - \frac{\partial c(e^\star(w,\theta,x))}{\partial\theta} = 0$

$\Leftrightarrow \frac{\partial w^\star(\theta,x)}{\partial\theta} + m(x,e^\star(w,\theta,x)) + \theta\frac{\partial m(x,e^\star(w,\theta,x))}{\partial e^\star(w,\theta,x)}\frac{\partial e^\star(w,\theta,x)}{\partial\theta} - \frac{\partial c(e^\star(w,\theta,x))}{\partial e^\star(w,\theta,x)}\frac{\partial e^\star(w,\theta,x)}{\theta} = 0$

$\Leftrightarrow \frac{\partial w^\star(\theta,x)}{\partial\theta} + m(x,e^\star(w,\theta,x)) + \Big[\underbrace{\theta\frac{\partial m(x,e^\star(w,\theta,x))}{\partial e^\star(w,\theta,x)} - \frac{\partial c(e^\star(w,\theta,x))}{\partial e^\star(w,\theta,x)}}_{=0\text{ (follows from first order condition of: }\frac{\partial U(w,\theta,x,e)}{\partial e})}\Big]\frac{\partial e^\star(w,\theta,x)}{\theta} = 0$

$\Leftrightarrow \frac{\partial w^\star(\theta,x)}{\partial\theta} + m(x,e^\star(w,\theta,x)) = 0$

3.3.1.1 Pre-study

The aim of the pre-study was to identify companies that could serve as a treatment manipulation in our main experiment. In particular, we were interested in identifying companies that vary in the perceived social commitment of their interns but are as equal as possible in their competitiveness to be accepted and the importance certain skills to work at these companies. This selection ensures that any observed treatment differences in the main experiment are driven by the difference in the prosociality that is attributed to the applicants and not by differences in their attributed skills.

3.3.1.1.1 Design and procedure

The procedure of our pre-study was as follows (see Section 3.6.2.1 for screenshots of the experiment). After reading the instructions, subjects were successively shown five questionnaires. The questionnaires all included the same questions but differed with respect to the company under consideration. On each questionnaire, subjects were randomly shown one company from the following list of pairs: (1) Münchner Bank eG vs. Umweltbank AG, (2) RWE AG vs. Innogy SE, (3) Bestseller A/S vs. armedangels GmbH, (4) Porsche AG vs. Tesla Inc., and (5) Interessenvertretung DEBRIV Bundesverband Braunkohle vs. Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V.[4] For each company, subjects were shown a brief description of the company's industry and mission (see Section 3.6.1.3 for the company descriptions). Next, they were asked to imagine a student who, out of several offers, decided to conduct an internship at the respective company. Based on this information, subjects were asked to answer the following questions (the corresponding variable names for our analysis are given below the questions).

- "Does this student rather prefer a high salary or to make a positive contribution to society?"
  Five-point Likert scale, ranging from 0 ("This student prefers money much more than most others") to 4 ("This student prefers making a positive contribution to society much more than most others").
  - *Prosociality*

[4]Subjects were only shown one company from each pair to avoid experimenter demand effects, resulting from the direct comparison of similar companies (e. g., Münchner Bank eG vs. Umweltbank AG).

- "How difficult is it in your opinion for a third-year TUM BWL student, with an average grade of 2.0, to be accepted for an internship at [company name]?" Five-point Likert scale, ranging from 0 ("Very easy") to 4 ("Very difficult").
  - *Competitiveness*
- "Please evaluate how important the following skills and competencies are for an internship at [company name]?"
  Five-point Likert scale, ranging from 0 ("Unimportant") to 4 ("Very important").
  - *Communication skills*
  - *Capacity for teamwork*
  - *Intercultural competence*
  - *Ability to deal with conflicts*
  - *Commitment and focus on results*
  - *Self-confidence*
  - *Creativity, flexibility, and innovation capacity*

Subjects received €1 for each question, to which their answer was the same as the answer of the majority of all other subjects. The Krupka-Weber method ensures that subjects' answers correspond to their perception of what others think and not their personal opinion (Krupka and Weber, 2013). The maximum possible payoff per rated internship scenarios was €9. As pre-announced, only one of the five internship scenarios was randomly selected to be payoff relevant. In addition, subjects were paid a fixed participation fee of €2.

3.3.1.1.2 Procedural details and descriptive statistics

On June 3, 2019, we ran two experimental sessions at experimenTUM, the experimental laboratory at TUM School of Management. The experiments were conducted with oTree (Chen et al., 2016). In total, 64 subjects participated in the pre-study. The oldest subject was born in 1958 and the youngest in 2000. In terms of gender, 34 subjects were male, 27 female, and 3 diverse or did not wish to identify. The sessions lasted between 15-20 minutes, and subjects earned an average of €5.70.

3.3.1.2 Main experiment

In the main experiment, we tested our hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers with HR managers from companies in Germany.

3.3.1.2.1 Design and procedure

The procedure of our main experiment was as follows (see Section 3.6.2.2 for screenshots of the experiment). After reading the instructions, subjects were shown one of four possible résumés (see Section 3.6.1.2 for the four possible résumés). Next, subjects had to answer the following questions (the corresponding variable names for our analysis are given below the questions).

- "How do you perceive this candidate, is this more someone who cares about a high salary or someone who cares about making a positive contribution to society?"
  Five-point Likert scale, ranging from 0 ("This candidate prefers money much more than most others") to 4 ("This candidate prefers making a positive contribution to society much more than most others").
  - *Perceived prosociality*
- "Assuming that the candidate masters all the interviews and you would like to make him an offer, what would be the gross annual income (in thousands of euros) that you would offer him?"
  Scale from €0 to €150,000, in thousand-Euro increments.
  - *Wage offer*
- "What do you think is the minimum annual gross income (in thousands of euros) at which the applicant will accept your offer?"
  Scale from €0 to €150,000, in thousand-Euro increments.
  - *Reservation wage*
- "How well does the applicant fit your company?"
  Five-point Likert scale, ranging from 0 ("Very badly") to 4 ("Very well").
  - *Perceived fit of applicant*
- "If you are socially involved in clubs, initiatives, etc., how much time (in hours) do you spend on this on average per week?"

  - *Prosociality (HR)*
- "In which industry does your company operate?"
  The categories were: information- and communication, healthcare, energy, chemistry, automotive, retail, finance, personnel services, and other sectors. Selecting multiple options was possible.
  - *Industry controls*
- "How many employees does your company have?"
  The categories were: [0, 50], [51, 250], [251, 1,000], [1,001, 10,000], [10,000, infinity).
  - *Company controls*
- "How do you rate your company? Is it more a company that only cares about making a high profit, or is it also important to make a positive contribution to society?"
  Five-point Likert scale, ranging from 0 ("Publicly, the company is perceived to value profit way more than others.") to 4 ("Publicly, the company is perceived to value making a contribution to society way more than others.").
  - *Prosociality (company)*
- "What is your gender?"
  Feasible options were: male, female, or diverse.
  - *Female (HR)*
- "How many years do you already work in the HR department?"
  - *Work experience (HR)*

For our analysis, we constructed the dummy variable *Prosociality (résumé)* which takes the value 1 if the fictitious applicant conducted internships at Tesla Inc. and Innogy SE and 0 otherwise (internships at Porsche AG and RWE AG). Furthermore, the dummy variable *Female (résumé)* takes the value 1 if the fictitious applicant is female and 0 otherwise (male).

Based on our theoretical prediction, we expect a negative effect of *Prosociality (résumé)* on *Wage offer* and *Reservation wage*. Furthermore, we expect that *Female (résumé)* moderates this effect (Day and Devlin (1997) document that the wage premium of prosociality is lower for females than for males).

3.3.1.2.2 Procedural details and descriptive statistics

We collected data between August 21 and November 26, 2019. Using LinkedIn and email, we sent out a total of 384 questionnaires to HR managers in Germany (see Section 3.6.1.1 for the elicitation email).[5]

Table 3.1: Descriptive statistics of the main experiment.

	Count	Mean	Std	Min	50%	Max
Wage offer	146	47.64	8.53	25.0	48.0	75.0
Reservation wage	146	47.62	8.81	25.0	47.0	77.0
Prosociality (company)	146	2.18	1.01	0.0	2.0	4.0
Prosociality (HR)	146	2.03	4.38	0.0	1.0	40.0
Perceived fit of applicant	146	1.52	1.00	0.0	2.0	4.0
Work experience (HR)	146	8.30	6.58	1.0	6.0	30.0

Table 3.1 shows the descriptive statistics of our main experiment. Excluding invalid entries, a total of 146 subjects participated in the questionnaire.[6] Among the subjects, 93 identified as female (*Female (HR) = 1*), and 53 (48 as male and 5 as diverse) as not female (*Female (HR) = 0*).

Table 3.2: Treatment overview by gender and prosociality of the fictitious applicant.

	Prosociality (résumé) = 1	Prosociality (résumé) = 0	Total
Female (résumé) = 1	41	34	75
Female (résumé) = 0	37	34	71
Total	78	68	146

Table 3.2 shows the distribution of résumés by gender and prosociality of the fictitious applicant. Out of all subjects, 78 (68) received a résumé of an applicant with work experiences at Tesla Inc. and Innogy SE (Porsche AG and RWE AG). Moreover, 75 (71) received a résumé of a female (male) applicant.

[5]The questionnaire for this paper was generated using Qualtrics software, Version *June 2019* of Qualtrics. Copyright © 2020 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com)

[6]We excluded subjects who did not finish the questionnaire, who did not provide entries for our two dependent variables (*Wage offer* and *Reservation wage*), or who stated that they spend more than 40 hours weekly on prosocial activities (*Prosociality (HR)*), which we consider unrealistic.

3.3.2 Laboratory experiment

In the laboratory experiment, we tested our hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers in a controlled environment.

3.3.2.1 Design and procedure

The design and procedure of our laboratory experiment are illustrated in Figure 3.1 and was as follows (see Section 3.6.3 for screenshots of the experiment).

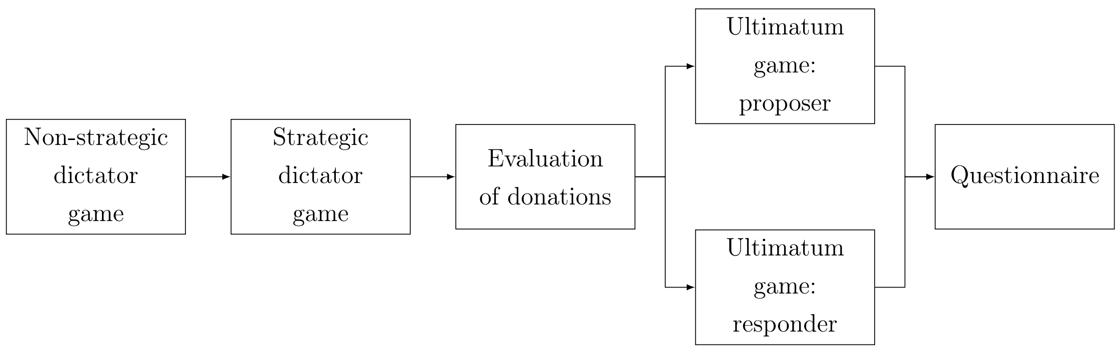


Figure 3.1: Design and procedure of the laboratory experiment.

In stage one, we measured subjects' true (i. e., non-strategic) prosociality by eliciting their charity preferences in a non-strategic dictator game. We call this dictator game non-strategic as subjects did not have an incentive to allocate any amount other than their preferred amount to the charity. In this dictator game, subjects received an endowment of €12 and had to allocate it between themselves and a charity that they could select from a given list.[7] To create an incentive to donate in the experiment, we implemented an efficiency gain of 10%. This means that for every Euro that was donated in the experiment, we donated an additional €0.1. At the time of the non-strategic dictator game, subjects did not know any details about the remainder of the experiment.

In stage two, we measured subjects' signaled (i. e., strategic) prosociality by eliciting their charity preferences in a strategic dictator game. We call this dictator game strategic as subjects had incentives to deviate from their preferred allocation

[7]Subjects could select one of the following charities: SOS-Kinderdörfer, WWF, Brot für die Welt, Aktion Mensch, Deutscher Tierschutzbund, Amnesty International, Reporter ohne Grenzen, UNO Flüchtlingshilfe, Bundesverband Deutsche Tafel, and Oxfam. We chose these charities based on their popularity and made sure that they represent a wide range of purposes so that every subject could find a cause they deem worthy of support.

amount. Again, all subjects received an endowment of €12 and had to allocate it between themselves and a charity that they could select from the same list as before. All donations were again subject to an efficiency gain of 10%. In contrast to the non-strategic dictator game, this time, we gave subjects the additional information that later in the experiment, they would play an ultimatum game, in which the roles of the proposer and responder would be randomly assigned. After the random role assignment, each proposer would be paired with a responder. Before making an offer in the ultimatum game, proposers would learn how their matched responder allocated the endowment in the strategic dictator game.[8] Proposers could, in principle, use this information as a signal on the responders' prosociality and adjust their offers accordingly. For responders who anticipated the signaling effect of their donation, it could create an incentive to deviate from their donation amount in the non-strategic dictator game. Depending on subjects' beliefs, they might allocate a greater amount to charity, hoping that this would be reciprocated with higher offers. Alternatively, they could allocate less to the charity, in order to be perceived as someone to whom money is important, in the hope that this will lead to higher offers by the proposers.

In stage three, we asked subjects to evaluate all possible donation amounts of the strategic dictator game in terms of what they reveal about a subject's prosociality. The reason for this stage was to make the signal of the donation amount in the strategic dictator game more salient and to elicit the information as a manipulation check, i. e., to see whether subjects perceive that higher donations are positively linked to prosociality. To this end, subjects had to indicate, for each possible donation amount ($d \in [0, 12]$), on a five-point Likert scale how they perceive a subject who donated this amount.[9] We incentivized subjects' evaluations with the Krupka-Weber method (Krupka and Weber, 2013). For each donation amount for which subjects chose the same answer as the mode of the experimental session, they received an additional €2. At the end of the experiment, one of the 13 possible donation amounts was selected at random to become payoff relevant.

In stage four, subjects played an ultimatum game. They were randomly assigned to the role of the proposer or responder, and each proposer was paired with a responder. Furthermore, proposer-responder pairs were randomly assigned to the

[8]In contrast, responders would not learn about the donation amount of their proposers.

[9]The Likert scale ranged from 0 (*"I perceive this person as someone to whom money is much more important than to most other people"*) to 4 (*"I perceive this person as someone to whom social commitment is much more important than to most other people"*).

treatment (*Donation treatment = 1*) or control (*Donation treatment = 0*) group. In both groups, we informed proposers about their matched responder's charitable donation amount in the strategic dictator game.[10] Proposers received an endowment of €20 (x) and had to make an offer w ∈ [0, 18] to their matched responder. Responders had to state their minimum acceptance threshold for which they would accept the offer of their matched proposer. We used the strategy method to elicit responders' minimum acceptance thresholds so that the responders' decisions are not influenced by the signal that proposers send through their offer (Selten, 1967). If proposers' offers were higher or equal to the minimum acceptance thresholds of responders, then the payoffs shown in Table 3.3 were realized; otherwise, all payoffs were zero.

Table 3.3: Payoffs by group in the laboratory experiment.

	Donation treatment = 1	**Donation treatment = 0**
$\pi_{Principal}$	x - w - 2	x - w
π_{Agent}	w	w
$\pi_{Charity}$	2	0

Notes: Payoffs (π) with x = 20 and w ∈ [0, 18].

Table 3.3 shows, for both groups, the payoffs from the ultimatum game. The only difference between the two groups is whether or not the acceptance of the offer triggered an automatic charitable donation of €2 (*Donation treatment = 1*) or not (*Donation treatment = 0*).[11] [12]

In stage five, we asked subjects to fill out a questionnaire in which we elicited information on their gender (*Female*), risk preference (*Risk preference*), and other control variables.[13] [14]

In the instructions, we informed subjects that, at the end of the experiment, either the strategic dictator game, the non-strategic dictator game, or the ultimatum game would be randomly selected to be payoff relevant. In addition to the payoff

[10] We did not inform responders about their matched proposer's donation amount.

[11] *Donation treatment = 0*: If the proposer's offered amount was, for instance, €8 and the responder's minimum acceptance threshold was equal or below, the responder received €8 and the proposer kept €20 − €8 = €12.

[12] *Donation treatment = 1*: If the proposer's offered amount was, for instance, €8 and the responder's minimum acceptance threshold was equal or below, the responder received €8, the charity €2, and the proposer kept €20 − €8 − €2 = €10.

[13] *Female* is a dummy variable and takes the value 1 if the subject was female and 0 otherwise.

[14] *Risk preference* ranges from 0 (very high risk-aversion) to 10 (very high risk-seeking).

from the randomly selected game, subjects, if eligible, received the bonus of €2 from the donation assessment of stage three.

For our analysis, we constructed the dependent variable *Offered* as a measure for the wage offer of the employer. It is a percentage value given by the offered amount (w) over the maximum achievable amount for the proposer (20 if *Donation treatment = 0* and 18 if *Donation treatment = 1*). We calculated it as a percentage value since the remaining amount for the proposers differed between the two treatments, and therefore the results of the two treatments would otherwise not be comparable. Likewise, we used the dependent variable *Accepted* as a measure for the reservation wage of the applicants. It is a percentage value given by the responder's minimum acceptance threshold over the maximum achievable amount for their matched proposer. *Non-strategic donation (responder)*/*Non-strategic donation (proposer)* is the donation by the responder/proposer in the non-strategic dictator game. The variable *Strategic donation (responder)* is the donation by the responder in the strategic dictator game. *Donation treatment* is a dummy variable that takes the value 1 if the proposer-responder pair was in the treatment group and 0 otherwise.

Based on our theoretical prediction, we expect that *Strategic donation (responder)* has a negative effect on *Offered* and *Accepted*. Furthermore, we expect that *Donation treatment* has a negative effect on *Offered* (Cassar (2019) and Nyborg and Zhang (2013) document a negative relationship between prosocial incentives and wages) and *Accepted* (Burbano (2016) and Ariely et al. (2008) document a negative relationship between prosocial incentives and reservation wages).

3.3.2.2 Procedural details and descriptive statistics

Between November 6 and November 11, 2019, we conducted 14 experimental sessions. The experiments took place at experimenTUM, the experimental laboratory of TUM, and were programmed with oTree (Chen et al., 2016). Among the 308 subjects who participated in the experiment, 175 identified as male, 132 as female, and 1 as diverse. The youngest subject was born in 2002, and the oldest was born in 1951.

Table 3.4: Descriptive statistics of the laboratory experiment.

	Count	Mean	Std	Min	50%	Max
Non-strategic donation	308	2.88	2.91	0.0	2.00	12.0
Strategic donation	308	3.13	2.69	0.0	2.00	12.0
Offered	154	0.42	0.12	0.0	0.45	0.7
Accepted	154	0.33	0.19	0.0	0.39	1.0
Risk preference	308	5.69	2.01	0.0	6.00	10.0

Table 3.4 shows the descriptive statistics of our laboratory experiment. As can be seen, subjects' donation amounts in the non-strategic (*Non-strategic donation*) and strategic (*Strategic donation*) dictator game range from 0 to 12, with an average donation amount of 2.88 and 3.13, respectively. Furthermore, the offered shares (*Offered*) in the ultimatum game range from 0% to 70%, and the accepted shares (*Accepted*) range from 0% to 100%. Subjects' risk preferences (*Risk preference*) range from 0 to 10 with a mean of 5.96.

3.4 Results

In this section, we answer the question if employers wage-discriminate against applicants based on their signaled prosociality. In the following, we present the results of our field and laboratory experiment.

3.4.1 Field experiment

In this section, we first present the results of our pre-study before presenting the results of our main experiment.

3.4.1.1 Pre-study

The results of our pre-study show that there are two company pairs, namely Porsche AG vs. Tesla Inc. and RWE AG vs. Innogy SE, that meet our requirements. On the one hand, the companies within these pairs differ significantly in terms of the perceived prosociality of the students who decided to conduct an internship there. On the other hand, they are (mostly) equal in terms of the importance of the elicited skills when doing an internship at the respective company.

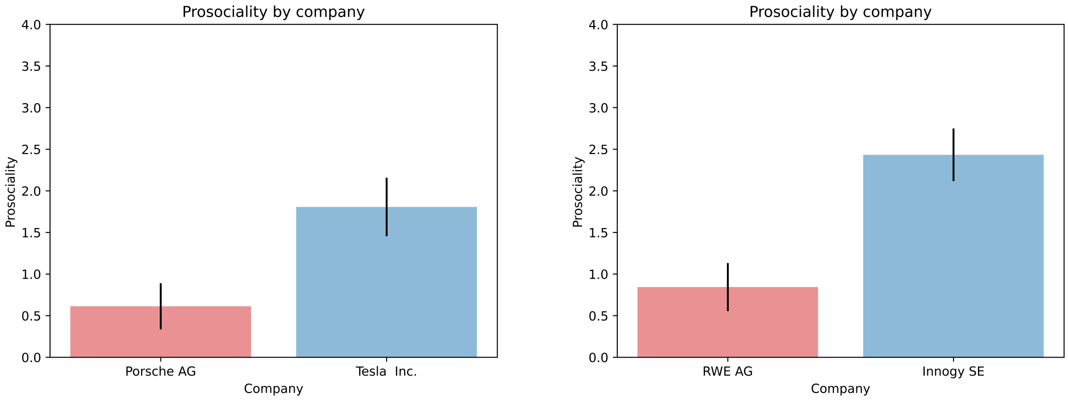


Figure 3.2: Perceived prosociality of the students by company.

Notes: Subjects were asked to rate on a Likert-scale, ranging from 0 ("This person prefers money much more than most others") to 4 ("This person prefers making a positive contribution to society much more than most others"), how they perceive the prosociality of students who decided to do an internship at the respective company. The error bars indicate 95% confidence intervals.

Figure 3.2 shows the perceived prosociality of students who decided to do an internship at the respective company. For Porsche AG vs. Tesla Inc. we find that the perceived prosociality (*Prosociality*) of the students who decided to do an internship at the respective company is significantly different, with values of 0.61 for Porsche AG vs. 1.81 for Tesla Inc. (two-sided Mann-Whitney U test: $p<0.001$). For RWE AG vs. Innogy SE, we also find that the perceived prosociality (*Prosociality*) of students who decided to do an internship at the respective company is significantly different, with values of 0.84 for RWE AG vs. 2.43 for Innogy SE (two-sided Mann-Whitney U test: $p<0.001$).

Table 3.5: Assessment of the competitiveness to be accepted and the importance of certain skills for doing an internship at Porsche AG vs. Tesla Inc.

	Porsche AG	**Tesla Inc.**	**p**
Competitiveness	2.81	2.94	0.635
Communication skills	0.81	0.58	0.414
Capacity for teamwork	0.74	0.65	0.848
Intercultural competence	1.13	1.16	0.830
Ability to deal with conflicts	1.00	1.23	0.254
Commitment and focus on results	0.35	0.29	0.429
Self-confidence	0.68	0.77	0.730
Creativity, flexibility, and innovation capacity	0.65	0.42	0.468

Notes: Subjects were asked to rate on a Likert-scale, ranging from 0 ("very easy") to 4 ("very difficult"), how difficult it would be for a third-year student at the Technical University of Munich (TUM), who is enrolled in the Management and Technology program and has an average grade of 2.0, to be accepted for an internship at the respective company (*Competitiveness*). Furthermore, subjects were asked to rate on a Likert-scale, ranging from 0 ("unimportant") to 4 ("very important") how important the respective skills (*Communication skills*, *Capacity for teamwork*, *Intercultural competence*, *Ability to deal with conflicts*, *Commitment and focus on results*, *Self-confidence*, and *Creativity, flexibility, and innovation capacity*) are for doing an internship at the company. p gives the respective p-value of a two-sided Mann-Whitney U test.

Table 3.5 shows the results for the control categories of Porsche AG vs. Tesla Inc. As can be seen, there is no significant difference in the competitiveness (*Competitiveness*) to obtain an internship at these companies, nor in the required skills for doing an internship at these companies.

Table 3.6: Assessment of the competitiveness to be accepted and the importance of certain skills for doing an internship at RWE AG vs. Innogy SE.

	RWE AG	**Innogy SE**	**p**
Competitiveness	1.69	1.70	0.893
Communication skills	1.38	0.93	0.047
Capacity for teamwork	1.19	0.87	0.082
Intercultural competence	1.84	1.57	0.333
Ability to deal with conflicts	1.19	1.33	0.616
Commitment and focus on results	0.59	0.63	0.808
Self-confidence	0.94	1.00	0.768
Creativity, flexibility, and innovation capacity	1.72	0.73	0.001

Notes: Subjects were asked to rate on a Likert-scale, ranging from 0 ("very easy") to 4 ("very difficult"), how difficult it would be for a third-year student at the Technical University of Munich (TUM), who is enrolled in the Management and Technology program and has an average grade of 2.0, to be accepted for an internship at the respective company (*Competitiveness*). Furthermore, subjects were asked to rate on a Likert-scale, ranging from 0 ("unimportant") to 4 ("very important") how important the respective skills (*Communication skills*, *Capacity for teamwork*, *Intercultural competence*, *Ability to deal with conflicts*, *Commitment and focus on results*, *Self-confidence*, and *Creativity, flexibility, and innovation capacity*) are for doing an internship at the company. p gives the respective p-value of a two-sided Mann-Whitney U test.

Table 3.6 shows the results for the control categories of RWE AG vs. Innogy SE. As can be seen, there is no significant difference in the competitiveness *Competitiveness*) to obtain an internship at these companies. Furthermore, there is also no significant difference in the importance of the majority of skills for doing an internship at these companies (*Intercultural competence*, *Ability to deal with conflicts*, *Commitment and focus on results*, and *Self-confidence*). We only find a significant difference for three skills (*Communication skills*, *Capacity for teamwork*, and *Creativity, flexibility, and innovation capacity*). However, we argue that in the energy sector, these skills are not that relevant. Hence, the differences should not affect wage offers too much.

Based on these findings, we used the internship pair Tesla Inc. and Innogy SE to signal high (*Prosociality (résumé) = 1*) and the internship pair Porsche AG and RWE AG to signal low (*Prosociality (résumé) = 0*) prosociality on our fictitious résumés in the main experiment.

3.4.1.2 Main experiment

In the following, we study the question of whether employers wage-discriminate against applicants based on their signaled prosociality by analyzing the data from our main experiment.

Table 3.7: OLS: Dependent variable is *Wage offer*.

	(1) Wage offer	(2) Wage offer	(3) Wage offer
Prosociality (résumé)	1.131	-5.487	-1.020
	(2.038)	(3.597)	(2.225)
Female (résumé)	1.971	-0.398	-0.276
	(2.081)	(2.116)	(2.190)
Prosociality (résumé)			
× Female (résumé)	-1.265	2.465	2.797
	(2.849)	(2.931)	(2.969)
Prosociality (company)		-0.884	
		(1.004)	
Prosociality (HR)			-0.080
			(0.344)
Prosociality (résumé)			
× Prosociality (company)		2.139	
		(1.398)	
Prosociality (résumé)			
× Prosociality (HR)			0.028
			(0.388)
Perceived fit of applicant		0.827	0.834
		(0.746)	(0.751)
Female (HR)		-1.689	-1.356
		(1.484)	(1.485)
Work experience (HR)		0.363***	0.352***
		(0.111)	(0.112)
Constant	46.382***	43.185***	41.045***
	(1.471)	(3.630)	(3.068)
Company controls	No	Yes	Yes
Industry controls	No	Yes	Yes
Adj. R-Squared	-0.01	0.14	0.12
N	146	146	146

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable *Wage offer* gives the wage offer of HR managers in thousand-Euro increments. *Prosociality (résumé)* is a dummy variable that takes the value 1 if the résumé included internships at Tesla Inc. and Innogy SE, and 0 otherwise (Porsche AG and RWE AG). *Female (résumé)* is a dummy variable that takes the value 1 if the résumé came from a female applicant, and 0 otherwise (male applicant).

Table 3.7 shows OLS regression results for the dependent variable *Wage offer*. The insignificant coefficients of *Prosociality (résumé)* in columns (1), (2), and (3) show that, in contrast to our hypothesis, applicants who signal a high level of prosociality are not discriminated against by receiving lower wage offers from HR managers.[15]

Result 1. *Signals on applicants' prosociality do not affect their wage offers.*

In line with this finding, we also observe that signals on applicants' prosociality do not affect their perceived reservation wages (see Section 3.6.1.4 for the results). Additionally, in contrast to the widespread prevalence of a gender pay gap (Bishu and Alkadry, 2017), we do not find evidence for wage discrimination against female applicants (see the insignificant coefficients of *Female (résumé)* in columns (1), (2), and (3)), nor any interaction of the applicants' prosociality with their gender (see the insignificant coefficients of *Prosociality (résumé)* $\times$ *Female (résumé) in columns (1), (2), and (3))*.[16]

The insignificant results above raise the question of whether our treatment manipulation, which we established in the laboratory, was also effective in the field. To confirm this assumption so that we can meaningfully interpret our results, we run a manipulation check.

[15] The average wage offer for applicants with *Prosociality (résumé) = 1* and *Prosociality (résumé) = 0* is 47.88 and 47.37 respectively (two-sided Mann-Whitney U test: p=0.744).

[16] The average wage offer for applicants with *Female (résumé) = 0* and *Female (résumé) = 1* is 46.97 and 48.28 respectively (two-sided Mann-Whitney U test: p=0.325).

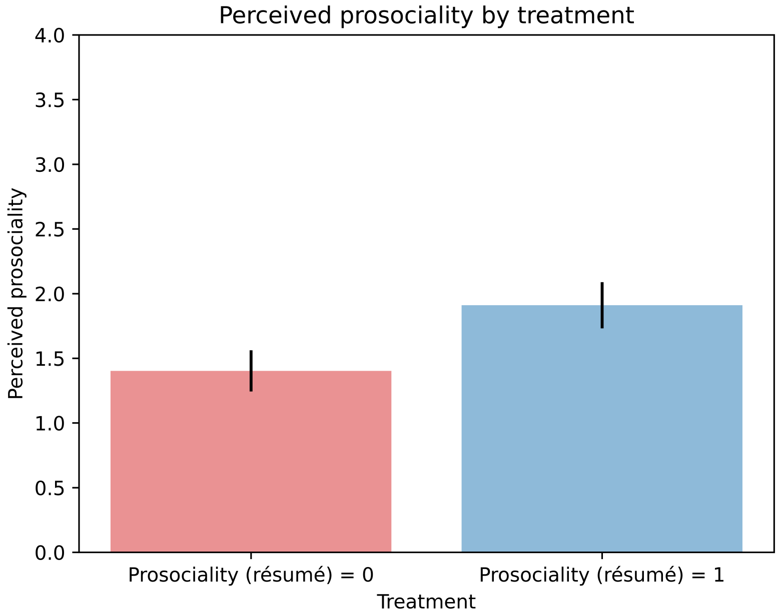


Figure 3.3: Perceived prosociality of the fictitious applicants by treatment.

Notes: Subjects were asked to rate on a Likert-scale, ranging from 0 ("This person prefers money much more than most others") to 4 ("This person prefers making a positive contribution to society much more than most others"), how they perceive the prosociality of the fictitious applicant (*Perceived prosociality*). *Prosociality (résumé)* is a dummy variable that takes the value 1 if the résumé included internships at Tesla Inc. and Innogy SE, and 0 otherwise (Porsche AG and RWE AG). The error bars show 95% confidence intervals.

Figure 3.3 shows how the HR managers perceived the prosociality of the fictitious applicants, depending on the treatment variable *Prosociality (résumé) = 0* vs. *Prosociality (résumé) = 1*. We find a statistically significant difference between our two treatments. The perceived prosociality of our *Prosociality (résumé) = 0* treatment is 1.40, and the perceived prosociality of our *Prosociality (résumé) = 1* treatment is 1.91 (two-sided Mann-Whitney U test: $p<0.001$). Hence, our treatment manipulation was highly effective, meaning that we successfully replicated the results from our pre-study outside the laboratory with HR managers.

Overall, our results from the field experiment do not confirm our theoretical prediction that the higher the signaled prosociality of applicants, the lower their wage offers. In contrast to this prediction, we do not find empirical evidence for wage-discrimination based on applicants' signaled prosociality. Furthermore, in line with this finding, we do not find that applicants' prosociality is negatively associated with their reservation wages.

3.4.2 Laboratory experiment

In the following, we study the question of whether employers wage-discriminate against applicants based on their signaled prosociality by analyzing the data from our laboratory experiment.

Table 3.8: OLS: Dependent variable is *Offered*.

	(1) Offered	(2) Offered	(3) Offered
Strategic donation (responder)	0.002	0.002	0.001
	(0.003)	(0.005)	(0.003)
Non-strategic donation (proposer)		0.007	0.007
		(0.005)	(0.006)
Strategic donation (responder) × Non-strategic donation (proposer)		-0.000	
		(0.001)	
Donation treatment		-0.003	-0.001
		(0.020)	(0.027)
Non-strategic donation (proposer) × Donation treatment			-0.001
			(0.007)
Female		0.017	0.017
		(0.020)	(0.020)
Risk preference		-0.008	-0.008
		(0.005)	(0.005)
Constant	0.410***	0.433***	0.433***
	(0.015)	(0.044)	(0.044)
Adj. R-Squared	-0.00	0.01	0.01
N	154	154	154

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable *Offered* is a percentage value given by the offered amount (w) over the maximum achievable amount for the proposer (20 if *Donation treatment = 0* or 18 if *Donation treatment = 1*). *Strategic donation (responder)* gives the donation amount of the responder in the strategic dictator game. *Non-strategic donation (proposer)* gives the donation amount of the proposer in the non-strategic dictator game. *Donation treatment* is a dummy that takes the value 1 if the proposer was in the treatment group and 0 otherwise (control group).

Table 3.8 shows OLS regression results for the dependent variable *Offered*. The insignificant coefficients of *Strategic donation (responder)* in columns (1), (2), and (3) indicate that proposers' offers do not depend on the amount donated by responders in the strategic dictator game. We interpret this as evidence that signaled prosociality does not affect wage offers.

Result 2. *Signals on applicants' prosociality do not affect their wage offers.*

Additionally, the insignificant coefficients of *Non-strategic donation (proposer)* in columns (2) and (3), which capture proposers' true prosociality, shows that wage offers by proposers do not depend on their true prosociality. Furthermore, the insignificant coefficient of *Strategic donation (responder)* × *Non-strategic donation (proposer)* in column (2) shows that there are no interaction effects between the true prosociality of the proposer and the signaled prosociality of the responder. The insignificant coefficients of *Donation treatment* in columns (2) and (3) show that the fact that a donation hinges on the acceptance of the offer does not affect offers by proposers. Lastly, the insignificant coefficient of *Non-strategic donation (proposer)* × *Donation treatment* in column (3) shows that the effect of the *Donation treatment* does not depend on the true prosociality of the proposer.

Our insignificant result raises the question of how effectively responders' signaled prosociality (i. e., their donation amounts in the strategic dictator game) can predict their true prosociality (i. e., their donation amounts in the non-strategic dictator game). Comparing their donation amounts in the strategic and non-strategic dictator game allows us to answer this question.

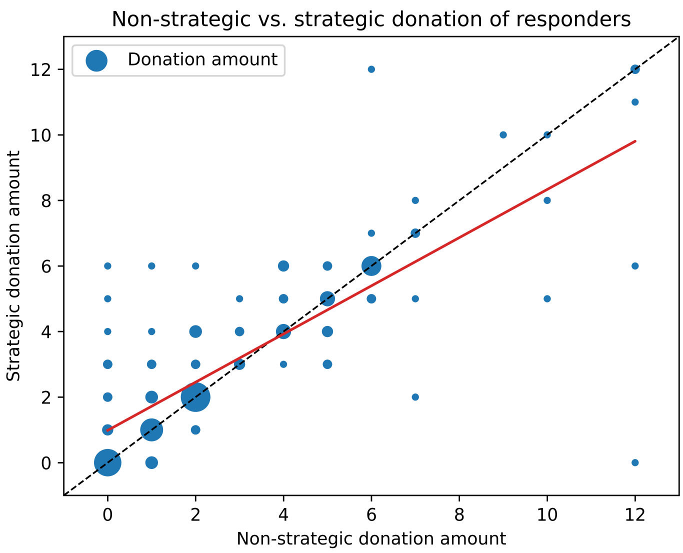


Figure 3.4: Donation amounts by responders in the non-strategic vs. strategic dictator game.

Notes: The size of the circles is proportional to the number of observations. The red line shows the best least squares fitted line.

Figure 3.4 shows the responders' donation amounts in the non-strategic vs. strategic dictator game. As can be seen, the results show a positive correlation between the donation amounts in the non-strategic and strategic dictator game (Pearson correlation coefficient: $r=0.795$; $p<0.001$). This indicates that the responders' signaled prosociality possesses a high predictive power for their true prosociality. Furthermore, a comparison between the average donation amounts shows whether responders preferred to be perceived as prosocial or not. The average donation amount in the strategic dictator game and the non-strategic dictator game was €3.14 and €2.94 respectively (two-sided Mann-Whitney U test: $p=0.244$).[17] This insignificant difference suggests that responders were not concerned about the perception of their donation amount in the strategic dictator game.[18]

In order to study whether there is a positive relationship between the perceived prosociality of responders and their donation amounts, we examine the evaluation of responders' donation amounts by proposers.

[17] In total, 94 responders donated exactly the same amount in the strategic and non-strategic dictator game. While 39 responders donated more in the strategic dictator game, 21 responders donated more in the non-strategic dictator game.

[18] Please note that we cannot rule out the possibility that order or wealth effects may have affected this finding.

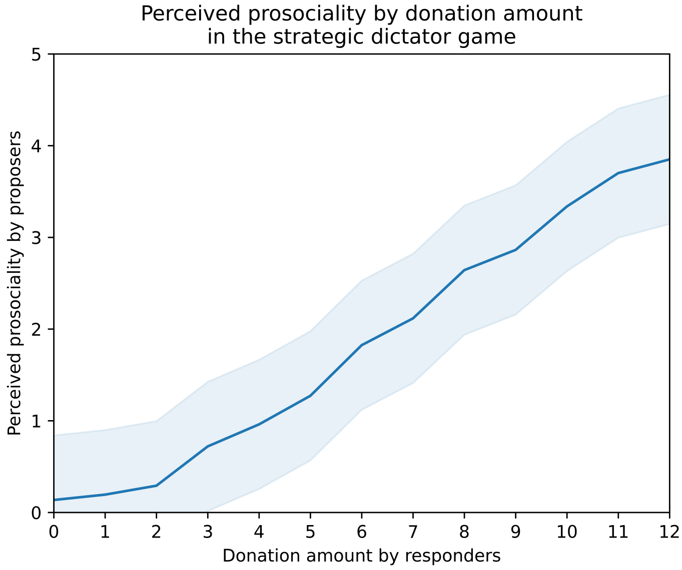


Figure 3.5: Perceived prosociality of responders, based on their donation amount in the strategic dictator game, by proposers.

Notes: Proposers were asked to rate on a Likert-scale, ranging from 0 ("This person prefers money much more than most others") to 4 ("This person prefers making a positive contribution to society much more than most others"), how they perceive specific donation amounts by responders in the strategic dictator game. The shaded area shows 95% confidence intervals.

Figure 3.5 shows the evaluation of responders' donation amounts in the strategic dictator game by proposers. The monotonic upward-sloping trend of the perceived prosociality indicates a positive relationship between the donation amount by responders and how their prosociality is perceived by proposers. This serves as a successful manipulation check, which allows us to interpret the results from the ultimatum game as intended.

Lastly, we answer the question of whether the true prosociality of responders affects their likelihood of accepting an offer.

Table 3.9: OLS: Dependent variable is *Accepted*.

	(1) Accepted	(2) Accepted
Non-strategic donation (responder)	0.009*	0.011
	(0.005)	(0.007)
Donation treatment		0.051
		(0.042)
Non-strategic donation (responder) $\times$ Donation treatment		-0.004
		(0.011)
Female		0.042
		(0.032)
Risk preference		0.018**
		(0.007)
Constant	0.304***	0.161***
	(0.021)	(0.054)
Adj. R-Squared	0.01	0.04
N	154	154

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable *Accepted* is a percentage value given by the responder's minimum acceptance threshold over the maximum achievable amount for their matched proposer (20 if *Donation treatment = 0* or 18 if *Donation treatment = 1*). *Non-strategic donation (responder)* gives the donation amount of the responder in the non-strategic dictator game. *Donation treatment* is a dummy that takes the value 1 if the responder was in the treatment group and 0 otherwise (control group).

Table 3.9 shows OLS regression results for the dependent variable *Accepted*. The significant coefficient of *Non-strategic donation (responder)* in column (1) suggests that the minimum accepted shares depend positively on the true type of the responder. However, the respective coefficient in column (2), the model that includes all control variables, is no longer significant. Hence, the minimum accepted shares do not depend on the true prosociality of the responder. We interpret this as evidence that applicants' reservation wages do not depend on their true prosociality. The insignificant coefficient of *Donation treatment* shows that the minimum accepted shares also do not depend on whether the acceptance of the offer triggers a charity donation or not. Hence, applicants are not willing to accept lower wage offers if their employer makes charitable donations. Furthermore, there is also no interaction effect between the true prosociality of the responder and the fact whether the acceptance of the offer triggers a charity donation (see the insignificant coefficient of *Non-strategic donation (responder)* $\times$ *Donation treatment* in column (2)). The positive and significant coefficient of *Risk preference* in column (2) shows that the responders' minimum accepted shares increase in their risk-seeking.

Overall, our results from the laboratory experiment do not confirm our theoretical prediction that the higher the signaled prosociality of applicants, the lower their wage offers. In contrast to this prediction, we do not find empirical evidence for wage-discrimination based on applicants' signaled prosociality. Furthermore, in line with this finding, we do not find that applicants' prosociality is negatively associated with their reservation wages.

3.5 Conclusion

In this study, we answer the question if employers wage-discriminate against applicants based on their signaled prosociality. Based on a worker's utility function by Kesternich et al. (2020), we derive the testable hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers.

We test this hypothesis in a field and laboratory experiment. In contrast to our theoretical prediction, neither of the two settings provides empirical evidence for wage discrimination based on applicants' signaled prosociality. Furthermore, our findings suggest that applicants' reservation wages do not depend on their true prosociality.

These findings do not imply that such wage discrimination does not exist but only that it did not materialize in our specific settings. In the field experiment, the hypothetical setup and the non-binding consequences might have led HR managers to simply state average wages at their company instead of taking the individual characteristics of the applicants into account. In the laboratory experiment, we assume that the donation in the strategic dictator game solely signals the responders' prosociality. However, the signal also gives an indication of the fairness preferences of the responders. This second signal indication works in the opposite direction to the effect that interests us, which is that higher donations in the strategic dictator game lead to lower offers in the ultimatum game. Hence, our experiment provides a harder test, which might be the reason why we obtain a null result. Lastly, in both experiments, the social desirability bias works in the opposite direction than the effect that interests us, which might be the reason why we obtain a null result. Social desirability bias is defined as "[...] the tendency of research subjects to choose responses they believe are more socially desirable or acceptable rather than choosing responses that are reflective of their true thoughts or feelings." (Grimm, 2010, p. 1). In both experiments, the social desirability bias causes employers to offer higher wages to prosocial applicants since the social norm states that prosocial behavior should be positively reciprocated. Future studies should take these considerations into account when investigating the effect of applicants' signaled prosociality on their wage offers.

3.6 Appendix

3.6.1 Main experiment

3.6.1.1 Email

Betreff: Einladung zu einer wissenschaftlichen Studie

Sehr geehrte /-r Frau/Herr ...,

Im Rahmen eines Forschungsprojekts an der Technischen Universität München sind wir an Ihrer Einschätzung zu einem Bewerber interessiert, der gerade sein Studium beendet hat und sich auf eine Stelle bei Ihnen im Unternehmen bewirbt. Die Befragung richtet sich an Personaler verschiedener Unternehmen in Deutschland. Mit Ihrer Teilnahme an diesem Fragebogen helfen Sie uns, ein umfassendes Bild zur Wahrnehmung verschiedener Absolventen zu erhalten.

Sämtliche Angaben werden anonym erhoben, vertraulich behandelt und ausschließlich für wissenschaftliche Zwecke verwendet. Die Bearbeitungszeit des Fragebogens beträgt 5-10 Minuten. Den Fragebogen finden Sie unter folgendem Link:

Falls Sie Fragen oder Anmerkungen haben, können Sie sich gerne an survey@wi.tum.de wenden.

Wir bedanken uns sehr herzlich für Ihre Teilnahme!

Mit freundlichen Grüßen

Prof. Dr. Michael Kurschilgen

—————

Prof. Dr. Michael Kurschilgen
Professur für Managerial Economics
Technische Universität München
TUM School of Management

Arcisstrasse 21
80333 München
T: 089 289 25394

E: m.kurschilgen@tum.de
W: www.mecon.wi.tum.de

Figure 3.6: Elicitation email of the field experiment as it was sent out with the questionnaire.

3.6.1.2 Résumés

SCHNEIDER Lukas

Adresse:	Blütenstraße 14
	80799 München
Emailadresse:	lukas.schneider@gmail.com
Staatsangehörigkeit:	Deutsch
Geburtsdaten:	15.04.1995 in München

Praktische Erfahrungen

04/2018 - 07/2018 **Porsche AG** - *Praktikant im Bereich Entwicklung* Stuttgart, Deutschland

- Unterstützung des Verbrennungsmotoren-Teams bei der Entwicklungsarbeit
- Statistische Auswertung von antriebsrelevanten Problemen bei Verbrennungsfahrzeugen

08/2015 - 10/2015 **RWE AG** - *Praktikant im Bereich Valuation* Essen, Deutschland

- Unterstützung bei der Bewertung von M&A-Projekten verschiedener fossiler Erzeugungstechnologien wie Braunkohle, Steinkohle und Erdöl
- Aufbereitung entscheidungsrelevanter Informationen für das Management der Sparte Fossil Fuels

Akademische Ausbildung

10/2016 - 05/2019 **Technische Universität München** München, Deutschland
Masterstudium: Technologie- und Managementorientierte BWL
Abschluss: Master of Science

09/2016 - 12/2016 **Universitat de València** Valencia, Spanien
Auslandssemester

10/2013 - 09/2016 **Technische Universität München** München, Deutschland
Bachelorstudium: Technologie- und Managementorientierte BWL
Abschluss: Bachelor of Science

2005-2013 **Wittelsbacher-Gymnasium** München, Deutschland
Abitur

Qualifikationen

Sprachen:	Deutsch (Muttersprache), Englisch (verhandlungssicher in Wort und Schrift), Spanisch (Grundkenntnisse), Latein (Großes Latinum)
EDV-Kenntnisse:	MS Office, Python

Hobbies

Fußball, Gitarre

Figure 3.7: Screenshot of résumé: *Female (résumé) = 0* and *Prosociality (résumé) = 0.*

SCHNEIDER Lukas

Adresse:	Blütenstraße 14
	80799 München
Emailadresse:	lukas.schneider@gmail.com
Staatsangehörigkeit:	Deutsch
Geburtsdaten:	15.04.1995 in München

Praktische Erfahrungen

04/2018 - 07/2018	**Tesla Inc.** - *Praktikant im Bereich Entwicklung*	Prüm, Deutschland

- Unterstützung des Elektromotoren-Teams bei der Entwicklungsarbeit
- Statistische Auswertung von antriebsrelevanten Problemen bei Elektrofahrzeugen

08/2015 - 10/2015	**Innogy SE -** *Praktikant im Bereich Valuation*	Essen, Deutschland

- Unterstützung bei der Bewertung von M&A-Projekten verschiedener regenerativer Erzeugungstechnologien wie Wind Onshore, Wind Offshore und Photovoltaik
- Aufbereitung entscheidungsrelevanter Informationen für das Management der Sparte Renewables

Akademische Ausbildung

10/2016 - 05/2019	**Technische Universität München** *Masterstudium: Technologie- und Managementorientierte BWL* Abschluss: Master of Science	München, Deutschland
09/2016 - 12/2016	**Universitat de València** *Auslandssemester*	Valencia, Spanien
10/2013 - 09/2016	**Technische Universität München** *Bachelorstudium: Technologie- und Managementorientierte BWL* Abschluss: Bachelor of Science	München, Deutschland
2005-2013	**Wittelsbacher-Gymnasium** Abitur	München, Deutschland

Qualifikationen

Sprachen:	Deutsch (Muttersprache), Englisch (verhandlungssicher in Wort und Schrift), Spanisch (Grundkenntnisse), Latein (Großes Latinum)
EDV-Kenntnisse:	MS Office, Python

Hobbies

Fußball, Gitarre

Figure 3.8: Screenshot of résumé: *Female (résumé) = 0* and *Prosociality (résumé) = 1.*

SCHNEIDER Anna

Adresse:	Blütenstraße 14
	80799 München
Emailadresse:	anna.schneider@gmail.com
Staatsangehörigkeit:	Deutsch
Geburtsdaten:	15.04.1995 in München

Praktische Erfahrungen

04/2018 - 07/2018 — **Porsche AG** - *Praktikantin im Bereich Entwicklung* — Stuttgart, Deutschland

- Unterstützung des Verbrennungsmotoren-Teams bei der Entwicklungsarbeit
- Statistische Auswertung von antriebsrelevanten Problemen bei Verbrennungsfahrzeugen

08/2015 - 10/2015 — **RWE AG -** *Praktikantin im Bereich Valuation* — Essen, Deutschland

- Unterstützung bei der Bewertung von M&A-Projekten verschiedener fossiler Erzeugungstechnologien wie Braunkohle, Steinkohle und Erdöl
- Aufbereitung entscheidungsrelevanter Informationen für das Management der Sparte Fossil Fuels

Akademische Ausbildung

10/2016 - 05/2019 — **Technische Universität München** — München, Deutschland
Masterstudium: Technologie- und Managementorientierte BWL
Abschluss: Master of Science

09/2016 - 12/2016 — **Universitat de València** — Valencia, Spanien
Auslandssemester

10/2013 - 09/2016 — **Technische Universität München** — München, Deutschland
Bachelorstudium: Technologie- und Managementorientierte BWL
Abschluss: Bachelor of Science

2005-2013 — **Wittelsbacher-Gymnasium** — München, Deutschland
Abitur

Qualifikationen

Sprachen:	Deutsch (Muttersprache), Englisch (verhandlungssicher in Wort und Schrift), Spanisch (Grundkenntnisse), Latein (Großes Latinum)
EDV-Kenntnisse:	MS Office, Python

Hobbies

Fußball, Gitarre

Figure 3.9: Screenshot of résumé: *Female (résumé) = 1* and *Prosociality (résumé) = 0.*

SCHNEIDER Anna

Adresse:	Blütenstraße 14
	80799 München
Emailadresse:	anna.schneider@gmail.com
Staatsangehörigkeit:	Deutsch
Geburtsdaten:	15.04.1995 in München

Praktische Erfahrungen

04/2018 - 07/2018 **Tesla Inc.** - *Praktikantin im Bereich Entwicklung* Prüm, Deutschland

- Unterstützung des Elektromotoren-Teams bei der Entwicklungsarbeit
- Statistische Auswertung von antriebsrelevanten Problemen bei Elektrofahrzeugen

08/2015 - 10/2015 **Innogy SE -** *Praktikantin im Bereich Valuation* Essen, Deutschland

- Unterstützung bei der Bewertung von M&A-Projekten verschiedener regenerativer Erzeugungstechnologien wie Wind Onshore, Wind Offshore und Photovoltaik
- Aufbereitung entscheidungsrelevanter Informationen für das Management der Sparte Renewables

Akademische Ausbildung

10/2016 - 05/2019 **Technische Universität München** München, Deutschland
Masterstudium: Technologie- und Managementorientierte BWL
Abschluss: Master of Science

09/2016 - 12/2016 **Universitat de València** Valencia, Spanien
Auslandssemester

10/2013 - 09/2016 **Technische Universität München** München, Deutschland
Bachelorstudium: Technologie- und Managementorientierte BWL
Abschluss: Bachelor of Science

2005-2013 **Wittelsbacher-Gymnasium** München, Deutschland
Abitur

Qualifikationen

Sprachen:	Deutsch (Muttersprache), Englisch (verhandlungssicher in Wort und Schrift), Spanisch (Grundkenntnisse), Latein (Großes Latinum)
EDV-Kenntnisse:	MS Office, Python

Hobbies

Fußball, Gitarre

Figure 3.10: Screenshot of résumé: *Female (résumé) = 1* and *Prosociality (résumé) = 1*.

3.6.1.3 Company descriptions

The original company descriptions of our pre-study were as follows:

- Münchner Bank eG: die älteste Genossenschaftsbank Bayerns
- Umweltbank AG: Deutschlands grünste Bank, die mit ihren Kundeneinlagen ausschließlich ökologische Kreditprojekte fördert
- RWE AG: ein Energieunternehmen mit Energieproduktion durch konventionelle fossile Kohle- und Gaskraftwerke
- Innogy SE: ein Energieunternehmen, das Energie aus erneuerbaren Quellen erzeugt, Verteilnetze betreibt und Energie vertreibt
- Bestseller A/S: ein Textileinzelhandelsunternehmen, wozu u.a. die Marken Vero Moda, Only, Jack & Jones gehören
- armedangels GmbH: ein Modelabel, das nachhaltige Mode kreiert und produziert, die sowohl ethisch als auch modisch ist
- Porsche AG: ein deutscher Kraftfahrzeughersteller, der vor allem Sportwagen produziert
- Tesla Inc.: ein Unternehmen, das Elektroautos sowie Stromspeicher- und Photovoltaikanlagen produziert
- Interessenvertretung DEBRIV Bundesverband Braunkohle: der Branchenverband der deutschen Braunkohlewirtschaft
- Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V.: eine nichtstaatliche Umwelt- und Naturschutzorganisation, die für den Schutz unserer Natur und Umwelt eintritt.

3.6.1.4 Reservation wage

Table 3.10: OLS: Dependent variable is *Reservation wage.*

	(1) Reservation wage	(2) Reservation wage	(3) Reservation wage
Prosociality (résumé)	-0.232	-5.691	-2.390
	(2.105)	(3.826)	(2.351)
Female (résumé)	1.559	-1.548	-1.044
	(2.149)	(2.251)	(2.315)
Prosociality (résumé) × Female (résumé)	-2.076	2.009	2.111
	(2.941)	(3.118)	(3.138)
Prosociality (company)		-0.006	
		(1.068)	
Prosociality (HR)			-0.311
			(0.364)
Prosociality (résumé) × Prosociality (company)		1.569	
		(1.487)	
Prosociality (résumé) × Prosociality (HR)			0.145
			(0.410)
Perceived fit of applicant		-0.396	-0.437
		(0.793)	(0.794)
Female (HR)		-0.384	0.137
		(1.579)	(1.569)
Work experience (HR)		0.224*	0.227*
		(0.119)	(0.119)
Constant	47.529***	49.342***	48.737***
	(1.519)	(3.862)	(3.243)
Company controls	No	Yes	Yes
Industry controls	No	Yes	Yes
Adj. R-Squared	-0.01	0.09	0.08
N	146	146	146

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable *Reservation wage* gives the estimated reservation wage of HR managers in thousand-Euro increments. *Prosociality (résumé)* is a dummy variable that takes the value 1 if the résumé included internships at Tesla Inc. and Innogy SE, and 0 otherwise (Porsche AG and RWE AG). *Female (résumé)* is a dummy variable that takes the value 1 if the résumé came from a female applicant, and 0 otherwise (male applicant).

Table 3.10 shows OLS regression results for the dependent variable *Reservation wage.* The insignificant coefficients of *Prosociality (résumé)* in columns (1), (2), and (3) show that the perceived reservation wages are not contingent on the prosociality of the applicant.[19] This finding explains why HR managers do not wage-discriminate against prosocial applicants. Since highly prosocial applicants are not perceived to have a lower reservation wage, HR managers do not offer them lower wages than less prosocial applicants.

3.6.2 Screenshots field experiment

3.6.2.1 Screenshots pre-study

Instruktionen

Herzlich Willkommen!

Bitte sprechen Sie während des Experiments nicht mit anderen Teilnehmern. Wenn Sie Fragen haben, heben Sie bitte die Hand. Einer der Experimentleiter wird dann zu Ihnen an den Platz kommen und Ihre Fragen persönlich beantworten. Wenn Sie sich nicht an diese Regeln halten, behalten wir es uns vor Sie von diesem Experiment auszuschließen. In diesem Fall erhalten Sie auch keine Vergütung.

Bitte lesen Sie sich die folgende Anleitung gründlich durch.

Anleitung:

Im Folgenden werden Ihnen nacheinander 5 verschiedene Praktika angezeigt, zu denen Sie jeweils 9 Fragen beantworten müssen.

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Bitte gedulden Sie sich einen Moment, das Experiment beginnt in Kürze.

Figure 3.11: Screenshot of the pre-study: page 1.

[19] The average attributed reservation wage for applicants with *Prosociality (résumé) = 1* and *Prosociality (résumé) = 0* is 47.03 and 48.31 respectively (two-sided Mann-Whitney U test: p=0.479).

Seite 1: Praktikum bei der **Umweltbank AG**

Erinnerung:

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Stellen Sie sich einen TUM BWL Studenten vor, der in den Semesterferien ein freiwilliges Praktikum absolvieren möchte. Er hatte mehrere Angebote und hat sich für ein Praktikum bei der **Umweltbank AG** entschieden. **Die Umweltbank AG ist Deutschlands grünste Bank, die mit ihren Kundeneinlagen ausschließlich ökologische Kreditprojekte fördert**.

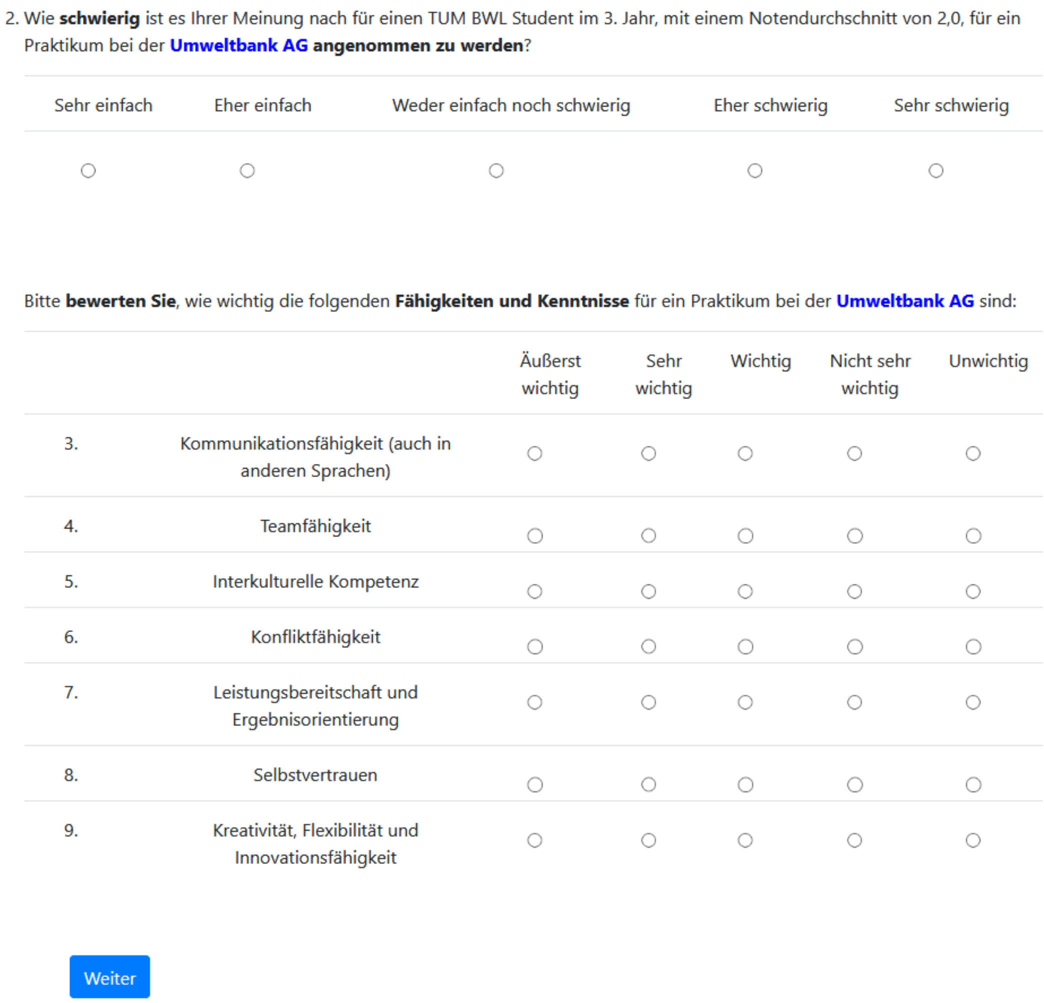
1. Wie schätzen Sie diesen Studenten ein, ist dies eher jemand, dem ein **hohes Gehalt** wichtig ist, oder dem es wichtig ist einen **positiven Beitrag** für die Gesellschaft zu leisten? Diesem Student ist ...

... **Gehalt viel wichtiger** als den meisten anderen.	... **Gehalt etwas wichtiger** als den meisten anderen.	... **beides** (Gehalt und Beitrag für die Gesellschaft) **ungefähr gleich wichtig** wie den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **etwas wichtiger** als den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **viel wichtiger** als den meisten anderen.
○	○	○	○	○

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Umweltbank AG angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Figure 3.12: Screenshot of the pre-study: page 2 (part 1).



2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Umweltbank AG angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Bitte **bewerten Sie**, wie wichtig die folgenden **Fähigkeiten und Kenntnisse** für ein Praktikum bei der **Umweltbank AG** sind:

		Äußerst wichtig	Sehr wichtig	Wichtig	Nicht sehr wichtig	Unwichtig
3.	Kommunikationsfähigkeit (auch in anderen Sprachen)	○	○	○	○	○
4.	Teamfähigkeit	○	○	○	○	○
5.	Interkulturelle Kompetenz	○	○	○	○	○
6.	Konfliktfähigkeit	○	○	○	○	○
7.	Leistungsbereitschaft und Ergebnisorientierung	○	○	○	○	○
8.	Selbstvertrauen	○	○	○	○	○
9.	Kreativität, Flexibilität und Innovationsfähigkeit	○	○	○	○	○

Weiter

Figure 3.13: Screenshot of the pre-study: page 2 (part 2).

Seite 2: Praktikum bei der **Innogy SE**

Erinnerung:

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Stellen Sie sich einen TUM BWL Studenten vor, der in den Semesterferien ein freiwilliges Praktikum absolvieren möchte. Er hatte mehrere Angebote und hat sich für ein Praktikum bei der **Innogy SE** entschieden. **Die Innogy SE ist ein Energieunternehmen, das Energie aus erneuerbaren Quellen erzeugt, Verteilnetze betreibt und Energie vertreibt**.

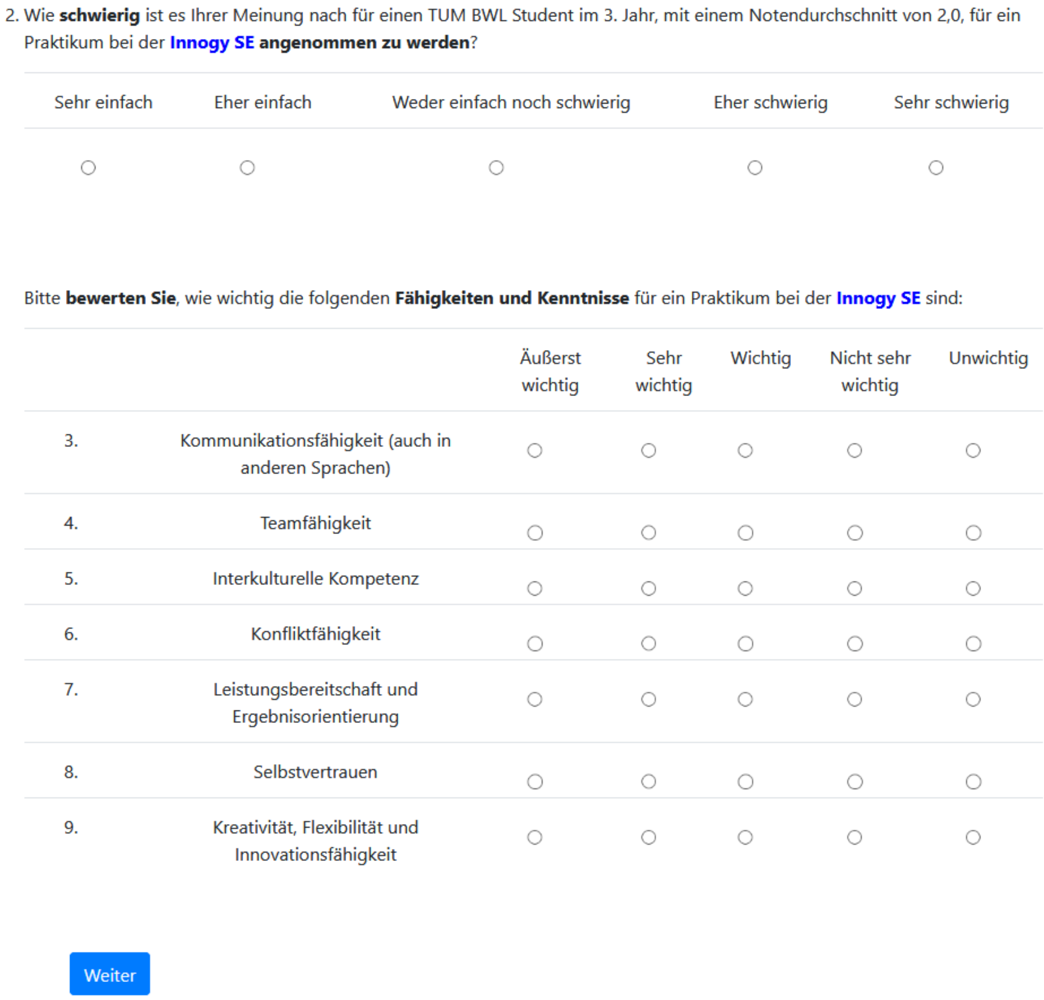
1. Wie schätzen Sie diesen Studenten ein, ist dies eher jemand, dem ein **hohes Gehalt** wichtig ist, oder dem es wichtig ist einen **positiven Beitrag** für die Gesellschaft zu leisten? Diesem Student ist ...

... **Gehalt viel wichtiger** als den meisten anderen.	... **Gehalt etwas wichtiger** als den meisten anderen.	... **beides** (Gehalt und Beitrag für die Gesellschaft) **ungefähr gleich wichtig** wie den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **etwas wichtiger** als den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **viel wichtiger** als den meisten anderen.
○	○	○	○	○

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Innogy SE angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Figure 3.14: Screenshot of the pre-study: page 3 (part 1).



2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Innogy SE** **angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Bitte **bewerten Sie**, wie wichtig die folgenden **Fähigkeiten und Kenntnisse** für ein Praktikum bei der **Innogy SE** sind:

		Äußerst wichtig	Sehr wichtig	Wichtig	Nicht sehr wichtig	Unwichtig
3.	Kommunikationsfähigkeit (auch in anderen Sprachen)	○	○	○	○	○
4.	Teamfähigkeit	○	○	○	○	○
5.	Interkulturelle Kompetenz	○	○	○	○	○
6.	Konfliktfähigkeit	○	○	○	○	○
7.	Leistungsbereitschaft und Ergebnisorientierung	○	○	○	○	○
8.	Selbstvertrauen	○	○	○	○	○
9.	Kreativität, Flexibilität und Innovationsfähigkeit	○	○	○	○	○

Weiter

Figure 3.15: Screenshot of the pre-study: page 3 (part 2).

Seite 3: Praktikum bei der **armedangels GmbH**

Erinnerung:

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Stellen Sie sich einen TUM BWL Studenten vor, der in den Semesterferien ein freiwilliges Praktikum absolvieren möchte. Er hatte mehrere Angebote und hat sich für ein Praktikum bei der **armedangels GmbH** entschieden. **Die armedangels GmbH ist ein Modelabel, das nachhaltige Mode kreiert und produziert, die sowohl ethisch als auch modisch ist**.

1. Wie schätzen Sie diesen Studenten ein, ist dies eher jemand, dem ein **hohes Gehalt** wichtig ist, oder dem es wichtig ist einen **positiven Beitrag** für die Gesellschaft zu leisten? Diesem Student ist ...

... **Gehalt viel wichtiger** als den meisten anderen.	... **Gehalt etwas wichtiger** als den meisten anderen.	... **beides** (Gehalt und Beitrag für die Gesellschaft) **ungefähr gleich wichtig** wie den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **etwas wichtiger** als den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **viel wichtiger** als den meisten anderen.
○	○	○	○	○

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **armedangels GmbH angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Figure 3.16: Screenshot of the pre-study: page 4 (part 1).

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **armedangels GmbH** **angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Bitte **bewerten Sie**, wie wichtig die folgenden **Fähigkeiten und Kenntnisse** für ein Praktikum bei der **armedangels GmbH** sind:

		Äußerst wichtig	Sehr wichtig	Wichtig	Nicht sehr wichtig	Unwichtig
3.	Kommunikationsfähigkeit (auch in anderen Sprachen)	○	○	○	○	○
4.	Teamfähigkeit	○	○	○	○	○
5.	Interkulturelle Kompetenz	○	○	○	○	○
6.	Konfliktfähigkeit	○	○	○	○	○
7.	Leistungsbereitschaft und Ergebnisorientierung	○	○	○	○	○
8.	Selbstvertrauen	○	○	○	○	○
9.	Kreativität, Flexibilität und Innovationsfähigkeit	○	○	○	○	○

Weiter

Figure 3.17: Screenshot of the pre-study: page 4 (part 2).

Seite 4: Praktikum bei der **Porsche AG**

Erinnerung:

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Stellen Sie sich einen TUM BWL Studenten vor, der in den Semesterferien ein freiwilliges Praktikum absolvieren möchte. Er hatte mehrere Angebote und hat sich für ein Praktikum bei der **Porsche AG** entschieden. **Die Porsche AG ist ein deutscher Kraftfahrzeughersteller, der vor allem Sportwagen produziert**.

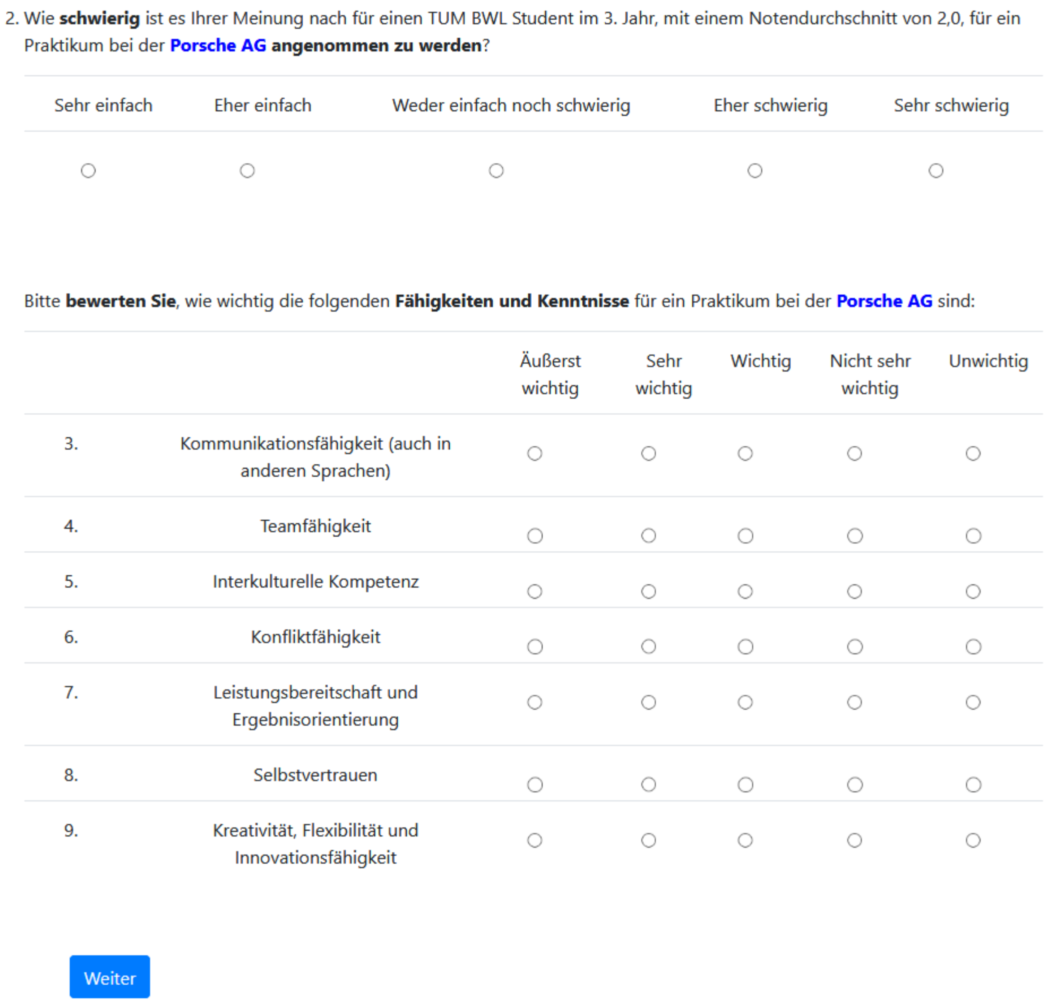
1. Wie schätzen Sie diesen Studenten ein, ist dies eher jemand, dem ein **hohes Gehalt** wichtig ist, oder dem es wichtig ist einen **positiven Beitrag** für die Gesellschaft zu leisten? Diesem Student ist ...

... **Gehalt viel wichtiger** als den meisten anderen.	... **Gehalt etwas wichtiger** als den meisten anderen.	... **beides** (Gehalt und Beitrag für die Gesellschaft) **ungefähr gleich wichtig** wie den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **etwas wichtiger** als den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **viel wichtiger** als den meisten anderen.
○	○	○	○	○

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Porsche AG angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Figure 3.18: Screenshot of the pre-study: page 5 (part 1).


2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Porsche AG angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Bitte **bewerten Sie**, wie wichtig die folgenden **Fähigkeiten und Kenntnisse** für ein Praktikum bei der **Porsche AG** sind:

		Äußerst wichtig	Sehr wichtig	Wichtig	Nicht sehr wichtig	Unwichtig
3.	Kommunikationsfähigkeit (auch in anderen Sprachen)	○	○	○	○	○
4.	Teamfähigkeit	○	○	○	○	○
5.	Interkulturelle Kompetenz	○	○	○	○	○
6.	Konfliktfähigkeit	○	○	○	○	○
7.	Leistungsbereitschaft und Ergebnisorientierung	○	○	○	○	○
8.	Selbstvertrauen	○	○	○	○	○
9.	Kreativität, Flexibilität und Innovationsfähigkeit	○	○	○	○	○

Weiter

Figure 3.19: Screenshot of the pre-study: page 5 (part 2).

Seite 5: Praktikum bei der **Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V.**

Erinnerung:

Für Ihre Teilnahme an diesem Experiment bekommen Sie in jedem Fall **pauschal 2,00 Euro**. Je nach Ihren Entscheidungen und denen der anderen Teilnehmer können Sie noch zusätzlich Geld verdienen.

Am Ende des Experiments wird eines der 5 Praktika zufällig ausgewählt. Für dieses Praktikum ermitteln wir, welche Einschätzung von den meisten Teilnehmern bei den jeweiligen Fragen angegeben wurde. **Pro Frage**, bei der Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben, erhalten Sie **zusätzlich noch 1,00 Euro**.

Stellen Sie sich einen TUM BWL Studenten vor, der in den Semesterferien ein freiwilliges Praktikum absolvieren möchte. Er hatte mehrere Angebote und hat sich für ein Praktikum bei der **Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V.** entschieden. **Die Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V. ist eine nichtstaatliche Umwelt- und Naturschutzorganisation, die für den Schutz unserer Natur und Umwelt eintritt.**.

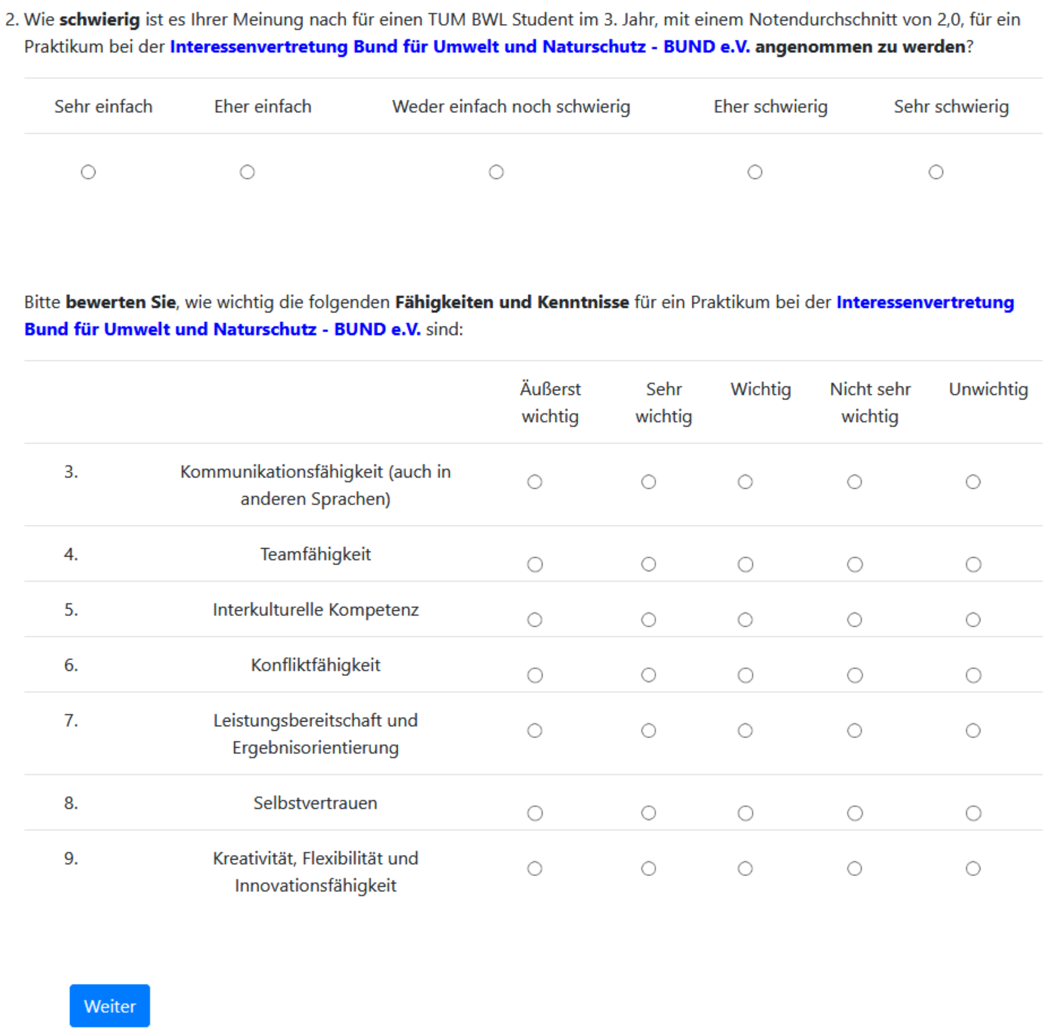
1. Wie schätzen Sie diesen Studenten ein, ist dies eher jemand, dem ein **hohes Gehalt** wichtig ist, oder dem es wichtig ist einen **positiven Beitrag** für die Gesellschaft zu leisten? Diesem Student ist ...

... **Gehalt viel wichtiger** als den meisten anderen.	... **Gehalt etwas wichtiger** als den meisten anderen.	... **beides** (Gehalt und Beitrag für die Gesellschaft) **ungefähr gleich wichtig** wie den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **etwas wichtiger** als den meisten anderen.	... einen **Beitrag** für die Gesellschaft zu leisten **viel wichtiger** als den meisten anderen.
○	○	○	○	○

2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V. angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig

Figure 3.20: Screenshot of the pre-study: page 6 (part 1).



2. Wie **schwierig** ist es Ihrer Meinung nach für einen TUM BWL Student im 3. Jahr, mit einem Notendurchschnitt von 2,0, für ein Praktikum bei der **Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V. angenommen zu werden**?

Sehr einfach	Eher einfach	Weder einfach noch schwierig	Eher schwierig	Sehr schwierig
○	○	○	○	○

Bitte **bewerten Sie**, wie wichtig die folgenden **Fähigkeiten und Kenntnisse** für ein Praktikum bei der **Interessenvertretung Bund für Umwelt und Naturschutz - BUND e.V.** sind:

		Äußerst wichtig	Sehr wichtig	Wichtig	Nicht sehr wichtig	Unwichtig
3.	Kommunikationsfähigkeit (auch in anderen Sprachen)	○	○	○	○	○
4.	Teamfähigkeit	○	○	○	○	○
5.	Interkulturelle Kompetenz	○	○	○	○	○
6.	Konfliktfähigkeit	○	○	○	○	○
7.	Leistungsbereitschaft und Ergebnisorientierung	○	○	○	○	○
8.	Selbstvertrauen	○	○	○	○	○
9.	Kreativität, Flexibilität und Innovationsfähigkeit	○	○	○	○	○

Weiter

Figure 3.21: Screenshot of the pre-study: page 6 (part 2).

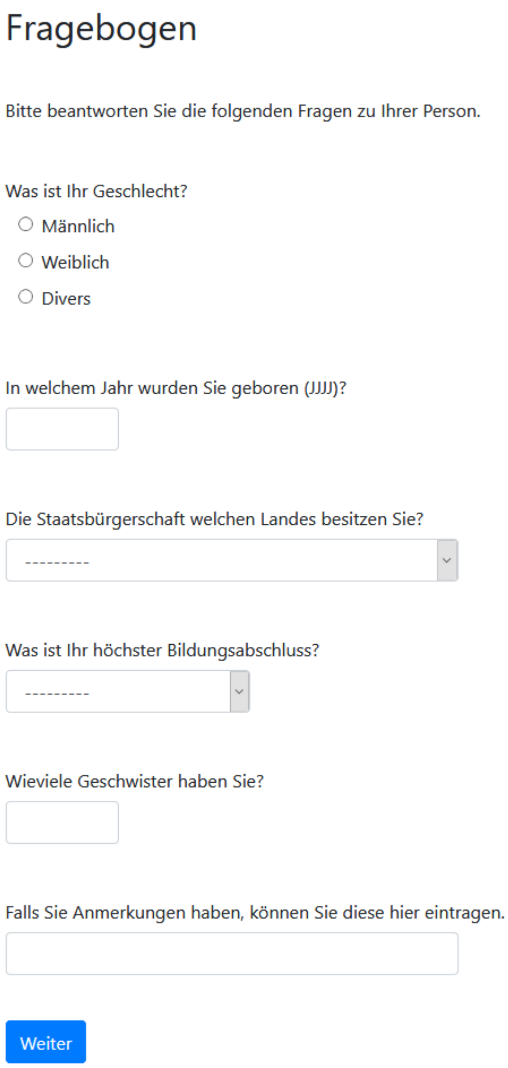


Figure 3.22: Screenshot of the pre-study: page 7.

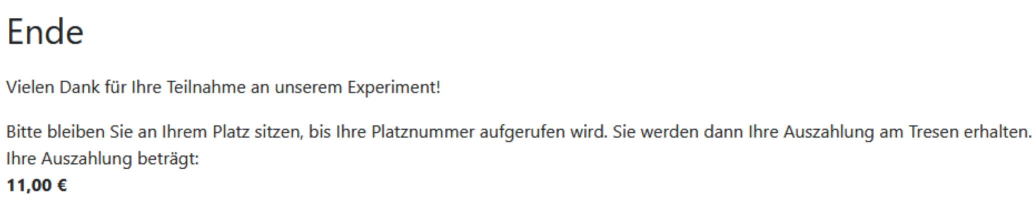
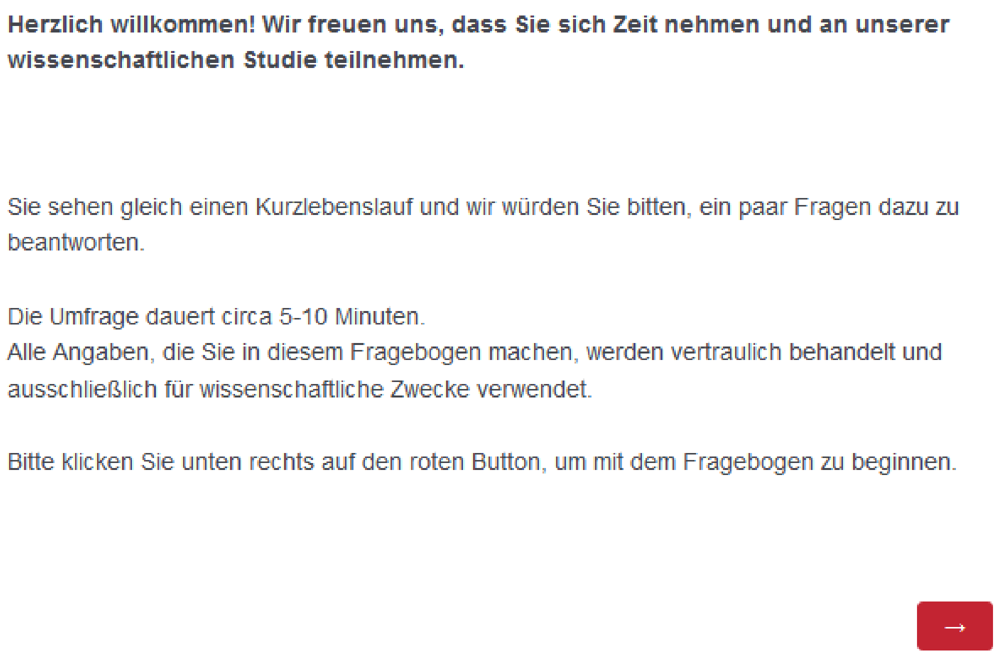


Figure 3.23: Screenshot of the pre-study: page 8.

3.6.2.2 Screenshots main experiment



Herzlich willkommen! Wir freuen uns, dass Sie sich Zeit nehmen und an unserer wissenschaftlichen Studie teilnehmen.

Sie sehen gleich einen Kurzlebenslauf und wir würden Sie bitten, ein paar Fragen dazu zu beantworten.

Die Umfrage dauert circa 5-10 Minuten.
Alle Angaben, die Sie in diesem Fragebogen machen, werden vertraulich behandelt und ausschließlich für wissenschaftliche Zwecke verwendet.

Bitte klicken Sie unten rechts auf den roten Button, um mit dem Fragebogen zu beginnen.

→

Figure 3.24: Screenshot of the main experiment: page 1.

Bitte lesen Sie sich den Kurzlebenslauf aufmerksam durch. Im Anschluss bitten wir Sie um Ihre Einschätzung zu diesem Kandidaten.

SCHNEIDER Anna

Adresse:	Blütenstraße 14
	80799 München
Emailadresse:	anna.schneider@gmail.com
Staatsangehörigkeit:	Deutsch
Geburtsdaten:	15.04.1995 in München

Praktische Erfahrungen

04/2018 - 07/2018 **Tesla Inc.** - *Praktikantin im Bereich Entwicklung* Prüm, Deutschland

- Unterstützung des Elektromotoren-Teams bei der Entwicklungsarbeit
- Statistische Auswertung von antriebsrelevanten Problemen bei Elektrofahrzeugen

08/2015 - 10/2015 **Innogy SE** – *Praktikantin im Bereich Valuation* Essen, Deutschland

- Unterstützung bei der Bewertung von M&A-Projekten verschiedener regenerativer Erzeugungstechnologien wie Wind Onshore, Wind Offshore und Photovoltaik
- Aufbereitung entscheidungsrelevanter Informationen für das Management der Sparte Renewables

Akademische Ausbildung

10/2016 - 05/2019 **Technische Universität München** München, Deutschland
Masterstudium: Technologie- und Managementorientierte BWL
Abschluss: Master of Science

09/2016 - 12/2016 **Universitat de València** Valencia, Spanien
Auslandssemester

10/2013 - 09/2016 **Technische Universität München** München, Deutschland
Bachelorstudium: Technologie- und Managementorientierte BWL
Abschluss: Bachelor of Science

2005-2013 **Wittelsbacher-Gymnasium** München, Deutschland
Abitur

Qualifikationen

Sprachen:	Deutsch (Muttersprache), Englisch (verhandlungssicher in Wort und Schrift), Spanisch (Grundkenntnisse), Latein (Großes Latinum)
EDV-Kenntnisse:	MS Office, Python

Hobbies

Fußball, Gitarre

Figure 3.25: Screenshot of the main experiment: page 2 (part 1).

Wie schätzen Sie diesen Kandidaten hinsichtlich seiner Leistungsbereitschaft und Ergebnisorientierung ein?

Sehr schlecht	Eher schlecht	Weder gut noch schlecht	Eher gut	Sehr gut

Wie schätzen Sie diesen Kandidaten hinsichtlich seiner interkulturellen Kompetenz ein?

Sehr schlecht	Eher schlecht	Weder gut noch schlecht	Eher gut	Sehr gut

Wie schätzen Sie diesen Kandidaten ein, ist dies eher jemand, dem ein hohes Gehalt wichtig ist, oder jemand, dem es wichtig ist einen positiven Beitrag für die Gesellschaft zu leisten? Diesem Kandidaten ist ...

Gehalt viel wichtiger als den meisten anderen.	Gehalt etwas wichtiger als den meisten anderen.	beides (Gehalt und Beitrag für die Gesellschaft) ungefähr gleich wichtig wie den meisten anderen.	einen Beitrag für die Gesellschaft zu leisten etwas wichtiger als den meisten anderen.	einen Beitrag für die Gesellschaft zu leisten viel wichtiger als den meisten anderen.

Stellen Sie sich jetzt bitte vor, dass sich dieser Kandidat (Kurzlebenslauf s.o.) auf eine Stelle in Ihrem Unternehmen bewirbt.

Für welche Jobprofile in Ihrem Unternehmen kommt dieser Kandidat in Frage? (Mehrfachauswahl möglich)

Assistenz der Geschäftsführung
Personalabteilung (HR)
Buchhaltung, Rechnungswesen

Figure 3.26: Screenshot of the main experiment: page 2 (part 2).

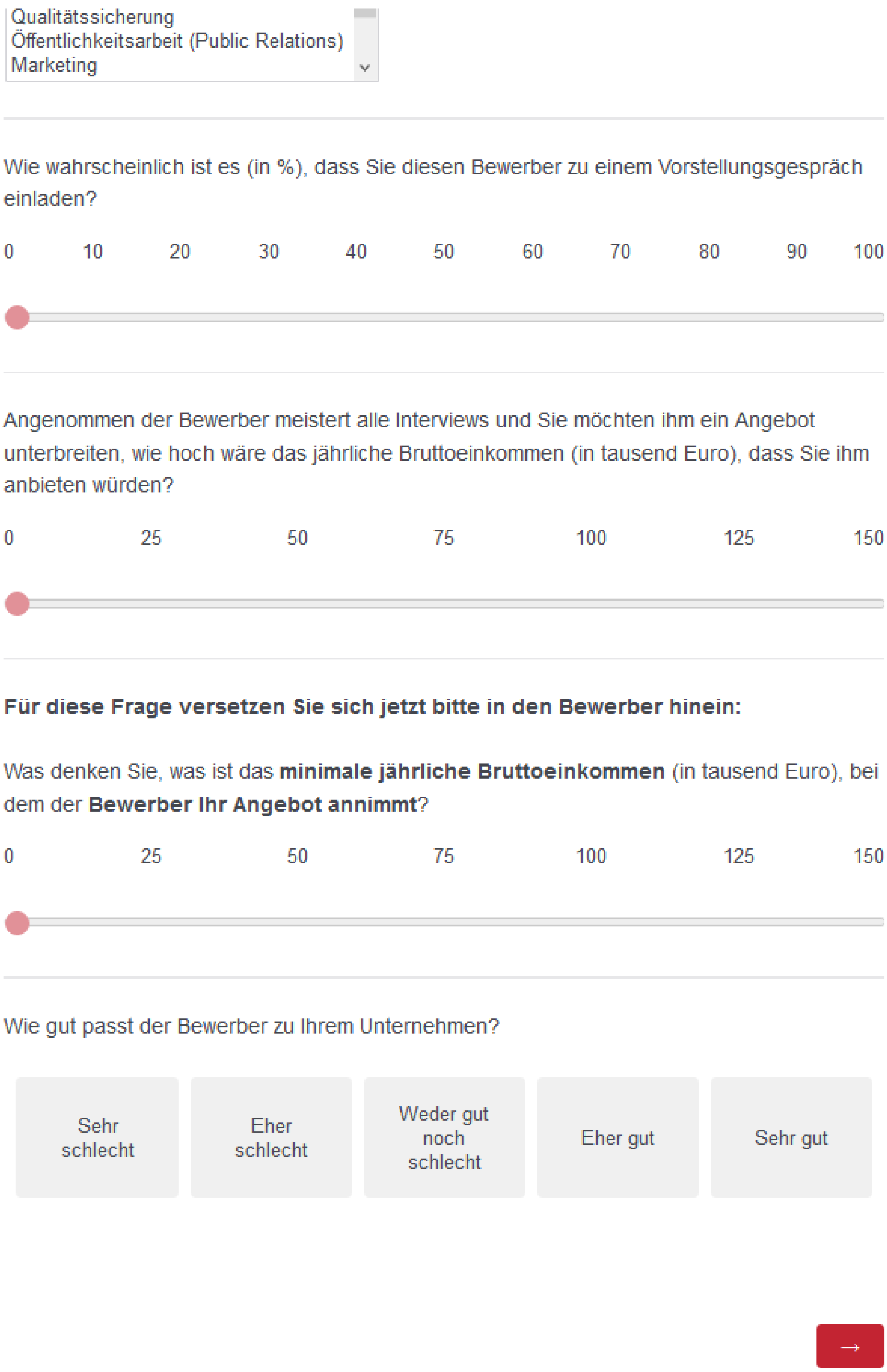


Figure 3.27: Screenshot of the main experiment: page 2 (part 3).

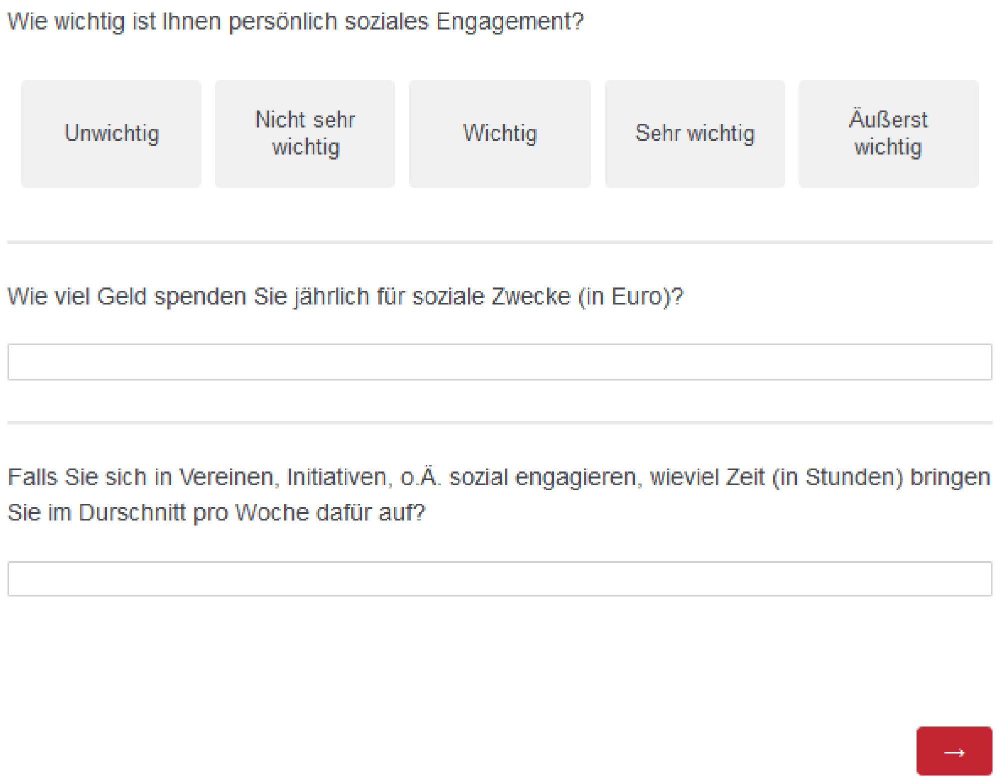


Figure 3.28: Screenshot of the main experiment: page 3.

In welcher Branche ist Ihr Unternehmen tätig?

Informations- und Kommunikationstechnologie
Gesundheitswesen
Energieversorger, Chemie
Automobilindustrie
Einzelhandel
Industrie
Finanzen
Personaldienstleistung
Andere Branche:

Wie viele Mitarbeiter beschäftigt Ihr Unternehmen?

< 50	51 bis 250	251 bis 1.000	1.001 bis 10.000	> 10.000

Wie schätzen Sie Ihr Unternehmen ein, handelt es sich eher um ein Unternehmen, dem nur ein hoher Gewinn wichtig ist, oder dem es auch wichtig ist einen positiven Beitrag für die Gesellschaft zu leisten?
Dem Unternehmen ist...

ein hoher Gewinn viel wichtiger als den meisten anderen.	ein hoher Gewinn etwas wichtiger als den meisten anderen.	beides (Gewinn und Beitrag für die Gesellschaft) ungefähr gleich wichtig wie den meisten anderen.	einen Beitrag für die Gesellschaft zu leisten etwas wichtiger als den meisten anderen.	einen Beitrag für die Gesellschaft zu leisten viel wichtiger als den meisten anderen.

Wie wird Ihr Unternehmen in der Öffentlichkeit wahrgenommen? Eher als ein Unternehmen, dem nur ein hoher Gewinn wichtig ist, oder dem es auch wichtig ist einen positiven Beitrag für die Gesellschaft zu leisten?
Das Unternehmen wird in der Öffentlichkeit eher als ein Unternehmen wahrgenommen, dem...

Figure 3.29: Screenshot of the main experiment: page 4 (part 1).

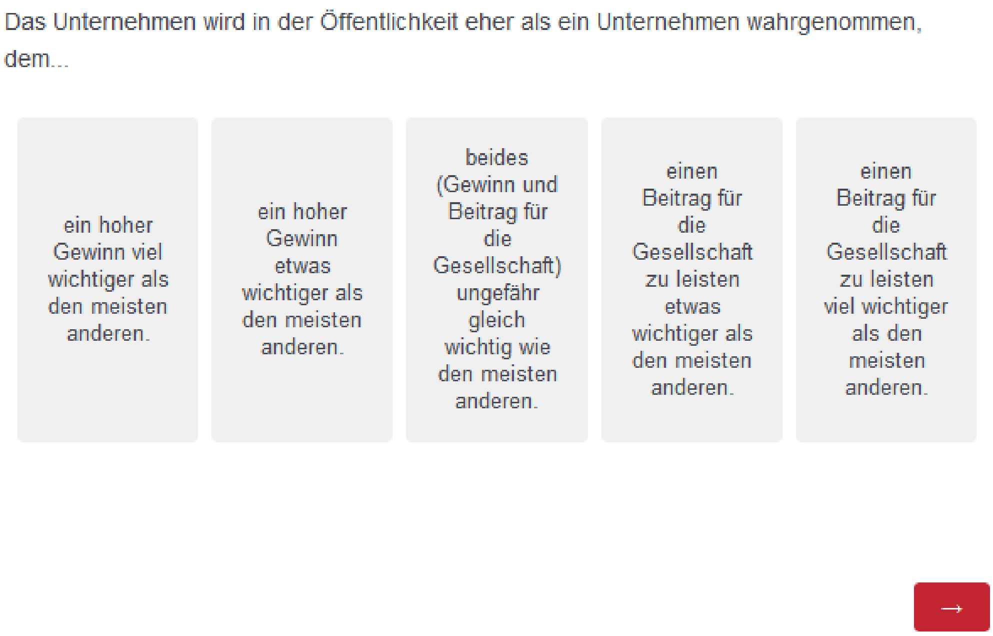


Figure 3.30: Screenshot of the main experiment: page 4 (part 2).

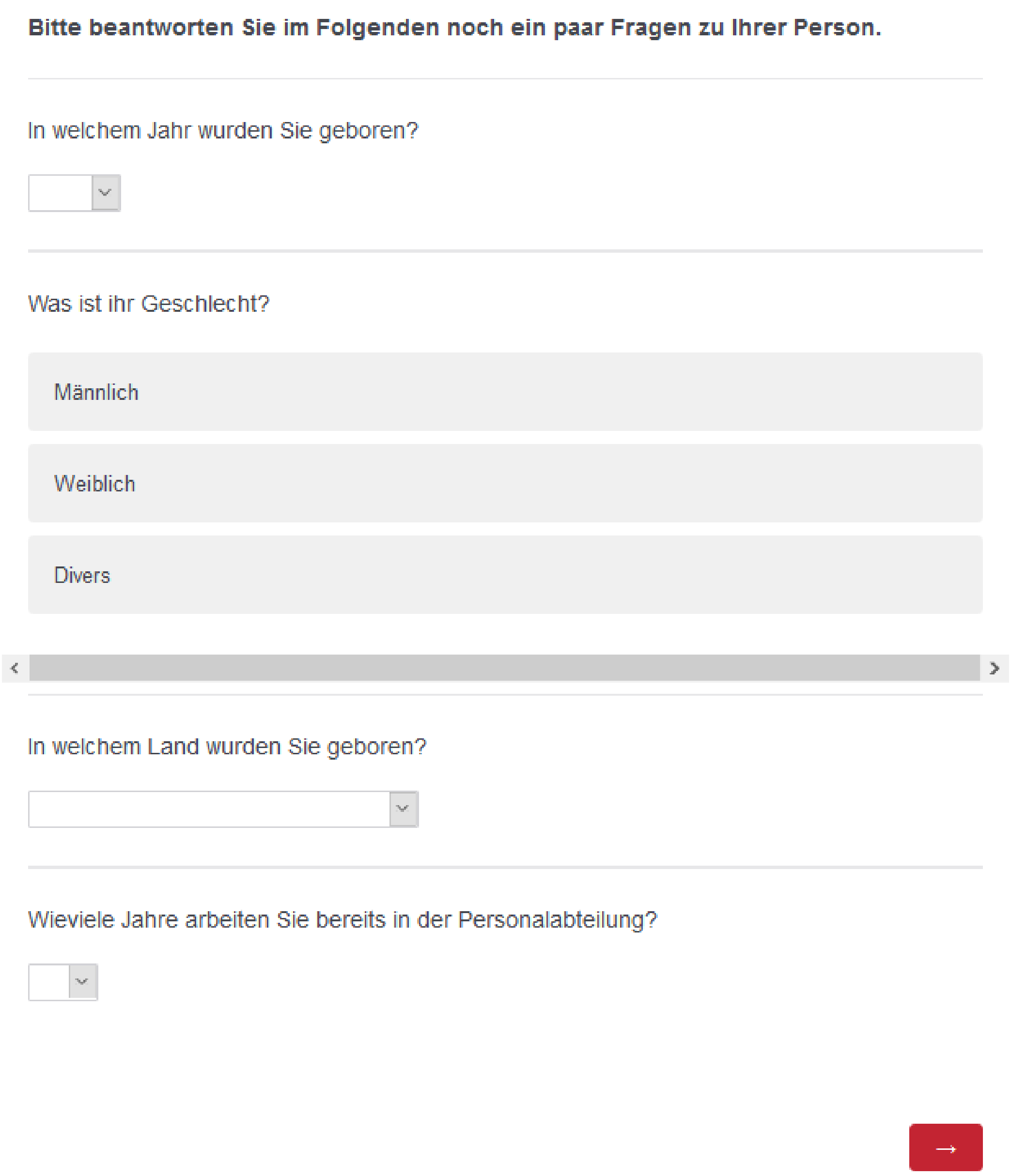


Figure 3.31: Screenshot of the main experiment: page 5.

3.6.3 Screenshots laboratory experiment

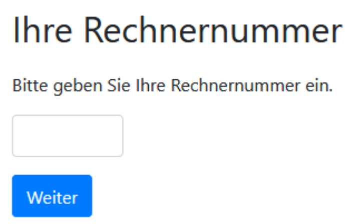


Figure 3.32: Screenshot of the laboratory experiment: page 1.

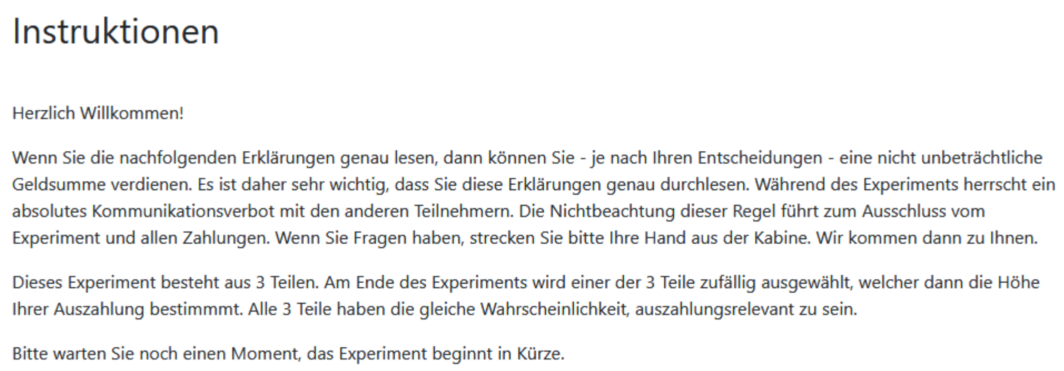


Figure 3.33: Screenshot of the laboratory experiment: page 2.

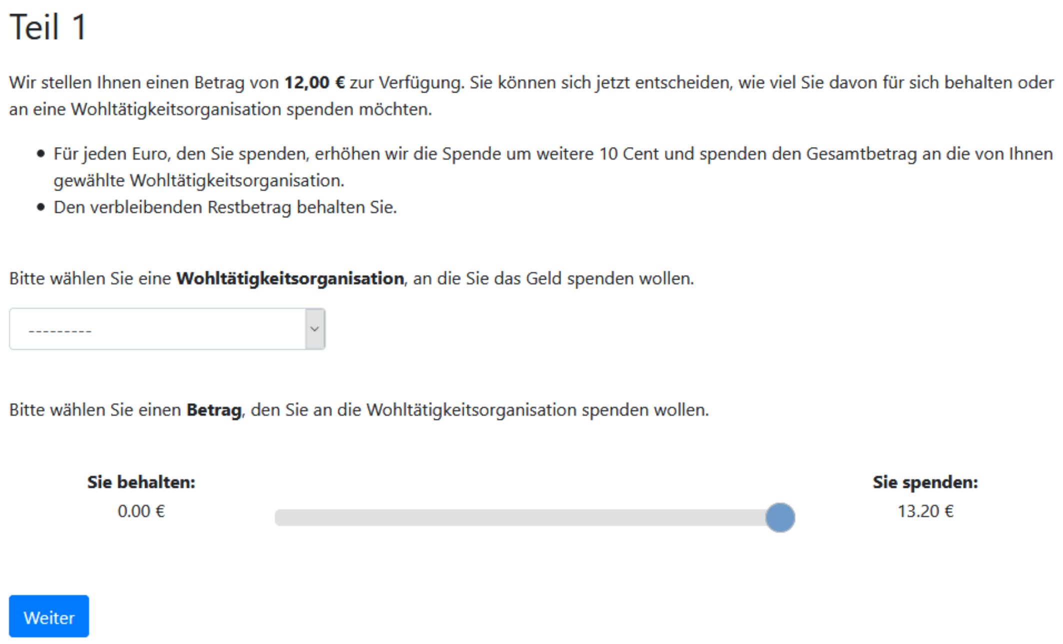
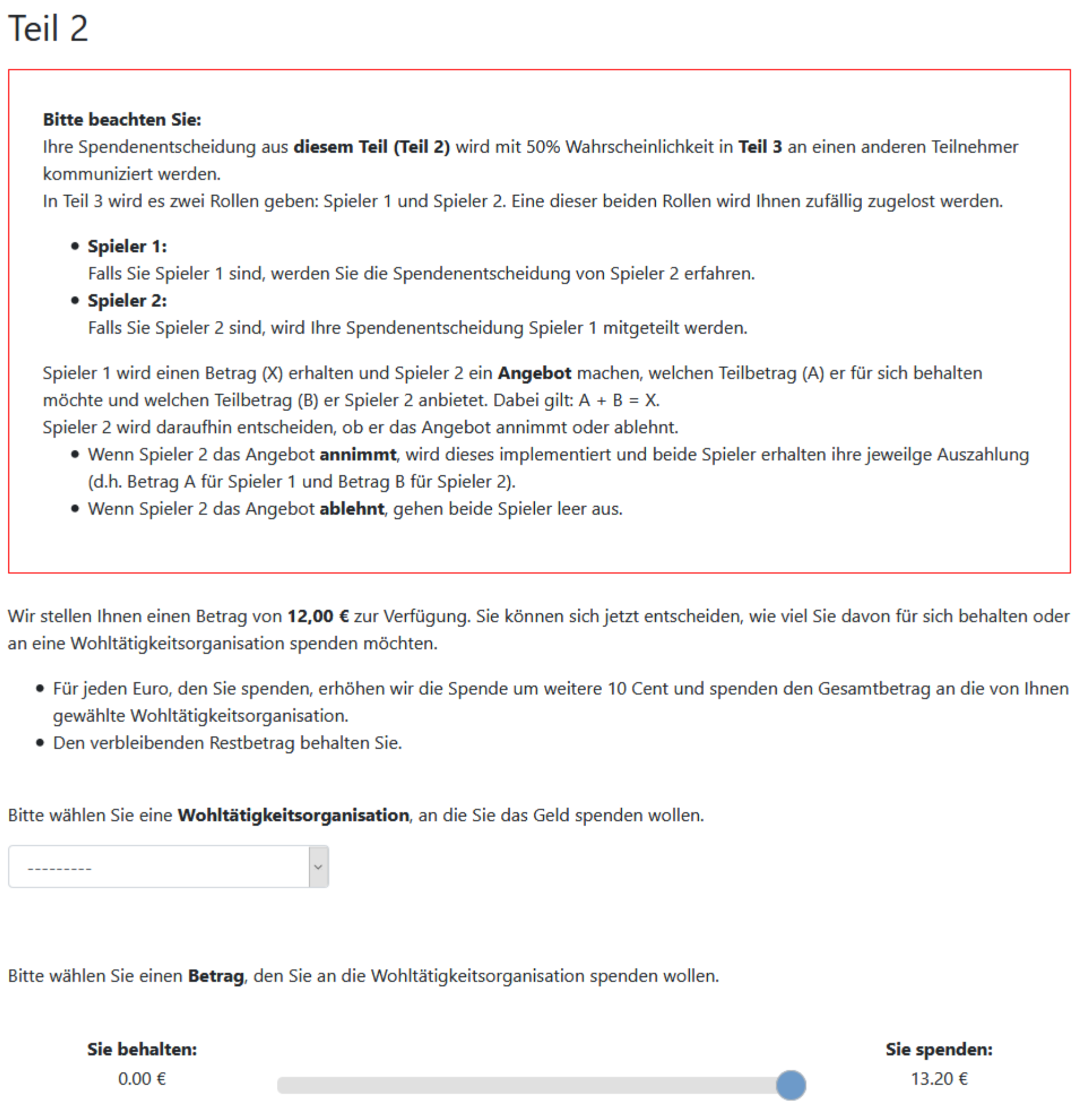


Figure 3.34: Screenshot of the laboratory experiment: page 3.



Teil 2

Bitte beachten Sie:
Ihre Spendenentscheidung aus **diesem Teil (Teil 2)** wird mit 50% Wahrscheinlichkeit in **Teil 3** an einen anderen Teilnehmer kommuniziert werden.
In Teil 3 wird es zwei Rollen geben: Spieler 1 und Spieler 2. Eine dieser beiden Rollen wird Ihnen zufällig zugelost werden.

- **Spieler 1:**
  Falls Sie Spieler 1 sind, werden Sie die Spendenentscheidung von Spieler 2 erfahren.
- **Spieler 2:**
  Falls Sie Spieler 2 sind, wird Ihre Spendenentscheidung Spieler 1 mitgeteilt werden.

Spieler 1 wird einen Betrag (X) erhalten und Spieler 2 ein **Angebot** machen, welchen Teilbetrag (A) er für sich behalten möchte und welchen Teilbetrag (B) er Spieler 2 anbietet. Dabei gilt: A + B = X.
Spieler 2 wird daraufhin entscheiden, ob er das Angebot annimmt oder ablehnt.

- Wenn Spieler 2 das Angebot **annimmt**, wird dieses implementiert und beide Spieler erhalten ihre jeweilge Auszahlung (d.h. Betrag A für Spieler 1 und Betrag B für Spieler 2).
- Wenn Spieler 2 das Angebot **ablehnt**, gehen beide Spieler leer aus.

Wir stellen Ihnen einen Betrag von **12,00 €** zur Verfügung. Sie können sich jetzt entscheiden, wie viel Sie davon für sich behalten oder an eine Wohltätigkeitsorganisation spenden möchten.

- Für jeden Euro, den Sie spenden, erhöhen wir die Spende um weitere 10 Cent und spenden den Gesamtbetrag an die von Ihnen gewählte Wohltätigkeitsorganisation.
- Den verbleibenden Restbetrag behalten Sie.

Bitte wählen Sie eine **Wohltätigkeitsorganisation**, an die Sie das Geld spenden wollen.

Bitte wählen Sie einen **Betrag**, den Sie an die Wohltätigkeitsorganisation spenden wollen.

Sie behalten: 0.00 €

Sie spenden: 13.20 €

Figure 3.35: Screenshot of the laboratory experiment: page 4.

Ihre Einschätzung

Wie schätzen Sie eine Person ein, die sich im **vorherigen Teil (Teil 2)** für folgende Aufteilungen entschieden hat?

Bewerten Sie diese Person auf einer Skala von *"Dieser Person ist **Geld viel wichtiger** als den meisten anderen"* bis *"Dieser Person is ein **positiver Beitrag für die Gesellschaft zu leisten** viel wichtiger als den meisten anderen"*.

Am Ende des Experiments wird eines der Szenarien ausgewählt. Für dieses Szenario ermitteln wir, welche Einschätzung die meisten Teilnehmern hatten. Sie erhalten zusätzlich 2,00 €, wenn Sie die **gleiche Einschätzung** wie die meisten anderen Teilnehmer angegeben haben.

	Geld viel wichtiger	Geld wichtiger	Beides ähnlich wichtig	Positiver Beitrag für die Gesellschaft wichtiger	Positiver Beitrag für die Gesellschaft viel wichtiger
Behalten: 0,00 € & Gespendet: 13,20 €	○	○	○	○	○
Behalten: 1,00 € & Gespendet: 12,10 €	○	○	○	○	○
Behalten: 2,00 € & Gespendet: 11,00 €	○	○	○	○	○
Behalten: 3,00 € & Gespendet: 9,90 €	○	○	○	○	○
Behalten: 4,00 € & Gespendet: 8,80 €	○	○	○	○	○
Behalten: 5,00 € & Gespendet: 7,70 €	○	○	○	○	○
Behalten: 6,00 € & Gespendet: 6,60 €	○	○	○	○	○
Behalten: 7,00 € & Gespendet: 5,50 €	○	○	○	○	○
Behalten: 8,00 € & Gespendet: 4,40 €	○	○	○	○	○
Behalten: 9,00 € & Gespendet: 3,30 €	○	○	○	○	○
Behalten: 10,00 € & Gespendet: 2,20 €	○	○	○	○	○

Figure 3.36: Screenshot of the laboratory experiment: page 5 (part 1).

Behalten: 10,00 € & Gespendet: 2,20 €	○	○	○	○	○
Behalten: 11,00 € & Gespendet: 1,10 €	○	○	○	○	○
Behalten: 12,00 € & Gespendet: 0,00 €	○	○	○	○	○

Weiter

Figure 3.37: Screenshot of the laboratory experiment: page 5 (part 2).

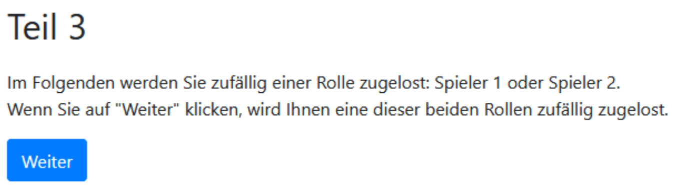


Figure 3.38: Screenshot of the laboratory experiment: page 6.

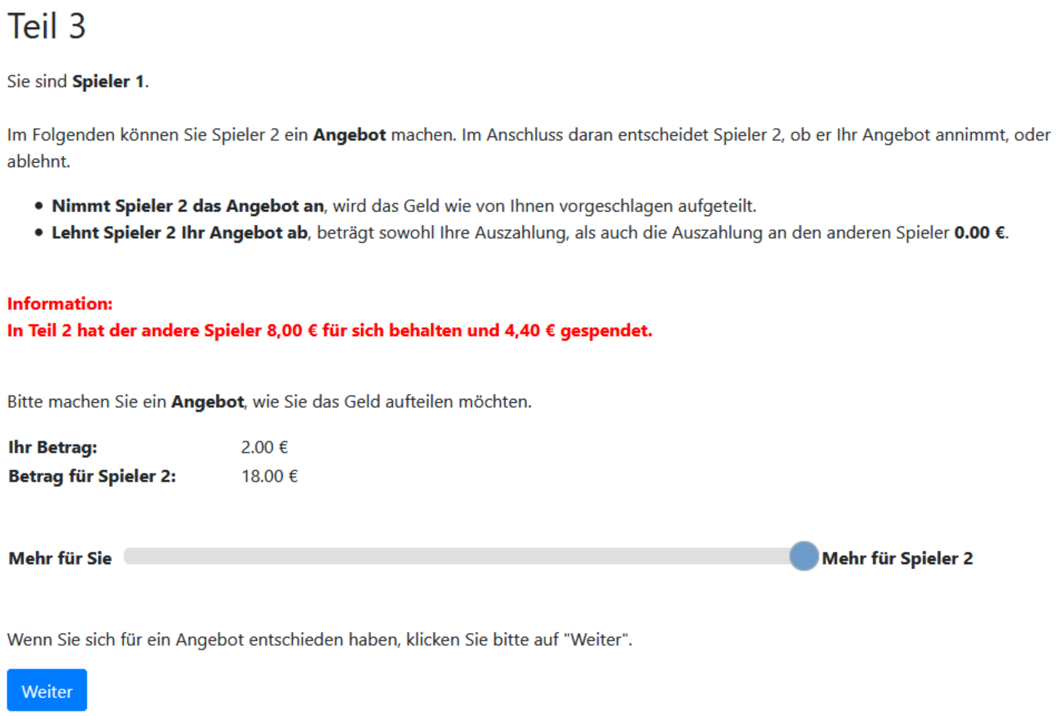
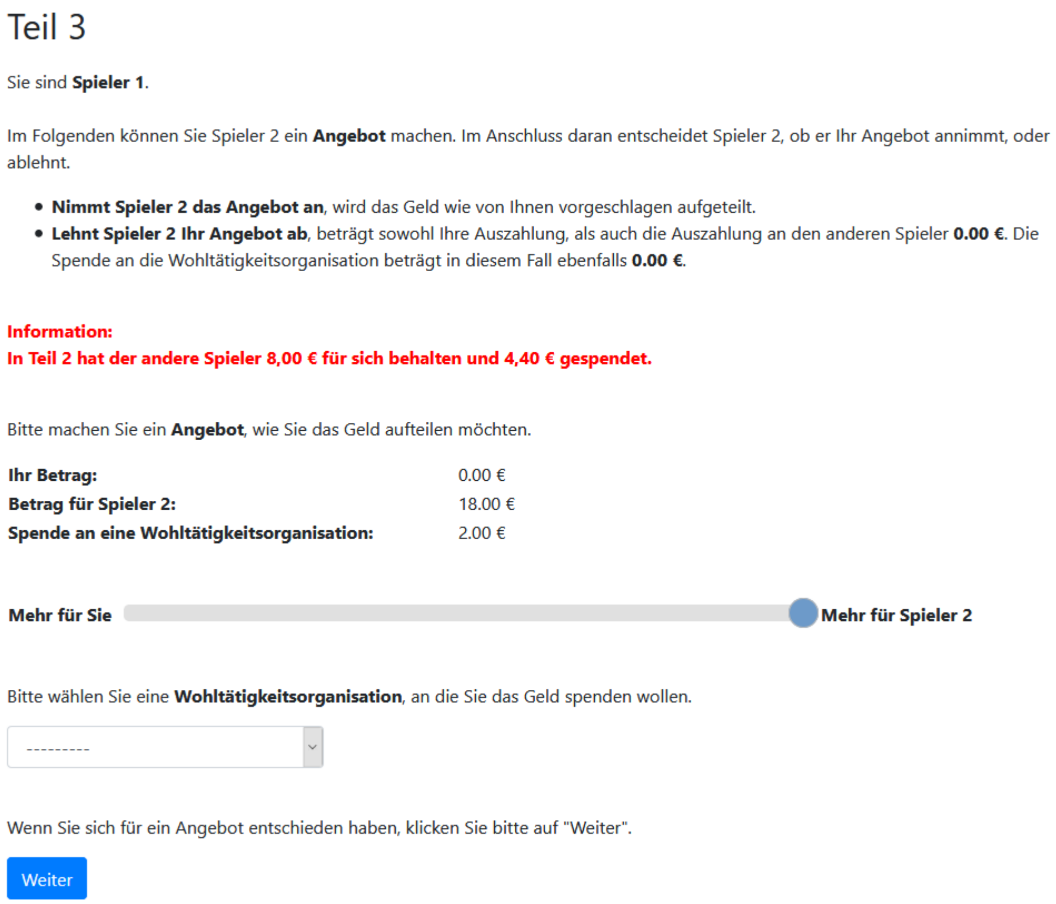


Figure 3.39: Screenshot of the laboratory experiment: page 7 (control).



Teil 3

Sie sind **Spieler 1**.

Im Folgenden können Sie Spieler 2 ein **Angebot** machen. Im Anschluss daran entscheidet Spieler 2, ob er Ihr Angebot annimmt, oder ablehnt.

- **Nimmt Spieler 2 das Angebot an**, wird das Geld wie von Ihnen vorgeschlagen aufgeteilt.
- **Lehnt Spieler 2 Ihr Angebot ab**, beträgt sowohl Ihre Auszahlung, als auch die Auszahlung an den anderen Spieler **0.00 €**. Die Spende an die Wohltätigkeitsorganisation beträgt in diesem Fall ebenfalls **0.00 €**.

Information:
In Teil 2 hat der andere Spieler 8,00 € für sich behalten und 4,40 € gespendet.

Bitte machen Sie ein **Angebot**, wie Sie das Geld aufteilen möchten.

Ihr Betrag:	0.00 €
Betrag für Spieler 2:	18.00 €
Spende an eine Wohltätigkeitsorganisation:	2.00 €

Mehr für Sie **Mehr für Spieler 2**

Bitte wählen Sie eine **Wohltätigkeitsorganisation**, an die Sie das Geld spenden wollen.

Wenn Sie sich für ein Angebot entschieden haben, klicken Sie bitte auf "Weiter".

Weiter

Figure 3.40: Screenshot of the laboratory experiment: page 7 (treatment).

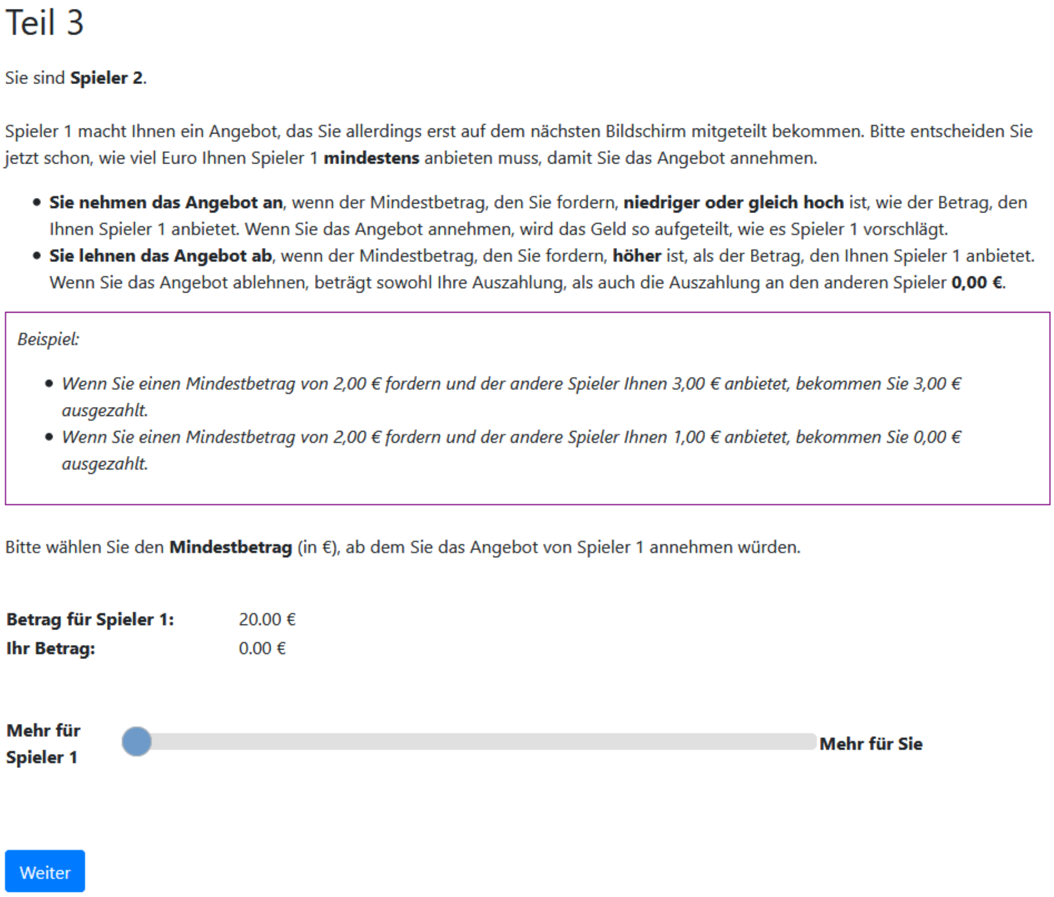


Figure 3.41: Screenshot of the laboratory experiment: page 8 (control).

Teil 3

Sie sind **Spieler 2**.

Spieler 1 macht Ihnen ein Angebot, das Sie allerdings erst auf dem nächsten Bildschirm mitgeteilt bekommen. Bitte entscheiden Sie jetzt schon, wie viel Euro Ihnen Spieler 1 **mindestens** anbieten muss, damit Sie das Angebot annehmen.

- **Sie nehmen das Angebot an**, wenn der Mindestbetrag, den Sie fordern, **niedriger oder gleich hoch** ist, wie der Betrag, den Ihnen Spieler 1 anbietet. Wenn Sie das Angebot annehmen, wird das Geld so aufgeteilt, wie es Spieler 1 vorschlägt. In diesem Fall werden **2,00 €** an eine Wohltätigkeitsorganisation gespendet.
- **Sie lehnen das Angebot ab**, wenn der Mindestbetrag, den Sie fordern, **höher** ist, als der Betrag, den Ihnen Spieler 1 anbietet. Wenn Sie das Angebot ablehnen, beträgt sowohl Ihre Auszahlung, als auch die Auszahlung an den anderen Spieler **0,00 €**. Die Spende an die Wohltätigkeitsorganisation beträgt in diesem Fall ebenfalls **0,00 €**.

Beispiel:

- *Wenn Sie einen Mindestbetrag von 2,00 € fordern und der andere Spieler Ihnen 3,00 € anbietet, bekommen Sie 3,00 € ausgezahlt.*
- *Wenn Sie einen Mindestbetrag von 2,00 € fordern und der andere Spieler Ihnen 1,00 € anbietet, bekommen Sie 0,00 € ausgezahlt.*

Bitte wählen Sie den **Mindestbetrag** (in €), ab dem Sie das Angebot von Spieler 1 annehmen würden.

Betrag für Spieler 1:	18.00 €
Ihr Betrag:	0.00 €
Spende an eine Wohltätigkeitsorganisation:	2.00 €

Mehr für Spieler 1 **Mehr für Sie**

Weiter

Figure 3.42: Screenshot of the laboratory experiment: page 8 (treatment).

Teil 3

Spieler 2 hat Ihr Angebot angenommen.

Die Auszahlungen für diesen Teil sind wie folgt:

Ihre Auszahlung:	14,00 €
Auszahlung an Spieler 2:	6,00 €

Falls diese Aufgabe als auszahlungsrelevant ausgewählt wird, erhalten Sie den oben genannten Betrag.

Weiter

Figure 3.43: Screenshot of the laboratory experiment: page 9 (control - part 1).

Teil 3

Sie haben das Angebot von Spieler 1 angenommen.

Die Auszahlungen für diesen Teil sind wie folgt:

Ihre Auszahlung:	6,00 €
Auszahlung an Spieler 1:	14,00 €

Falls diese Aufgabe als auszahlungsrelevant ausgewählt wird, erhalten Sie den oben genannten Betrag.

Weiter

Figure 3.44: Screenshot of the laboratory experiment: page 9 (control - part 2).

Teil 3

Spieler 2 hat Ihr Angebot angenommen.

Die Auszahlungen für diesen Teil sind wie folgt:

Ihre Auszahlung:	9,00 €
Auszahlung an Spieler 2:	9,00 €
Spende an die Wohltätigkeitsorganisation:	2,00 €

Falls diese Aufgabe als auszahlungsrelevant ausgewählt wird, erhalten Sie den oben genannten Betrag.

Weiter

Figure 3.45: Screenshot of the laboratory experiment: page 9 (treatment - part 1).

Teil 3

Sie haben das Angebot von Spieler 1 angenommen.

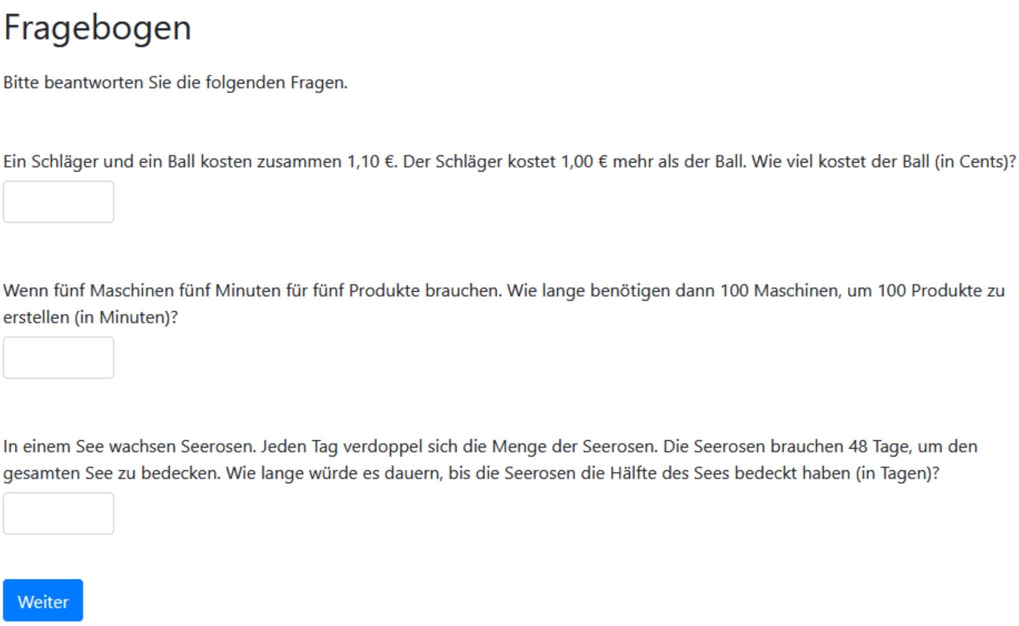
Die Auszahlungen für diesen Teil sind wie folgt:

Ihre Auszahlung:	9,00 €
Auszahlung an Spieler 1:	9,00 €
Spende an die Wohltätigkeitsorganisation:	2,00 €

Falls diese Aufgabe als auszahlungsrelevant ausgewählt wird, erhalten Sie den oben genannten Betrag.

Weiter

Figure 3.46: Screenshot of the laboratory experiment: page 9 (treatment - part 2).


Fragebogen

Bitte beantworten Sie die folgenden Fragen.

Ein Schläger und ein Ball kosten zusammen 1,10 €. Der Schläger kostet 1,00 € mehr als der Ball. Wie viel kostet der Ball (in Cents)?

Wenn fünf Maschinen fünf Minuten für fünf Produkte brauchen. Wie lange benötigen dann 100 Maschinen, um 100 Produkte zu erstellen (in Minuten)?

In einem See wachsen Seerosen. Jeden Tag verdoppel sich die Menge der Seerosen. Die Seerosen brauchen 48 Tage, um den gesamten See zu bedecken. Wie lange würde es dauern, bis die Seerosen die Hälfte des Sees bedeckt haben (in Tagen)?

Weiter

Figure 3.47: Screenshot of the laboratory experiment: page 10.

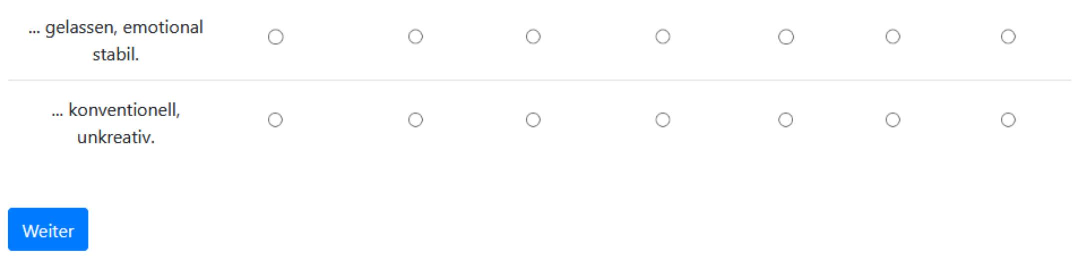
Fragebogen

Im Folgenden sehen Sie mehrere Charaktereigenschaften, die möglicherweise auf Sie zutreffen. Bitte geben Sie für jede Aussage an, wie sehr diese auf Sie zutrifft.

Ich sehe mich selbst als ...

	Trifft überhaupt nicht zu	Trifft größtenteils nicht zu	Trifft eher nicht zu	Weder zutreffend noch unzutreffend	Trifft eher zu	Trifft größtenteils zu	Trifft voll und ganz zu
... extrovertiert, begeistert.	○	○	○	○	○	○	○
... kritisch, streitsüchtig.	○	○	○	○	○	○	○
... zuverlässig, selbstdiszipliniert.	○	○	○	○	○	○	○
... ängstlich, leicht aus der Fassung zu bringen.	○	○	○	○	○	○	○
... offen für neue Erfahrungen, vielschichtig.	○	○	○	○	○	○	○
... zurückhaltend, still.	○	○	○	○	○	○	○
... verständnisvoll, warmherzig.	○	○	○	○	○	○	○
... unorganisiert, achtlos.	○	○	○	○	○	○	○
... gelassen, emotional stabil.	○	○	○	○	○	○	○

Figure 3.48: Screenshot of the laboratory experiment: page 11 (part 1).



... gelassen, emotional stabil.	○	○	○	○	○	○	○
... konventionell, unkreativ.	○	○	○	○	○	○	○

Weiter

Figure 3.49: Screenshot of the laboratory experiment: page 11 (part 2).

Fragebogen

Bitte beantworten Sie die folgenden Fragen.

	Unwichtig	Nicht sehr wichtig	Wichtig	Sehr wichtig	Äußerst wichtig
Wie wichtig ist Ihnen persönlich soziales Engagement?	○	○	○	○	○

Wie viel Geld spenden Sie jährlich für soziale Zwecke (in Euro)?

Falls Sie sich in Vereinen, Initiativen, o.Ä. engagieren, wie viel Zeit (in Stunden) bringen Sie im Durchschnitt pro Woche dafür auf?

Wie schätzen Sie sich persönlich ein:
Sind Sie im Allgemeinen ein risikobereiter Mensch oder versuchen Sie, Risiken zu vermeiden?
Antworten Sie bitte anhand der folgenden Skala, wobei der Wert 0 bedeutet: gar nicht risikobereit und der Wert 10: sehr risikobereit.
Mit den Werten dazwischen können Sie Ihre Einschätzung abstufen.

Weiter

Figure 3.50: Screenshot of the laboratory experiment: page 12.

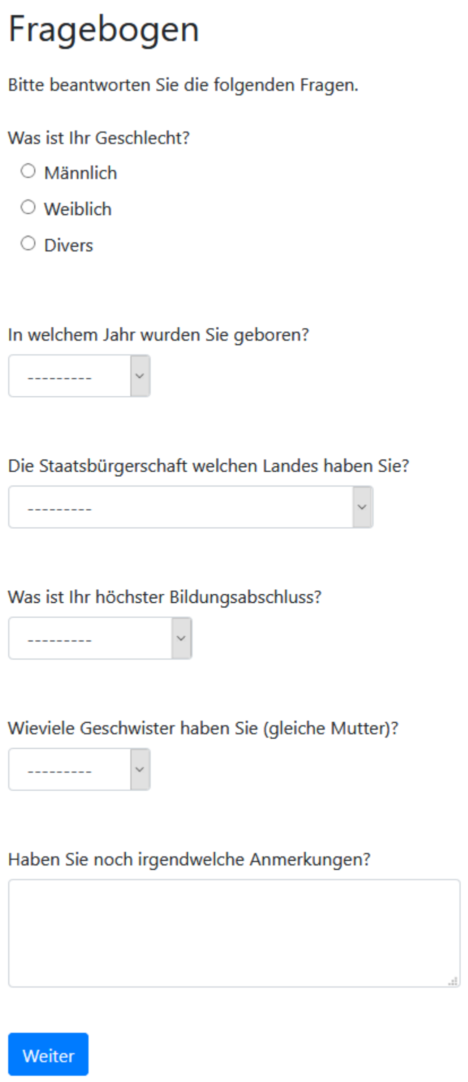


Figure 3.51: Screenshot of the laboratory experiment: page 13.

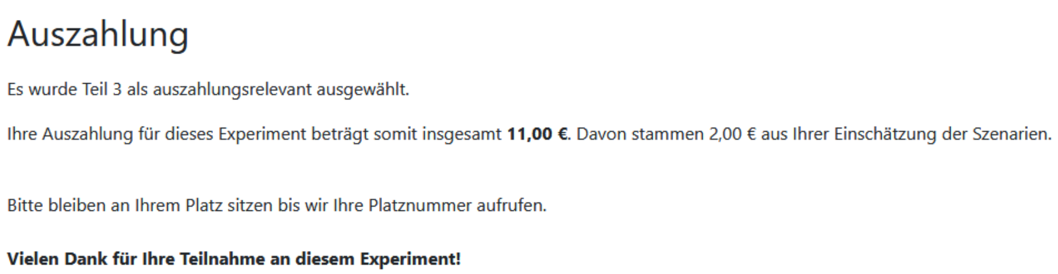


Figure 3.52: Screenshot of the laboratory experiment: page 14.

4. Does an increase in monitoring reduce sabotage? Empirical evidence from professional soccer

Abstract

The deterrence hypothesis posits that an increase in monitoring reduces sabotage. We test this conjecture with empirical data from a quasi-natural experiment caused by the introduction of the Video Assistant Referee (VAR) in German professional soccer. The introduction of the VAR in the 1. Bundesliga, but not in the 2. Bundesliga, allows us to study the effect of this increase in monitoring with a difference-in-differences approach. To measure sabotage, we introduce a novel proxy, namely substitutions that are due to an in-match injury. In contrast to the deterrence hypothesis, we do not find evidence that an increase in monitoring reduces sabotage. We discuss possible reasons for this result.

Keywords:	Contest; Home bias; Monitoring; Sabotage; Soccer; Video Assistant Referee
JEL Codes:	D74; K42; M52; Z28
Authors:	Thomas Daske, Julian Hackinger, and Magnus Strobel

4.1 Introduction

Many domains of life are deliberately designed as contests. A contest is defined as "[...] a situation in which individuals expend irretrievable resources to win valuable prize(s)." (Chowdhury and Gürtler, 2015, p. 135). Contests exist as promotion decisions at the workplace, championships in sports, and grant applications in academia. The appeal of contests (i. e., relative performance evaluations) as an incentive-scheme over absolute performance evaluations is that they are robust to common shocks (Nalebuff and Stiglitz, 1983) and workers' relative positions are typically less costly to observe than their absolute output (Lazear and Rosen, 1981).[1]

However, as first pointed out by Lazear (1989), a major problem of relative performance evaluations is that contestants can not only be successful by exerting productive effort but also by exerting destructive effort, i. e., sabotage. The author defines sabotage as "any (costly) actions that one worker takes that adversely affect output of another." (Lazear, 1989, p. 563).[2] While sabotage might prove beneficial for the saboteur, it adversely affects overall welfare. Reasons for this are the opportunity costs of sabotage for the saboteurs and the decrease in productive output of the sabotaged contestants (Chowdhury and Gürtler, 2015). Furthermore, the expectation of being sabotaged has a discouragement effect (Gürtler and Münster, 2013, 2010), and sabotage may even lead to adverse selection into the contest (Münster, 2007).

There are two main policy approaches to address the problem of sabotage, namely reducing the benefits of sabotage or increasing its costs (Chowdhury and Gürtler, 2015). Typically, contest designers resort to the latter by using deterrence through an increase in monitoring. The deterrence hypothesis posits that crime decreases in the certainty or the severity of punishment (Becker, 1968). Since sabotage is typically forbidden, it is difficult to observe it in the field. Therefore, researchers mostly use rule-breaking in sports contests as an alternative way to study sabotage. For instance, existing studies used cards in soccer (Deutscher and Schneemann, 2017; Bartling et al., 2015; del Corral et al., 2009; Garicano and Palacios-Huerta, 2005), shido in judo (Balafoutas et al., 2012), time penalties in ice hockey (Allen, 2016; Heckelman and Yates, 2003; Levitt, 2002) as proxies for sabotage. While the theoretical prediction of the deterrence hypothesis is unambiguous, empirical evidence

[1]See Connelly et al. (2014) for a literature review on contest theory.

[2]See Chowdhury and Gürtler (2015) for a literature review on sabotage in contests.

from the field is rather inconclusive. For instance, McCormick and Tollison (1984) found that an increase in the number of referees in college basketball from two to three leads to a reduction in fouls by 34%. Heckelman and Yates (2003) found that the addition of a second referee in the National Hockey League (NHL) does not have a significant deterrence effect. A contradictory result comes from Allen (2016) who found an increase in violent penalties in the NHL when a second referee is deployed.[3] The inconsistency of these findings might be due to the shortcomings of the existing proxies for sabotage, such as their dependence on the detection probability of rule violations and their bias caused by errors and favoritism of referees (see Section 4.2.2.2 for an extended discussion of these shortcomings).

To shed further light on the deterrence hypothesis, we provide novel empirical evidence from German professional soccer. Our data comprises information on all matches of the 1. and 2. Bundesliga for the seasons 2015/16 to 2018/19. To measure sabotage, we introduce a novel proxy, namely substitutions that are due to an in-match injury. We argue that in-match injuries are either self-inflicted or caused by foul-play of the opposing team, with the latter being what we define as sabotage. Furthermore, the introduction of the Video Assistant Referee (VAR) allows us to study the effect of an increase in monitoring. The VAR is an additional referee with access to video footage who reviews decisions in clearly defined categories and notifies the main referee in case of a clear and obvious error (The International Football Association Board, 2018b). In matches with the VAR, decision accuracy for reviewable categories increases from 93.0% to 98.9% (The International Football Association Board, 2018a). At the beginning of the season 2017/18, the VAR was introduced in the 1. Bundesliga, but not in the 2. Bundesliga. This created a quasi-natural experiment that allowed us to study whether this increase in monitoring reduces sabotage with a difference-in-differences approach.[4]

Overall, our results cast doubt on the deterrence hypothesis. In contrast to its prediction, we do not find empirical evidence that an increase in monitoring reduces sabotage. More precisely, the introduction of the VAR does not reduce the probability that a substitution that is due to an in-match injury takes place during

[3] Even though fouls which resulted in particularly violent penalties were likely to be already noticed by a single referee, the author cannot rule out a dominant apprehension effect as a driver for this result.

[4] We call it a quasi-natural experiment since the treatment assignment was not purely random Dinardo (2010).

a match. Furthermore, it does not reduce the number of substitutions that are due to an in-match injury. This holds for both home and away teams.

There could be several possible reasons for this result which we discuss successively. First, players may not take changes in monitoring into account when deciding whether to commit sabotage or not. Second, the increase in monitoring may be too small to be taken into account by players when deciding whether to commit sabotage or not. Third, players may not immediately adjust their sabotage effort in response to the rule change. Fourth, an increase in monitoring may actually reduce sabotage, but our proxy for sabotage may be too noisy. Fifth, due to the existence of the home bias, the introduction of the VAR may lead to two counteractive effects for home and away teams, which offset each other. Sixth, the increase in monitoring may change the players' perception of the situation in which they are involved, which causes an increase in their sabotage effort that offsets the deterrence effect.

The scientific contribution of this study is threefold. First, it contributes to the empirical literature on sabotage in contests by providing further empirical evidence on the deterrence hypothesis. So far, the results from this strand of literature are rather inconclusive (Allen, 2016; Heckelman and Yates, 2003; Levitt, 2002; McCormick and Tollison, 1984). While McCormick and Tollison (1984) found that an increase in monitoring reduces sabotage, Heckelman and Yates (2003) found no such effect, and Allen (2016) even found that it may increase sabotage. Second, this study contributes to the empirical literature on the home bias by studying how an increase in monitoring, and therefore a reduction in the wiggle room of referees, affects sabotage of home and away teams. The home bias describes favoritism of referees towards the home team, which is attributed to the presence (Pettersson-Lidbom and Priks, 2010) and number (Buraimo et al., 2012) of fans in the stadium and their proximity (Buraimo et al., 2012, 2010; Scoppa, 2008). The existence of a home bias is well-documented in the literature. For instance, the referee bias has been found to exist in the allocation of injury/extra/stoppage time (Riedl et al., 2015; Rocha et al., 2013; Dohmen, 2008; Scoppa, 2008; Garicano et al., 2005), the awarding of penalty kicks (Dohmen, 2008; Sutter and Kocher, 2004), goals (Dohmen, 2008), and yellow and red cards (Buraimo et al., 2012, 2010; Pettersson-Lidbom and Priks, 2010; Dawson et al., 2007). Through the introduction of the VAR and the associated reduction in the wiggle room of referees, this home bias is likely to be reduced. Empirical evidence for this conjecture comes from Albanese et al. (2020) who show that the introduction of two additional assistant referees in soccer is associated with lower referee bias in terms of home favoritism. Third, this study contributes methodologically by adding a novel proxy for sabotage in sports, namely substitutions that are due to an in-match injury. So far, the existing empirical literature has used cards, penalty kicks, and time penalties as proxies for sabotage (Deutscher and Schneemann, 2017; Allen, 2016; Bartling et al., 2015; Balafoutas et al., 2012; del Corral et al., 2009; Garicano and Palacios-Huerta, 2005; Heckelman and Yates, 2003; Levitt, 2002). In contrast to conventional proxies for sabotage, this novel proxy is not affected by changes in the detection probability of sabotage and therefore allows to directly measure the deterrence effect. Furthermore, it is not biased by errors and favoritism of referees (see Section 4.2.2.2 for an extended discussion of these advantages).

The remainder of this study is structured as follows. In Section 4.2, we present our data, variables, and estimation method. In Section 4.3, we present the results of our study and discuss possible reasons for our result. Section 4.4 concludes the study, and Section 4.5 includes supplementary materials.

4.2 Data, variables, and estimation method

In this section, we present our data, variables, and estimation method.

4.2.1 Data

We study how an increase in monitoring affects sabotage by using data from German professional soccer. Our data comprises information on all matches of the 1. and 2. Bundesliga for the seasons 2015/16 to 2018/19.[5] In total, we have data on 2,448 matches (2 leagues $\times$ 4 seasons $\times$ 306 matches per league and season). In the following, we provide a detailed description of the variables in our analysis.

4.2.2 Variables

4.2.2.1 Increase in monitoring: Video Assistant Referee

In matches without the VAR, there is one referee, two assistant referees, and one so-called fourth official. The referee's role is to enforce the *Laws of the Game* (i. e., the

[5]We obtained our data by using DataGorri (Hackinger, 2018) from Transfermarkt and Kicker which are both public websites that provide detailed soccer information.

codified rules of association football) as the final decision-making authority. The two assistant referees advise the referee in critical situations. The fourth official's role is to assist the referee with administrative tasks (e. g., to supervise the teams' technical areas and supervise the substitution procedures). This team of referees does not have the possibility to review any situation to make a better-informed decision. Therefore, matches without the VAR are prone to errors and favoritism of referees (Riedl et al., 2015; Rocha et al., 2013; Buraimo et al., 2012, 2010; Pettersson-Lidbom and Priks, 2010; Dohmen, 2008; Scoppa, 2008; Boyko et al., 2007; Dawson et al., 2007; Garicano et al., 2005; Sutter and Kocher, 2004).

In matches with the VAR, the referee is supported by an additional referee who has access to video footage of the match. According to the *Laws of the Game*, the definition of a VAR is as follows:

> "A video assistant referee (VAR) is a match official who may assist the referee to make a decision using replay footage only for a 'clear and obvious error' or 'serious missed incident' relating to a goal/no goal, penalty/no penalty, direct red card (not a second caution) or a case of mistaken identity when the referee cautions or sends off the wrong player of the offending team." (The International Football Association Board, 2018b)

During the match, the VAR constantly reviews all decisions that are made by the referee. If the VAR notices a clear and obvious error in one of the relevant categories, the referee is notified via a headset. The referee then decides to either follow the VAR's recommendation or to stop the match to review the video footage before making a decision. According to a study commissioned by the Fédération Internationale de Football Association (FIFA), the VAR increases the decision accuracy for reviewable categories from 93.0% to 98.9% (The International Football Association Board, 2018a). Hence, the introduction of the VAR corresponds to an increase in monitoring.

4.2.2.2 Proxy for sabotage: In-match injuries

Due to its adverse effects on the overall welfare, sabotage is typically forbidden in contests. Hence, it is difficult to study sabotage in the field. However, there are some field studies on sabotage, mostly from the realm of sports. In these studies, rule-breaking that is punished by referees serves as a proxy for sabotage (Deutscher and Schneemann, 2017; Allen, 2016; Bartling et al., 2015; Balafoutas et al., 2012; del Corral et al., 2009; Garicano and Palacios-Huerta, 2005; Heckelman and Yates, 2003; Levitt, 2002). However, this approach proves to be problematic when testing the deterrence hypothesis. The problem is that an increase in monitoring also affects the number of cards, shido, and time penalties. Hence, it becomes difficult to distinguish between the deterrence effect and the apprehension effect.[6] The apprehension effect includes two possibly counteractive sub-effects. First, an increase in monitoring also leads to less mistakenly awarded fouls (i. e., a decrease in false positives). Second, an increase in monitoring also leads to more previously overlooked fouls now being awarded (i. e., a decrease in false negatives). To get around this problem, the existing literature uses different approaches. For instance, Heckelman and Yates (2003) used an instrumental variable (IV) approach.[7] (Levitt, 2002) combined estimated parameters with a model to capture the deterrence effect. McCormick and Tollison (1984) found that in college basketball, more referees lead to fewer fouls awarded. Based on their finding, the authors conclude that the deterrence effect dominates the apprehension effect. In contrast, Allen (2016) found an increase in violent penalties in the NHL when a second referee is deployed, suggesting a dominant "apprehension effect" rather than a dominant "deterrence effect." While these approaches offer some insights on the deterrence effect, none of them is able to capture the effect itself.

To solve this problem, we propose a novel proxy for sabotage, namely substitutions that are due to an in-match injury. In contrast to conventional proxies for sabotage, such as cards, penalty kicks, and time penalties, our novel proxy has two major advantages. First, it is not affected by changes in the detection probability of sabotage. That is, if the detection probability changes, the same number of injured players remain on the pitch and the number of injured players who are substituted but not recorded as injured remains also unaffected (false negatives). Furthermore,

[6]Please note that other studies, refer to the "apprehension effect" as the "monitoring effect" (Heckelman and Yates, 2003). Due to the notational ambiguity, we use the term "apprehension effect."

[7]In the season 1999–2000, the NHL experimented with introducing a second referee in their matches. For this season, each of the 28 teams had to complete 50 two-referee games over the 82-game schedule. Since each team had to play the same number of games with the extra referee, Heckelman and Yates (2003) concluded that past history of referee assignment must be a valid predictor for referee assignment for the current game. Hence, they constructed the following two variables as instruments: (1) the percentage of the home team's previous games that had only one referee, and (2) the percentage of the opponent's games that had only one referee.

changes in the detection probability do not affect the number of uninjured players being substituted and recorded as injured (false positives). Hence, since our proxy is not prone to the apprehension effect, it allows us to directly measure the deterrence effect. Second, it is not biased by errors and favoritism of referees.

The appeal of our proxy rests on the following assumptions. First, in-match injuries are either self-inflicted or caused by foul-play of the opposing team. The latter is what we consider sabotage.[8] There is indeed ample anecdotal evidence that injuries result from (strategic) foul-play of the opposing team. For instance, many spectators of the Champions League final of 2018 still remember the foul-play of Real Madrid's Sergio Ramos against Mohamed Salah, one of Liverpool F.C.'s best players, which resulted in an in-match injury and subsequent substitution of the injured player. Second, all players who are injured get substituted, which is correctly recorded in the match report. In total, each team can make a maximum of three substitutions per match. Teams typically make use of all their substitutions. During our observation period, a total of 14,688 (306 matches per league and season $\times$ 4 seasons $\times$ 2 leagues $\times$ 2 teams per match $\times$ 3 substitutions per team per match) regular substitutions could have taken place out of which 14,163 (96.4%) were used. If a team already made all three substitutions, players can still be taken off the pitch without bringing other players on the pitch. For instance, if a player is injured, but the maximum number of substitutions has already been reached, the injured player can still be taken off the pitch. During our observation period, it happened in 14 matches (out of 2,448 matches) that more than 6 players left the pitch during a match. In all these matches, a total of 7 players left the pitch. Even though the incidence of more than 6 players leaving the pitch is quite rare, we argue that it rarely happens because coaches typically save one substitution until the end of the match to be able to substitute injured players, even in the final minutes of the match. Furthermore, during a regular season, the extra benefit of keeping an injured player on the pitch when all substitutions have already been made is typically outweighed by the risk of worsening the injury and therefore missing the player for an extended period of time. Third, the risk of suffering a self-inflicted in-match injury is not affected by an increase in monitoring. We argue that changes in the risk of suffering a self-inflicted in-match injury depend on players' age and fitness level, which are unlikely to be affected by an increase in monitoring.

[8]We consider injuries being caused by players of the own team as self-inflicted.

Based on this proxy, we construct two different dependent variables for our analysis, namely *Injury* and *Injuries*. *Injury* is a dummy variable that takes the value 1 if, during the match, any substitution due to an in-match injury took place, and 0 otherwise. *Injuries* gives the $\ln(y+1)$ transformation of the number of substitutions that are due to an in-match injury (given by y). We applied this transformation because the distribution of substitutions that are due to an in-match injury is highly skewed, with the majority of matches having no substitutions that are due to an in-match injury (see Section 4.5.2 for the distribution of the number of substitutions that are due to an in-match injury per match).

4.2.2.3 Control variables

In addition to our independent and dependent variables above, we used the following variables to control for factors that might affect the intensity of the match and thereby substitutions that are due to an in-match injury. As shown by Deutscher and Schneemann (2017); Berger and Nieken (2016), sabotage decreases with the heterogeneity of the match. To control for the heterogeneity of the match, we include the variable *Heterogeneity*. Following Frick et al. (2008), we calculate the heterogeneity of the match as the absolute value of the difference of the squared winning probabilities of the home and away team ($Heterogeneity = |P_H^2 - P_A^2|$, where P_H and P_A represent the implicit winning probabilities of the home and away teams, respectively).[9] The implicit winning probabilities are calculated by dividing the payout ratio by the payoffs associated with the respective outcomes. The payout ratio (π) is defined as follows.

[9]Please note that betting odds might entail an endogeneity problem (Deutscher et al., 2013). If bookmakers take the optimal sabotage efforts into account, the direction of causality between asymmetry measures based on betting odds and the amount of sabotage may not be clear. Hence, there could be reversed causality, which would bias the results. Typically, the argument goes that the betting odds, and therefore the match's heterogeneity, affect the sabotage of the teams. For instance, the worse the betting odds of a team relative to its opposing team, the more they want to sabotage since otherwise they do not stand a chance. However, bookmakers most likely take the team's expected sabotage into account when calculating the betting odds. For instance, if a team announces on the last press conference before the match that they will play very aggressively or that certain players, who have a track record of committing many fouls, will be part of the starting lineup, this will affect the calculated betting odds of the match. Hence, it is not entirely clear whether only the betting odds affect the sabotage of the teams or whether the expected sabotage of the teams also affects the betting odds.

$$\pi = \frac{1}{(1/\text{payoff home win}) + (1/\text{payoff draw}) + (1/\text{payoff away win})} \tag{4.1}$$

Rewriting the term of our heterogeneity variables gives *Heterogeneity* $= |(P_H - P_A)|(1 - P_D)$, where P_D gives the probability for a draw. This term accounts for two drivers of heterogeneity. First, the match's heterogeneity should be greater the higher the difference between the winning probabilities of the home and away team ($|(P_H - P_A)|$). Second, the heterogeneity of the match should be greater if, for a fixed difference in the winning probabilities between the home and away team, the probability of a draw decreases ($(1 - P_D)$.[10] To control for the number of spectators in the stadium, we include *Attendance*, which gives the natural logarithm of the number of spectators in the stadium. Another factor that might affect the number of injuries during a match is whether the match occurs during the first or second leg of the season. During the first leg of the season, players are typically less drained out than in the second leg, and therefore less prone to self-inflicted injuries. To account for that, we include the dummy variable *Second leg*, which takes the value 1 if the match took place during the second part of the season, and 0 otherwise. In derbies, teams typically play more aggressively than in regular matches. To account for that, we include the dummy variable *Derby*, which takes the value 1 if the match is considered as a derby and 0 otherwise (see Secion 4.5.3 for a list of all derbies). To control for effects caused by referees, we include *Referee effects*. This accounts for the fact that some referees might be more lenient than others in awarding fouls, which is likely to affect the match's intensity.

4.2.3 Estimation method

To study the effect of an increase in monitoring on sabotage, we used data from a quasi-natural experiment caused by the introduction of the VAR in professional soccer. At the beginning of the season 2017/18, the VAR was introduced in the 1. Bundesliga, but not in the 2. Bundesliga.[11] Due to their similarity in terms of the *Laws of the Game*, geography, weather conditions, number of teams, season

[10] For instance, a match with the probabilities for a win of the home team, a draw, and a win of the away team of 60%, 30%, and 10% should be more equal than a match with the probabilities of 70%, 10%, and 20%.

[11] At the beginning of the season 2019/20, so after the end of our observation period, the VAR was also introduced in the 2. Bundesliga.

schedule, and managing body, the 2. Bundesliga serves as a suitable control group for the 1. Bundesliga. The main differences between the two leagues are the skill-level of the players, the level of scrutiny of the public, and the teams' financial resources.[12] However, as argued below, the introduction of the VAR is unrelated to the number of substitutions that are due to an in-match injury, and therefore these differences do not matter. Hence, by using the 2. Bundesliga as our control group for the 1. Bundesliga, we can study the effect of the introduction of the VAR (i. e., an increase in monitoring) on the number of substitutions that are due to an in-match injury (i. e., sabotage) with a difference-in-differences approach. To this end, we used ordinary least-squares (OLS) estimations and included dummies for the seasons (*Season 2015/16*, *Season 2017/18*, and *Season 2018/19*), with the season 2016/17 (the last season prior to the introduction of the VAR) as the reference period. Furthermore, we included the dummy *1. Bundesliga*, which takes the value 1 if the fixture took place in the 1. Bundesliga, and 0 otherwise. Since in our observation period, the VAR was only introduced in the 1. Bundesliga, but never in the 2. Bundesliga, this dummy indicates the treatment group. By also interacting these dummy variables, we can study the effect of the introduction of the VAR over time.

Using a difference-in-differences approach rests on the following assumptions. First, the intervention has to be unrelated to the outcome. That is, the introduction of the VAR has to be independent of the number of in-match injuries in the two leagues. In the season prior to the introduction of the VAR (2016/17), there was no statistically significant difference in the number of substitutions that are due to an in-match injury between the 1. Bundesliga and 2. Bundesliga (two-sided Mann-Whitney U test: $p=0.170$).[13] This makes it plausible that the introduction of the VAR in the 1. Bundesliga, but not in the 2. Bundesliga is unrelated to the number of substitutions that are due to an in-match injury. Second, the composition of the

[12] To check whether there is a difference in the appraisal between the 1. Bundesliga and the 2. Bundesliga, we collected all match reports from Kicker. Kicker is Germany's leading sports magazine, focused primarily on soccer. A comparison of the polarity scores (polarity lies between [-1, 1], with -1 defining a negative sentiment and 1 defining a positive sentiment) of match reports for the season 2015/16 and 2016/17 between the 1. Bundesliga and 2. Bundesliga shows that they significantly differ, with an average score of 0.076 for the 1. Bundesliga and 0.062 for the 2. Bundesliga (two-sided Mann-Whitney U test: $p<0.001$). This suggests that there is more negative sentiment towards matches in the 2. Bundesliga than for matches in the 1. Bundesliga.

[13] Additionally, there is also no statistically significant difference in the number of substitutions that are due to an in-match injury between the 1. Bundesliga and 2. Bundesliga in the season 2015/16 (two-sided Mann-Whitney U test: $p=0.348$).

treatment and control group is stable over the studied time period. That is, there are no fixtures that occur in the 1. Bundesliga and the 2. Bundesliga. Since there are promotion and relegation in the Bundesliga, we excluded all fixtures, which during our observation period appeared in the 1. Bundesliga as well as in the 2. Bundesliga from our analyses. This reduced our dataset by 96 observations, from 2,448 observations to 2,352 observations. Third, there are no spillover effects. That is, the treatment status of any unit must not affect the outcomes of any other unit. This assumption is likely to be met since there are hardly any reasons why the availability of the VAR in one fixture should affect the number of in-match injuries in other fixtures. Fourth, the treatment and control group must have common trends in the outcome. Therefore, the trend in the number of substitutions that are due to an in-match injury must be the same for the 1. Bundesliga and the 2. Bundesliga. The insignificant coefficients of *Season 2015/16* $\times$ *1. Bundesliga* in columns (1), (2), and (3) of Table 4.2 show that this assumption is met.

4.3 Results and discussion

In this section, we empirically test the deterrence hypothesis and discuss potential reasons for our result.

4.3.1 Results

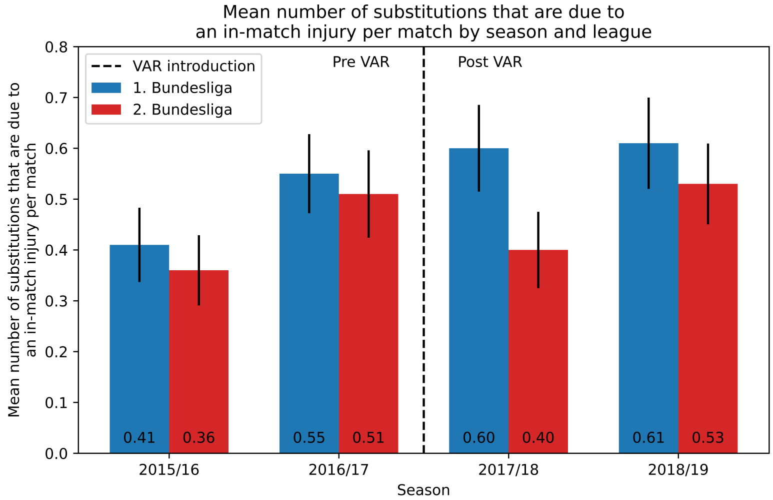


Figure 4.1: Mean number of substitutions that are due to an in-match injury per match by season and league.

Notes: Results are based on all 2,448 fixtures that took place between the season 2015/16 and 2018/19 in the 1. and 2. Bundesliga. The error bars indicate 95% confidence intervals.

Figure 4.1 illustrates the development of the mean number of substitutions that are due to an in-match injury per match from the season 2015/16 to 2018/19 for the 1. and 2. Bundesliga.[14] The number of substitutions that are due to an in-match injury follows a common trend in the 1. Bundesliga and 2. Bundesliga before the introduction of the VAR. However, after the introduction of the VAR, we see a relative increase in the mean number of substitutions that are due to an in-match injury per match in the 1. Bundesliga, relative to the 2. Bundesliga. In contrast to this observation, the deterrence hypothesis predicts that an increase in monitoring leads to a relative decrease of the mean number of substitutions that are due to an in-match injury between the 1. Bundesliga and 2. Bundesliga. In the following, we attempt to shed some light on this surprising observation.

[14]See Section 4.5.1 for a breakdown on the leg-level.

In the following, we use OLS difference-in-differences estimations to study the effect of the introduction of the VAR on our dependent variables *Injury* and *Injuries*. Throughout our analysis, we clustered standard errors at the fixture level.[15]

Table 4.1: OLS difference-in-differences: Dependent variable is *Injury*.

	(1) Injury	(2) Injury	(3) Injury
Season 2015/16	-0.0816**	-0.0818**	-0.0814**
	(0.037)	(0.042)	(0.045)
Season 2017/18	-0.0663*	-0.0668*	-0.0651*
	(0.088)	(0.089)	(0.094)
Season 2018/19	0.0408	0.0360	0.0392
	(0.301)	(0.368)	(0.342)
1. Bundesliga	0.0574	0.0446	0.0613
	(0.158)	(0.305)	(0.183)
Season 2015/16 × 1. Bundesliga	-0.0280	-0.0308	-0.0293
	(0.616)	(0.585)	(0.608)
Season 2017/18 × 1. Bundesliga	0.0814	0.0848	0.0751
	(0.149)	(0.137)	(0.194)
Season 2018/19 × 1. Bundesliga	-0.0215	-0.0194	-0.0246
	(0.707)	(0.735)	(0.681)
Heterogeneity		0.141*	0.128*
		(0.053)	(0.082)
Attendance		-0.00501	0.00709
		(0.824)	(0.761)
Second leg		-0.0327	-0.0308
		(0.120)	(0.142)
Derby		0.0114	0.0221
		(0.871)	(0.755)
Constant	0.378***	0.426*	0.296
	(0.000)	(0.054)	(0.197)
Referee effects	No	No	Yes
Adj. R-Squared	0.0127	0.0137	0.0135
N	2352	2329	2329

Notes: p-values in parentheses. Standard errors clustered at fixture level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Injury* is a dummy variable that takes the value 1 if, during the match, any substitution due to an in-match injury took place, and 0 otherwise.

Table 4.1 shows how the introduction of the VAR affects the probability that a substitution that is due to an in-match injury takes place during a match. The insignificant coefficients of *Season 2017/18 × 1. Bundesliga* and *Season 2018/19 × 1. Bundesliga* in column (1), (2), and (3) suggest that the introduction of the VAR

[15] A fixture is a pairing of two teams in a specific order. Hence, FC Bayern München vs. Borussia Dortmund is another fixture than Borussia Dortmund vs. FC Bayern München.

does not affect the probability that a substitution that is due to an in-match injury takes place during a match.

Table 4.2: OLS difference-in-differences: Dependent variable is *Injuries*.

	(1) Injuries	(2) Injuries	(3) Injuries
Season 2015/16	-0.0819**	-0.0807**	-0.0769**
	(0.012)	(0.016)	(0.023)
Season 2017/18	-0.0608*	-0.0608*	-0.0589*
	(0.065)	(0.068)	(0.076)
Season 2018/19	0.0274	0.0244	0.0291
	(0.417)	(0.475)	(0.411)
1. Bundesliga	0.0320	0.0197	0.0366
	(0.354)	(0.596)	(0.355)
Season 2015/16 × 1. Bundesliga	-0.00495	-0.00841	-0.00907
	(0.914)	(0.857)	(0.848)
Season 2017/18 × 1. Bundesliga	0.0848*	0.0862*	0.0749
	(0.073)	(0.072)	(0.125)
Season 2018/19 × 1. Bundesliga	0.0145	0.0152	0.00392
	(0.771)	(0.761)	(0.940)
Heterogeneity		0.0932	0.0816
		(0.134)	(0.196)
Attendance		0.00174	0.0124
		(0.928)	(0.531)
Second leg		-0.0177	-0.0161
		(0.319)	(0.364)
Derby		-0.00554	-0.00127
		(0.922)	(0.982)
Constant	0.311***	0.292	0.177
	(0.000)	(0.123)	(0.364)
Referee effects	No	No	Yes
Adj. R-Squared	0.0144	0.0139	0.00992
N	2352	2329	2329

Notes: p-values in parentheses. Standard errors clustered at fixture level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Injuries* gives the $\ln(y+1)$ transformation of the number of substitutions that are due to an in-match injury (given by y).

Table 4.2 shows how the introduction of the VAR affects the number of substitutions that are due to an in-match injury. The positive and significant coefficients of *Season 2017/18 × 1. Bundesliga* in columns (1) and (2) suggest that the introduction of the VAR increased the number of substitutions that are due to an in-match injury. However, the respective coefficient in column (3), which also controls for *Referee effects*, is no longer significant. Furthermore, the insignificant coefficients of *Season 2018/19 × 1. Bundesliga* in column (1), (2), and (3) show that there was no effect on the number of substitutions that are due to an in-match injury in the second season after the introduction of the VAR.

Overall, our results do not support the deterrence hypothesis. That is, we do not find empirical evidence that an increase in monitoring reduces sabotage.

4.3.2 Discussion

There are a plethora of reasons why the introduction of the VAR did not reduce the probability that a substitution due to an in-match injury takes place during a match, respectively the number of substitutions that are due to an in-match injury. In the following, we discuss possible reasons for this contradictory finding.

First, players may not take changes in monitoring into account when deciding whether to commit sabotage or not. If players act in the heat of the moment with very little reaction time, they might not make deliberated decisions by weighing all pros and cons of committing sabotage. Evidence against this reason comes from McCormick and Tollison (1984) who studied the effect of an increase in the number of referees in college basketball, from two to three. The authors found that the associated increase in the detection probability of fouls led to a reduction of fouls by 34%. Since basketball and soccer are both equally fast-paced games, players' decision-making in the two sports should be comparable. Therefore, it is very likely that soccer players also take changes in the detection probability into account.

Second, the increase in monitoring may be too small to be taken into account by players when deciding whether to commit sabotage or not. As a study commissioned by the The International Football Association Board (2018a) shows, the introduction of the VAR increases the decision accuracy for reviewable categories from 93.0% to 98.9%. While this 5.9% increase might not appear as much, the media attention which the introduction of the VAR caused makes it unlikely that affected players completely ignored it in their decision-making.[16]

Third, players may not immediately adjust their sabotage effort in response to the rule change. This reason is supported by Garicano and Palacios-Huerta (2005), who assume that soccer players do not immediately react to a rule change, which affects the incentive scheme to win the match. In their study, the authors analyzed the effect of a rule change from the 2-1-0 points scheme to the 3-1-0 points scheme in

[16] A Google search for the term "Videobeweis Fußball" returns about 51,900 news articles related to this topic.

the Spanish soccer league. To account for non-immediate behavioral responses, the authors allowed for a four-year adjustment period by using data from the seasons 1994/95 (the last season with the 2-1-0 points scheme) and 1998/99 (with the new 3-1-0 points scheme). Due to the irregularities in the season 2019/20 caused by the outbreak of the Coronavirus, we cannot address this concern by including it in our analysis. However, the two-season post-intervention trend suggests that "the effect" rather faded out over time than becoming stronger. Hence, it is unlikely that this reason is the cause of our result.

Fourth, an increase in monitoring may actually reduce sabotage, but our proxy for sabotage might be too noisy. For instance, there could be an effect for minor but not for major acts of sabotage. In contrast to the existing proxies of sabotage, our proxy requires a certain severity of the foul so that a substitution due to an in-match injury becomes necessary. Hence, we are not able to study changes in minor acts of sabotage with our proxy. Evidence for a difference between minor and major acts of sabotage comes from Allen (2016) who found an increase in violent penalties in the NHL when a second referee is deployed. However, the fact that the VAR can only intervene in relevant categories, which includes major fouls punished by red cards, makes this reason unlikely.

Fifth, due to the existence of the home bias, the introduction of the VAR may lead to two counteractive effects for home and away teams, which offset each other. The home bias can come in two ways, either favoritism of the home team or discrimination of the away team. Favoritism of the home team exists when the referee makes decisions for the home team that fall into the category of false negatives. That is, referees do not punish players even though it would have been the correct decision. Discrimination of away teams exists when the referee makes decisions for the away team that fall into the category of false positives. That is, referees punish players even though it is an incorrect decision. With the introduction of the VAR and the associated increase in the decision accuracy, the wiggle room of referees decreased. This could have led to a reduction of the home bias. Evidence that an increase in monitoring reduces the home bias comes from Albanese et al. (2020). The authors show the existence of a referee bias in matches with four referees, the main referee, and three assistant referees. However, this referee bias disappears with the introduction of two additional assistant referees. If the introduction of the VAR also leads to a reduction of the home bias, this would affect home and away teams differently. For home teams, the reduction of false negatives leads to a higher chance of punishment for sabotage, which, ceteris paribus, disincentivizes sabotage. In contrast, for away teams the reduction of false positives leads to less overall punishment, which, ceteris paribus, incentivizes sabotage. Hence, the introduction of the VAR could lead to a reduction of sabotage of home teams and an increase of sabotage of away teams. To study whether this is the case, we analyzed how the introduction of the VAR affects the number of substitutions that are due to an in-match injury of home teams (*Injuries (home)*) and away teams (*Injuries (away)*) separately.[17]

[17] Although we are not able to directly prove the existence of a home bias in our data, the positive mean (=0.132) of the differences in the implicit winning probabilities of home and away teams ($P_H - P_A$) indicates the existence of a home advantage in our data (Deutscher et al., 2013). As shown by Boyko et al. (2007), the home advantage is partially driven by the home bias of referees. Furthermore, Sutter and Kocher (2004) prove the existence of a home bias in the 1. Bundesliga for the season 2000/1 by showing that away teams are refused a regular penalty significantly more often.

Table 4.3: OLS difference-in-differences: Dependent variable is *Injury (home)/Injury (away)*.

	(1) Injury (home)	(2) Injury (home)	(3) Injury (away)	(4) Injury (away)
Season 2015/16	-0.0782**	-0.0712**	-0.0476	-0.0463
	(0.017)	(0.038)	(0.129)	(0.155)
Season 2017/18	-0.0525	-0.0509	-0.0455	-0.0439
	(0.117)	(0.145)	(0.147)	(0.160)
Season 2018/19	-0.00680	0.00240	0.0272	0.0220
	(0.841)	(0.947)	(0.437)	(0.541)
1. Bundesliga	0.0256	0.0270	0.0323	0.0385
	(0.472)	(0.494)	(0.356)	(0.338)
Season 2015/16 × 1. Bundesliga	0.00974	0.00157	-0.0346	-0.0319
	(0.838)	(0.974)	(0.442)	(0.488)
Season 2017/18 × 1. Bundesliga	0.0590	0.0504	0.0488	0.0412
	(0.227)	(0.322)	(0.293)	(0.386)
Season 2018/19 × 1. Bundesliga	0.0422	0.0240	-0.0307	-0.0312
	(0.410)	(0.655)	(0.547)	(0.555)
Heterogeneity		0.0365		0.0784
		(0.544)		(0.229)
Attendance		-0.00133		0.0192
		(0.946)		(0.340)
Second leg		-0.0295*		0.00630
		(0.092)		(0.716)
Derby		-0.00818		-0.00483
		(0.879)		(0.934)
Constant	0.231***	0.253	0.218***	0.00330
	(0.000)	(0.192)	(0.000)	(0.987)
Referee effects	No	Yes	No	Yes
Adj. R-Squared	0.00829	0.00721	0.00495	0.00497
N	2352	2329	2352	2329

Notes: p-values in parentheses. Standard errors clustered at fixture level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Injury (home)/Injury (away)* is a dummy variable that takes the value 1 if, during the match, any substitution due to an in-match injury took place for home/away teams, and 0 otherwise.

Table 4.3 shows how the introduction of the VAR affects the probability that a substitution that is due to an in-match injury takes place during a match for home and away teams. The insignificant coefficients of *Season 2017/18 × 1. Bundesliga* and *Season 2018/19 × 1. Bundesliga* in column (1) and (2) suggest that the introduction of the VAR does not affect the probability for home teams that a substitution that is due to an in-match injury takes place during a match. We interpret this as empirical evidence that an increase in monitoring does not affect the sabotage of away teams. Likewise, the respective insignificant coefficients in columns (3) and (4) show the same result for away teams. Similarly, we conclude that an increase in monitoring does not affect sabotage of home teams.

Table 4.4: OLS difference-in-differences: Dependent variable is *Injuries (home)/Injuries (away)*.

	(1) Injuries (home)	(2) Injuries (home)	(3) Injuries (away)	(4) Injuries (away)
Season 2015/16	-0.0649**	-0.0580**	-0.0330	-0.0320
	(0.011)	(0.029)	(0.151)	(0.177)
Season 2017/18	-0.0414	-0.0398	-0.0304	-0.0292
	(0.116)	(0.144)	(0.192)	(0.209)
Season 2018/19	-0.00529	0.00305	0.0244	0.0198
	(0.843)	(0.915)	(0.354)	(0.465)
1. Bundesliga	0.00614	0.00934	0.0234	0.0256
	(0.819)	(0.757)	(0.364)	(0.385)
Season 2015/16 × 1. Bundesliga	0.0174	0.00962	-0.0180	-0.0157
	(0.628)	(0.793)	(0.592)	(0.649)
Season 2017/18 × 1. Bundesliga	0.0512	0.0430	0.0450	0.0395
	(0.167)	(0.263)	(0.201)	(0.272)
Season 2018/19 × 1. Bundesliga	0.0453	0.0286	-0.0170	-0.0161
	(0.253)	(0.491)	(0.661)	(0.690)
Heterogeneity		0.0179		0.0548
		(0.694)		(0.279)
Attendance		0.000472		0.0162
		(0.974)		(0.289)
Second leg		-0.0198		0.00720
		(0.141)		(0.589)
Derby		0.00904		-0.0231
		(0.835)		(0.575)
Constant	0.177***	0.178	0.158***	-0.0205
	(0.000)	(0.224)	(0.000)	(0.891)
Referee effects	No	Yes	No	Yes
Adj. R-Squared	0.00830	0.00679	0.00495	0.00453
N	2352	2329	2352	2329

Notes: p-values in parentheses. Standard errors clustered at fixture level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Injuries (home)/Injuries (away)* gives the $\ln(y+1)$ transformation of the number of substitutions that are due to an in-match injury for home/away teams (given by y).

Table 4.4 shows how the introduction of the VAR affects the number of substitutions that are due to an in-match injury for home and away teams. The insignificant coefficients of *Season 2017/18 × 1. Bundesliga* and *Season 2018/19 × 1. Bundesliga* in column (1), (2), (3), and (4) corroborate our finding that an increase in monitoring does not affect sabotage of home and away teams. Hence, we rule out that two counteractive effects for home and away teams are the cause of our result.

Sixth, the increase in monitoring may change the players' perception of the situation in which they are involved, which causes an increase in their sabotage effort that offsets the deterrence effect. Similar to the effect reported by Gneezy and Rustichini (2000), an increase in monitoring could change the players' perception of the situation in which they are involved. Before the introduction of the VAR, not all sabotage got detected. Hence, the social norm dictated that it should be exerted with restrain. However, after the introduction of the VAR, sabotage gets detected with almost certainty.[18] This implies that players will almost certainly get punished for it. Hence, the players' perceptions might shift from a situation that is governed by a social norm ("do not sabotage") to a market setting in which players sabotage as much as they find convenient given the consequences. Since the benefit of sabotage sometimes outweighs its associated costs, this perceptual change might increase sabotage. Overall, this effect might offset the deterrence effect, leading to an insignificant overall effect of an increase in monitoring on sabotage. Due to the lack of data on players' perceptions of the situation in which they are involved, we cannot rule out this reason as the cause for our result.

4.4 Conclusion

In this study, we provide further empirical evidence on the deterrence hypothesis (Becker, 1968). It predicts that an increase in monitoring reduces sabotage. To test this conjecture, we used data from German professional soccer. Our data comprises information on all matches of the 1. and 2. Bundesliga for the seasons 2015/16 to 2018/19. To measure sabotage, we introduced a novel proxy, namely substitutions that are due to an in-match injury. Furthermore, the introduction of the VAR allowed us to study the effect of an increase in monitoring. At the beginning of the season 2017/18, the VAR was introduced in the 1. Bundesliga, but not in the

[18] With the VAR, the decision accuracy for reviewable categories is 98.9% (The International Football Association Board, 2018a).

2. Bundesliga. This created a quasi-natural experiment that allowed us to study the effect of the VAR on the number of substitutions that are due to an in-match injury in a difference-in-differences approach between the 1. Bundesliga and the 2. Bundesliga.

Overall, we do not find empirical evidence to support the deterrence hypothesis. While it predicts that an increase in monitoring reduces sabotage, we do not find empirical evidence for this conjecture in our data. More precisely, we find that the introduction of the VAR does not reduce the probability that a substitution that is due to an in-match injury takes place during a match, nor does it reduce the number of substitutions that are due to an in-match injury.

We discuss potential causes for this result. This includes the possibility of separate effects for home and away teams, with the two effects offsetting each other. However, a separate analysis of the effect for home and away teams does not provide evidence for this conjecture. Due to the lack of data, we cannot rule out the conjecture that the increase in monitoring changes the players' perceptions of the situation in which they are involved, which causes an increase in their sabotage effort that offsets the deterrence effect. In particular, we conjecture that similar to the effect reported by Gneezy and Rustichini (2000), an increase in monitoring could change the players' perceptions from a situation that is governed by the social norm "do not sabotage" to a market setting in which players sabotage as much as they find convenient given the consequences. This possibility calls for additional research to better understand how players' perceptions of the situation in which they are involved affect their sabotage effort.

4.5 Appendix

4.5.1 Mean number of substitutions that are due to an in-match injury per match by leg of season and league

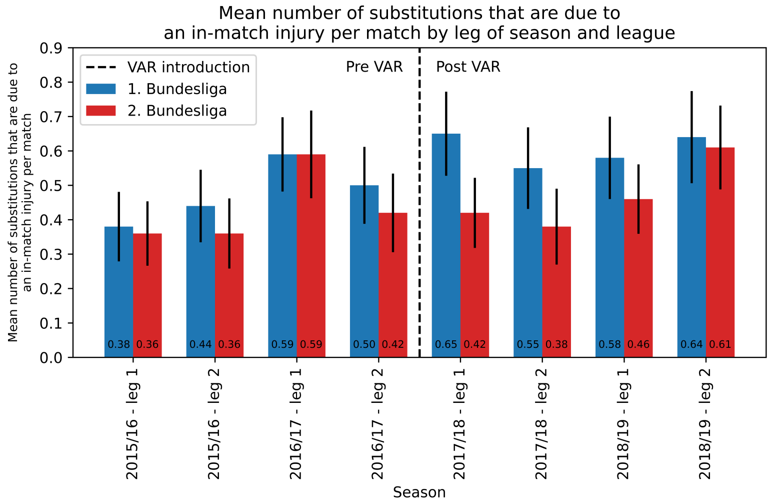


Figure 4.2: Mean number of substitutions that are due to an in-match injury per match by leg of season and league.

Notes: Results are based on all 2,448 fixtures that took place between the season 2015/16 and 2018/19 in the 1. and 2. Bundesliga. The error bars indicate 95% confidence intervals.

Figure 4.2 illustrates the development of the mean number of substitutions that are due to an in-match injury per match from the first leg of the season 2015/16 to the second leg of the season 2018/19 for the 1. Bundesliga and 2. Bundesliga. While there is no statistically significant difference in the number of substitutions that are due to an in-match injury between the four legs before the introduction of the VAR (two-sided Mann-Whitney U tests: p=0.976, p=0.202, p=0.551, p=0.179 (from left to right)), there is a statistically significant difference in the first two legs after the introduction of the VAR (two-sided Mann-Whitney U tests: p=0.006, p=0.013, (from left to right)).

4.5.2 Distribution of substitutions that are due to an in-match injury

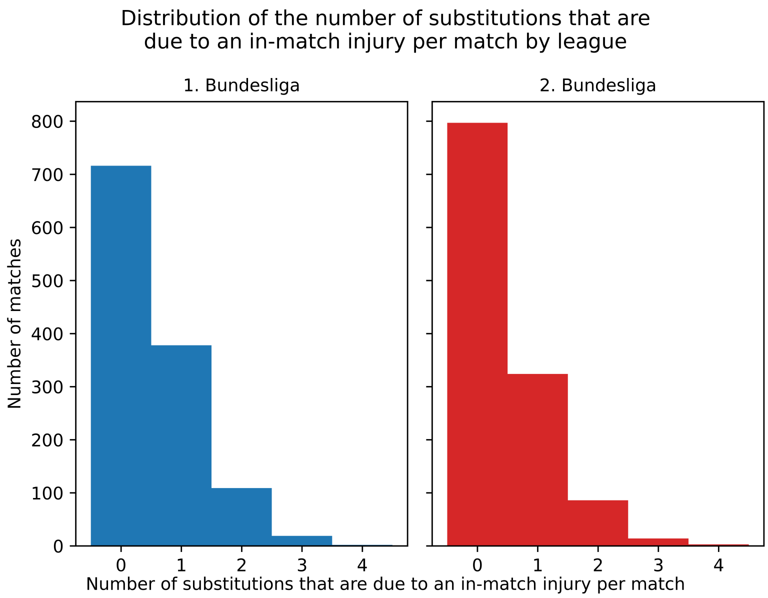


Figure 4.3: Distribution of substitutions that are due to an in-match injury per match by league.

Notes: Results are based on all 2,448 fixtures that took place between the season 2015/16 and 2018/19 in the 1. and 2. Bundesliga.

Figure 4.3 shows the distribution of substitutions that are due to an in-match injury per match for the 1. Bundesliga and 2. Bundesliga. Overall, the two distributions are very similar, with the majority of matches having no substitutions that are due to an in-match injury. The maximum number of substitutions that are due to an in-match injury is four per match.

4.5.3 List of derbies

The following list shows all the fixtures (we omitted the reversed order) that we considered as a derby. The list is based on www.derbys.org, which provides information on all derbies in German professional soccer.

- FC Bayern München vs. Borussia Dortmund
- Eintracht Frankfurt vs. 1. FC Kaiserslautern
- Hertha BSC vs. 1. FC Union Berlin
- 1. FC Heidenheim vs. VfR Aalen

- Fortuna Düsseldorf vs. Borussia Mönchengladbach
- FC Bayern München vs. 1. FC Nürnberg
- Eintracht Frankfurt vs. 1. FSV Mainz 05
- Bayer 04 Leverkusen vs. 1. FC Köln
- Fortuna Düsseldorf vs. 1. FC Köln
- Hamburger SV vs. SV Werder Bremen
- Hamburger SV vs. FC St. Pauli
- Karlsruher SC vs. VfB Stuttgart
- SpVgg Greuther Fürth vs. 1. FC Nürnberg
- TSV 1860 München vs. FC Bayern München
- Borussia Dortmund vs. FC Schalke 04
- Borussia Mönchengladbach vs. 1. FC Köln

5. Conclusion

5.1 Summary

Managerial decision-making is oftentimes quite complex. This complexity arises from human behavior, such as faking or sabotage, which causes allegedly optimal decisions to not result in the desired outcome. In this dissertation, I study three important situations that are frequently faced by managerial-decision makers in which the literature does not yet provide sufficient answers on how to deal with these challenges.

In Chapter 2, I study a situation in which managerial decision-makers try to identify the most suitable candidate among all applicants for an advertised job vacancy. To better understand this situation, I used data from an online experiment to study the question of how incentives to fake affect the predictive power of personality assessments. The results of this study show that in the absence of incentives to fake, classifiers that make predictions on subjects' cooperativeness based on personality scores fail to make significantly better than chance predictions. In contrast, classifiers that make predictions on subjects' cooperativeness based on linguistic scores achieve significantly better than chance predictions. Furthermore, the results show that, in the presence of incentives to fake, all classifiers fail to make significantly better than chance predictions, independently of the approach.

In Chapter 3, I shed light on a situation in which managerial decision-makers try to make applicants optimal wage offers. The difficulty of this task is to correctly estimate the applicants' reservation wages. Theory suggests that prosociality is negatively associated with reservation wages and therefore serves as a good proxy. To analyze whether this is the case, I used data from a field and laboratory experiment to study the question of whether employers wage-discriminate against applicants based on their signaled prosociality. The results of this study do not provide empirical evidence to support the hypothesis that the higher the signaled prosociality of applicants, the lower their wage offers. In the field experiment, the hypothetical wage offers by HR managers are not affected by the prosociality of the fictitious applicants, signaled by the work experiences on their résumés. Furthermore, I find that estimated reservation wages by HR managers are not affected by the prosociality of the fictitious applicants. Likewise, in the laboratory experiment, proposers' offers in the ultimatum game are not affected by responders' donation amounts in the strategic dictator game. In line with this finding, I also do not find that responders' reservation wages, measured by their minimum acceptance thresholds in the ultimatum game, are affected by their prosociality. Overall, these results provide empirical evidence that signaling prosociality does not backfire financially by leading to lower wage offers.

In Chapter 4, I research a situation in which managerial decision-makers try to prevent sabotage among employees. As suggested by the deterrence hypothesis, an increase in monitoring reduces sabotage. To test the deterrence hypothesis, I used publicly available data from professional soccer to study the question of whether an increase in monitoring reduces sabotage. Overall, the results of this study cast doubt on the deterrence hypothesis. In contrast to its prediction, the results do not provide empirical evidence that an increase in monitoring reduces sabotage. That is, the introduction of the VAR does not reduce the probability that a substitution that is due to an in-match injury takes place during a match, nor does it reduce the number of substitutions that are due to an in-match injury. This holds for both home and away teams.

The results of this dissertation help to better understand how the "human factor" affects the outcomes of managerial decision-making. Furthermore, they call for additional research to fully understand how to optimally account for these complexities in the decision-making process.

5.2 Avenues for future research

From the studies of this dissertation originate different avenues for future research. In particular, it might be fruitful for future studies to address the following questions.

The results of Chapter 2 show the promising nature of linguistic scores as a predictor for personality traits in the presence of faking. Therefore, future studies could refine this approach so that it becomes more successful. For instance, better results might be obtained by using spoken self-descriptions instead of written self-descriptions. This is because written responses allow for sufficient time to reflect, whereas spoken responses require immediate action, which yields spontaneous, undisguised responses. Hence, assessing applicants' personality traits based on transcribed spoken self-descriptions might yield better results. Better results could also be achieved by using tailored features for the prediction. The literature shows that linguistic characteristics are dependent on age and gender (Newman et al., 2008). [1] Hence, using linguistic features that are tailored to demographic characteristics might prove to yield even better results. A further improvement in the predictability of linguistic features could be obtained by extending this approach to other personality traits than cooperativeness which are more strongly correlated with linguistic features. For instance, the literature documents a strong link between extraversion and linguistic expression. In particular, in comparison to introverts, extroverts use more social words, words expressing positive emotions, and words that indicate external focus (Mairesse et al., 2007). Lastly, when it comes to data, it holds for machine learning methods that "more is better", so the results might already improve by simply enlarging the sample size. Thus, by accounting for these deficiencies of the first study, the prediction results of machine learning classifiers based on linguistic scores might improve considerably.

The results of Chapter 3 are in contrast to other studies which document a positive relationship between prosociality and wages (Cozzi et al., 2017; Sauer, 2015; Hackl et al., 2007; Prouteau and Wolff, 2006; Day and Devlin, 1998, 1997). These studies focus on volunteering as a proxy for prosociality. Therefore, they only provide estimates for the overall effect of prosociality on wages. Day and Devlin (1997) list three possible channels through which volunteering may affect wages. First, the human capital hypothesis suggests that applicants who do voluntary work may acquire additional skills and experience that make them more productive. Second, the screening hypothesis suggests that voluntary work may provide a signal to employers of an otherwise unobservable ability. Third, the networking hypothesis suggests that voluntary work may provide access to informal networks of contacts that may be useful when searching for employment opportunities. This calls for additional research to make a judgment about whether prosociality itself has an effect on wages and not the skills and competencies which are acquired through acts of prosociality (e. g., volunteering).

[1] See Pennebaker et al. (2003) for an overview.

The results of Chapter 4 call for additional research to better understand how situational characteristics affect the effectiveness of deterrence implemented through an increase in monitoring. In particular, it remains unclear whether an increase in monitoring can shift the affected individuals' perceptions from the social norm of "do not sabotage" to a market frame in which they sabotage as much as they find convenient given the consequences (Gneezy and Rustichini, 2000). Gaining a better understanding of such adverse effects will help to increase the effectiveness of deterrence as a means to reduce sabotage in contests.

The outlined avenues for future research show that it is still a long way until we fully understand all the complexities caused by the "human factor" in managerial decision-making. However, future studies will add new tesserae to the mosaic of research in this field so that it will eventually provide a clear picture and description of optimal managerial decision-making.

6. Bibliography

P. L. Ackerman and E. D. Heggestad. Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121(2):219–245, 1997. ISSN 0033-2909. doi: 10.1037/0033-2909.121.2.219.

A. Albanese, S. Baert, and O. Verstraeten. Twelve eyes see more than eight. Referee bias and the introduction of additional assistant referees in soccer. *PLOS One*, 15(2):e0227758, 2020. doi: 10.1371/journal.pone.0227758.

W. D. Allen. Crime, punishment, and recidivism. *Journal of Sports Economics*, 3 (1):39–60, 2016. ISSN 1527-0025. doi: 10.1177/1527002502003001004.

D. Ariely, E. Kamenica, and D. Prelec. Man's search for meaning: The case of Legos. *Journal of Economic Behavior & Organization*, 67(3-4):671–677, 2008. ISSN 01672681. doi: 10.1016/j.jebo.2008.01.004.

J. L. Baddeley and J. A. Singer. Telling losses: Personality correlates and functions of bereavement narratives. *Journal of Research in Personality*, 42(2):421–438, 2008. ISSN 00926566. doi: 10.1016/j.jrp.2007.07.006.

S. Baert and S. Vujić. Does it pay to care? Volunteering and employment opportunities. *Journal of Population Economics*, 31(3):819–836, 2018. ISSN 0933-1433. doi: 10.1007/s00148-017-0682-8.

L. Balafoutas, F. Lindner, and M. Sutter. Sabotage in tournaments: Evidence from a natural experiment. *Kyklos*, 65(4):425–441, 2012. ISSN 00235962. doi: 10.1111/kykl.12000.

M. R. Barrick, M. K. Mount, and R. Gupta. Meta-analysis of the relationship between the five-factor model of personality and Holland's occupational types. *Per-*

sonnel Psychology, 56(1):45–74, 2003. ISSN 0031-5826. doi: 10.1111/j.1744-6570.2003.tb00143.x.

B. Bartling, L. Brandes, and D. Schunk. Expectations as reference points: Field evidence from professional soccer. *Management Science*, 61(11):2646–2661, 2015. ISSN 0025-1909. doi: 10.1287/mnsc.2014.2048.

M. H. Bazerman and D. A. Moore. *Judgment in managerial decision making*. Wiley, Hoboken, NJ, 7 edition, 2009. ISBN 9780470049457. URL `http://www.loc.gov/catdir/enhancements/fy0814/2008008490-d.html`.

A. Becker, T. Deckers, T. Dohmen, A. Falk, and F. Kosse. The relationship between economic preferences and psychological personality measures. *Annual Review of Economics*, 4(1):453–478, 2012. ISSN 1941-1383. doi: 10.1146/annurev-economics-080511-110922.

G. S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217, 1968. ISSN 0022-3808. doi: 10.1086/259394.

J. Berger and P. Nieken. Heterogeneous contestants and the intensity of tournaments. *Journal of Sports Economics*, 17(7):631–660, 2016. ISSN 1527-0025. doi: 10.1177/1527002514538639.

S. G. Bishu and M. G. Alkadry. A systematic review of the gender pay gap and factors that predict it. *Administration & Society*, 49(1):65–104, 2017. ISSN 0095-3997. doi: 10.1177/0095399716636928.

G. D. Bond and A. Y. Lee. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3):313–329, 2005. ISSN 0888-4080. doi: 10.1002/acp.1087.

R. L. Boyd and J. W. Pennebaker. Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18:63–68, 2017. ISSN 23521546. doi: 10.1016/j.cobeha.2017.07.017.

R. H. Boyko, A. R. Boyko, and M. G. Boyko. Referee bias contributes to home advantage in English Premiership football. *Journal of Sports Sciences*, 25(11):1185–1194, 2007. doi: 10.1080/02640410601038576.

B. Buraimo, D. Forrest, and R. Simmons. The 12th man? Refereeing bias in English and German soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):431–449, 2010. ISSN 09641998. doi: 10.1111/j.1467-985X.2009.00604.x.

B. Buraimo, R. Simmons, and M. Maciaszczyk. Favoritism and referee bias in European soccer: Evidence from the Spanish League and the UEFA Champions League. *Contemporary Economic Policy*, 30(3):329–343, 2012. ISSN 10743529. doi: 10.1111/j.1465-7287.2011.00295.x.

V. C. Burbano. Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces. *Organization Science*, 27 (4):1010–1028, 2016. ISSN 1047-7039. doi: 10.1287/orsc.2016.1066.

G. Carlo, M. A. Okun, G. P. Knight, and M. R. T. de Guzman. The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences*, 38(6):1293–1305, 2005. ISSN 01918869. doi: 10.1016/j.paid.2004.08.012.

J. P. Carpenter and E. Gong. Motivating agents: How much does the mission matter? *Journal of Labor Economics*, 34(1):211–236, 2016.

A. Caspi, B. W. Roberts, and R. L. Shiner. Personality development: Stability and change. *Annual Review of Psychology*, 56:453–484, 2005. ISSN 0066-4308. doi: 10.1146/annurev.psych.55.090902.141913.

L. Cassar. Job mission as a substitute for monetary incentives: Benefits and limits. *Management Science*, 65(2):896–912, 2019. ISSN 0025-1909. doi: 10.1287/mnsc.2017.2903.

L. Cassar and S. Meier. Intentions for doing good matter for doing well: The (negative) signaling value of prosocial incentives. Technical report, National Bureau of Economic Research, 2017.

L. Cassar and S. Meier. Nonmonetary incentives and the implications of work as a source of meaning. *Journal of Economic Perspectives*, 32(3):215–238, 2018. ISSN 0895-3309. doi: 10.1257/jep.32.3.215.

D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013. ISSN 01672681. doi: 10.1016/j.jebo.2013.03.003.

G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014. ISSN 00457906. doi: 10.1016/j.compeleceng.2013.11.024.

D. L. Chen, M. Schonger, and C. Wickens. oTree – An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016. ISSN 22146350. doi: 10.1016/j.jbef.2015.12.001.

D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020. doi: 10.1186/s12864-019-6413-7.

S. M. Chowdhury and O. Gürtler. Sabotage in contests: A survey. *Public Choice*, 164(1-2):135–155, 2015. ISSN 0048-5829. doi: 10.1007/s11127-015-0264-9.

CIRCLE. Youth voting and civic engagement in America, 2018. URL `https://youthdata.circle.tufts.edu/?state=US&datayear=2018&ycevolunteering=1&customize_changeset_uuid`.

S. M. Clancy and S. J. Dollinger. Identity, self, and personality: I. Identity status and the five-factor model of personality. *Journal of Research on Adolescence*, 3 (3):227–245, 1993. ISSN 1050-8392. doi: 10.1207/s15327795jra0303_2.

B. L. Connelly, L. Tihanyi, T. R. Crook, and K. A. Gangloff. Tournament theory: Thirty years of contests and competitions. *Journal of Management*, 40(1):16–47, 2014. ISSN 0149-2063. doi: 10.1177/0149206313498902.

G. Cozzi, N. Mantovan, and R. M. Sauer. Does it pay to work for free? Negative selection and the wage returns to volunteer experience. *Oxford Bulletin of Economics and Statistics*, 79(6):1018–1045, 2017. ISSN 03059049. doi: 10.1111/obes.12183.

P. Dawson, S. Dobson, J. Goddard, and J. Wilson. Are football referees really biased and inconsistent? Evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in*

Society), 170(1):231–250, 2007. ISSN 09641998. doi: 10.1111/j.1467-985X.2006.00451.x.

K. M. Day and R. A. Devlin. Can volunteer work help explain the male-female earnings gap? *Applied Economics*, 29(6):707–721, 1997. ISSN 0003-6846. doi: 10.1080/000368497326642.

K. M. Day and R. A. Devlin. The payoff to work without pay: Volunteer work as an investment in human capital. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 31(5):1179, 1998. ISSN 00084085. doi: 10.2307/136465.

J. del Corral, J. Prieto-Rodríguez, and R. Simmons. The effect of incentives on sabotage: The case of Spanish football. *Journal of Sports Economics*, 11(3): 243–260, 2009. ISSN 1527-0025. doi: 10.1177/1527002509340666.

C. Deutscher and S. Schneemann. The impact of intermediate information on sabotage in tournaments with heterogeneous contestants. *Managerial and Decision Economics*, 38(2):222–237, 2017. ISSN 01436570. doi: 10.1002/mde.2771.

C. Deutscher, B. Frick, O. Gürtler, and J. Prinz. Sabotage in tournaments with heterogeneous contestants: Empirical evidence from the soccer pitch. *The Scandinavian Journal of Economics*, 115(4):1138–1157, 2013. ISSN 03470520. doi: 10.1111/sjoe.12036.

T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197.

J. Dinardo. Natural experiments and quasi-natural experiments. In S. N. Durlauf and L. E. Blume, editors, *Microeconometrics*, pages 139–153. Palgrave Macmillan UK, London, 2010. ISBN 978-0-230-23881-7. doi: 10.1057/9780230280816_18.

T. J. Dohmen. The influence of social forces: Evidence from the behavior of football referees. *Economic Inquiry*, 46(3):411–424, 2008. ISSN 00952583. doi: 10.1111/j.1465-7295.2007.00112.x.

S. A. Dwight and J. J. Donovan. Do warnings not to fake reduce faking? *Human Performance*, 16(1):1–23, 2003. ISSN 0895-9285. doi: 10.1207/S15327043HUP1601_1.

A. Eberl and G. Krug. When and how does volunteering influence wages? – Evidence from panel data. *Acta Sociologica*, 2020. ISSN 0001-6993. doi: 10.1177/0001699320902809.

J. E. Ellingson, P. R. Sackett, and L. M. Hough. Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology*, 84(2):155–166, 1999. ISSN 0021-9010. doi: 10.1037/0021-9010.84.2.155.

L. A. Fast and D. C. Funder. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, 94(2):334–346, 2008. ISSN 1939-1315. doi: 10.1037/0022-3514.94.2.334.

U. Fischbacher, S. Gächter, and E. Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001. ISSN 01651765. doi: 10.1016/S0165-1765(01)00394-9.

B. Frick, O. Gürtler, and J. Prinz. Anreize in Turnieren mit heterogenen Teilnehmern – Eine empirische Untersuchung mit Daten aus der Fußball-Bundesliga. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, 60(4):385–405, 2008. ISSN 0341-2687. doi: 10.1007/BF03372799.

S. Gächter and E. Renner. The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3):364–377, 2010. ISSN 1386-4157. doi: 10.1007/s10683-010-9246-4.

L. Garicano and I. I. Palacios-Huerta. Sabotage in tournaments: Making the beautiful game a bit less beautiful. 2005.

L. Garicano, I. Palacios-Huerta, and C. Prendergast. Favoritism under social pressure. *Review of Economics and Statistics*, 87(2):208–216, 2005. ISSN 0034-6535. doi: 10.1162/0034653053970267.

L. Gerhards. The incentive effects of missions – Evidence from experiments with NGO employees and students. *European Economic Review*, 79:252–262, 2015. ISSN 00142921. doi: 10.1016/j.euroecorev.2015.07.012.

U. Gneezy and A. Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000. ISSN 0047-2530. doi: 10.1086/468061.

L. R. Goldberg. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990. ISSN 1939-1315. doi: 10.1037/0022-3514.59.6.1216.

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003. ISSN 00926566. doi: 10.1016/S0092-6566(03)00046-1.

P. Grimm. Social desirability bias. In J. Sheth and N. Malhotra, editors, *Wiley International Encyclopedia of Marketing*. John Wiley & Sons, Ltd, Chichester, UK, 2010. ISBN 9781405161787. doi: 10.1002/9781444316568.wiem02057.

O. Gürtler and J. Münster. Sabotage in dynamic tournaments. *Journal of Mathematical Economics*, 46(2):179–190, 2010. ISSN 03044068. doi: 10.1016/j.jmateco.2009.11.003.

O. Gürtler and J. Münster. Rational self-sabotage. *Mathematical Social Sciences*, 65(1):1–4, 2013. ISSN 01654896. doi: 10.1016/j.mathsocsci.2012.06.008.

J. Hackinger. DataGorri: A tool for automated data collection of tabular web content. *NETNOMICS: Economic Research and Electronic Networking*, 19(1-2): 31–41, 2018. ISSN 1385-9587. doi: 10.1007/s11066-018-9125-2.

F. Hackl, M. Halla, and G. J. Pruckner. Volunteering and income? The fallacy of the good samaritan? *Kyklos*, 60(1):77–104, 2007. ISSN 00235962. doi: 10.1111/j.1467-6435.2007.00360.x.

J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007. doi: 10.1080/01638530701739181.

C. Harbring, B. Irlenbusch, M. Kräkel, and R. Selten. Sabotage in corporate contests – An experimental analysis. *International Journal of the Economics of Business*, 14(3):367–392, 2007. ISSN 1357-1516. doi: 10.1080/13571510701597445.

V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19(4):307–342, 2015. doi: 10.1177/1088868314556539.

P. C. L. Heaven and S. Bucci. Right-wing authoritarianism, social dominance orientation and personality: An analysis using the IPIP measure. *European Journal of Personality*, 15(1):49–56, 2001. ISSN 0890-2070. doi: 10.1002/per.389.

J. C. Heckelman and A. J. Yates. And a hockey game broke out: Crime and punishment in the NHL. *Economic Inquiry*, 41(4):705–712, 2003. ISSN 00952583. doi: 10.1093/ei/cbg038.

E. D. Heggestad, M. Morrison, C. L. Reeve, and R. A. McCloy. Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91(1):9–24, 2006. ISSN 0021-9010. doi: 10.1037/0021-9010.91.1.9.

J. B. Hirsh and J. B. Peterson. Personality and language use in self-narratives. *Journal of Research in Personality*, 43(3):524–527, 2009. ISSN 00926566. doi: 10.1016/j.jrp.2009.01.006.

J. Hogan, P. Barrett, and R. Hogan. Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92(5):1270–1285, 2007. ISSN 0021-9010. doi: 10.1037/0021-9010.92.5.1270.

R. R. Holden, L. L. Wood, and L. Tomashewski. Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology*, 81(1):160–169, 2001. ISSN 1939-1315. doi: 10.1037/0022-3514.81.1.160.

A. Imas. Working for the "warm glow": On the benefits and limits of prosocial incentives. *Journal of Public Economics*, 114:14–18, 2014. ISSN 00472727. doi: 10.1016/j.jpubeco.2013.11.006.

D. N. Jackson, V. R. Wroblewski, and M. C. Ashton. The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, 13 (4):371–388, 2000. ISSN 0895-9285. doi: 10.1207/S15327043HUP1304_3.

B. A. Jacob and S. D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3): 843–877, 2003. ISSN 0033-5533. doi: 10.1162/00335530360698441.

L. A. Jensen-Campbell, R. Adams, D. G. Perry, K. A. Workman, J. Q. Furdella, and S. K. Egan. Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality*, 36 (3):224–251, 2002. ISSN 00926566. doi: 10.1006/jrpe.2002.2348.

T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick. The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, 52(3):621–652, 1999. ISSN 0031-5826. doi: 10.1111/j.1744-6570.1999.tb00174.x.

E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014. ISSN 0261-927X. doi: 10.1177/0261927X13502654.

J. Kagel and P. McGee. Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters*, 124(2):274–277, 2014. ISSN 01651765. doi: 10.1016/j.econlet.2014.05.034.

H. H. Kassarjian. Personality and consumer behavior: A review. *Journal of Marketing Research*, 8(4):409–418, 2018. ISSN 0022-2437. doi: 10.1177/002224377100800401.

I. Kesternich, H. Schumacher, B. Siflinger, and S. Schwarz. Work meaning and labor supply. *FEB Research Report Department of Economics*, 2020.

K. Konrad. Sabotage in rent-seeking contests. *Journal of Law, Economics, and Organization*, 16(1):155–165, 2000. ISSN 14657341. doi: 10.1093/jleo/16.1.155.

S. L. Koole, W. Jager, A. E. van den Berg, C. A. J. Vlek, and W. K. B. Hofstee. On the social nature of personality: Effects of extraversion, agreeableness, and feedback about collective resource use on cooperation in a resource dilemma. *Personality and Social Psychology Bulletin*, 27(3):289–301, 2016. ISSN 0146-1672. doi: 10.1177/0146167201273003.

M. Kosfeld, S. Neckermann, and X. Yang. The effects of financial and recognition incentives across work contexts: The role of meaning. *Economic Inquiry*, 55(1): 237–247, 2017. ISSN 00952583. doi: 10.1111/ecin.12350.

E. L. Krupka and R. A. Weber. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524, 2013. ISSN 15424766. doi: 10.1111/jeea.12006.

L. M. Larson, P. J. Rottinghaus, and F. H. Borgen. Meta-analyses of big six interests and big five personality factors. *Journal of Vocational Behavior*, 61(2):217–239, 2002. ISSN 00018791. doi: 10.1006/jvbe.2001.1854.

E. P. Lazear. Pay equality and industrial politics. *Journal of Political Economy*, 97 (3):561–580, 1989. ISSN 0022-3808. doi: 10.1086/261616.

E. P. Lazear and S. Rosen. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864, 1981. ISSN 0022-3808. doi: 10.1086/261010.

J. Levashina and M. A. Campion. A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, 14(4):299–316, 2006. ISSN 0965075X. doi: 10.1111/j.1468-2389.2006.00353.x.

S. D. Levitt. Testing the economic model of crime: The National Hockey League's two-referee experiment. *Contributions in Economic Analysis & Policy*, 1(1), 2002. ISSN 1538-0645. doi: 10.2202/1538-0645.1014.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007. doi: 10.1613/jair.2349.

S. Matz, Y. W. F. Chan, and M. Kosinski. Models of personality. In M. Tkalčič, B. de Carolis, M. de Gemmis, A. Odić, and A. Košir, editors, *Emotions and Personality in Personalized Services*, Human–Computer Interaction Series, pages 35–54. Springer International Publishing, Cham, 2016. ISBN 978-3-319-31411-2. doi: 10.1007/978-3-319-31413-6_3.

R. E. McCormick and R. D. Tollison. Crime on the court. *Journal of Political Economy*, 92(2):223–235, 1984. ISSN 0022-3808. doi: 10.1086/261221.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. ISSN 0033-3123. doi: 10.1007/BF02295996.

M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862–877, 2006. ISSN 1939-1315. doi: 10.1037/0022-3514.90.5.862.

T. Meier, R. Boyd, J. Pennebaker, M. Mehl, M. Martin, M. Wolf, and A. Horn. DE-LIWC2015, 2015.

F. P. Morgeson, M. A. Campion, R. L. Dipboye, J. R. Hollenbeck, K. Murphy, and N. Schmitt. Reconsiderung the use of personality tests in personnel selection contexts. *Personnel Psychology*, 60(3):683–729, 2007. ISSN 0031-5826. doi: 10.1111/j.1744-6570.2007.00089.x.

J. Münster. Selection tournaments, sabotage, and participation. *Journal of Economics & Management Strategy*, 16(4):943–970, 2007. ISSN 1058-6407. doi: 10.1111/j.1530-9134.2007.00163.x.

B. J. Nalebuff and J. E. Stiglitz. Prizes and incentives: Towards a general theory of compensation and competition. *The Bell Journal of Economics*, 14(1):21, 1983. ISSN 0361915X. doi: 10.2307/3003535.

A. Neal, G. Yeo, A. Koy, and T. Xiao. Predicting the form and direction of work role performance from the big 5 model of personality traits. *Journal of Organizational Behavior*, 33(2):175–192, 2012. ISSN 08943796. doi: 10.1002/job.742.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003. ISSN 0146-1672. doi: 10.1177/0146167203029005010.

M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236, 2008. doi: 10.1080/01638530802073712.

K. Nyborg and T. Zhang. Is corporate social responsibility associated with lower wages? *Environmental and Resource Economics*, 55(1):107–117, 2013. ISSN 0924-6460. doi: 10.1007/s10640-012-9617-8.

D. J. Ozer and V. Benet-Martínez. Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57:401–421, 2006. ISSN 0066-4308. doi: 10.1146/annurev.psych.57.102904.190127.

S. V. Paunonen and M. C. Ashton. Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, 81(3):524–539, 2001. ISSN 1939-1315. doi: 10.1037/0022-3514.81.3.524.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999. ISSN 1939-1315. doi: 10.1037/0022-3514.77.6.1296.

J. W. Pennebaker and T. C. Lay. Language use and personality during crises: Analyses of mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, 36(3):271–282, 2002. ISSN 00926566. doi: 10.1006/jrpe.2002.2349.

J. W. Pennebaker and L. D. Stone. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301, 2003. ISSN 1939-1315. doi: 10.1037/0022-3514.85.2.291.

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language. use: Our words, our selves. *Annual Review of Psychology*, 54: 547–577, 2003. ISSN 0066-4308. doi: 10.1146/annurev.psych.54.101601.145041.

J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver. When small words foretell academic success: The case of college admissions essays. *PLOS One*, 9(12):e115844, 2014. doi: 10.1371/journal.pone.0115844.

J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis. Linguistic inquiry and word count: LIWC2015, 2015.

I. B. Penzel, M. R. Persich, R. L. Boyd, and M. D. Robinson. Linguistic evidence for the failure mindset as a predictor of life span longevity. *Annals of Behavioral Medicine*, 51(3):348–355, 2017. doi: 10.1007/s12160-016-9857-x.

P. Pettersson-Lidbom and M. Priks. Behavior under social pressure: Empty italian stadiums and referee bias. *Economics Letters*, 108(2):212–214, 2010. ISSN 01651765. doi: 10.1016/j.econlet.2010.04.023.

M. Piopiunik, G. Schwerdt, L. Simon, and L. Woessmann. Skills, signals, and employability: An experimental investigation. *European Economic Review*, 123: 103374, 2020. ISSN 00142921. doi: 10.1016/j.euroecorev.2020.103374.

S. D. Pressman and S. Cohen. Use of social words in autobiographies and longevity. *Psychosomatic Medicine*, 69(3):262–269, 2007. doi: 10.1097/PSY.0b013e31803cb919.

L. Prouteau and F.-C. Wolff. Does volunteer work pay off in the labor market? *The Journal of Socio-Economics*, 35(6):992–1013, 2006. ISSN 10535357. doi: 10.1016/j.socec.2005.11.021.

Queens University of Charlotte. Communicating in the modern workplace: How millenials and their managers compare, 2021. URL `https://online.queens.edu/resources/infographic/communicating-in-the-workplace/`.

D. G. Rand, G. Kraft-Todd, and J. Gruber. The collective benefits of feeling good and letting go: Positive emotion and (dis)inhibition interact to predict cooperative behavior. *PLOS One*, 10(1):e0117426, 2015. doi: 10.1371/journal.pone.0117426.

D. Riedl, B. Strauss, A. Heuer, and O. Rubner. Finale furioso: Referee-biased injury times and their effects on home advantage in football. *Journal of Sports Sciences*, 33(4):327–336, 2015. doi: 10.1080/02640414.2014.944558.

R. W. Robins, A. Caspi, and T. E. Moffitt. It's not just who you're with, it's who you are: Personality and relationship experiences across multiple relationships. *Journal of Personality*, 70(6):925–964, 2002. ISSN 0022-3506. doi: 10.1111/1467-6494.05028.

R. L. Robinson, R. Navea, and W. Ickes. Predicting final course performance from students' written self-introductions. *Journal of Language and Social Psychology*, 32(4):469–479, 2013. ISSN 0261-927X. doi: 10.1177/0261927X13476869.

B. Rocha, F. Sanches, I. Souza, and J. Carlos Domingos da Silva. Does monitoring affect corruption? Career concerns and home bias in football refereeing. *Applied Economics Letters*, 20(8):728–731, 2013. ISSN 1350-4851. doi: 10.1080/13504851.2012.736938.

B. D. Rosso, K. H. Dekas, and A. Wrzesniewski. On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior*, 30:91–127, 2010. ISSN 01913085. doi: 10.1016/j.riob.2010.09.001.

G. Saucier. Isms and the structure of social attitudes. *Journal of Personality and Social Psychology*, 78(2):366–385, 2000. ISSN 1939-1315. doi: 10.1037/0022-3514.78.2.366.

R. M. Sauer. Does it pay for women to volunteer? *International Economic Review*, 56(2):537–564, 2015. ISSN 00206598. doi: 10.1111/iere.12114.

H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS One*, 8(9):e73791, 2013. doi: 10.1371/journal.pone.0073791.

V. Scoppa. Are subjective evaluations biased by social factors or connections? An econometric analysis of soccer referee decisions. *Empirical Economics*, 35(1):123–140, 2008. ISSN 0377-7332. doi: 10.1007/s00181-007-0146-1.

R. Selten. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. *Beiträge zur experimentellen Wirtschaftsforschung*, pages 136–168, 1967.

SHRM. SHRM/ACT: Job analysis activities, 2014. URL `https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Pages/2014-job-analysis-activities.aspx`.

R. B. Slatcher, S. Vazire, and J. W. Pennebaker. Am "I" more important than "we"? Couples' word use in instant messages. *Personal Relationships*, 15(4):407–424, 2008. ISSN 13504126. doi: 10.1111/j.1475-6811.2008.00207.x.

L. D. Stone and J. W. Pennebaker. Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology*, 24(3):173–183, 2002. ISSN 0197-3533. doi: 10.1207/S15324834BASP2403_1.

M. Sutter and M. G. Kocher. Favoritism of agents – The case of referees' home bias. *Journal of Economic Psychology*, 25(4):461–469, 2004. ISSN 01674870. doi: 10.1016/S0167-4870(03)00013-8.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. ISSN 0261-927X. doi: 10.1177/0261927X09351676.

The International Football Association Board. Information on the video assistant referee (VAR) experiment, incl. provisional results, 2018a. URL `https://static-3eb8.kxcdn.com/documents/639/165902_220118_IFAB_Media_Package_ABM2017_all_media_FINAL.pdf`.

The International Football Association Board. Laws of the game 2018/19, 2018b. URL `https://img.fifa.com/image/upload/khhloe2xoigyna8juxw3.pdf`.

M. Tonin and M. Vlassopoulos. Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science*, 61(8):1795–1811, 2015. ISSN 0025-1909. doi: 10.1287/mnsc.2014.1985.

T. J. Trull and K. J. Sher. Relationship between the five-factor model of personality and axis I disorders in a nonclinical sample. *Journal of Abnormal Psychology*, 103 (2):350–360, 1994. ISSN 0021-843X. doi: 10.1037/0021-843X.103.2.350.

A. van Hiel, I. Mervielde, and F. de Fruyt. The relationship between maladaptive personality and right wing ideology. *Personality and Individual Differences*, 36 (2):405–417, 2004. ISSN 01918869. doi: 10.1016/S0191-8869(03)00105-3.

C. Viswesvaran and D. S. Ones. Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2): 197–210, 1999. ISSN 0013-1644. doi: 10.1177/00131649921969802.

S. Volk, C. Thöni, and W. Ruigrok. Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, 50(6):810–815, 2011. ISSN 01918869. doi: 10.1016/j.paid.2011.01.001.

S. Wallrodt and L. Thieme. The role of sports volunteering as a signal in the job application process. *European Sport Management Quarterly*, 20(3):255–275, 2020. ISSN 1618-4742. doi: 10.1080/16184742.2019.1598457.

K. E. Walton and B. W. Roberts. On the relationship between substance use and personality traits: Abstainers are not maladjusted. *Journal of Research in Personality*, 38(6):515–535, 2004. ISSN 00926566. doi: 10.1016/j.jrp.2004.01.002.

S. Wiltsey Stirman and J. W. Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4), 2001. URL `https://journals.lww.com/psychosomaticmedicine/Fulltext/2001/07000/Word_Use_in_the_Poetry_of_Suicidal_and_Nonsuicidal.1.aspx`.

T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, 2010. ISSN 00926566. doi: 10.1016/j.jrp.2010.04.001.

A. Yavas and D. Vandegrift. An experimental test of sabotage in tournaments. *Journal of Institutional and Theoretical Economics*, 166(2):259, 2010. ISSN 0932-4569. doi: 10.1628/093245610791343012.

L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1):81–106, 2004. ISSN 0926-2644. doi: 10.1023/B:GRUP.0000011944.62889.6f.

M. Ziegler, D. Bensch, U. Maaß, V. Schult, M. Vogel, and M. Bühner. Big five facets as predictor of job training performance: The role of specific job demands. *Learning and Individual Differences*, 29:1–7, 2014. ISSN 10416080. doi: 10.1016/j.lindif.2013.10.008.

Publisher:
Magnus Emanuel Strobel
Schwindstr. 13
80798 Munich
magnus.strobel@tum.de