Essays in Applied Microeconomics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschaftswissenschaften

durch

die Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Chui Yee Ho

aus Negeri Sembilan

2024

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Lorenz Götte
Zweitreferent:	Prof. Dr. Thomas Dohmen
Tag der mündlichen Prüfung:	22. May 2024

Acknowledgements

I would firstly like to thank my supervisors Lorenz Götte and Thomas Dohmen, who provided me with incredible joint project opportunities, as well as funding throughout my years at the BGSE, and who always found the time to provide invaluable feedback on research projects.

I am secondly grateful to all my friends who made my time at the BGSE and Germany enjoyable. In no particular order: Matthias for being the best office mate and go-to mover, Alina for the many sports sessions and walks that kept me fit throughout my PhD; Ximi, whose optimism and can-do spirit made the impossible possible; Timo, another best office mate with whom I've had many eye-opening conversations; Luca, who is the definition of Willkommenskultur; Radost, whose combination of firm pressure and kind reassurances ensured that this thesis was submitted in 2024 and not in 2025, and for taking a chance on me; Anna for the many swimming and sports sessions, for putting up with my third paper/job search stress, and for taking the time to coach me on interviews; Daniela for helping out with foreigner issues; Jing Jing for being one of my first friends in Bonn; and Karsten who ensured that we both won the thesis hand-in competition.

Last but not least I am indebted to my family for their unwavering support throughout the long years, and whose weekly dose of humor always lightened up my day, as well as my host family for being a family away from home.

Chui Yee Ho
Bonn, January 2024

Contents

List of Figures

List of Tables

Introduction

Standard economic models typically assume that individuals behave independently and only care about their own interests. However, individuals often consider the impact of their actions on others, and care about how their own actions are perceived by others in making decisions. Indeed, aspects of individual behavior that at first glance appear puzzling when viewed through the lens of standard economic models can be explained by reputational or image concerns, social sanctions, coordination with others' actions, or concern for the well-being of others. This thesis consists of three chapters, where I explore in each chapter how different social considerations may influence individual actions.

In Chapter 1 (a joint work with Ximeng Fang), we study how individuals react to increased observability of their actions. More specifically, we study the effect of a transparency reform on performance evaluation by judge panels in professional figure skating. Prior to the reform, individual judges' scores were only published anonymously after each competition. However, from the 2016-17 season onwards, these scores were published openly. Using a difference-in-differences design, we show that the within-panel dispersion of artistic scores (but not of the more objective technical score) decreased significantly in response to higher transparency, indicating a larger degree of conformity. This effect is stronger for high-profile competitions that attract larger public attention. However, we find no evidence for a reduction in nationalistic favoritism following the reform. Our results are consistent with a beauty-contest model in which transparency influences evaluation decisions through increased conformity concerns.

In Chapter 2 (a joint work with Ximeng Fang, Timo Freyer, Lorenz Goette, and Zihua Chen), we explore how individuals' consideration of the consequences of their actions on others can have a tangible impact on real-world problems such as the COVID-19 pandemic. The spread of COVID-19 induces a social dilemma: engaging in preventive health behaviors is costly for individuals but generates benefits that accrue to society at large. The extent to which individuals internalize the social impact of their actions may depend on their prosociality, i.e. the willingness to behave in a way that mostly benefits other people. We conduct a nationally representative online survey in Germany ($n = 5{,}843$) to investigate the role of prosociality in reducing the spread of COVID-19 during the second coronavirus wave. At the individual

level, higher prosociality is strongly positively related to compliance with public health behaviors such as mask wearing and social distancing. A one standard deviation (SD) increase in prosociality is associated with a 0.3 SD increase in compliance ($p < 0.01$). At the regional (NUTS-2) level, a one SD higher average prosociality is associated with an 11% lower weekly incidence rate ($p < 0.01$), and a 2%p lower weekly growth rate ($p < 0.01$) of COVID-19 cases, controlling for a host of demographic and socio-economic factors. This association is driven by higher compliance with public health behaviors in regions with higher prosociality. Our correlational results thus support the common notion that voluntary behavioral change plays a vital role in fighting the pandemic and, more generally, that social preferences may determine collective action outcomes of a society.

Lastly, in Chapter 3 (a joint work with Thomas Dohmen), we study how changes in institutions can influence individual's preferred work hours. The number of hours individuals prefer to supply can be influenced by many factors, including social norms pertaining to the socially accpetable number of hours to work, and the number of hours others in their community choose to work. We use data from the SOEP (waves 1985 to 2017) to investigate how widespread reforms affect the preferred labor supply choices of civil servants and public sector employees. We estimate linear probability models and find that a decrease in the number of hours individuals are contractually required to work in a week (the standard workweek) is followed by an immediate 13%p increase in the fraction of individuals preferring to work the same number of hours as the new standard workweek, and a 21%p decrease in the fraction of individuals preferring the old standard workweek. In contrast, increases to the standard workweek are followed by weaker changes in the composition of desired hours— we find a gradual 12.2%p increase in the fraction of individuals preferring the new standard workweek, and a 8.2%p decrease in the fraction of individuals preferring the old standard workweek. While the effect for workweek reductions is mainly driven by individuals directly switching from preferring the old to the new standard workweek, this is not the case for the workweek increases.

Chapter 1

The effect of transparency on performance evaluation in committees – evidence from professional figure skating

Joint with Ximeng Fang

1.1 Introduction

High-stakes decisions and evaluations are often delegated to groups of experts, as opposed to a single individual. This includes, among many other examples, the recommendation and implementation of government policies through specialized committees, judicial rulings by panels of jurors or judges, hiring decisions in the labor market, and performance evaluation in professional sports. Drawing on the views of multiple evaluators can improve the accuracy and precision of the final decision or recommendation by collecting and aggregating information (à la Condorcet), while simultaneously mitigating the influence of idiosyncratic preferences and biases.

However, the effectiveness of aggregating multiple evaluations depends crucially on the institutional design and the (strategic) incentives generated by it. One important feature is whether the votes and opinions of each individual are made public or kept secret. On the one hand, higher transparency of the decision-making process allows the public to hold individual evaluators accountable, who may in turn try to stay more impartial and put in more effort in acquiring and communicating relevant information. On the other hand, transparency may expose evaluators to undesired influences (such as outside pressure), and it can also cause excessive conformity or conservatism, i.e., members becoming hesitant in expressing contro-

versial opinions or deviating from a norm or consensus.[1] This may be particularly relevant in the absence of truly objective benchmarks for ex post validation. Thus, the effects of higher transparency on subjective decision-making can be theoretically ambiguous and nuanced (e.g., Levy, 2007; Gersbach and Hahn, 2012; Fehrler and Hughes, 2018; Mattozzi and Nakaguma, 2019; Fehrler and Janas, 2021). Yet, with a few notable exceptions (e.g., Meade and Stasavage, 2008; Benesch, Bütler, and Hofer, 2018; Hansen, McMahon, and Prat, 2018), causal evidence on the effects of transparency in real-world evaluation contexts remains scarce, mainly due to lack of suitable data and other empirical challenges.

In this paper, we study the effect of transparency on performance evaluation in the context of competitive figure skating. Figure skating is an inherently subjective sport, since the quality of an athlete's performance is partially derived from artistic aspects such as music interpretation and choreography. Hence, skaters' performances are independently evaluated by a panel of (typically nine) expert judges. Prior to the 2016-17 season, judges' scores in many competitions were published anonymously, meaning that only the distribution of scores and the identities of judges on the panel were known, but the two could not be linked to each other. In 2016, following allegations of biased evaluations due to nationalistic favoritism, a major transparency reform was implemented, so each judges' scores were published openly from the 2016-17 season onwards. We examine the effects of this transparency reform on judges' performance evaluation behavior in a difference-in-differences design, using as control group a subset of events (Junior Grand Prix competitions) in which individual judges' scores were already published openly pre-reform.

This setting allows us to overcome several empirical challenges. First, we observe a large number of comparable decisions by professional evaluators in a high-stakes context, both under anonymous and transparent disclosure regimes. Second, the aggregation mechanism is common knowledge and we observe all inputs that contribute to the overall decision. Third, we can rule out joint deliberation and strategic agreements within the committee, as figure skating judges are not allowed to communicate with each other when awarding scores. Finally, the difference-in-differences setup allows us to control for general time trends unrelated to the reform, thus helping us to isolate the effect of higher transparency.

Individuals have generally been found to shift their behavior more towards the socially acceptable norm when (feeling) observed by others.[2] Accordingly, if judges want to appear competent and impartial in the public eye, then higher transparency

1. The famous experiment by Asch (1951) is a classical example of how group conformity overrules reason. Similarly, it has been argued that the wisdom-of-crowds phenomenon may not hold when the aggregated judgements are not independent but exposed to social influence (Lorenz et al., 2011).

2. For example, students tend to reduce (visible) schooling investments when their rankings are revealed to their classmates (Bursztyn and Jensen, 2015), grocery store workers work harder when observed by more productive co-workers (Mas and Moretti, 2009), individuals are more likely to vote if

could trigger judges' image and reputation concerns and thereby induce them to report more accurate evaluations (see, e.g., Suurmond, Swank, and Visser, 2004; Bar-Isaac, 2012; Gersbach and Hahn, 2012; Hansen, McMahon, and Prat, 2018; Mattozzi and Nakaguma, 2019; Swank and Visser, 2021). This may be of particular importance in the presence of significant subjective bias and favoritism in evaluation decisions, which has been well documented in figure skating and beyond.[3] However, there is no completely objective metric in figure skating against which judges' evaluation decisions can be validated against, i.e., the "accurate" score is never truly revealed — which is the very reason why performances are evaluated by a panel of expert judges in the first place. Thus, subjective performance evaluation includes elements of a credence good (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006). In such situations, a natural benchmark for evaluations of individual panel members is the comparison to evaluations by the other members.[4] This can create strategic incentives for judges to become more "conformist", i.e., to report scores that are closer to the scores that (they think) other judges will report. This can encourage higher individual effort to determine what would be objectively fair, but it could also lead to a loss of information value (Prendergast, 1993; Prat, 2005).

To explore the potential effects of transparency more formally, we present a theoretical model based on a beauty contest framework à la Morris and Shin (2002) with endogenous information acquisition. Judges are partially motivated by a truth-telling motive, but they also have a distortion motive due to subjective biases (such as favoritism toward compatriot athletes). Additionally, reputation-concerned judges have a conformity motive, i.e., they want to award scores that are similar to those of their fellow judges. We interpret higher transparency through the publishing of individual scores as an exogenous increase in this conformity motive. The model highlights three key mechanisms through which transparency can affect judge evaluation behavior. Firstly, judges exert higher effort to generate more precise signals, as a reduction in noise will generally lead to higher correlation of signals within the panel. Secondly, judges become more cautious and conservative in their scores, e.g. by anchoring towards a common prior, thus leading them to place lower weight on

---

they believe that their voting status would be revealed to their neighbors (Gerber, Green, and Larimer, 2008).

3. Systematic biases, especially in the form of nationalistic favoritism, has been documented in figure skating (Campbell and Galbraith, 1996; Zitzewitz, 2006; Lee, 2008; Litman and Stratmann, 2018) as well as in other professional sports where performance is evaluated by judge panels (see e.g. Sandberg, 2018). Relatedly, there is evidence for home team bias and racial bias in refereeing decisions (Garicano, Palacios-Huerta, and Prendergast, 2005; Price and Wolfers, 2010; Parsons et al., 2011). Subjective biases are also prevalent in the evaluation of academic research (see, e.g., Li, 2017; Huber et al., 2022).

4. Indeed, committee members are frequently evaluated by comparing them to their peers. This is based on the rationale that evaluations that are more accurate will generally be more strongly (positively) correlated with each other. In figure skating, large deviations from average scores can lead to disciplinary actions against judges.

their private signal than they would under anonymous scoring. Lastly, transparency can induce judges to curb the expression of their idiosyncratic biases towards certain skaters; paradoxically, this may not lead to lower *aggregate* bias in the panel, as conformity concerns create the perverse incentive for judges to match the expected biases of other judges on the panel.

Several testable predictions arise. Above all, the model unambiguously predicts that the dispersion of scores across judges for a given performance will decrease after the transparency reform. This consensus effect is expected to be larger the more difficult it is to observe an objective score — implying in our context that conformity should be stronger for the artistic elements, rather than the technical elements of the performance —, the higher public attention on the performance is, and the stronger preconceived biases are (e.g., due to nationalistic favoritism). The model also predicts that, contrary to the aim of the reform, *aggregate* nationalistic bias will not necessarily decrease under greater transparency. To examine the effects of the transparency reform empirically, we analyze scores from almost 17,000 figure skating performances across 127 competitions organized by the International Skating Union (ISU) between 2013 and 2020. Our empirical identification strategy compares changes in the distribution of judge scores after the 2016 transparency reform between JGP (Junior Grand Prix) events, which were not affected by the reform, and Non-JGP events, which were.

Our empirical results are in line with the theoretical predictions. Importantly, we find that individual judges' scores for a given performance become more similar to each other after the transparency reform takes effect. In particular, the dispersion of artistic scores within the judge panel drops sharply for Non-JGP events, relative to JGP events. The consensus effect in artistic scores is both statistically significant and quantitatively sizable — constituting approximately 9% of the pre-reform average and 29% of the pre-reform standard deviation of within-panel score dispersion — and it is mainly driven by the reduction of large outliers, so judges' scores become more tightly packed around the mean. It is also particularly pronounced for high-profile events, which arguably garner greater public attention, thus supporting the notion that the effects of transparency on judge evaluations are mediated by image and reputation concerns. However, we observe no consensus effect for the more objective technical score, which covers aspects like difficulty and execution of technical elements (jumps, spins, etc.). Moreover, there is no evidence that the reform led to a decrease in *aggregate* nationalistic bias, as measured by the average score advantage a skater receives when he or she has a compatriot judge on the panel. Although surprising given the reform's original intentions, this is consistent with our theoretical predictions.

Our theoretical framework highlights three mechanisms that can generate our empirical findings: higher effort, implicit coordination on common priors or signals, and conformity in biases. We find no evidence that judges give more similar scores the longer they have been evaluating together in the same panel, which speaks

against implicit coordination through social learning. Furthermore, there is only weak evidence that the conformity effect is stronger for performances with a compatriot judge on the panel, and quantitatively it cannot fully explain the average decrease in score dispersion across judges. This suggests that a significant part of the consensus effect may be driven by more precise evaluations through higher effort or attention. To provide suggestive evidence for this, we analyze the sub-scores for different artistic components (e.g., choreography, music interpretation, transitions, ...) that sum up to the overall artistic score. We first document that within-judge consistency of sub-scores across artistic components could be interpreted as proxy for accuracy, as higher consistency is associated with other markers of evaluation quality at the individual judge level. Second, we document that the consistency of artistic (but not technical) sub-scores increases significantly post-reform, which could thus be interpreted as marker for higher effort when awarding scores. As a robustness check, we verify that the transparency reform did not induce a different selection of judges into committees based on observable characteristics. Yet eventually, as we cannot determine an objective score for a performance without using the judge panel scores, we are not able to fully distinguish between these different mechanisms empirically.

Our paper contributes firstly to the literature on the consequences of transparency in committee decision-making. Theoretical models typically study how members' reputation concerns, i.e. their desire to appear competent, determine how they respond to transparency. Although transparency may under some circumstances induce anti-conformism to signal individual competence (Levy, 2007), committees may also have a preference for showing a united front in the public, in particular if true states cannot be observed ex post (Visser and Swank, 2007; Swank, Swank, and Visser, 2008; Swank and Visser, 2021). Higher transparency can also lead to more pre-decision information acquisition (Gersbach and Hahn, 2012; Swank and Visser, 2021). One difference to our setting is that these theoretical papers typically study a binary decision, whereas scores in our setting are awarded on a scale and aggregated by averaging.[5] Empirical evidence on the effect of transparency on committee decision-making is relatively scarce. Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2019) provide laboratory evidence on the role of different transparency regimes on information aggregation in groups. With regard to real-world committees, several studies examine how monetary policy deliberations responded to a reform that resulted in transcripts of FOMC meetings being made public after Fall 1993. Meade and Stasavage (2008) find that members are less likely to voice disagreement with the Committee Chairman post-reform; using computational linguistics tools, Hansen, McMahon, and Prat (2018) find that

5. Rosar (2015) studies committee decision rules with continuous reporting and decision spaces and shows how this gives rise to incentives for strategic exaggeration.

FOMC members tend to give more similar statements and engage less in back-and-forth dialogue post-reform, but also that especially rookie members seem to be better prepared with quantitative information on a diverse set of topics. Benesch, Bütler, and Hofer (2018) study a transparency reform in the Upper House of the Swiss parliament and show that, post-reform, legislators exhibit greater party discipline. Though we also find a conformity effect, there are several noteworthy differences in our setting. Firstly, the report space in our setting is continuous, which allows for strategies that do not exist under a binary report space. Secondly, and more importantly, the lack of a deliberation or discussion stage in the current setup implies that the result we find is not due to (direct) coercion or coordination with other judges. Thus, this paper thus adds to this literature by demonstrating a conformity effect under greater transparency even in the absence of information exchange, thus providing stronger evidence for the way social image concerns can affect behavior of committee members.

A large number of previous studies have utilized large-scale publicly available data from professional sports contexts to investigate, among others, determinants of performance (e.g. Dohmen, 2008a; Lichter, Pestel, and Sommer, 2017; Jiang, 2020), systematic decision errors (e.g. Bruine de Bruin, 2006; Pope and Schweitzer, 2011), gender differences (e.g. Böheim, Lackner, and Wagner, 2020), as well as favoritism (e.g. Garicano, Palacios-Huerta, and Prendergast, 2005; Zitzewitz, 2006; Sandberg, 2018; Fernando and George, 2021) and racial biases (e.g. Price and Wolfers, 2010; Parsons et al., 2011; Pope, Price, and Wolfers, 2018). Two closely related papers to ours are by Zitzewitz (2014) and Lee (2008), who study a set of reforms in figure skating (following a vote trading scandal at the 2002 Winter Olympics) that in fact introduced the anonymous scoring regime that was eventually reversed in 2016. Zitzewitz (2014) finds a slight but statistically insignificant increase in the compatriot score advantage after the reform, and Lee (2008) finds an increase in the standard deviation of judges' scores under anonymized publication. However, a number of other major reforms were implemented at that time, including an increase in the size of the judging panel and random dropping of judges' scores from the calculation of the final score, followed by another extensive series of reforms two years later. Our current setting using the 2016 reform allows for a cleaner attribution of changes in judge scoring behavior to increased transparency of judges' decisions, and our use of JGP events as control group in a difference-in-differences design further tightens the empirical identification by controlling for counterfactual time trends.

We also contribute to the literature studying whether changes in information structures could reduce discrimination. In recent years, a variety of reforms have been implemented at a large-scale (e.g. quotas, increased minority representation on

selection committees, blind applications, pay transparency etc) to mixed results.[6] We provide a new empirical case study on the efficacy (or lack thereof) of a transparency-based method to counter favoritism/discrimination. Our results show that there is no evidence for any reduction in nationalistic favoritism following the publication of individual judge scores in figure skating. This could be due to several reasons. First, fairness norms might not be strong enough or offset by opposing loyalty norms induced by judges' home audience. Second, the group structure of committees could interact with conformity concerns, so that judges aim to give more similar scores to their peers by matching their biases, or alternatively, that the non-compatriot judges might skew their scores slightly upwards when one of their peers has the same nationality as the skater.[7] Third, the bias-correcting properties of aggregating multiple votes reduces the scope for reducing the aggregate bias.

The remainder of the paper is organized as follows. Section 1.2 gives a brief overview of our empirical context. In Section 1.3, we discuss how transparency can lead to changes in behavior through the lens of a theoretical model. We describe our data and provide summary statistics in Section 1.4. The empirical strategy is outlined in Section 1.5. In Section 1.6, we present our main empirical results, and Section 1.7 shows additional results to explore the underlying mechanisms. Section 1.8 concludes.

1.2 Context

Figure skating is a sport in which athletes (individuals or pairs) skate on ice and perform a choreographed sequence of jumps, spins, and dance moves to a musical track. There are four main disciplines in figure skating: Men's Singles, Women's Singles, Pairs Skating, and Ice Dance. In this paper, we focus on official international events recognized by the International Skating Union (ISU). Some of the most prestigious ISU events include the World Championships, the Grand Prix Series and Finals, and the quadrennial Olympics Winter Games. Each event typically consists of four competitions, one for each discipline. Within each competition, skaters skate twice, once in the Short Program and once in the Long Program. The skater's final placement in the competition is determined by the sum of total scores in each program.

6. See, e.g., Bertrand et al. (2018) and Maida and Weber (2019) for evidence on quotas, Bagues and Esteve-Volart (2010) and Bagues, Sylos-Labini, and Zinovyeva (2017) for evidence on the effectiveness of gender representation on selection committees, Krause, Rinne, and Zimmermann (2012) and Behaghel, Crépon, and Le Barbanchon (2015) on blind applications, Mas (2017) and Baker et al. (2019) on pay transparency.

7. Bagues and Esteve-Volart (2010) also hint at strategic dependencies between committee members leading to worse outcomes for female candidates paired with academic committees with greater female representation, as male committee members became less favorable when there were more female members on the committee.

1.2.1 Scoring of figure skating performances

Within the ISU Judging System, skaters are evaluated by a panel of (typically) 9 judges, who watch the performance and award scores to indicate its technical and artistic quality. Judges are not allowed to confer with each other while grading the performance.Scores consist of two main parts: the Technical Elements Score (TES), which evaluates the difficulty and execution of technical elements, and the Program Component Score (PCS), which evaluates the artistic value of the performance. The Total Score (TS) for a skating performance is given by the sum of the TES and the PCS, minus any potential deductions (e.g., due to rule violations). Throughout the paper, we will often refer to the TES as the "technical score" and to the PCS as the "artistic score".

The TES is determined as follows. Skaters perform a number of technical elements (jumps, spins, etc.) in their performance, and each element receives a score from the judge panel. This score is computed based on the Base Value, which increase in the difficulty level of the element, and the Grade of Execution (GoE), which is assigned by each member of the judge panel and indicates how cleanly the element was executed.[8] This GoE is then scaled according to the difficulty of the element and added to its Base Value, with more difficult technical elements receiving higher GoE scaling factor. To hinder manipulation and reduce the impact of outliers, the highest and lowest GoEs for each technical element in the judge panel are dropped. The overall TES of a performance is obtained by calculating the (trimmed) average scores for all technical element across judges and summing them up.

In contrast to the TES, the artistic scores that determine the PCS are awarded after the end of the performance. Each judge assigns a score to the artistic components of performance, which include the interpretation of music, skating skills, transitions between technical elements, composition, and performance. Each component can be marked on a range from 0.25 to 10 in quarter-point increments. Again, the highest and lowest scores in the judge panel for each component are dropped. The PCS is obtained by calculating the (trimmed) average scores for all components across judges and summing them up.

1.2.2 Transparency reform in 2016

Each season, there are around 20 ISU events, including the European Championships, Four Continents Championships, World Championships, Olympics Winter Games, the Grand Prix Series and Final, the Junior World Championships, and Junior Grand Prix (JGP) Series and Final. After each event, the ISU publishes detailed scoring information for all performances, including the individual judge scores that

8. The GoE ranges between -3 and +3, with increments of 1. From the 2018-19 season onwards, the range of the GoE was increased, to span from -5 to +5.

make up the final score, its official website. Prior to the 2016-17 season, with the exception of Junior Grand Prix (JGP) Series events, these individual scores were published anonymously. That is, while the identities of the judges on the panel were known, the individual scores are published in random order, so that they cannot be linked to an individual judge.[9]

This lack of transparency meant that judges could not be held accountable for their decisions, which led to accusations of biased judging by the public. Such allegations came to a head with the scoring of the 2014 Olympics Ladies competition, where Russian competitor Alina Zagitova was awarded gold ahead of the South Korean competitor Kim Yu-Na. Indeed, public outrage over the scoring reached such a point that the International Skating Union (ISU) considered abolishing judge anonymity in their General Meeting in 2014. While the proposal failed narrowly, it was brought up once again two years later (in 2016) and passed, so that from 2016-17 onwards, judges' scores from all competitions were published openly. Though other reforms were implemented at the 2016-17 meeting, these reforms were not explicitly aimed at reducing nationalistic judging, and mostly affect both JGP (Control) and Non-JGP (Treatment) events.[10]

Because JGP events already published scores openly prior to the transparency reform, they were unaffected by the reform and thus serve as a control group. JGP events follow the same scoring format and criteria as Non-JGP events and, to a certain extent, share the same pool of judges as Non-JGP events— over the study period of 2013-2020, half of the judges have judged in at least one JGP event and Non-JGP event. The core difference between these two groups of events lies in the level of prestige and exclusivity. JGP events are typically less prestigious and exclusive than Non-JGP events, so that scores from JGP events tend to be lower.

1.3 Theoretical Framework

The main consequence of the transparency reform is that individual judges' evaluations become perfectly observable, with the aim of encouraging more accurate and less biased judge evaluations through reputational incentives. Thus, the idea is that career-concerned judges will want to appear competent and impartial in the face of public scrutiny. However, there is often no truly objective yardstick against which an individual judge's evaluation accuracy can be compared against. This is clearly the case in the context of competitive figure skating, as the subjective nature of the

9. See Figure 1.A.1 for an example of a published score sheet.

10. Other reforms are mostly concerned with changes in required technical elements and updated scoring guidelines, which are typically implemented every two years (when a General Meeting is held). A few rule changes are specific to Senior events; however, these are mostly specific to the technical elements.

sport is the very reason why athletes' performances are evaluated by aggregating multiple individual scores from panel of expert judges.

A natural and intuitive approach to evaluate the marking accuracy and impartiality of individual judges is to compare their scores against the scores awarded by the other expert judges on the panel (Heiniger and Mercier, 2021). Outlier judges who express very different opinions from those of their peers may be perceived as being incompetent, inattentive, or biased, whereas judges who are close to the median might be perceived as competent and impartial.Therefore, the transparency reform plausibly generates stronger incentives for judges to report scores that are more similar to those of others. Note that it is not possible (and not allowed) for judges to deliberate together or coordinate their scores, but judges could potentially react to transparency by exerting more effort into marking accurately, by curbing their biases toward certain skaters (e.g., of the same nationality), or by anchoring conservatively towards a common prior.

To formalize these intuitions and to derive predictions for how transparency could affect the distribution of scores within the judge panel, we present a theoretical model of judges' performance evaluation behavior that is based on the well-studied beauty contest framework introduced by Morris and Shin (2002), and extended by Colombo and Femminis (2008) to incorporate costly information acquisition.

1.3.1 Basic setup

Skater i performs in a competition. Judges $j = 1, ..., N$ sit on the panel and evaluate the quality of the performance by each reporting a score π_{ji} without joint deliberation. These individual scores $\pi_{1i}, ...\pi_{Ni}$ are then aggregated to an overall average score $\pi_i = \frac{1}{N}\sum_j \pi_{ji}$. For simplicity, we abstract from the trimming of the highest and lowest scores.

The common prior of performance quality θ_i for skater i follows a normal distribution with mean μ_i and (non-zero) variance σ_i^2. Judges may reasonably have different priors about, e.g., a consistently world-class skater compared to a capricious rookie, so both μ_i and σ_i can differ across skaters. As there is a strong artistic aspect to figure skating and thus no simple objective criterion for evaluating a performance, the "true" realized quality θ_i is imperfectly observable ex post. However, by watching the performance, each judge receives a private signal of the performance quality:

$$x_{ji} = \theta_i + \varepsilon_{ji}, \tag{1.1}$$

which can be thought of as reflecting the judge's own personal assessment.[11] The signal is unbiased but contains an idiosyncratic noise term ε_{ji} that is independent of

11. We simplify the Morris and Shin (2002) framework by not including a public signal y_i that is the main focus of their paper and of much of the literature it spurred. However, the skater-specific

θ_i and that follows a normal distribution with mean 0 and variance σ_i^2/τ_{ji}, where $\tau_{ji} \in (1, \infty)$ denotes the precision of judge j's signal for skater i. We assume that the private signal after oberving the performance is always more informative than the prior ($\tau_{ji} > 1$), but never so informative that θ_i is perfectly observed ($\tau_{ji} < \infty$). This offers a rationale for assigning final scores by aggregating the (independent) opinions of multiple judges in order to reduce the influence of idiosyncratic tendencies and judgement errors. However, ε_{ji} can be heteroscedastic. For example, an experienced and attentive judge may be able to evaluate the quality of a performance more reliably than a judge who is inexperienced or inattentive. Similarly, a performance that is excellent all around is arguably easier to evaluate than a mediocre performance with highs and lows.

1.3.2 Simplified model

To build intuition, we will first present a stripped-down version of our model in which judges behave non-strategically and in which signal precision τ_{ji} is given exogenously. We assume that judges are partially motivated to give a genuinely accurate assessment of the performance quality when reporting their scores, but that they can additionally be biased towards rewarding systematically higher or lower scores to skater i. This bias may reflect favoritism, e.g. due to same nationality or a preferred skating style (Zitzewitz, 2006; Litman and Stratmann, 2018), but it could in principle also reflect stable differences in judges' general strictness or leniency, if the bias is invariant to the skater's identity. We model these two elements through the following payoff function:

$$u_j(\pi_{ji}, b_{ji}, \theta_i) \quad = \quad -(\pi_{ji} - \theta_i - b_{ji})^2 \,. \tag{1.2}$$

b_{ji} is the (fixed) bias of judge j towards skater i. Judges choose π_{ji} to maximize their expected utility. The quadratic loss formulation leads to a classical signal extraction problem, and the optimal non-strategic report $\tilde{\pi}_{ji}$ can be obtained using Bayes' rule:

$$\tilde{\pi}_{ji} \quad = \quad E[\theta_i | x_{ji}, y_i] + b_{ji} \quad = \quad \frac{1}{1+\tau_{ji}}\mu_i + \frac{\tau_{ji}}{1+\tau_{ji}}x_{ji} + b_{ji} \,. \tag{1.3}$$

The first component $E[\theta_i | x_{ji}, y_i]$ is a linear combination of the private signal x_{ji} and the common posterior μ_i and represents the actual posterior belief about performance quality θ_i that the judge forms. The more accurately a judge is able evaluate the performance, i.e. the higher τ_{ji}, the more weight will be put on his or her actual signal. The second component b_{ji} creates a distortion in the reported score due to the judge's bias towards skater i. Depending on how the biases are distributed

prior with mean μ_i and variance σ_i^2 could be interpreted implicitly as the interim posterior distribution conditional on public information about ex ante obervable characteristics of skater i, such as their previous performance scores.

across judges in the panel, they may not completely average out when scores are aggregated, so some skaters may have an unfair advantage compared to others, if it so happens that the panel is tilted in favor of them, e.g., if a compatriot judge sits on the panel.

Assuming homogenous precision $\tau_{ji} = \tau_i$ for all judges, the expectation and variance of scores across judges in the panel conditional on the performance θ_i are

$$E[\tilde{\pi}_{ji}|\theta_i] = \theta_i + \frac{1}{1+\tau_i}(\mu_i - \theta_i) + E[b_{ji}], \tag{1.4}$$

$$Var[\tilde{\pi}_{ji}|\theta_i] = \frac{\tau_i}{(1+\tau_i)^2}\sigma_i^2 + Var[b_{ji}]. \tag{1.5}$$

The overall score can be ex post biased from two sources. First, the reported scores are conservative, i.e., slanted towards the common prior expectation μ_i, because judges can only observe θ_i with noise. Hence, hypothetically, the identical performance delivered by a famous world-class skater may be awarded a higher score than if delivered by an unknown rookie skater — this is sometimes referred to as the Matthew effect (Merton, 1968; Kim and King, 2014; Huber et al., 2022). Second, a skater will receive systematically higher or lower scores if there is asymmetry in judges' biases, for example if one judge exhibits strong nationalistic favoritism and the other judges in the panel are unbiased. While public focus often lies on bias and favoritism, a reduction in noise can be equally important in ensuring the validity of a decision making process (Kahneman, Sobony, and Sunstein, 2021). The expected variance of scores decreases with higher signal precision τ_i and with lower bias heterogeneity $Var[b_{ji}]$ across judges.

1.3.3 Full model

Our full model extends the non-strategic setup from above with two elements. First, judges are reputation-concerned, meaning that they want to appear competent in the way they award scores to a skating performance. As performance quality is not perfectly observable even ex post, especially with regard to the more artistic aspects, one straightforward way to evaluate a judges' score is to compare it to the score of other judges. Therefore, we model image concerns in a way that they lead to a motive for conforming with other judges, i.e. by not deviating too far from their scores. Second, we allow judges to endogenously adjust their signal precision τ_{ji} through costly information acquisition, which could be interpreted as level of effort or attentiveness when observing the performance.The judge's payoff function is

$$u_j(\pi_i, \tau_{ji}, \theta_i) = -\left(\pi_{ji} - \theta_i - b_{ji}\right)^2 - \eta\left(\pi_{ji} - \tfrac{1}{N-1}\sum_{l\neq j}\pi_{li}\right)^2 - C(\tau_{ji}), \tag{1.6}$$

where $\eta \in (0,1)$ captures the strength of the conformity motive relative to the truthfulness motive, and $C(\tau_{ji})$ is the effort cost necessary to achieve precision level

τ_{ji}. Following Colombo and Femminis (2008), we assume a linear cost function $C(\tau_{ji}) = c\tau_{ji}$. The unit "price" of precision is $c \in (0, \bar{c})$, with upper limit $\bar{c} = \frac{\sigma_i^2}{4(1+\eta)}$ to ensure that agents choose signal precisions τ_{ji} that are not implausibly low.[12] Note that there is now a strategic aspect to reporting behavior, since judge *j*'s expected utility depends on the scores of the other judges, and vice versa. As a solution concept, we compute the symmetric Bayesian Nash equilibrium, in which each judge makes inferences about the distribution of other judges' signals based on her own signal and then awards her optimal scores in response to other judges' reporting strategy. The individual rationality condition requires that for all $j = 1, ..., N$ and $l \neq j$,

$$\begin{aligned} \pi_{ji} &= \frac{1}{1+\eta}\left(E[\theta_i|x_{ji}, y_i] + b_{ji}\right) + \frac{\eta}{1+\eta} E[\pi_{li}|x_{ji}, y_i] \\ &= \frac{1}{1+\eta}\,\tilde{\pi}_{ji} + \frac{\eta}{1+\eta} E[\pi_{li}|x_{ji}, y_i]\,. \end{aligned} \tag{1.7}$$

As already observed by Morris and Shin (2002), a symmetric equlibrium implies that we can plug in the best response π_{li} from equation (1.7) for all $l \neq j$, leading to a feedback loop of higher-order beliefs that converges to a unique social equilibrium in which every judge *j* reports

$$\pi_{ji} = \frac{1+\eta}{1+\eta+\tau_{ji}}\,\mu_i + \frac{\tau_{ji}}{1+\eta+\tau_{ji}}\,x_{ji} + \frac{1}{1+\eta}\,b_{ji} + \frac{\eta}{1+\eta} E[b_i]\,. \tag{1.8}$$

This equilibrium condition has to be true regardless of the level of precision τ_{ji} that judges choose. Holding constant τ_{ji}, the optimal strategic report π_{ji} is more conservative than the non-strategic report $\tilde{\pi}_{ji}$, i.e., it is attenuated more strongly towards the common prior expectation μ_i. Hence, it resembles a tacit coordination of judges to deviate from their true posterior beliefs of performance quality and move their scores closer towards an uncontroversial benchmark. Interestingly, the desire to appear more in line with other judges also leads to conformity in biases, as judges now realign their bias partially towards the expected bias $E[b_i]$.

Next, we need to find the equilibrium level of effort τ_{ji}. Let all judges $l \neq j$ follow the same strategy, with report π_{ji} from equation (1.8) and homogeneous effort level $\tau_{li} = \tau_i$. Judge *j* takes this as given and and seeks to determine his or her individual effort level τ_{ji}. Adapting the results from Colombo and Femminis (2008), the optimal signal precision for all judges *j* in a symmetric equilibrium can be shown to be

$$\tau_{ji} = \tau_i = \sqrt{1+\eta} \cdot \frac{\sigma_i}{\sqrt{c}} - (1+\eta)\,. \tag{1.9}$$

12. As we will later see, this condition on *c* implies that $\tau_{ji} > 1 + \eta$ and ensures that judges will always place more weight on their private signal than on the common posterior when reporting their score, which is arguably a reasonable assumption. This also ensures that the variance of scores always decreases in signal precision, because when judges placed a higher weight on the common posterior than the private signal, scores would become very uniform.

Notice that this term is increasing in the conformity concern η for all $c \in (0, \bar{c}]$. Hence, transparency can be used as reputational incentive mechanism for inducing higher judge effort when evaluating skater performances.

Conditional on θ_i, the expectation and variance of performance scores across judges look as follows when taking into account conformity concerns and endogenous signal precision:

$$\begin{aligned} E[\pi_{ji}|\theta_i] &= \theta_i + \frac{1+\eta}{1+\eta+\tau_i}(\mu_i - \theta_i) + E[b_{ji}] \\ &= \theta_i + \frac{\sqrt{(1+\eta)c}}{\sigma_i}(\mu_i - \theta_i) + E[b_{ji}], \end{aligned} \tag{1.10}$$

$$\begin{aligned} Var[\pi_{ji}|\theta_i] &= \frac{\tau_i}{(1+\eta+\tau_i)^2}\sigma_i^2 + \frac{1}{(1+\eta)^2}Var[b_{ji}] \\ &= \frac{\sqrt{c}\,\sigma_i}{2(1+\eta)^{\frac{3}{2}}} - c + \frac{1}{(1+\eta)^2}Var[b_{ji}]. \end{aligned} \tag{1.11}$$

The distribution of judge scores still follows similar properties as in the simple model. Judges' scores exhibit conservatism towards the prior expectation and scores are further distorted through the average bias toward skater i in the judge panel. The more precisely judges can observe the performance quality, the less conservative and the less noisy the scores become. On top of that, the full model also allows us to study how the score distribution is affected by the conformity motive μ, which is arguably affected by whether judging is transparent or anonymous. In the following, we will use the results in equations (1.10) and (1.11) to derive testable predictions for the effects of the transparency reform.

1.3.4 Predicted effects of the transparency reform

Under anonymous scoring, the public cannot observe which judge gave which score. Hence, judges do not have to worry much about appearing incompetent or biased when the score they award is discrepant from the other judges' scores. In contrast, when scoring becomes transparent, judges may start worrying more about their social image and their desire to appear competent. In our model, we therefore interpret scoring under transparency as an increase in η compared to scoring under anonymity. Conducting comparative statics with regard to η then allows us to derive a number of testable predictions for how the transparency reform affects judges' scores, which we list below.

(1) Lower score dispersion for a given performance. — If transparency leads to stronger conformity concerns, the variance of scores across judges in the panel for a given performance decreases:

$$\frac{\partial}{\partial \eta} Var[\pi_{ji}|\theta_i] < 0. \tag{1.12}$$

There are three reasons for this lower score dispersion. First, stronger conformity concerns result in scores that are more conservative in the sense that they are attenuated towards the common posterior z_i, which means that judges place less weight on their idiosyncratic information. Second, increasing effort in η leads to less noise in judges' private signals. Third, dispersion can further decrease due to judges adjusting their individual biases more towards the average bias in the panel, which implies that the impact of transparency would be stronger if $Var[b_{ji}]$ is high, meaning that judges are very polarized in their biases towards a skater.

(2) Effect on score dispersion increases in subjectivity. — Skaters are evaluated both on the technical aspects and the artistic aspects of their performance. The latter is arguably much more subjective than the former, which implies that judges may have a harder time trying to award the artistic score as accurately as possible. We therefore look at another comparative static, which is how the effect of transparency on dispersion of scores is affected by an increase in the level of subjectivity/noise σ_i when judging performance quality. It is straightforward to show that

$$\frac{\partial^2}{\partial \eta \, \partial \sigma_i} Var[\pi_{ji}|\theta_i] < 0. \tag{1.13}$$

This implies that the reduction in score dispersion in prediction (1) is more pronounced if objective performance evaluation is more difficult. In particular, we would expect to see a larger reduction in dispersion for the artistic score than for the technical score.

Note that the same would hold if we replaced σ_i with the cost of information acquisition c_i. Further rationales for expecting smaller effects for the technical score is that conformity to other judges may play less of a role (i.e. η is lower), because its relative objectivity makes it more important for reputation-concerned judges to give their most accurate assessment, or because technical scores are awarded almost instantaneously and judges may not have time to consider other judges' behavior.

(3) No decrease in aggregate bias. — Perhaps surprisingly, our model suggests that, on average, higher transparency may leave the *aggregate* bias $B_i = \sum_j b_{ji}$ of the panel towards skater i unchanged, as the bias component in equation (1.10) is invariant to η:

$$\frac{\partial^2}{\partial \eta \, \partial E[b_{ji}]} E[\pi_{ji}|\theta_i] = 0. \tag{1.14}$$

The reason is that with conformity concerns, judges also incorporate beliefs about other judges' biases $E[b_i]$ in order to match their scores more closely. This prediction is consistent with the results in Sandberg (2018), who finds that judges in dressage competitions favor athletes of the same nationality as other judges on the panel.

In our context, one may therefore also expect conformity effects to be particularly strong when judge biases can be easily inferred, such as when there are matching nationalities.

1.3.5 Further potential channels of transparency

Transparency may also affect judge behavior through other mechanisms that are not explicitly included in our model. In the following, we will briefly discuss some of these mechanisms and how they may affect our theoretical predictions.

Appealing to the home constituency. Public monitoring generally induces individuals to behave more in accordance to prevailing norms and expectations, but these might not necessarily encourage impartiality. For example, audiences in the judge's home countries and the national federation that appointed the judge may in fact expect him or her to favor compatriot skaters and discriminate against rival skaters (Zitzewitz, 2006).[13] If this was the case, we would expect transparency to lead to an increase in nationalistic judging and an increase in score dispersion for performances with a compatriot judge on the panel, contrary to the predictions of our model.

Exaggeration and counterexaggeration. When there is a potentially biased judge on the panel, other judges can in fact react to this strategically by biasing their scores in the opposite direction if they have fairness concerns for the aggregate score awarded to skaters (Li, Rosen, and Suen, 2001; Rausser, Simon, and Zhao, 2015). Transparency could potentially break such feedback loops of bias and counterbias, which would also predict a decrease in score dipersion for a given performance, though mostly concentrated on performances where the presumed biases are particularly strong, e.g. when there is a compatriot judge on the panel. Note, however, that some previous studies on the behavior of sports judge panels find that non-compatriot judges may in fact move their scores closer towards those of the compatriot judge instead of the opposite (Zitzewitz, 2006; Sandberg, 2018).

Vote trading and rigging. Transparency can also facilitate corruption, e.g. by rigging or vote trading, because potential bribers can now verify whether the bribed judge actually followed through, and colluding judges can better monitor each others' behavior and implement repeated game strategies.[14] However, assuming that

13. Dohmen (2008b), for instance, finds that football referees exhibit home team favoritism, in particular when the physical distance of the public crowd to the field is smaller, and when the crowd consists of supporters of the home team. Benesch, Bütler, and Hofer (2018) find greater party discipline after the transparency reform in the Swiss Upper House, even though this is not necessarily in line with the preferences of the median cantonal voter. Stasavage (2007) finds that in a model with biased and unbiased experts, unbiased experts only vote truthfully under public voting if reputational concerns are sufficiently weak.

14. In fact, anonymous voting was first introduced by the ISU in 2002 precisely in response to a vote trading scandal at the Salt Lake City Olympics, where a French judge admitted (though later

vote trading strategies need to be sophisticated enough that they are not easily detectable, it is difficult to predict how observed scoring patterns would be affected. Since collusion and cheating are risky endeavors with uncertain success chances, given the limited impact of individual judges, it seems unlikely that this would cause strong universal changes in observed judging behavior.

1.4 Data and Descriptive Statistics

To study how the 2016 transparency reform affected performance scoring by judges, we obtain from the ISU website information on skaters' performances at all official ISU competitions from the 2013-14 season to the 2019-20 season. Thus, our sample includes three pre-reform seasons under the anonymous scoring regime and four post-reform seasons under the transparency regime.[15] This information includes all scores awarded by judges on the panel towards each technical element and artistic program component of the performance, as well as the identities and nationalities of the skater and of the judges.

In total, our sample comprises 16,821 skating performances by 1,905 different skaters across 127 events. A figure skating event (e.g., 2018 Winter Olympics) can typically be further broken down into four competitions, one in each of the four disciplines (Men's Singles, Women's Singles, Pairs Skating, Ice Dance), and two rounds per competition (Short Program and Free Skating).[16] Within each round, the judge panel stays constant, so all skaters performing in the same round are evaluated by the same judges. Table 1.1 further breaks our sample down into observation categories according to our difference-in-differences identification strategy. We observe a comparable sample of performances in both treated Non-JGP events and untreated JGP events, although the number of observations is slightly lower for JGP events. Furthermore, as we include four post-reform and three pre-reform seasons, we have slightly more observations under transparency than under anonymity. We restrict

recanted) to having been pressured by her national federation to rank the Russian pair first in the pairs' competition, in exchange for higher votes to a French couple that would perform in the ice dance competition a few days later.

15. Though data is available until the 2005-06 season, the main presented results are restricted to observations from the 2013-14 season onwards. This is firstly due to a number of changes in event formats in the 2010-11 and 2011-12 seasons (e.g. the Compulsory Dance and Original Dance segments were replaced with the Short Dance segment; instead of holding a Preliminary Qualification Round in Senior events, qualifications were done based on scores from the Short Program after the 2011-12 season.), so that it is not possible to control for discipline × segment. Secondly, JGP (Control) skaters typically do not have long careers, so these skaters are no longer in the dataset after a few years; results with skater FEs are mainly identified from performances close to the reform period. Results using the full dataset (without skater FEs or discipline × segment controls) are presented in the Appendix.

16. Note that the number of rounds is not 8 times the number of events in our sample, because some events hold more than one competitions per discipline, whereas some (JGP) events do not hold a competition for each discipline.

Table 1.1. Number of Observations

		JGP (control)		Non-JGP (treated)	
	full sample	pre-reform	post-reform	pre-reform	post-reform
# Performances	16821	3103	4340	3994	5384
# Events	127	21	28	34	44
# Rounds	1028	152	200	292	384
# Skaters	1905	711	954	617	730
# Judges	563	333	379	323	338

Notes: This table shows the number of observations in our sample, split by JGP events and Non-JGP events before and after the 2016 reform, respectively. An event typically consists of 4 competitions, one for each discipline (Men's Singles, Women's Singles, Pairs Skating, Ice Dance), and each competition consists of 2 rounds (Short Program and Free Skating). However, some JGP events do not include a Pairs Skating competition, and some other events hold more than one competition per discipline. We exclude 520 performances for which the panel included fewer than 9 judges.

the dataset to performances from competitions where there was a full panel of 9 judges.[17]

Table 1.2 presents descriptive statistics for the performance scores in our sample. The average Program Component Score (PCS), i.e., the artistic score, is about 38.08 over all performances, and the average Technical Elements Score (TES) is about 39.16. The average Total Score is somewhat lower than the sum of both, as skaters are sometimes punished with score deductions for rule violation. In general, scores in JGP events tend to be somewhat lower compared to Non-JGP events, reflecting the lower level of prestige and hence lower average quality of performances. Furthermore, there seems to be an upward time trend for all event types, so average post-reform scores tend to be higher the average pre-reform scores.

Judges are not unanimous in their evaluation decisions. As measure of disagreement about a performance in the panel we calculate the within-panel standard deviation (Panel SD), i.e., the score dispersion across judges for any given performance: $\sigma_p = \sqrt{\frac{1}{9}\sum_{j=1}^{9}(\pi_{pj} - \bar{\pi}_p)^2}$, where π_{pj} is the score awarded by judge j towards performance p. From Table 1.2, we can see that the mean Panel SD is about 1.75 for the PCS and 1.33 for the TES over all performances, reflecting the subjective nature of the sport. Another way to illustrate the magnitude of dispersion is by the calculating the gap between the highest and the lowest score in the judge panel for the same performance: this gap is 5.73 points for the PCS and 4.31 points for the TES. Notice that there is generally less disagreement on the more objective technical score compared to the artistic score. Notice also that the mean Panel SD of artistic scores

17. Due to budget constraints, some competitions (typically JGP) have panels with fewer than 9 judges. Nonetheless, such panels are uncommon, consisting only of 520 performances. Including these observations does not lead to in any significant changes in results.

Table 1.2. Descriptive Statistics

		JGP (control)		Non-JGP (treated)	
	full sample	pre-reform	post-reform	pre-reform	post-reform
Program Component Score (PCS)					
Average score	38.08	30.95	33.09	41.06	44.00
Mean Panel SD	1.75	1.83	1.84	1.78	1.62
Compatriot mean	40.46	31.74	34.50	43.06	46.24
Technical Elements Score (TES)					
Average score	39.16	31.09	33.72	42.08	46.04
Mean Panel SD	1.33	1.03	1.18	1.40	1.56
Compatriot mean	41.61	32.02	35.56	43.78	48.16
Total Score					
Average score	76.75	61.42	66.19	82.73	89.67
Mean Panel SD	3.13	2.98	3.13	3.20	3.17
Compatriot mean	81.62	63.18	69.52	86.42	94.04
% Compatriot	61	54	52	66	68

Notes: This table shows the number of observations in our sample, split by JGP events and Non-JGP events before and after the 2016 reform, respectively.

drops from 1.78 to 1.62 in Non-JGP events after the transparency reform was introduced, whereas it stayed nearly unchanged in JGP events that were not affected by the reform.

Finally, Table 1.2 also shows the mean scores for compatriot performances, defined as performances for which there is at least one judge on the panel who has the same nationality as the skater. This is true for about 61% of performances in our full sample. In general, we observe that compatriot performances tend to be receive higher score relative to non-compatriot performances. Naturally, this compatriot score gap alone is no evidence for nationalistic favoritism. Countries that are traditionally strong in figure skating (such as China, Russia, USA, and Japan) are also overrepresented on judge panels, since judges are often former competitive figure skaters themselves, so a positive correlation between compatriot performances and scores is to be expected.

1.5 Empirical Strategy

1.5.1 Identification

We use a difference-in-differences approach to empirically identify the effects of the transparency reform on judges' performance evaluation behavior, using perfomances

in JGP events as control group, since deanonymized scores were already published before the 2016 reform for these events. The main identification assumption is that performance scores in treated Non-JGP events and in untreated JGP events would have followed the same counterfactual time trend in absence of the transparency reform. While JGP events are notably less prestigious than Non-JGP events, any level differences in performance score statistics between these events are not problematic as long as the common trends assumption holds. Moreover, we need to assume that the reform does not affect skaters' performance per se (in an unobservable way), but only the way judges award scores for these performances. This seems plausible given that for skaters, nothing changes about how and when they learn about their scores.

Ideally, we would study deanonymized judge scores both before and after the reform, for example to evaluate how behavior changes for a compatriot judge on the panel compared to non-compatriot judges, or how the same judge behaves under different publication regimes. Unfortunately, it is precisely the anonymization of individual judges' scores that prevents any analyses that require scores to be matched to judge identity before the reform. Therefore, we will mainly focus on judge panel-level statistics such as the aggregate score or the within-panel score dispersion as outcome variables. This implies that we are not able to identify the extent of favoritism by the compatriot judge him-/herself prior to the reform for Non-JGP events. Instead, we will investigate the *aggregate* net bias of a skaters' score when there is a compatriot judge on the panel, which may also include potential favoritism by non-compatriot judges, e.g. due to bloc-voting, as well as strategic counter-exaggerations.

1.5.2 Estimating effects on score dispersion

In our baseline specification, we estimate the following difference-in-differences model using judge score data at the performance-level:

$$\begin{aligned} \sigma_{isrp} \;=\; & \alpha + \beta_1 \cdot NonJGP_p + \beta_2 \cdot NonJGP_p \times Post_s \\ & + \delta' x_{isrp} + \varphi_s + \varepsilon_{isrp}\,, \end{aligned} \tag{1.15}$$

where σ_{isrp} is the within-panel standard deviation of scores for performance p by skater i in round r and season s. $NonJGP_p$ is an indicator variable for performances at Non-JGP events. φ_s represents season fixed effects that capture any changes in score statistics over time. The main independent variable of interest is $NonJGP \times Post_s$, which is the interaction of the Non-JGP indicator with an indicator for post-reform events (season 2016-17 onwards). Hence, β_2 is the estimated average effect of the transparency reform on the outcome of interest. We include a number of control variables such as the skater's current ISU world rank.[18] Importantly, we control for

18. Skaters' world ranks are updated by the ISU after every event, and are computed based on the skater's highest/second highest placements at various sanctioned competitions from the previous

a quadratic polynomial of the median score in the panel, as differences in score levels may be linked to higher or lower dispersion across judges, for example due to ceiling effects at the upper score bound.[19] To further test robustness, we also estimate additional specifications with skater fixed effects α_i.

1.5.3 Estimating effects on nationalistic bias

Identifying biases in performance evaluation is not a straightforward task when scores are anonymized. It is commonly suspected that figure skating judges tend to be positively biased toward skaters with the same nationality, but all we can do without knowledge of individual judges' scores is to compare the aggregate scores for performances by skaters with a compatriot judge on the panel with scores for performances by skaters whose nation is not represented on the panel. Conceptually, this gives us a measure of the *aggregate* bias in the panel that combines behavior by compatriot judges and potential responses by the non-compatriot judges.

The main complication with this comparison is that the presence (or absence) of a compatriot judge on the panel is generally also positively correlated with the skater's skill, because countries with traditionally strong figure skating athletes also tend to be overrepresented in judge panels — judges usually being former competitive skaters themselves. To identify nationalistic bias, we therefore exploit that, from the skater's point of view, the composition of the panel can be regarded as quasi-random. Thus, by including skater fixed effects, we compare scores for the same skater depending on whether he or she performs with a compatriot judge on the panel or not. To hold constant the judge panel and the general performance level of the competitors, we further include skating round fixed effects. The statistical model is then the following:

$$\pi_{irp} = \alpha_i + \beta_1 \cdot Comp_{irp} + \varphi_r + \delta' x_{irp} + \varepsilon_{irp}, \tag{1.16}$$

where π_{irp} is the artistic (technical) score a skater i received for performance p in round r, which is calculated as trimmed average score of all judges in the panel. The main regressor of interest here is the indicator variable $Comp_{irp}$, which takes the value 1 if the panel for performance p includes a judge with the same nationality as the performing skater i, and 0 otherwise. Hence, β_1 gives us an estimate of the baseline score gap. α_i and φ_r represent skater and round fixed effects, respectively. In additional specifications, we also control for a vector of other objective skater

two seasons and the current season. Some skaters are not ranked, because they placed too low in previous competitions or because they are new. To account for this, we create an indicator variable for being unranked. Communication No. 1629 (International Skating Union, 2010) provides details regarding rank point distributions.

19. We use the median rather than the (trimmed) mean score because it is more robust to outliers, which could themselves affect the standard deviation. That said, the correlation is more than 99.8%.

and performance characteristics x_{irp}, such as skaters' world rank (at the time of performance) and a home event dummy, indicating whether the event took place in a skater's home country, as well as the Base Value, which gives us a performance-level measure that sums up the difficulty of technical elements the skater chose to include in the choreography. Our most stringent specification replaces α_i with skater-season fixed effects α_{is}, thereby accounting for variation in a skater's performance levels throughout the career.[20]

To facilitate interpretation and make scores comparable across a wide range of different events, π_{irp} is normalized across rounds so that its unit is the standard deviation of scores across all performances in round r. This has the additional intuitive appeal that a one-point increase in absolute score is much more impactful for the final rankings when skaters are in a neck-to-neck competition with each other than when their scores are highly dispersed.

After estimating the net degree of nationalistic favoritism in the full sample, we ask whether the transparency reform led to reduction in bias, using the difference-in-differences approach that compares post-reform changes for Non-JGP events relative to JGP events:

$$\begin{aligned}\pi_{irp} \;=\; & \alpha_i + \beta_1 \cdot Comp_{irp} + \beta_2 \cdot Comp \times NonJGP_{irp} \\ & + \beta_3 \cdot Comp \times Post_{irp} + \beta_4 \cdot Comp \times NonJGP \times Post_{irp} \\ & + \varphi_r + \delta' x_{irp} + \varepsilon_{irp}\,.\end{aligned} \tag{1.17}$$

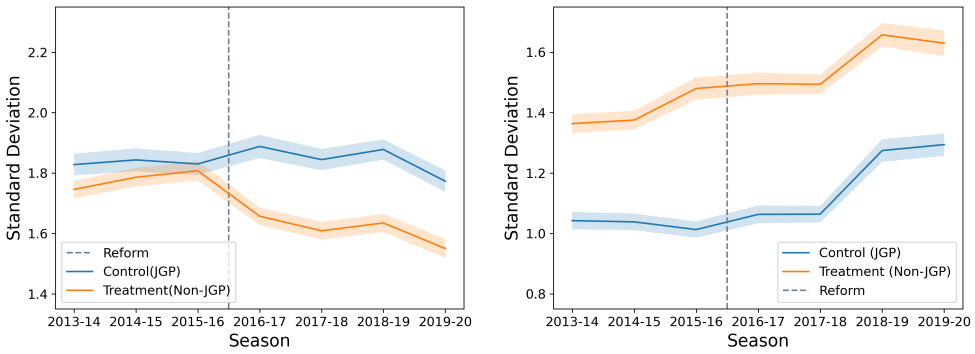
Compared to equation 1.16, we further interact the compatriot performance indicator with an indicator for Non-JGP events ($Comp \times NonJGP_{irp}$), to control for time-invariant differences between the level of favoritism between JGP and Non-JGP events, and with an indicator for post-reform events ($Comp \times Post_{irp}$), to control for common time trends. Crucially, the triple-interaction term $Comp \times NonJGP \times Post_{irp}$ allows us to estimate how the transparency reform affects the compatriot score advantage.

1.6 Main Empirical Results

1.6.1 Effects on average score dispersion

First, we examine whether the transparency reform affected the dispersion of scores across judges for the same performance. Figure 1.1 plots the average season-by-season within-panel standard deviations of the artistic score and the technical score, respectively, separately for Non-JGP and JGP performances. Reassuringly, the within-panel standard deviations seem to follow parallel trends both in the pre-reform

20. Note that this can heavily affect the implicit weights of observations when identifiying the compatriot score advantage, as for some skaters we observe few or no performances at all with/without a compatriot judges on the panel in a given season.



(a) SD of artistic score **(b)** SD of technical score

Figure 1.1. Within-panel standard deviation of scores in JGP (control) and Non-JGP (treated) events

Notes: Each point indicates the average panel standard deviation for a season, for JGP (Control, blue) and Non-JGP (Treated, orange) performances. The dashed line indicates the implementation of the transparency reform in 2016; error bars indicate 95% confidence intervals.

seasons and in the post-reform seasons.[21] But strikingly, there is a sharp drop in the artistic score dispersion for Non-JGP performances after the introduction of the transparency reform in 2016 relative to JGP performances, and this gap persists over time. This provides some first descriptive evidence that judges within a panel award more similar scores for the same performance under transparency than under anonymity. However, we observe no analogous effect for the technical score. While the pre-reform gap between treatment and control performances is much starker, the difference remains more or less constant post-reform. The general increase in the technical score dispersion from season 2018-19 onwards is likely due to a scoring reform that increases the range of possible GOEs that judges can assign from 7 points (-3 to 3 in one-point increments) to 11 points (-5 to 5 in one-point increments).

Table 1.3 presents the formal difference-in-differences estimates based on regression equation 1.15. In general, the extent of disagreement among judges follow an inverse-U shaped pattern with regard to the quality of the performance, proxied by the median score — within-panel score dispersion is highest in the middle ranges, whereas scores become more uniform when the performance was either very good or very poor. In contrast, technical score dispersion generally increases with performance quality, because grades are scaled proportionally to the difficulty of the executed elements. Additionally, we observe that the presence of a compatriot judge (with the same nationality as the skater) on the panel is associated with a small but

21. To further examine the plausibility of the parallel trend assumption, we plot in the Figure 1.A.2 season-by-season panel standard deviations (as in Figure 1.1), but with an extended pre-reform period, starting from the 2005-06 season, which is the first season under the current ISU scoring system.

Table 1.3. Effect of de-anonymized publication on standard deviation of panel scores.

	SD of Artistic Score		SD of Technical Score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.014	-0.033	0.008	-0.018	-0.009
	(0.041)	(0.043)	(0.020)	(0.021)	(0.020)
Post × Non-JGP	-0.121***	-0.103**	-0.025	-0.034	-0.009
	(0.045)	(0.049)	(0.028)	(0.028)	(0.029)
Compatriot	0.038**	0.039**	0.028***	0.025**	0.024*
	(0.015)	(0.015)	(0.008)	(0.010)	(0.012)
Median score	0.709***	0.694***	0.395***	0.368***	0.272***
	(0.053)	(0.097)	(0.025)	(0.033)	(0.032)
Median score squared	-0.099***	-0.107***	0.011***	0.015***	0.025***
	(0.007)	(0.011)	(0.004)	(0.005)	(0.005)
Constant	3.479***	3.662***	1.201***	1.209***	1.109***
	(0.111)	(0.195)	(0.034)	(0.038)	(0.042)
Skater FEs	—	Yes	—	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	1.840	1.840	1.115	1.115	1.044
Observations	16821	16764	16821	16764	12119
R^2	0.141	0.301	0.551	0.615	0.615

Notes: Estimates of equation (1.15), with standard deviation of panel scores as dependent variable. World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. Standard errors clustered at event level (e.g. Olympics 2018). Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

statistically significant increase in score dispersion by around 2%, which is hints at potential score inflation by the compatriot judge due to nationalistic favoritism.

The main coefficient of interest is $Post \times Non-JGP$, which is the indicator for treated events after the transparency reform. The estimates confirm the pattern we observe in Figure 1.1. Column (1) shows that this coefficient is negative and highly significant for the artistic score, implying that different judges award more similar performance scores in response to the reform. The coefficient of -0.121 ($p = 0.008$) is quantitatively meaningful, corresponding to an effect size of about 21% of a pre-reform standard deviation (across performances) in panel score dispersion. This decrease in score dispersion that we estimate is also robust to the inclusion of skater fixed effects in column (2), although the coefficient drops slightly to -0.103 ($p =$

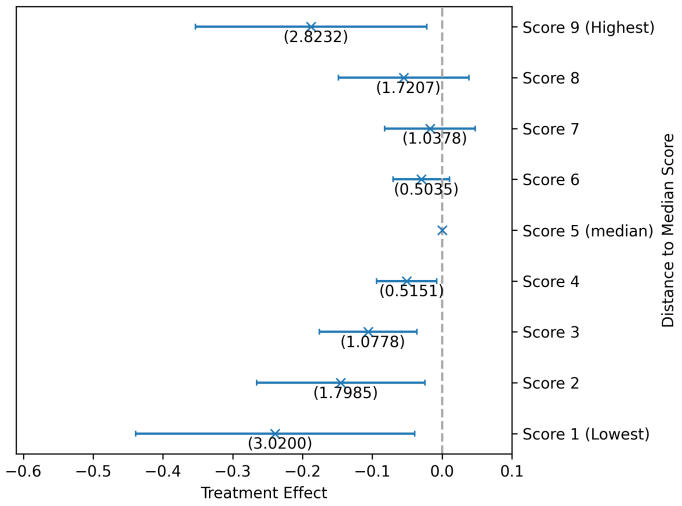


Figure 1.2. Estimated effect of transparency on distance to the median score, by ranked order

Notes: Each point plots the coefficient on Non-JGP×Post, obtained from estimating Equation (1.15) with the distance of the k-th highest(lowest) score on the panel to the median score as the dependent variable. Controls for discipline×segment, panel median score, panel median score squared and season fixed effects are included. Whiskers indicate 95% confidence intervals (adjusted for clustering at event level); figures in parentheses indicate pre-reform means for Non-JGP (Treat) performances.

0.035). In contrast, there is no effect on the within-performance standard deviation of the technical elements score. While the coefficients are always negative, indicating a decrease in score dispersion, they are quantitatively much smaller and statistically insignificant. This null result stays the same when we only include performances until season 2017-18 in column (5), due to the change in grading scales for the technical score starting from season 2018-19.

We can further break down the score compression effect of the transparency reform into effects across the full distribution of individual performance scores in the panel. To do so, we rank the nine individual scores for any given performance from lowest (1st) to the highest (9th) and calculate their distance to the median score (5th) in the panel. We then use these score distances as dependent variable to estimate the difference-in-differences model on the performance-judge level, i.e., seperately for the lowest score, second-lowest score, and so on. If, for example, a reduction in nationalistic bias was the main driver of lower average score dispersion, we may expect a disproportionate effect at the higher end of the score distribution, which is presumably where compatriot judges are likely to fall into.

Figure 1.2 plots the estimated coefficients. We can see that after the reform, scores generally becoming more closely packed around the median (for Non-JGP relative to JGP performances). Particularly the extreme scores at either end of the distribution move much closer to the center, implying a reduction in large outliers. Interestingly, the compression pattern is asymmetric, with lower scores on average

moving more upwards than higher scores move downwards. The asymmetry is not driven by large outliers and ceiling effects. If the within-panel standard deviation dropped due to a decrease in nationalistic favoritism under transparency, we would expect the opposite, namely an overproportionate effect on positive outliers rather than negative outliers.

Overall, the results in this section show that, in response to the transparency reform, judges award more similar evaluations to their peers' with regard to artistic aspects of a performance, but not with regard to the more objective technical score. This is in line with what our theoretical framework in Section 1.3 predicts. When facing greater public visibility, reputation concerns can make skaters averse to appearing incompetent or biased when their scores are too out-of-line with fellow judges, in particular in the absence of objective standards against which the public can gauge the accuracy of a judge's scores. As judges cannot communicate with each other and explicitly coordinate their scores, the question thus becomes how the conformity effect comes about. The theoretical framework suggests that higher effort exertion or collective conservatism, i.e., anchoring more towards a common prior, could be potential channels. Another potential channel is that judges curb their idiosyncratic biases toward skaters, with the most prominent source of bias being nationality. In the following, we will explore nationalistic favoritism in judge evaluations and how it was impacted by the transparency reform.

1.6.2 Effects on nationalistic bias

Next, we look at nationalistic favoritism and how the transparency reform affected the compatriot score advantage, as measured by how much higher the score is for skaters with a compatriot judge on the panel, compared to similar skaters without a compatriot judge on the panel. To make the outcome variable more comparable across rounds, we normalize scores such that one unit corresponds to the standard deviation of scores across skaters within the respective round, and the average performance in each round takes the value 0. This is intuitively appealing, as even a small positive bias in a skater's absolute scores is can result in a sizable relative advantage for the final ranking when all competitors are very close to each other, whereas it would be of little consequence when the competitors' scores are far apart from each other.

1.6.2.1 Documenting nationalistic bias

We first document a robust and statistically significant score advantage for skaters who have a compatriot judge on the panel and argue that it is likely indicative of nationalistic favoritism in performance evaluation. Table 1.2 showed that without including controls for ability and other characteristics, skaters with a compatriot judge on the panel receive on average more than 2 points higher raw score in both the artistic and technical domain, compared to their peers without a compatriot on

Table 1.4. Estimated compatriot score advantage in the full sample

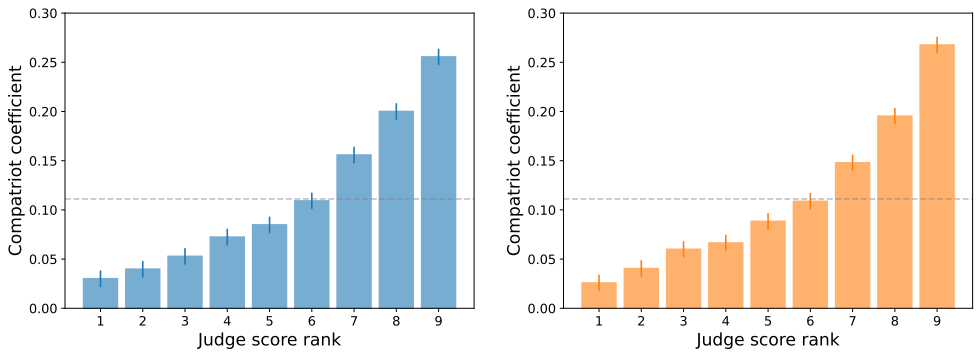
	Artistic score (std.)			Technical score (std.)		
	(1)	(2)	(3)	(4)	(5)	(6)
Compatriot	0.066***	0.046***	0.050***	0.044***	0.014**	0.020***
	(0.010)	(0.009)	(0.008)	(0.014)	(0.007)	(0.007)
Home event	–	0.084***	0.074***	–	0.067***	0.061***
		(0.018)	(0.017)		(0.013)	(0.014)
Base value (std.)	–	0.204***	0.133***	–	0.732***	0.706***
		(0.008)	(0.007)		(0.007)	(0.008)
World rank controls	–	–	Yes	–	–	Yes
Skater × Season FEs	–	–	Yes	–	–	Yes
Skater FEs	Yes	Yes	–	Yes	Yes	–
Round FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	16764	16764	16589	16764	16764	16589
R^2	0.868	0.891	0.937	0.709	0.911	0.933

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

the panel. However, this score gap could be driven by higher performance quality, as judges are more likely recruited from countries that are traditionally strong in figure skating. To control for this, we estimate equation 1.16, using skater fixed effects to adjust for differences in skater skill, as well as round fixed effects to compare between skaters who compete in the same round and are evaluated by the same panel of judges.

Table 1.4 columns 1 and 4 show that, once controlling for round and skater fixed effects, the estimated compatriot score advantage in our full sample is about 6.6% of a round-level SDs ($p < 0.001$) for the artistic score and 4.4% for the technical score ($p = 0.002$). When adding flexible controls for the skaters' current world rank at the time of competition and the Base Value, which is an objective measure of the performance difficulty, the compatriot effect drops to about 4.6% of a within-round SD for the artistic score and 1.4% for the technical score, but remains highly statistically significant. These estimates stay unchanged when using a stricter specification with skater × season fixed effects that allows us to explain more than 93% of the within-round variation in skaters' performance scores.[22] Our estimates for the aggregate nationalistic bias are quantitatively almost identical to those reported by Zitzewitz (2014).

22. Differences in average scores across rounds in themselves already explain about 85% (71%) of the variation in raw artistic (technical) scores across all skating performances.



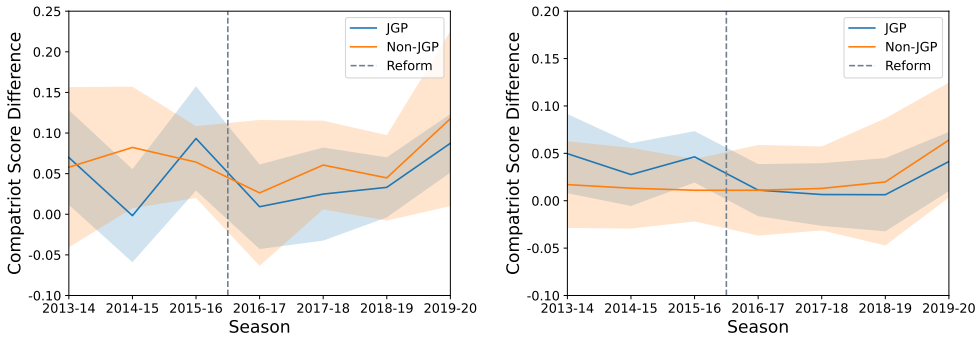
(a) Artistic score **(b)** Technical score

Figure 1.3. Distribution of compatriot score rankings towards compatriot performances

Notes: Each bar plots the coefficient from the regression of a binary variable of a particular judge score rank (1 = lowest score, 9 = highest score) against a binary variable indicating whether a judge is a compatriot judge using performance × judge level dataset, with performance fixed effects. Error bars indicate 95% confidence intervals.

To further confirm that this residual compatriot score advantage is likely driven by nationalistic bias rather than higher (unobserved) performance quality, we analyze behavior by individual judges on the panel. This restricts our sample to performances under the transparent judging regime, i.e., JGP events and post-reform Non-JGP events. Figure 1.3 plots the post-reform distribution of judges' score rankings within the panel when they evaluate performances by skaters of the same nationality as themselves. If the compatriot judge was not more likely to award higher scores to a compatriot skater, relative to other judges on the panel, the probability of each score ranking should be 1/9. However, this is clearly not the case. The distribution is heavily left-skewed for both the artistic and technical score, implying that compatriot judge often award unusually generous scores compared to the non-compatriot peers. Indeed, compatriot judges are almost four times as likely to award a score above the panel median than they are to award a below-median score.

Appendix Table 1.A.2 shows that, compared to the non-compatriot judges, a compatriot judge awards a 1.15 points higher overall artistic score and a 1.14 points higher overall GOE score on average for the same performance. Unlike Sandberg (2018), we find no evidence that skaters with a compatriot judge on the panel are evaluated more favorably even by the non-compatriot judges, but there is also no evidence for compensating fairness through strategic counter-exaggeration. Note that judges' evaluations are more impactful for the artistic compared to the technical score, as the letter is determined both by the GOE, awarded by judges, and the objective Base Value, which reflects the difficulty of the performed technical elements.



(a) Artistic score **(b)** Technical score

Figure 1.4. Compatriot score advantage for JGP (Control) and Non-JGP (Treated) events

Notes: Lines indicates the average within-round compatriot score differential by season, separately for JGP (Control) and Non-JGP (Treated) events. We regress the within-round normalized artistic (technical) score on compatriot×season dummies, including round and skater fixed effects, and controlling for home event, within-round normalized base value, squared base value, within-round normalized deductions and squared deductions. Standard errors clustered at event level. The dashed line indicates the implementation of the transparency reform; error bars are 95% confidence intervals.

1.6.2.2 Effects of higher transparency

Having documented a statistically significant and robust compatriot score advantage that is suggestive of nationalistic bias in performance evaluation, we next turn to the question of whether this score advantage was reduced by the transparency reform, which arguably allowed closer public scrutiny of compatriot judge behavior. As first descriptive evidence, Figure 1.4 plots the evolution of estimated (within-round) compatriot score differentials over time, separately for JGP and Non-JGP events. Despite some fluctuations in the order of magnitude that is statistically to be expected, JGP and Non-JGP events do seem to follow roughly similar pre-trends in the three seasons before the reform in our data, thus corroborating our difference-in-difference identification strategy. However, the visual patterns do not show any evidence for a decreasing compatriot score advantage in treated events (Non-JGP) following the transparency reform compared to non-treated events (JGP).

Table 1.5 presents our formal regression results that implement the estimation strategy described in equation 1.17. For the artistic score, we find no significant pre-reform difference in the compatriot score advantage between JGP and Non-JGP events, despite individual judges' scores from JGP events already being published openly. For the technical score, we find that the pre-reform bias is slightly stronger for Non-JGP events, if anything. Importantly, we find no evidence for a decrease in the average compatriot bias for treated Non-JGP events relative to JGP events after the reform in 2016. The estimated coefficient of 0.014 for the artistic score is statistically insignificant and goes in the opposite direction. Based on the coefficients in column 2, the implied estimate for the post-reform compatriot bias at Non-JGP events is positive (0.067) and remains statistically different from zero ($p < 0.001$).

Table 1.5. Effect of the transparency reform on compatriot score advantage

	Artistic score (std.)		Technical score (std.)	
	(1)	(2)	(3)	(4)
Compatriot	0.070***	0.035*	0.037***	0.032**
	(0.019)	(0.019)	(0.012)	(0.012)
Compatriot × Non-JGP	-0.006	0.018	-0.032*	-0.022
	(0.026)	(0.030)	(0.018)	(0.018)
Compatriot × Post	-0.042*	0.000	-0.035**	-0.024
	(0.024)	(0.023)	(0.015)	(0.018)
Compatriot × Post × Non-JGP	0.040	0.015	0.050**	0.046*
	(0.036)	(0.035)	(0.024)	(0.025)
Home event	0.072***	0.075***	0.065***	0.061***
	(0.017)	(0.017)	(0.013)	(0.014)
Base value (std.)	0.213***	0.133***	0.733***	0.706***
	(0.008)	(0.006)	(0.007)	(0.008)
World rank controls	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16589	16764	16589
R^2	0.885	0.937	0.911	0.933

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

For the technical score, the point estimate is also positive (0.046) and marginally statistically significant at the 10% level. Thus, it seems that the transparency reform was unsuccessful in achieving one of its main objective, i.e. to reduce nationalistic favoritism.

The absence of any decrease in the aggregate compatriot score advantage is consistent with our theoretical model from section 1.3, which predicts that a reduction in individual judges' favoritism may be offset in the aggregate score by conformity motives of other judges. However, due to the anonymity of judges' scores prior to the reform, we cannot, unfortunately, directly investigate how much individual judges' behavior changed due to the transparency reform. Another explanation could be that transparency triggers opposing motives for judges evaluations. For example, public scrutiny and fairness norms would push biased judges to curb their tendencies for favoritism, whereas audiences in the home country as well as national associations

that appoint the judges may in fact expect that judges behave in a biased way by skewing scores for their compatriot skaters upwards. For example, Zitzewitz (2006) provides suggestive evidence that national associations tend to appoint judges who are more rather than less biased, which can create perverse incentives for judges to favor compatriot athletes as a signal to their national association.

1.6.3 The mediating role of public attention

In the theoretical framework from Section 1.3, we assumed that the channel through which transparency affects judge evaluation behavior is through reputational concerns. This implies that the effects of the transparency reform should be particularly pronounced in highly prestigious events that generate large public attention. To test this, we extend the baseline difference-in-differences model from equation 1.15 by including interactions of the post-reform Non-JGP indicator with prestige of the competition. We proxy prestige by the average world rank of skater's performing in round r. Thus, we estimate the following regression equation:

$$\begin{aligned}\sigma_{isp} = \alpha &+ \beta_1 \cdot NonJGP_{isp} + \beta_2 \cdot NonJGP \times Post_{isp} \\ &+ \gamma_1 \cdot RoundQ \times NonJGP_{isp} + \gamma_2 \cdot RoundQ \times NonJGP \times Post_{isp} \\ &+ \sum_{k=1}^{2} \delta_k \tilde{\pi}_p^k + \varphi_s + \varepsilon_{isp}\,,\end{aligned} \tag{1.18}$$

where *RoundQ* is our proxy measure for round quality, computed using the average rank of skaters performing in the the round and, for ease of interpretation, normalized to mean 0 and standard deviation 1 for Non-JGP events. We interact *RoundQ* with the Non-JGP indicator and the post-reform Non-JGP indicator, respectively. The main coefficient of interest here is γ_2, which measures how much the treatment effect of transparency on within-panel score dispersion changes for a one standard deviation increase in round quality. Note that this is not a full triple-differences model. We notably omit the main effects for *RoundQ*, because JGP events, which serve as our control group, are generally less exclusive and prestigious than Non-JGP events; hence, the effect of higher round quality is not comparable between these classes of events, as the complete overlap condition is not fulfilled.

Table 1.6 presents the results on treatment effect heterogeneity for the within-panel dispersion of both the artistic scores and of the technical scores. We can see from columns (1) and (2) that higher event prestige indeed leads to stronger conformity in judges' artistic scores in response to the transparency reform. A one standard deviation increase in round quality is associated with an additional reduction of the within-panel standard deviation by about 0.08 points post-reform, which corresponds to around two-thirds of the effect at the mean. There is no such pattern with regard to the technical score. Overall, the patterns of heterogeneity we observe are consistent with the hypothesis that the higher degree of conformity, in the form of

Table 1.6. Heterogeneous effects on score dispersion by round prestige

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.001	-0.006	0.014	-0.025	-0.027
	(0.038)	(0.041)	(0.021)	(0.025)	(0.024)
Post × Non-JGP	-0.119***	-0.140***	-0.024	-0.032	-0.015
	(0.043)	(0.046)	(0.028)	(0.030)	(0.032)
Round quality × Non-JGP	0.071***	0.063***	0.000	-0.012	-0.016
	(0.015)	(0.017)	(0.012)	(0.014)	(0.015)
Round quality × Non-JGP × Post	-0.080***	-0.087***	0.018	0.008	-0.009
	(0.021)	(0.025)	(0.015)	(0.017)	(0.018)
Compatriot	0.035**	0.037**	0.026***	0.026***	0.028**
	(0.015)	(0.015)	(0.009)	(0.010)	(0.012)
Median score (std)	0.700***	0.620***	0.396***	0.367***	0.268***
	(0.052)	(0.094)	(0.025)	(0.033)	(0.032)
Median score (std) squared	-0.098***	-0.098***	0.011***	0.016***	0.026***
	(0.006)	(0.011)	(0.004)	(0.005)	(0.005)
Skater FEs	—	Yes	—	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.142	0.301	0.550	0.615	0.615

Notes: Estimates of Equation (1.18), with standard deviation of panel scores as dependent variable. Standard errors clustered at event level (e.g. Olympics 2018). Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

lower dispersion of (artistic) scores within the panel, is driven by stronger reputation concerns when each judge's score is published openly.

While we found no evidence in Section 1.6.2 for a decrease in the compatriot score advantage on average following the transparency reform, it is conceivable that publishing individual judges' scores also has differential effects on nationalistic judging depending on how prestigious the event is and how much public attention it thus generates. However, using average world rank of skaters as proxy for public attention as before, we do not find that the aggregate compatriot score advantage in rounds with higher prestige decreases more strongly in response to the reform (see Appendix Table 1.A.4).

1.7 Investigating Potential Mechanisms

In the previous section, we have found that the transparency reform led to a decrease in the artistic score dispersion within the judge panel. Why is this the case, especially given that judges are not allowed to communicate and coordinate with each other? The theoretical framework suggests several ways through which judges can adjust their scoring behavior to this effect, namely through effort, conservatism, or bias-matching. Which of these mechanisms is at play can lead to diametrically opposed implications for whether the reform improved or worsened the accuracy of overall scores. In this section, we present additional empirical results to further explore these mechanisms. Although we are eventually not able to isolate any specific mechanims, we will explore some of the empirical implications of each mechanism.Lastly, we show that the results are unlikely driven by selection effects due to changes in the composition of judge panels after the reform.

1.7.1 Consistency as proxy for accuracy

Judges' scores becoming more aligned with each other after the transparency reform could be an indicator for more effort and less noise, but it could also be driven by deliberate attempts to match other judges' scores in an attempt to signal competence in the absence of objectively verifiable yardsticks. Therefore, we explore another potential marker of evaluation accuracy that is arguably less salient as public signal, namely how internally consistent judges are in their evaluations. As described in Section 1.2, the artistic score (i.e., the program component score) awarded by judges is calculated from subscores for (five) different components of the performance, e.g. skating skills, interpretation of music. Likewise, the technical score is calculated from grades of execution for each technical element (e.g jump, spin) performed by the skater. Using performance-judge-level data, we can thus compute the standard deviation of the artistic (technical) subscores for each judge's evaluation of a given skater performance. A low standard deviation implies a high score consistency, which could be interpreted as confidence in judgement, whereas large variability across subscores could be an indicator for incertitude or arbitrariness.

Using the same difference-in-differences approach as for the main empirical analyses, we test whether the transparency reform lead to a decrease in subscore dispersion at the performance-judge level. Table 1.7 presents the results. We find that judges indeed become more consistent in their evaluations for artistic score components, but not the technical score components. Columns (1) and (2) show that after the transparency reform, the standard deviation of artistic components drops by 0.016 for Non-JGP performances compared to JGP performances. This effect is statistically significant at the 1% level. However, we find no effect of transparency on within-judge consistency of GOEs awarded for the different technical elements. Hence, our results on within-judge consistency are analogous to the previous find-

Table 1.7. Effect of transparency on within-judge consistency of scores

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.017***	0.012***	0.021	-0.027*	-0.026*
	(0.004)	(0.004)	(0.014)	(0.014)	(0.015)
Post × Non-JGP	-0.016***	-0.017***	0.005	-0.007	0.009
	(0.005)	(0.004)	(0.018)	(0.016)	(0.016)
Median score	0.003	-0.024***	-0.034***	-0.075***	-0.088***
	(0.002)	(0.004)	(0.003)	(0.004)	(0.005)
Median score squared	-0.001***	0.000	-0.007***	-0.005***	-0.006**
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	0.219	0.219	1.034	1.034	1.051
Observations	150458	150458	150431	150431	108675
R^2	0.041	0.090	0.233	0.360	0.342

Notes: Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

ings on the score dispersion across judges in a panel, in that we only find effects for the more subjective and more deliberately assigned artistic scores, but not for the more objective and more spontaneously assigned technical scores. Furthermore, we also find similar heterogeneity patterns as before, with effects of transparency being more pronounced for events that draw higher public attention (see Appendix Table 1.A.7).

However, some ambiguity remains as to whether more consistent scores are indeed an indicator for more accurate performance evaluations. Similarity in subscores could understate the true degree of a performance's variation in the artistic merit across different components. It could even be a mark of laziness, for example if the judge awards the same grade for every artistic subscore — although we note that this happens extremely rarely (0.11% in our sample). Finally, judges may simply use higher consistency as a cheap signaling tool to feign the appearance of competence and thoughtful evaluations (Falk and Zimmermann, 2017).[23]

23. Note that in their laboratory experiment, response consistency plausibly signals skills because consistent answers across tasks actually corresponds to the correct answers. In our context, the validity of consistency as a signal of skills would depend on how correlated (the audience perceives) the individual score components are.

To argue that the increase in score consistency is likely driven by higher accuracy, we relate it to a number of other proxies for the quality of a judge's evaluations. First, Appendix Table 1.A.8 shows that within-performance, i.e., holding constant the "actual" consistency of the skater's delivery, lower variation across artistic subscores is strongly positively related to how close a judge's score is to the median score in the panel, which is a natural evaluation benchmark for individual judges' scores.[24] This relation appears already in the baseline sample of events with anonymous score reporting, and it is more or less unaffected by the transparency reform (see Appendix Table 1.A.9). Importantly, it is not of purely mechanical nature, as consistency of *artistic* subscores also predicts closeness of the *technical* score to the panel median. Second, higher score consistency is associated with a lesser reliance on the heuristic use of whole numbers — although each artistic component can be rated on a scale from 0.25 to 10.00 in quarter-point increments, almost half (47.96%) of the actual reported subscores have integer values, pointing toward an overuse of integers as cognitive shortcut. We find that a one SD increase in artistic score consistency predicts a 7.6% reduction in the frequency of integer subscores.Third, we use the subsample of JGP events and post-reform Non-JGP events — where individual scores can be linked to judge identity — to show that more experienced judges tend to award scores with higher component consistency (see Appendix Table 1.A.11). This result is partly driven by selection effects rather than pure experience effects, i.e., selective appointment of judges to panels based on prior judging behavior.

Overall, these patterns suggest that within-judge consistency of subscores could plausibly be interpreted as rough proxy for accuracy and confidence in judgement. The transparency reform may thus have partially reduced score dispersion across judges due to genuinely higher effort and evaluation quality.

1.7.2 Conformity through social learning?

Apart from higher effort toward more accurate evaluations, another mechanism through which scores could become more similar to each other is conservatism, meaning that judges award scores that are anchored more towards a presumed consensus score (e.g., a common prior), at the potential loss of signal value from personal assessments. In practice, the question is how judges would be able to form accurate beliefs about a potential consensus score without being able to communicate with each other during performances. One possible answer is that judges can in principle observe and learn about fellow judges' tendencies over time, as the panel remains together throughout a competition round and the aggregate scores are dis-

24. While the general increase in score conformity across panels may in principle result from implicit coordination on a common prior, the current argumentation hinges on the assumption that when evaluating individual judges within the panel, it is the judges who are closer to the median that have likely been more accurate in their scoring.

played after each performance. The median (average) round includes 12 (16.4) skating performances, which gives judges a reasonable sample to receive feedback about how their own scores compared to the aggregate score. Thus, if transparency induces judges to try to move closer to each other by anticipating and guessing which scores the other judges would report, we should observe that conformity increases the later a performances occurs in a round.

This would be straightforward to test if the order of skating was random. It is, however, not — well-performing skaters tend to skate later in the round. Typically, skaters are placed into starting groups based on their world rank or their placement in the short program, with those who ranked or placed better being assigned to later groups.[25] To generate quasi-exogenous variation, we exploit that the order of performance is randomly determined within the skating groups, and thus plausibly uncorrelated to a skaters' ability, conditional on the group. Grand Prix Series and Final events form an exception, because skating orders are usually determined completely based on previous ranking or placement, so we exclude these events from our analyses in this subsection.

Thus, to test the hypothesis of conformity via social learning over time, we take the difference-in-differences specification from equation 1.15 and add interactions with skaters' starting number as well as skating group fixed effects:

$$\begin{aligned}\sigma_{irgp} \;=\; & \alpha + \beta_1 \cdot Stnr_{irp} + \beta_2 \cdot Stnr_{irp} \times Post_r \\ & + \beta_3 \cdot Stnr_{irp} \times NonJGP_r + \beta_4 \cdot Stnr_{irp} \times NonJGP_r \times Post_r \\ & + \delta' x_{irp} + \varphi_{rg} + \varepsilon_{igp}\,,\end{aligned} \tag{1.19}$$

Where $Stnr_{irp}$ is the starting number of skater i in round r, and φ_{rg} represents fixed effects for each skating group g in round r. All else is defined as before. As starting order may have an influence on the generosity of scores (Bruine de Bruin, 2006), we control for the median performance score and its square, as before. The relevant coefficient of interest here is β_4, which estimates whether the conformity effect in response to the transparency reform is stronger or weaker for performances later in a round. If the results in Section 1.6.1 are driven by social learning of judges, we should expect β_4 to be negative, indicating a larger decrease in the panel standard deviation for late performances.

Table 1.8 presents the results for both artistic and technical score. Prior to the reform, the within-panel artistic score dispersion of Non-JGP performances (but not of JGP events) tends to decrease as the round proceeds. Scaling by the average number of skaters in a starting order group, the estimates in column (1) would imply

25. The typical size of a skating group varies. Pooling short- and long-program rounds, starting-order groups tend to be larger for JGP rounds (14), compared to Non-JGP rounds (6.5). This is because JGP short program rounds have completely randomized starting numbers. Draw group sizes are similar for the long program (3.9 for both JGP and Non-JGP rounds).

Table 1.8. Heterogeneous effects on score dispersion by starting order

	SD of Artistic Score		SD of Technical Score	
	(1)	(2)	(3)	(4)
Starting number	0.001	-0.001	0.000	0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Starting number × Post	-0.003	-0.001	0.001	0.001
	(0.003)	(0.002)	(0.002)	(0.002)
Starting number × Non-JGP	-0.019***	-0.015***	-0.002	-0.000
	(0.006)	(0.005)	(0.003)	(0.004)
Starting number × Non-JGP × Post	0.020**	0.015**	0.005	0.003
	(0.008)	(0.007)	(0.005)	(0.005)
Median score	0.093***	0.082***	0.028***	0.025***
	(0.010)	(0.014)	(0.001)	(0.002)
Median score squared	-0.002***	-0.002***	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.584***	0.913***	0.244***	0.303***
	(0.150)	(0.212)	(0.028)	(0.047)
Skater FEs	—	Yes	—	Yes
Skating group FEs	Yes	Yes	Yes	Yes
Observations	12861	12788	12861	12788
R^2	0.412	0.552	0.739	0.787

Notes: Estimates of Equation (1.19), with standard deviation of panel scores as dependent variable. Standard errors clustered at event level (e.g. Olympics 2018). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

a decrease of 0.078 from the first to the last skater in the group. This could potentially be due to social learning even under anonymous scoring, as judges acquire panel-specific information on scoring with each additional skater, but alternative explanations are also possible — for example, evaluations may become less noisy when judges see more performances that they can use as reference points.

Importantly, we find no evidence of progressively stronger reductions of score conformity when the transparency reform is introduced. Indeed, the estimate on $Non-JGP \times Post \times StNr$ for the artistic score (columns 1 and 2) is positive, and quantitatively similar in absolute value to the estimated coefficient on $Non-JGP \times StNr$. Hence, we find that the tendency to award more similar scores towards later performances in Non-JGP rounds disappears post-reform. Columns (3) and (4) show that the standard deviation of the technical score does not seem to be affected by starting order in any form whatsoever.

We conclude that, for skaters of ex-ante comparable skill, the conformity effect does not seem to vary with starting number, as predicted by social learning. Instead, we find the reversed order effect for the artistic score, which may point to other mechanisms. For example, it is possible that prior to the reform, judges become more deft in their evaluations over time, as they build a reference base of comparable performances against which they can benchmark the current performance. The transparency reform could thus have induced judges to exert greater effort in evaluating earlier performances, so that the observed panel standard deviation becomes more uniform throughout the round.

1.7.3 Presence of compatriot judges

Anchoring to other judges' scores may not actually require learning and adapting over multiple performances. For example, as discussed in section 1.3, a decrease in panel score dispersion may be partly driven by judges matching the biases of other judges on the panel. This may be well anticipated ex ante, e.g. in the case of nationalistic favoritism, and therefore do not require any learning over the round. Conformity would create pressure for compatriot judges to adjust their scores downwards, and for the non-compatriot judge to move their score slightly upwards toward the biased judge, so that overall, the score dispersion decreases more for compatriot performances. This mechanism would be consistent with Sandberg (2018), who finds that judges for dressage competitions have a bias towards athletes of the same nationality as other judges on the panel. Alternatively, there might be strategic exaggeration and counter-exaggeration motives among panel judges, for example if judges with fairness concerns want to compensate for favoritism by a compatriot judge on the panel by counter-biasing. Transparency could mitigate such motives, in which case we would observe an even larger drop in the standard deviation of scores of performances with a compatriot judge on the panel. Finally, compatriot performances may simply draw larger public scrutiny, which would lend further support to the notion that reform works by triggering reputation concerns.

Table 1.9 presents results from fixed effects regressions of within-panel standard deviation on interactions between the treatment status dummies and an indicator for compatriot performances. For all specifications, we include round fixed effects, so that estimates compare skaters of similar skill and facing the same judge panel. With regard to the artistic score, we find some weak evidence to support our hypotheses that scores for compatriot performances become more uniform in response to the transparency reform. The point estimates for the compatriot triple-interaction with Non-JGP and post-reform are negative, indicating an additional conformity effect of transparency in artistic scores of compatriot performances. The coefficient is statistically insignificant, although it becomes weakly significant when including skater fixed effects. Quantitatively, it is smaller than the average treatment effect estimates in Table 1.3, so it compatriot performances alone cannot explain the aver-

Table 1.9. Heterogeneous effects on score dispersion by presence of compatriot judges

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.019	0.018	0.026**	0.017	0.014
	(0.027)	(0.031)	(0.011)	(0.017)	(0.017)
Compatriot × Non-JGP	0.066*	0.066*	0.010	0.026	0.023
	(0.036)	(0.038)	(0.015)	(0.022)	(0.022)
Compatriot × Post	-0.005	0.029	0.005	0.017	0.007
	(0.034)	(0.040)	(0.014)	(0.020)	(0.021)
Compatriot × Post × Non-JGP	-0.042	-0.087*		-0.022	-0.010
	(0.047)	(0.049)		(0.030)	(0.033)
Median score	0.091***	0.076***	0.026***	0.021***	0.018***
	(0.007)	(0.012)	(0.001)	(0.002)	(0.002)
Median score squared	-0.002***	-0.002***	0.000	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Skater FEs	—	Yes	—	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.315	0.448	0.641	0.693	0.690

Notes: Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

age score conformity effects, given that a compatriot judge is present for about 67% of Non-JGP performances and is generally even higher for very prestigious events.[26] Overall, we find some weak suggestive evidence that the effects of the transparency reform may be amplified by the presence of compatriot judges on the panel, which could be explained by bias-matching or by larger perceived public scrutiny for these types of performances.

1.7.4 Composition of judge panels

Finally, we test whether our results on the effect of higher transparency could be explained by changes in the composition of judge panel following the reform, as opposed to changes in the scoring behaviour by individual judges. The process of selecting and appointing judges to a panel is not random and not uniform across

26. Recall that the score conformity effect also tends to be stronger per se, as we have shown in Table 1.6. Additional results controlling for skater's relative rank within the round in Appendix Table 1.A.3 show that the point estimates for the compatriot skater interaction remain similar.

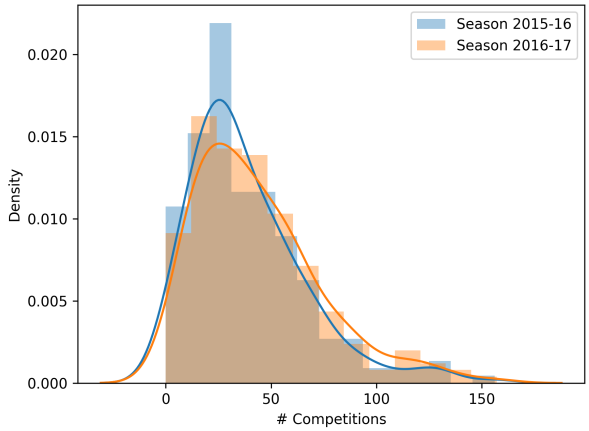


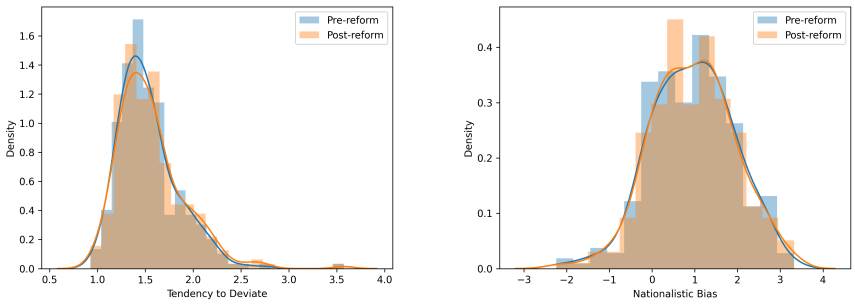
Figure 1.5. Distribution of Non-JGP judge experience around the transparency reform.

Notes: Judge experience in a season is measured as the number of competitions he or she has judged at, from season 2005-06 up to the previous season.

events. For JGP events and a small subset of Non-JGP events (the Grand Prix Series), judges are selected by the organizing country.[27] For all other Non-JGP events, judges are selected in a two-step procedure. In the first step, each national skating federation nominates a judge from their country to serve in a particular competition; next, the ISU randomly draws the required number of judges from the pool of proposed candidates. Note that under anonymization, a judge's past scores are also concealed from national skating federations and organizing countries, so that evaluations in JGP competitions were the only objective source of information that federations could use to select judges before the 2016 season.

The observed decrease in score dispersion could be caused by changes in the selection criteria of organizing countries (JGP and GP Series) or national skating federations (all other Non-JGP events) — for instance if under transparent scoring, countries or federations feel compelled to propose judges that are more experienced, less biased, or that have proven more capable in the past. Similarly, potential judges who doubt their own ability may become less willing to serve in panels when they know that their scores will be publicly disclosed. While selection effects can in general be important and meaningful consquences of a transparency reform, we provide several pieces of evidence that speak against these mechanisms.

27. Selection is subject to the restrictions that judges must come from a pool of qualified individuals ('International Judges') and that no more than one judge from their country is allowed to serve in a given competition. As the Grand Prix Series only feature very few skaters, these events only account for a small fraction of observations in our sample. Our results are robust to dropping these observations.



(a) Score accuracy proxy by judge **(b)** Nationalistic bias proxy by judge

Figure 1.6. Distribution of baseline judge-level scoring proxies

Notes: Score accuracy is proxied by a judge's average absolute deviation from the scores given by other judges on the panel. Nationalistic bias is proxied by the difference in the average deviation from other judges' scores for compatriot skater performance relative to non-compatriot performances. Both measures are based on JGP data from the seasons 2005-06 to 2012-13.

First, we check if countries become more likely to select more experienced judges, where we construct a proxy for experience using the number of competitions since the 2005-06 season (the earliest season we can observe) in which a judge has served in a panel. Figure 1.5 compares histograms of judge experience in the last pre-reform season (2015-16) to the first post-reform season (2016-17). There is no evidence that the distribution changes significantly from pre-reform to post-reform (p-value of Kolmogorov-Smirnov test = 0.1397), and inspecting the distribution of judge experience across all seasons in our sample (Appendix Figure 1.A.4) does not reveal any major upward shifts either.

Next, we investigate whether judges selected after the transparency reform differ in their revealed baseline scoring behavior. As the pre-reform Non-JGP results are anonymized, we use data from JGP events over the 2005-06 season to the 2012-13 season, where scores were transparent even before the reform. This allows us to construct individual-level judging measures for about 80% of the judges in our sample. As proxy measure of a judge's scoring accuracy, we calculate the average absolute deviation of a judge's scores from scores by the fellow judges on the panel (see, e.g., Heiniger and Mercier, 2021). As proxy measure of a judge's impartiality regarding nationalistic judging, we calculate the average deviation of a judge's scores from other scores in the panel for performances where the skater is a compatriot, relative to the average deviation in performances where the skater is not a compatriot. Figure 1.6 shows that, comparing the last pre-reform to the first post-reform season, there do not appear to be significant shifts in the distribution of judges, neither based on baseline score accurace nor on baseline bias.[28]

28. For histograms of judge scoring behaviour across all seasons in our estimation sample, see Appendix Figures 1.A.5 and 1.A.6.

Finally, we directly examine potential opting-out of Non-JGP events after the introduction of transparent scoring by following the "careers" of judges who have served in Non-JGP event prior to the reform — which also includes judges who are not represented in the previous analysis. Appendix Tables 1.A.13 and 1.A.14 show that there is no significant extensive or intensive margin decrease in judges' propensity to serve in Non-JGP event following the transparency reform. Thus, we find little overall evidence that the conformity effect induced by the transparency reform could be plausibly driven by selection effects rather than effects on individual judging behavior.

1.8 Conclusion

In this paper, we studied the effect of transparency on performance evaluation in committees in a high-stakes, professional context. Specifically, we evaluated a reform implemented in the sport of figure skating that increased the visibility of judges' decisions. Prior to the reform, judges' scores were published anonymously, thus shielding the judge from public censure or supervision. While this prevents judges from being swayed by public opinion and coerced into collusion by their fellow judges, this opacity also made it was relatively easy for judges to engage in nationalistic favoritism, so that, following accusations of nationalistic judging in the 2014 Sochi Olympics, the ISU de-anonymized result publication for all events.

To illustrate how increased visibility might impact judges' scoring behavior, we proposed a theoretical framework à la Morris and Shin (2002) with potentially biased and conformist judges, in which the transparency reform enters as an increase in conformist concerns. In line with the predictions of the model, we find that the within-performance score dispersion for artistic scores decreases sharply post-reform, indicating that judges tend to award more similar scores. In further support of a conformity-based explanation, we also see that this effect is stronger in settings with greater public attention, where judges might feel higher pressure to conform. Lastly, we find that skaters are scored higher when they have a compatriot judge on the panel, and that this compatriot advantage does not decrease post-reform. This is, at first glance, perhaps surprising, given that the reform was implemented precisely to address such concerns. However, this finding is compatible with our model's predictions, and highlights the limited impact that greater transparency can have on aggregate biases in committee decisions.

Though the sharp increase in scoring similarity is in line with previous research in different contexts, the inability of judges to communicate with each other in our setting rules out informational exchange or persuasion as mechanisms driving the conformity effect we see. Similarly, we do not find any evidence of social learning in our setting. Our model instead suggests two potential sources for this result— increased effort leading to higher signal precision, or herding on a common prior—

with largely different welfare consequences. The former leads to less arbitrary and random scoring, whereas the latter has the opposite effect, and could over time lead to a more entrenched system where performances by rookie skaters are insufficiently rewarded. We ultimately cannot distinguish between these channels with our data, and leave this as a potential avenue to explore in future research.

In general, transparency, by activating social image concerns, is a powerful tool that can be used to align individual behavior with public norms and expectations. Whether this can be successfully utilized to achieve desirable committee outcomes, however, likely depends on a variety of factors. These include, among others, the prevailing norms in the society, the degree of subjectivity of the decision, and the composition of the committee, which influence the quality of decisions made under transparency. Thus, policy makers should carefully consider the context when implementing transparency policies. However, one advantage of higher transparency is hardly disputable: it generates publicly available data for third parties like journalists and researchers and thereby potentially long-term value.

Appendix 1.A Supplementary Figures and Tables

ISU European Championships 2014

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	Javier FERNANDEZ	ESP	20	175.55	88.19	87.36	0.00

#	Executed Elements	Info	Base Value		GOE	The Judges Panel (in random order)									Ref	Scores of Panel
1	4T		10.30		-0.43	-1	0	-1	-1	1	-2	0	2	-1		9.87
2	4S+3T<	<	13.40		-0.43	0	0	-1	-1	0	-1	0	1	-1		12.97
3	3A		8.50		1.71	1	1	2	0	2	1	3	3	2		10.21
4	CSSp4		3.00		0.57	1	1	1	1	1	1	2	2	1		3.57
5	StSq3		3.30		0.79	2	1	2	1	1	2	2	2	1		4.09
6	4S		11.55	x	-2.00	-2	-2	-2	-2	-2	-2	-2	-1	-2		9.55
7	2Lz+2T		3.74	x	0.04	0	0	0	0	0	0	1	1	0		3.78
8	3Lo		5.61	x	0.80	1	1	2	1	1	1	1	2	1		6.41
9	3F+1Lo+3S		11.00	x	0.50	1	1	0	0	1	0	2	2	0		11.50
10	FCCoSp4		3.50		0.29	1	0	0	0	1	1	1	1	0		3.79
11	ChSq1		2.00		1.50	2	2	3	2	2	2	3	2	2		3.50
12	3S		4.62	x	0.40	1	0	0	0	1	0	1	2	1		5.02
13	CCoSp4		3.50		0.43	1	1	1	1	1	0	1	1	0		3.93
			84.02													**88.19**
	Program Components				Factor											
	Skating Skills				2.00	8.75	8.75	8.50	8.25	8.25	8.50	9.25	8.75	7.75		8.54
	Transition / Linking Footwork				2.00	9.50	8.75	8.75	8.25	8.75	8.75	8.75	8.00	8.00		8.57
	Performance / Execution				2.00	9.00	9.00	9.00	8.50	9.00	8.50	9.00	8.50	9.00		8.86
	Choreography / Composition				2.00	8.75	9.00	9.00	8.50	8.50	8.75	9.50	8.25	8.50		8.71
	Interpretation				2.00	9.50	9.25	9.25	8.50	9.00	9.00	9.50	8.50	8.25		9.00
	Judges Total Program Component Score (factored)															**87.36**
	Deductions:															**0.00**

< Under-rotated jump x Credit for highlight distribution, base value multiplied by 1.1

(a) Pre-reform

ISU European Figure Skating Championships 2017

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	**Javier FERNANDEZ**	**ESP**	**22**	**190.59**	**98.29**	**93.30**	**1.00**

#	Executed Elements	Info	Base Value		GOE	J1	J2	J3	J4	J5	J6	J7	J8	J9	Ref	Scores of Panel
1	4T		10.30		2.71	2	3	2	2	3	3	3	3	3		13.01
2	4S+2T		11.80		-0.20	-2	1	0	0	0	0	1	-1	-1		11.60
3	3A+3T		12.80		0.86	1	1	1	1	0	1	2	-1	1		13.66
4	CSSp3		2.60		0.43	1	1	1	1	0	1	2	0	1		3.03
5	ChSq1		2.00		1.50	2	2	1	2	2	3	3	2	2		3.50
6	4S		11.55	x	-4.00	-3	-3	-2	-3	-3	-3	-3	-3	-3		7.55
7	3A		9.35	x	-0.86	-2	0	-1	-1	-1	-1	2	-1	-1		8.49
8	3Lz		6.60	x	1.10	1	2	1	1	2	1	3	2	2		7.70
9	3F+1Lo+3S	!	11.22	x	0.00	0	0	0	0	-1	0	1	0	0		11.22
10	FCCoSp4		3.50		0.36	1	1	0	1	0	1	1	1	0		3.86
11	3Lo		5.61	x	-0.80	-2	-1	-1	-1	-2	-1	-1	-1	-1		4.81
12	StSq4		3.90		1.60	3	2	1	2	2	3	3	2	2		5.50
13	CCoSp4		3.50		0.86	2	1	1	2	1	2	2	2	2		4.36
			94.73													**98.29**
	Program Components				Factor											
	Skating Skills				2.00	9.50	9.25	8.75	9.25	9.00	9.00	9.50	9.50	9.50		9.29
	Transitions				2.00	9.75	9.00	8.75	9.25	8.75	9.00	9.50	9.25	9.50		9.18
	Performance				2.00	9.75	9.00	9.00	9.50	9.00	8.00	9.50	9.25	9.25		9.21
	Composition				2.00	9.50	9.50	9.25	9.50	9.25	9.25	10.00	9.50	9.50		9.43
	Interpretation of the Music				2.00	10.00	9.25	8.50	9.50	9.50	9.50	10.00	9.50	9.50		9.54
	Judges Total Program Component Score (factored)															**93.30**
	Deductions				Falls: -1.00 (1)											**-1.00**

x Credit for highlight distribution, base value multiplied by 1.1 ! Not clear edge

(b) Post-reform

Figure 1.A.1. Online publication of results for Non-JGP (Treat) events pre- and post-reform.

Notes: Notice that the order of panel judges is not revealed in panel (a), while it is revealed in panel (b). This order can be linked back to the individual judges on the panel.

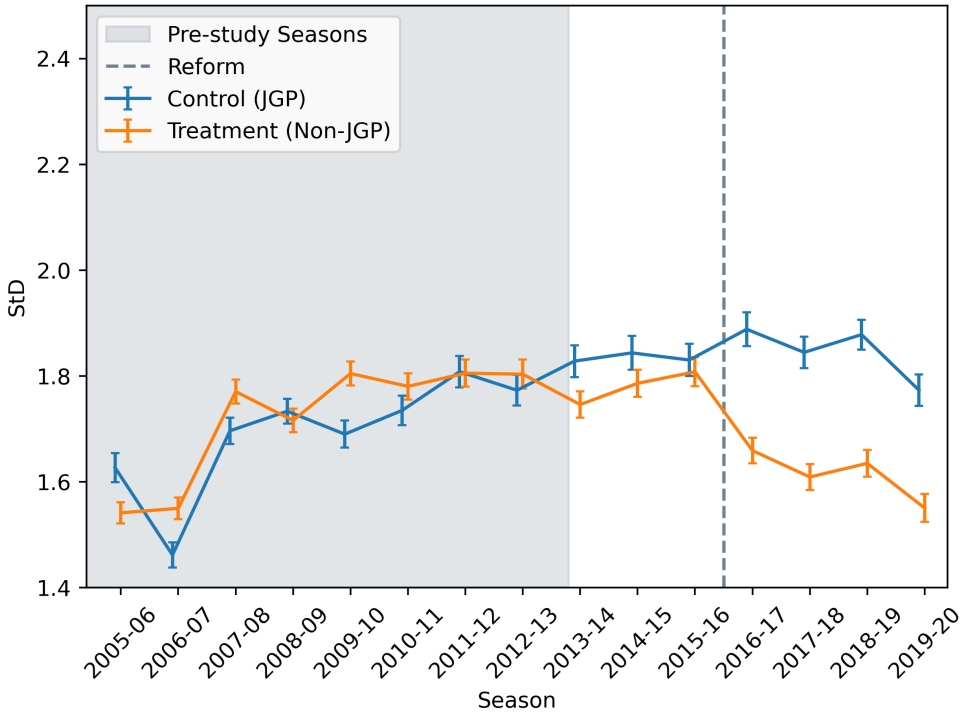
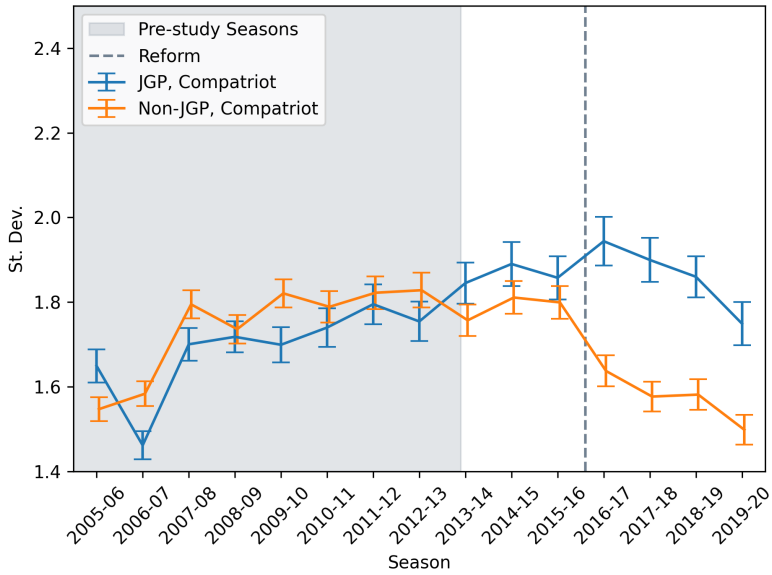
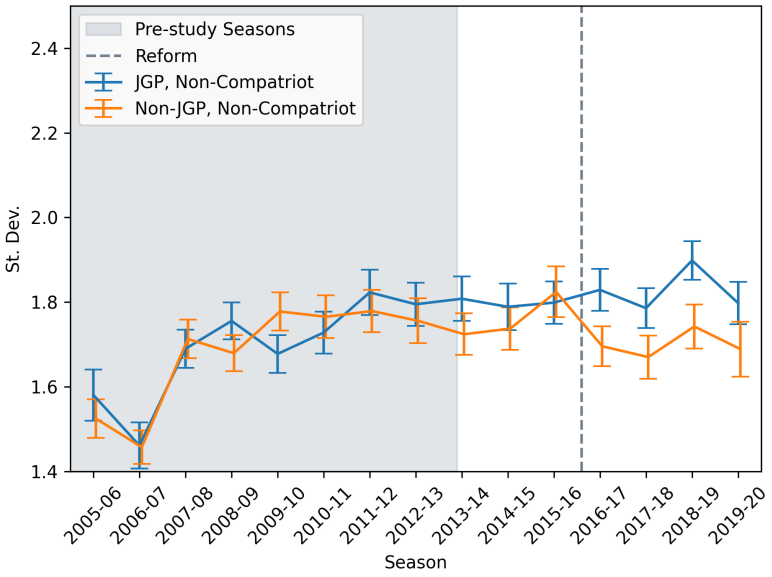


Figure 1.A.2. Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2005-06 to 2019-20

Notes: Each orange(blue) point plots the average panel standard devation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.



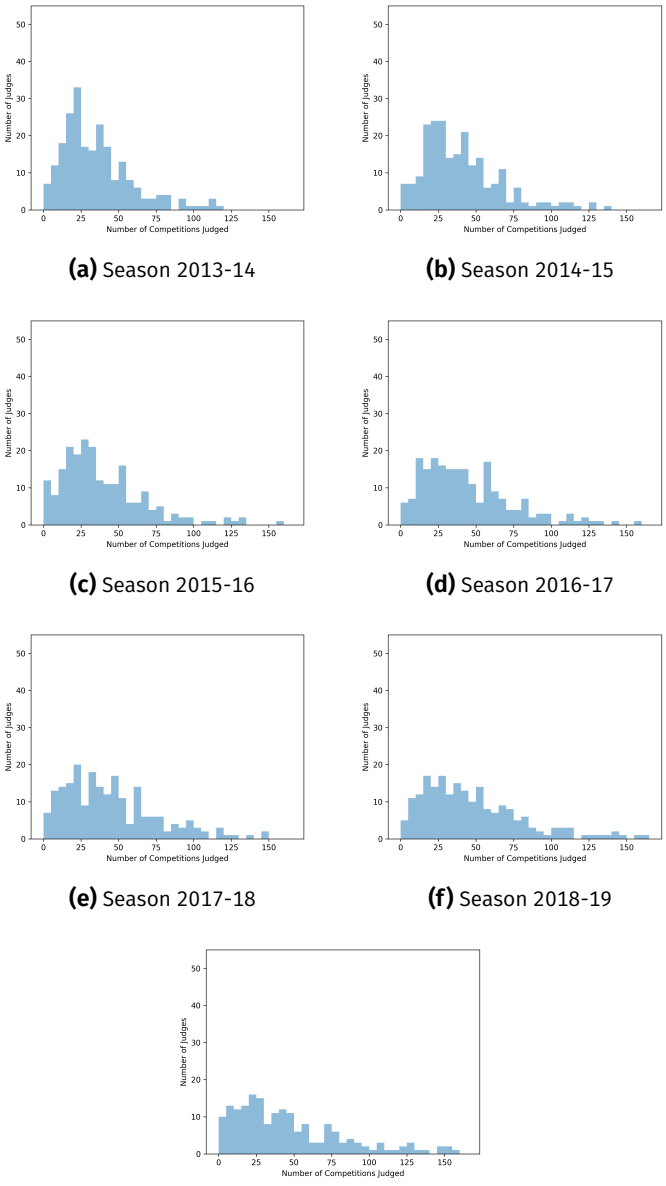
(a) Compatriot



(b) Non Compatriot

Figure 1.A.3. Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2005-06 to 2019-20, split by presence of compatriot judge on panel.

Notes: Each orange(blue) point plots the average panel standard deviation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.



(a) Season 2013-14

(b) Season 2014-15

(c) Season 2015-16

(d) Season 2016-17

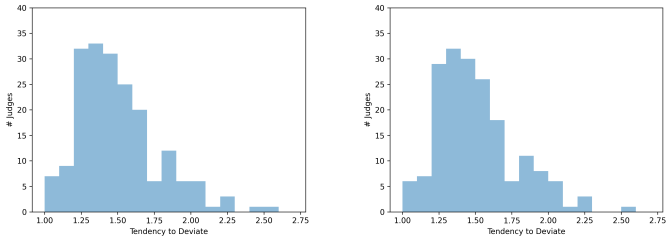
(e) Season 2017-18

(f) Season 2018-19

(g) Season 2019-20

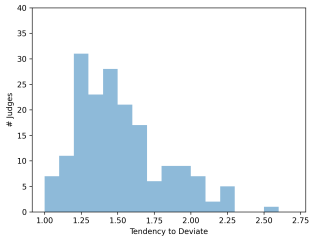
Figure 1.A.4. Distributions of Non-JGP (Treat) judge experience by season, from seasons 2013-14 to 2019-20.

Notes: Judge experience in a season is computed as the number of competitions he/she has judged at up until that season.

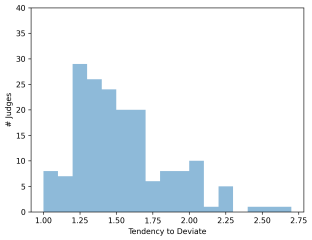


(a) Season 2013-14

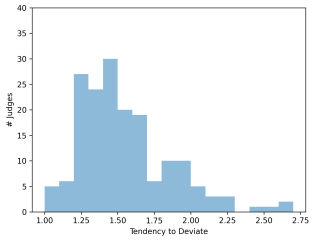
(b) Season 2014-15



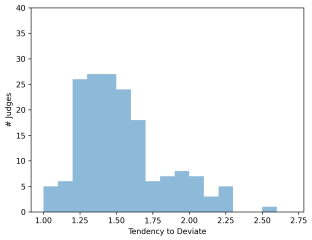
(c) Season 2015-16



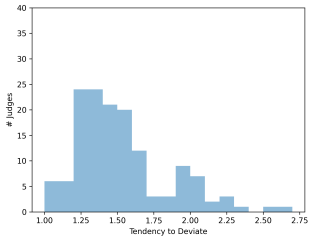
(d) Season 2016-17



(e) Season 2017-18



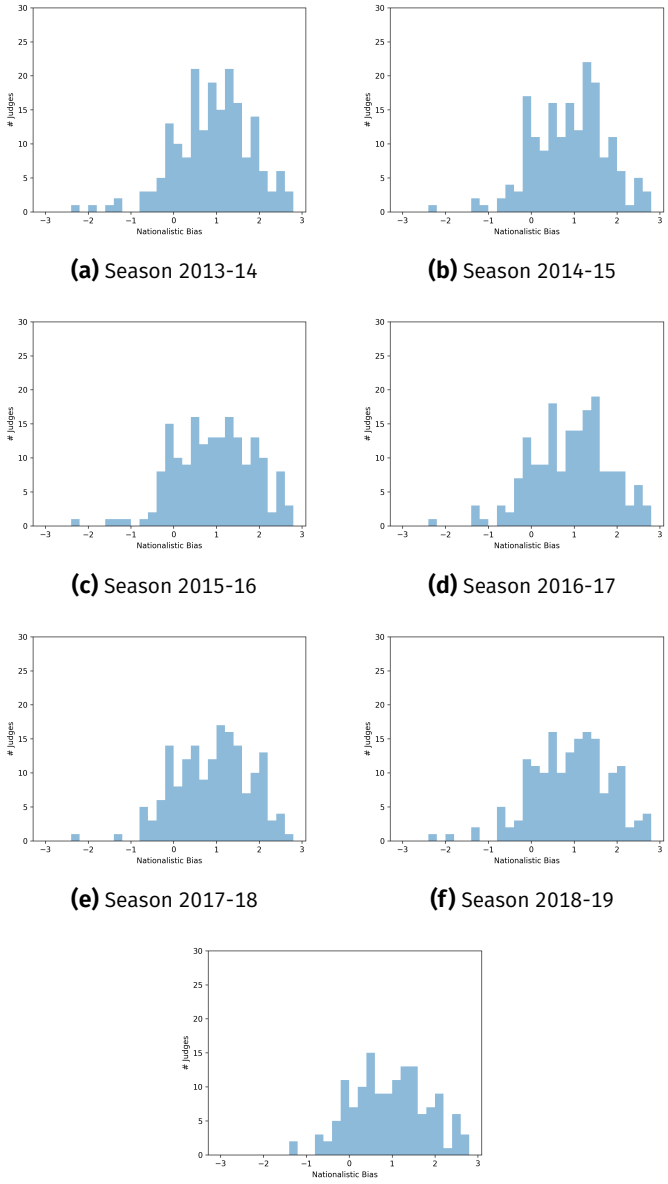
(f) Season 2018-19



(g) Season 2019-20

Figure 1.A.5. Distributions of Non-JGP score accuracy by season, from seasons 2013-14 to 2019-20.

Notes: For each judge, his/her measure of deviation is the average deviation of all performances where he/she has judged in, where his/her deviation in a performance is calculated as the absolute value of his score from that of the leave-one-out panel mean.



(a) Season 2013-14

(b) Season 2014-15

(c) Season 2015-16

(d) Season 2016-17

(e) Season 2017-18

(f) Season 2018-19

(g) Season 2019-20

Figure 1.A.6. Distributions of Non-JGP nationalistic bias by season, from seasons 2013-14 to 2019-20.

Notes: For each judge, his/her measure of (nationalistic) impartiality is the average deviation from the leave-one-out panel mean when the skater is compatriot, minus the the average deviation from the leave-one-out panel mean when the skater is non-compatriot.

Table 1.A.1. Estimated compatriot score advantage in the full sample

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot	0.052***	0.052***	0.033**	0.033**
	(0.009)	(0.009)	(0.014)	(0.014)
Home event	0.083***	0.083***	0.115***	0.109***
	(0.018)	(0.018)	(0.024)	(0.024)
World rank controls	—	Yes	—	Yes
Skater × Season FEs	Yes	Yes	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes
Observations	16589	16589	16589	16589
R^2	0.931	0.931	0.794	0.795

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.2. Compatriot score advantage

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot Judge	1.156***	1.156***	1.142***	1.142***
	(0.030)	(0.031)	(0.043)	(0.044)
Compatriot	0.068	0.055*	0.063	0.052
	(0.042)	(0.031)	(0.066)	(0.078)
Home event	0.472***	0.488***	0.504***	0.425***
	(0.070)	(0.056)	(0.112)	(0.102)
Base Value	0.135***	0.089***	1.137***	1.133***
	(0.004)	(0.003)	(0.010)	(0.011)
Controls for current world rank	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Judge × Round FEs	Yes	Yes	Yes	Yes
Observations	109296	109296	109296	109296
R^2	0.936	0.950	0.977	0.981

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.3. Heterogeneous effects within rounds

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.019	0.019	0.018	0.018	0.015
	(0.028)	(0.031)	(0.011)	(0.017)	(0.017)
Compatriot × Non-JGP	0.063*	0.061	0.022	0.022	0.020
	(0.036)	(0.038)	(0.022)	(0.022)	(0.022)
Compatriot × Post	-0.004	0.028	0.014	0.014	0.004
	(0.034)	(0.039)	(0.016)	(0.020)	(0.021)
Compatriot × Post × Non-JGP	-0.038	-0.080	-0.009	-0.014	-0.003
	(0.046)	(0.048)	(0.030)	(0.030)	(0.034)
Relative rank	0.046	-0.047	0.091***	-0.023	-0.054
	(0.055)	(0.055)	(0.030)	(0.036)	(0.036)
Relative rank × Non-JGP	0.031	0.143**	-0.027	0.075*	0.069
	(0.068)	(0.069)	(0.040)	(0.044)	(0.044)
Relative rank × Post	-0.016	0.050	0.029	0.068	0.067
	(0.065)	(0.072)	(0.035)	(0.048)	(0.057)
Relative rank × Non-JGP × Post	-0.039	-0.139	-0.158**	-0.143**	-0.120
	(0.084)	(0.088)	(0.062)	(0.063)	(0.081)
Median score	0.092***	0.075***	0.024***	0.021***	0.018***
	(0.007)	(0.012)	(0.001)	(0.002)	(0.002)
Median score squared	-0.002***	-0.002***	0.000*	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Skater FEs	—	Yes	—	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.315	0.448	0.643	0.694	0.690

Notes: Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.4. Heterogeneous effects on compatriot score advantage

	Artistic score (std.)		Technical score (std.)	
	(1)	(2)	(3)	(4)
Compatriot	0.053***	0.034*	0.035***	0.031**
	(0.020)	(0.018)	(0.012)	(0.013)
Comp. × Non-JGP	0.010	0.025	-0.031*	-0.020
	(0.031)	(0.032)	(0.018)	(0.018)
Comp. × Post	-0.034	0.002	-0.034**	-0.024
	(0.025)	(0.023)	(0.015)	(0.018)
Comp. × Post × Non-JGP	0.030	0.010	0.051**	0.047*
	(0.039)	(0.037)	(0.025)	(0.026)
Comp. × Round quality × Non-JGP	0.009	0.029	-0.004	0.008
	(0.018)	(0.021)	(0.012)	(0.012)
Comp. × Round qual. × Non-JGP × Post	0.002	-0.020	0.011	0.003
	(0.022)	(0.024)	(0.016)	(0.015)
Home event	0.084***	0.074***	0.066***	0.061***
	(0.018)	(0.017)	(0.013)	(0.014)
Base value (std.)	0.206***	0.136***	0.733***	0.707***
	(0.008)	(0.007)	(0.007)	(0.008)
Controls for current world rank	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16589	16764	16589
R^2	0.891	0.937	0.911	0.933

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.5. Effect of de-anonymized publication on variance of panel scores

	Artistic score	Technical score
	(1)	(2)
Compatriot	0.131	0.039
	(0.138)	(0.051)
Compatriot × Non-JGP	0.208	0.073
	(0.168)	(0.074)
Compatriot × Post	0.140	0.100
	(0.187)	(0.071)
Compatriot × Post × Non-JGP	-0.455**	-0.007
	(0.229)	(0.137)
Home event	-0.113*	-0.008
	(0.065)	(0.060)
Controls for current world rank	Yes	Yes
Skater FEs	Yes	Yes
Round FEs	Yes	Yes
Observations	16764	16764
R^2	0.421	0.623

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.6. Effect of de-anonymized publication on compatriot score advantage

	Artistic score	Technical score
	(1)	(2)
Compatriot	0.207*	0.321***
	(0.118)	(0.089)
Compatriot × Non-JGP	0.092	-0.331**
	(0.145)	(0.134)
Compatriot × Post	-0.122	-0.283*
	(0.147)	(0.150)
Compatriot × Post × Non-JGP	0.051	0.414*
	(0.185)	(0.238)
Home event	0.600***	0.453***
	(0.076)	(0.096)
Controls for current world rank	Yes	Yes
Skater FEs	Yes	Yes
Round FEs	Yes	Yes
Observations	16106	11568
R^2	0.962	0.984

Notes: Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.7. Heterogeneity of effects on within-judge consistency of subscores

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.029***	0.026***	0.054***	0.018	0.015
	(0.005)	(0.004)	(0.016)	(0.016)	(0.017)
Post × Non-JGP	-0.020***	-0.022***	0.003	0.017	0.026
	(0.005)	(0.005)	(0.018)	(0.019)	(0.020)
Round quality × Non-JGP	0.011***	0.009***	0.034***	0.015*	0.015*
	(0.002)	(0.002)	(0.009)	(0.009)	(0.009)
Round quality × Post × Non-JGP	-0.006***	-0.005**	0.002	0.012	0.006
	(0.002)	(0.002)	(0.010)	(0.010)	(0.011)
Median score	0.005**	-0.024***	0.079***	0.095***	0.054***
	(0.002)	(0.004)	(0.013)	(0.018)	(0.018)
Median score squared	-0.001***	0.000	-0.008***	-0.013***	-0.009***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.002)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Control group mean	0.216	0.216	1.018	1.018	1.038
Observations	150458	150458	150431	150431	108675
R^2	0.037	0.088	0.236	0.365	0.348

Notes: Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.8. Association between score consistency and score distance to the median judge

	Distance to the median judge					
	Artistic score			Technical score		
	(1)	(2)	(3)	(4)	(5)	(6)
SD of artistic subscores	1.219***		1.209***	0.407***		0.394***
	(0.082)		(0.082)	(0.097)		(0.098)
SD of technical subscores		0.102***	0.041		0.075*	0.056
		(0.026)	(0.025)		(0.042)	(0.042)
Constant	1.154***	1.320***	1.113***	1.698***	1.711***	1.644***
	(0.018)	(0.027)	(0.032)	(0.022)	(0.043)	(0.047)
Performance FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	102068	102041	102041	102068	102041	102041
R^2	0.128	0.122	0.128	0.173	0.173	0.173

Notes: Only observations under anonymous scoring are included, i.e. Non-JGP events before the 2016-17 season. Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.9. Association between score consistency and score distance to the median judge

	Artistic score	Technical score
	(1)	(2)
Artistic subscore SD	1.319***	0.572***
	(0.121)	(0.144)
Artistic subscore SD × Post	-0.255	-0.136
	(0.169)	(0.238)
Artistic subscore SD × Non-JGP	-0.055	-0.293
	(0.197)	(0.212)
Artistic subscore SD × Post × Non-JGP	0.341	0.303
	(0.255)	(0.333)
Constant	1.103***	1.817***
	(0.014)	(0.019)
Performance FEs	Yes	Yes
Observations	150458	150458
R^2	0.133	0.182

Notes: Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.10. Association between score consistency and the use of integer score

	Share of integer values in the PCS		
	(1)	(2)	(3)
SD of artistic subscores	0.076***		0.076***
	(0.006)		(0.006)
SD of technical subscores		0.002	-0.002
		(0.002)	(0.002)
Constant	0.463***	0.478***	0.465***
	(0.001)	(0.002)	(0.003)
Performance FEs	Yes	Yes	Yes
Observations	150458	150431	150431
R^2	0.122	0.120	0.122

Notes: Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.11. Effect of judge experience on within-judge consistency of scores

	SD of artistic subscores		SD of technical subscores	
	(1)	(2)	(3)	(4)
log(judge experience)	-0.0038**	-0.0027	-0.0050**	0.0046
	(0.0017)	(0.0026)	(0.0019)	(0.0032)
Constant	0.2230***	0.2197***	1.0323***	1.0046***
	(0.0047)	(0.0073)	(0.0059)	(0.0093)
Judge FEs	—	Yes	—	Yes
Performance FEs	Yes	Yes	Yes	Yes
Observations	113728	113728	113701	113701
R^2	0.198	0.381	0.850	0.861

Notes: Experience is measured as the number of competitions in which a judge has judge at, from season 2005-06 up to the previous season. Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.12. Statistics on pool of countries submitting judges to Non-GP treatment events.

Event Type	# Country	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
European Championships	Outgoing	3	3	1	4	3	3	N.A.
	From Previous Season	N.A.	23	24	25	25	24	24
	Incoming	N.A.	4	2	4	2	3	4
	Total	26	27	26	29	27	27	28
Four Continents	Outgoing	7	11	8	7	8	9	N.A.
	From Previous Season	N.A.	20	19	20	19	20	20
	Incoming	N.A.	10	9	6	9	9	6
	Total	27	30	28	26	28	29	26
World Juniors	Outgoing	7	5	7	5	5	7	N.A.
	From Previous Season	N.A.	23	25	24	25	27	23
	Incoming	N.A.	7	6	6	7	3	6
	Total	30	30	31	30	32	30	29
World Championships	Outgoing	4	5	5	3	6	9	N.A.
	From Previous Season	N.A.	25	23	21	23	23	21
	Incoming	N.A.	3	3	5	6	7	8
	Total	29	28	26	26	29	30	29
Total	Outgoing	21	24	21	19	22	28	N.A.
	From Previous Season	N.A.	91	91	90	92	94	88
	Incoming	N.A.	24	20	21	24	22	24
	Total	112	115	111	111	116	116	112

Table 1.A.13. Proportion of Non-JGP (Treatment) judges remaining next season.

Season	# Judges	% Remaining Next Season	Difference Next Season	T-test p-value
2005-06	245	0.706	0.046	0.257
2006-07	238	0.752	-0.11	0.009
2007-08	240	0.642	0.054	0.228
2008-09	207	0.696	-0.052	0.248
2009-10	230	0.643	0.019	0.682
2010-11	216	0.662	0.044	0.332
2011-12	214	0.706	0.056	0.189
2012-13	222	0.761	-0.045	0.277
2013-14	229	0.716	-0.069	0.116
2014-15	218	0.647	0.028	0.545
2015-16	215	0.674	0.049	0.268
2016-17	210	0.724	-0.085	0.06
2017-18	216	0.639	-0.043	0.366
2018-19	208	0.596	N.A.	N.A.

Table 1.A.14. Number of Competitions by Non-JGP (Treatment) judges Who remain in next season.

Season	# Competitions Season	# Competitions Season + 1	Difference	T-test p-value
2005-06	5.734	4.965	-0.769	0.057
2006-07	5.067	5.017	-0.050	0.889
2007-08	5.286	5.143	-0.143	0.731
2008-09	5.118	5.201	0.083	0.853
2009-10	5.297	4.642	-0.655	0.124
2010-11	4.937	5.238	0.301	0.465
2011-12	5.060	4.589	-0.470	0.245
2012-13	4.219	4.941	0.722	0.085
2013-14	4.817	4.207	-0.610	0.152
2014-15	4.482	4.447	-0.035	0.935
2015-16	4.566	4.821	0.255	0.549
2016-17	4.724	5.493	0.770	0.110
2017-18	4.775	4.638	-0.138	0.774
2018-19	4.815	4.540	-0.274	0.566

Table 1.A.15. Share of judges for which we could construct the nationalistic bias proxy

Total	Not Found	Found	Percent Found
228	33	195	0.855263
218	35	183	0.839450
215	35	180	0.837209
209	31	178	0.851675
214	40	174	0.813084
208	36	172	0.826923
188	44	144	0.765957

Table 1.A.16. Share of judges for which we could construct the score accuracy proxy

Total	Not Found	Found	Percent Found
228	32	196	0.859649
218	33	185	0.848624
215	35	180	0.837209
209	31	178	0.851675
214	40	174	0.813084
208	35	173	0.831731
188	44	144	0.765957

References

Asch, Solomon E. 1951. "Effects of group pressure upon the modification and distortion of judgment." In *Groups, leadership and men; research in human relations,* edited by H. Guetzkow, 177–90. Pittsburgh: Carnegie Press. [4]

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review* 107 (4): 1207–38. https://doi.org/10.1257/aer.20151211. [9]

Bagues, Manuel F., and Berta Esteve-Volart. 2010. "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment." *Review of Economic Studies* 77 (4): 1301–28. https://doi.org/10.1111/j.1467-937X.2009.00601.x. [9]

Baker, Michael, Yosh Halberstam, Kory Kroft, Alexandre Mas, and Derek Messacar. 2019. *Pay Transparency and the Gender Gap: Working Paper,* 25834. National Bureau of Economic Research. https://doi.org/10.3386/w25834. [9]

Bar-Isaac, Heski. 2012. "Transparency, Career Concerns, and Incentives for Acquiring Expertise." *The B.E. Journal of Theoretical Economics* 12 (1). https://doi.org/10.1515/1935-1704.1796. [5]

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. "Unintended Effects of Anonymous Résumés." *American Economic Journal: Applied Economics* 7 (3): 1–27. https://doi.org/10.1257/app.20140185. [9]

Benesch, Christine, Monika Bütler, and Katharina E. Hofer. 2018. "Transparency in parliamentary voting." *Journal of Public Economics* 163: 60–76. https://doi.org/10.1016/j.jpubeco.2018.04.005. [4, 8, 18]

Bertrand, Marianne, Sandra E. Black, Sissel Jensen, and Adriana Lleras-Muney. 2018. "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway." *Review of Economic Studies* 86 (1): 191–239. https://doi.org/10.1093/restud/rdy032. [9]

Böheim, René, Mario Lackner, and Wilhelm Wagner. 2020. "Raising the Bar: Causal Evidence on Gender Differences in Risk-Taking from a Natural Experiment." *IZA Discussion Paper No. 12946.* [8]

Bruine de Bruin, Wändi. 2006. "Save the last dance II: unwanted serial position effects in figure skating judgments." *Acta psychologica* 123 (3): 299–311. https://doi.org/10.1016/j.actpsy.2006.01.009. [8, 38]

Bursztyn, Leonardo, and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics* 130 (3): 1329–67. https://doi.org/10.1093/qje/qjv021. [4]

Campbell, Bryan, and John W. Galbraith. 1996. "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *The Statistician* 45 (4): 521–26. [5]

Colombo, Luca, and Gianluca Femminis. 2008. "The social value of public information with costly information acquisition." *Economics Letters* 100 (2): 196–99. https://doi.org/10.1016/j.econlet.2008.01.009. [12, 15]

Darby, Michael R., and Edi Karni. 1973. "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics* 16 (1): 67–88. [5]

Dohmen, Thomas J. 2008a. "Do professionals choke under pressure?" *Journal of Economic Behavior & Organization* 65 (3-4): 636–53. [8]

Dohmen, Thomas J. 2008b. "The Influence of Social Forces: Evidence From The Behavior of Football Referees." *Economic Inquiry* 46 (3): 411–24. https://doi.org/10.1111/j.1465-7295.2007.00112.x. [18]

Dulleck, Uwe, and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." *Journal of Economic Literature* 44 (1): 5–42. https://doi.org/10.1257/002205106776162717. [5]

Falk, Armin, and Florian Zimmermann. 2017. "Consistency as a Signal of Skills." *Management Science* 63 (7): 2197–210. https://doi.org/10.1287/mnsc.2016.2459. [36]

Fehrler, Sebastian, and Niall Hughes. 2018. "How Transparency Kills Information Aggregation: Theory and Experiment." *American Economic Journal: Microeconomics* 10 (1): 181–209. https://doi.org/10.1257/mic.20160046. [4, 7]

Fehrler, Sebastian, and Moritz Janas. 2021. "Delegation to a Group." *Management Science* 67 (6): 3714–43. https://doi.org/10.1287/mnsc.2020.3665. [4]

Fernando, A. Nilesh, and Siddharth Eapen George. 2021. "Debiasing Discriminators: Evidence from the Introduction of Neutral Umpires." *Working paper,* https://doi.org/10.2139/ssrn.3620043. [8]

Garicano, Luis, Ignacio Palacios-Huerta, and Canice Prendergast. 2005. "Favoritism Under Social Pressure." *Review of Economics and Statistics* 87 (2): 208–16. https://EconPapers.repec.org/RePEc:tpr:restat:v:87:y:2005:i:2:p:208-216. [5, 8]

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102 (1): 33–48. [5]

Gersbach, Hans, and Volker Hahn. 2012. "Information acquisition and transparency in committees." *International Journal of Game Theory* 41 (2): 427–53. https://doi.org/10.1007/s00182-011-0295-5. [4, 5, 7]

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *Quarterly Journal of Economics* 133 (2): 801–70. https://doi.org/10.1093/qje/qjx045. [4, 5, 7]

Heiniger, Sandro, and Hugues Mercier. 2021. "Judging the judges: evaluating the accuracy and national bias of international gymnastics judges." *Journal of Quantitative Analysis in Sports* 0 (0). https://doi.org/10.1515/jqas-2019-0113. [12, 43]

Huber, Jürgen, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith. 2022. "Nobel and novice: Author prominence affects peer review." *Proceedings of the National Academy of Sciences of the United States of America* 119 (41): e2205779119. https://doi.org/10.1073/pnas.2205779119. [5, 14]

International Skating Union. 2010. *Communication No. 1629.* https://www.isu.org/figure-skating/rules/fsk-communications/1625-1629-world-standing-sandp-id/file. [23]

Jiang, Lingqing. 2020. "Splash with a teammate: Peer effects in high-stakes tournaments." *Journal of Economic Behavior & Organization* 171: 165–88. [8]

Kahneman, Daniel, Olivier Sobony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement.* New York: Little, Brown Spark. [14]

Kim, Jerry W., and Brayden G. King. 2014. "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring." *Management Science* 60 (11): 2619–44. https://doi.org/10.1287/mnsc.2014.1967. [14]

Krause, Annabelle, Ulf Rinne, and Klaus F. Zimmermann. 2012. "Anonymous job applications of fresh Ph.D. economists." *Economics Letters* 117 (2): 441–44. https://doi.org/10.1016/j.econlet.2012.06.029. [9]

Lee, Jungmin. 2008. "Outlier Aversion in Subjective Evaluation: Evidence From World Figure Skating Championships." *Journal of Sports Economics* 9 (2): 141–59. https://doi.org/10.1177/1527002507299203. [5, 8]

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–68. [4, 7]

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics* 9 (2): 60–92. https://doi.org/10.1257/app.20150421. [5]

Li, Hao, Sherwin Rosen, and Wing Suen. 2001. "Conflicts and Common Interests in Committees." *American Economic Review* 91 (5): 1478–97. [18]

Lichter, Andreas, Nico Pestel, and Eric Sommer. 2017. "Productivity effects of air pollution: Evidence from professional soccer." *Labour Economics* 48: 54–66. [8]

Litman, Cheryl, and Thomas Stratmann. 2018. "Judging on thin ice: the effects of group membership on evaluation." *Oxford Economic Papers* 70 (3): 763–83. https://doi.org/10.1093/oep/gpx054. [5, 13]

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the National Academy of Sciences of the United States of America* 108 (22): 9020–25. https://doi.org/10.1073/pnas.1008636108. [4]

Maida, Agata, and Andrea Weber. 2019. "Female leadership and gender gap within firms: Evidence from an italian board reform." *ILR Review,* 0019793920961995. [9]

Mas, Alexandre. 2017. "Does Transparency Lead to Pay Compression?" *Journal of Political Economy* 125 (5): 1683–721. https://doi.org/10.1086/693137. [9]

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–45. https://doi.org/10.1257/aer.99.1.112. [4]

Mattozzi, Andrea, and Marco Y. Nakaguma. 2019. "Public versus Secret Voting in Committees." *Working paper.* [4, 5, 7]

Meade, Ellen E., and David Stasavage. 2008. "Publicity of Debate and the Incentive to Dissent: Evidence from the US Federal Reserve." *Economic Journal* 118 (528): 695–717. https://doi.org/10.1111/j.1468-0297.2008.02138.x. [4, 7]

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159 (3810): 56–63. [14]

Morris, Stephen, and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review* 92 (5): 1521–34. [5, 12, 15, 44]

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101 (4): 1410–35. https://doi.org/10.1257/aer.101.4.1410. [5, 8]

Pope, Devin G., Joseph Price, and Justin Wolfers. 2018. "Awareness reduces racial bias." *Management Science* 64 (11): 4988–95. [8]

Pope, Devin G., and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review* 101 (1): 129–57. [8]

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (3): 862–77. https://doi.org/10.1257/0002828054201297. [5]

Prendergast, Canice. 1993. "A Theory of "Yes Men"." *American Economic Review* 83 (4): 757–70. [5]

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees." *Quarterly Journal of Economics* 125 (4): 1859–87. https://doi.org/10.1162/qjec.2010.125.4.1859. [5, 8]

Rausser, Gordon C., Leo K. Simon, and Jinhua Zhao. 2015. "Rational exaggeration and counter-exaggeration in information aggregation games." *Economic Theory* 59 (1): 109–46. [18]

Rosar, Frank. 2015. "Continuous decisions by a committee: Median versus average mechanisms." *Journal of Economic Theory* 159: 15–65. https://doi.org/10.1016/j.jet.2015.05.010. [7]

Sandberg, Anna. 2018. "Competing Identities: A Field Study of In-group Bias Among Professional Evaluators." *Economic Journal* 128 (613): 2131–59. https://doi.org/10.1111/ecoj.12513. [5, 8, 17, 18, 30, 40]

Stasavage, David. 2007. "Polarization and Publicity: Rethinking the Benefits of Deliberative Democracy." *Journal of Politics* 69 (1): 59–72. https://doi.org/10.1111/j.1468-2508.2007.00494.x. [18]

Suurmond, Guido, Otto H. Swank, and Bauke Visser. 2004. "On the bad reputation of reputational concerns." *Journal of Public Economics* 88 (12): 2817–38. https://doi.org/10.1016/j.jpubeco.2003.10.004. [5]

Swank, Job, Otto H. Swank, and Bauke Visser. 2008. "How Committees of Experts Interact with the outside World: Some Theory, and Evidence from the Fomc." *Journal of the European Economic Association* 6 (2-3): 478–86. [7]

Swank, Otto H., and Bauke Visser. 2021. "Committees as Active Audiences: Reputation Concerns and Information Acquisition." *Working paper.* [5, 7]

Visser, Bauke, and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics* 112 (1): 337–72. [7]

Zitzewitz, Eric. 2006. "Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making." *Journal of Economics & Management Strategy* 15 (1): 67–99. https://doi.org/10.1111/j.1530-9134.2006.00092.x. [5, 8, 13, 18, 33]

Zitzewitz, Eric. 2014. "Does Transparency Reduce Favoritism and Corruption? Evidence From the Reform of Figure Skating Judging." *Journal of Sports Economics* 15 (1): 3–30. https://doi.org/10.1177/1527002512441479. [8, 29]

Chapter 2

Prosociality predicts individual behavior and collective outcomes in the COVID-19 pandemic*

Joint with Ximeng Fang, Timo Freyer, Zihua Chen, and Lorenz Götte

2.1 Introduction

To curb the COVID-19 pandemic, individuals have to engage in costly preventive behaviors such as reducing social contacts, wearing face masks, or using contact tracing apps. However, the benefits from a lower rate of transmission accrue to society at large and thus constitute a public good. This results in a social dilemma, where "the maximization of short-term self-interest yields outcomes leaving all participants worse off than feasible alternatives." (Ostrom, 1998, p.1). In this sense, the pandemic is comparable to other collective action problems such as civic engagement or the fight against climate change.

Which factors determine the success of groups or societies in overcoming collective action problems has been a long-standing question in the social sciences. One plausible determinant is the extent to which individual members are prosocial, i.e., how willing they are to behave in a way that primarily benefits other people or society at large. Prosocial individuals may help their groups in achieving more beneficial

* We would like to thank Peter Andre, Felix Chopra, Thomas Dohmen, Luca Henkel, Sven Heuser, Sebastian Kube, Anna Schulze-Tilling, as well as participants of the IAME Applied Micro Coffee for helpful comments and suggestions already at early stages of this project. Financial support by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B07) and Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

This paper is published as Fang, X., Freyer, T., Ho, C. Y., Chen, Z., & Goette, L., 2022. "Prosociality predicts individual behavior and collective outcomes in the COVID-19 pandemic." *Social Science & Medicine*, 308, 115192. https://doi.org/10.1016/j.socscimed.2022.115192.

outcomes in the face of social dilemmas, both by contributing more to a common cause themselves and by increasing cooperation rates among other members — for example through establishing and enforcing corresponding social norms (Fehr and Gächter, 2002; Fehr and Fischbacher, 2003; Fischbacher and Gächter, 2010; Albrecht, Kube, and Traxler, 2018; Fehr and Schurtenberger, 2018). Previous studies have documented associations between (pro-)social preferences and, amongst others, pro-environmental behavior (Andre et al., 2021; Fuhrmann-Riebel, D'Exelle, and Verschoor, 2021; Lades, Laffan, and Weber, 2021), donation and volunteering decisions (Falk et al., 2018), redistributive voting (Epper, Fehr, and Senn, 2020), as well as labor market outcomes (Dohmen et al., 2008; Burks, Carpenter, and Goette, 2009; Kosse and Tincani, 2020). However, combining data of both individual- and group-level behavior and outcomes under collective action problems in real-world contexts remains challenging.

In this paper, we examine the relationship between prosociality and individual behavior as well as collective health outcomes in the context of the COVID-19 pandemic. When fighting the pandemic, governments and public health experts have recurringly appealed to people's altruistic motivations to protect others from getting infected by embracing voluntary behavioral changes. More prosocial individuals may be more likely to respond to (and propagate) such norms and appeals, and they may generally be more inclined to internalize the health externalities that their behavior imposes on others. Consistent with this, studies have found that more prosocial individuals tend to follow social distancing and hygiene guidelines more stringently (van Hulsen, Rohde, and van Exel, 2020; Campos-Mercade et al., 2021; Müller and Rau, 2021). One implication is that regions with higher average levels of prosociality in the population might be more successful in slowing the spread of the virus. This is also proposed theoretically in recent susceptible-infected-recovered (SIR) models with endogenous behavior (Alfaro et al., 2021b; Farboodi, Jarosch, and Shimer, 2021; Quaas et al., 2021). Indeed, some empirical studies provide evidence that proxies for social (or civic) capital are related to mobility flows and COVID-19 incidence rates at the subnational level (Bartscher et al., 2020; Alfaro et al., 2021a; Barrios et al., 2021; Durante, Guiso, and Gulino, 2021; Makridis and Wu, 2021), but they do not combine regional-level associations with individual-level data.

We study the role of prosociality in the COVID-19 pandemic by employing data from a representative online survey in Germany ($n = 5,843$) that we conducted during the second coronavirus wave, between mid-November and mid-December 2020. This period was characterized by steeply increasing incidence rates and a relatively lenient "lockdown light". To measure individuals' public health behavior (PHB) during that time, we included a series of questions about the extent to which they engage in physical distancing, mask-wearing, precautionary hygiene measures, self-quarantining, etc., which we then combine into a single index variable of PHB by means of a factor analysis. Although imperfect, self-reported PHB measures such as ours have been shown to be good indicators of actual behavior in the pandemic

(Jensen, 2020; Gollwitzer et al., 2021). We further use experimentally-validated survey measures by Falk et al. (2016) to elicit different components of individuals' prosocial preferences and beliefs — altruism, trust, positive reciprocity, and indirect (negative) reciprocity — and collapse them into single summary measure of "prosociality".

Our data confirms that prosociality is strongly positively related to compliance with recommended social distancing and hygiene measures. Due to the large sample size, we can further aggregate our survey measures to regional-level averages across NUTS-2 regions in Germany and link them to official statistical data on COVID-19 incidence and deaths reported by the Robert-Koch-Institut (RKI), the federal government agency and research institute responsible for disease control and prevention in Germany. Our focus on within-country variation has the advantage that policy mandates and regulations in response to the pandemic remain largely similar. We find that the individual-level relation between prosociality and PHB translates into better health outcomes at the regional level — the spread of Sars-CoV-2 is slower in regions where average prosociality in the population is high. This relationship is mediated by compliance with public health measures, which supports our suggested pathway of prosociality leading to greater PH compliance, which in turn leads to lower incidence rates.

2.2 Theoretical Predictions

The rates of social contact and disease transmission are key parameters in epidemiological models, namely the susceptible-infected-recovered (SIR) model and its various modifications (Kermack and McKendrick, 1927; Keeling and Rohani, 2011), but they are typically determined exogenously and do not respond to voluntary behavioral adaptation by individuals in a pandemic.

Canonical SIR models can be extended by endogenizing behavioral responses of forward-looking agents who face a trade-off between utility from social contacts and disutility from increased risk of getting infected (e.g., Bauch and Earn, 2004; Fenichel et al., 2011; Jones, Philippon, and Venkateswaran, 2021). To protect themselves, individuals may choose to engage in preventive health behaviors even in the absence of government restrictions. However, individuals' actions also impose health externalities on others, and social costs of infections can exceed private costs significantly — e.g. for young and healthy individuals in the COVID-19 pandemic. Hence, behavioral adaption due to purely self-interested motives (i.e., avoiding to get infected) only flattens the infection trajectory to a limited extent.

Recent theoretical studies have explicitly incorporated prosocial motives in SIR models with endogenous behavior (Alfaro et al., 2021b; Farboodi, Jarosch, and Shimer, 2021; Quaas et al., 2021). Agents in these models are not only concerned about their own health, but also about other people's health. Thus, they partially

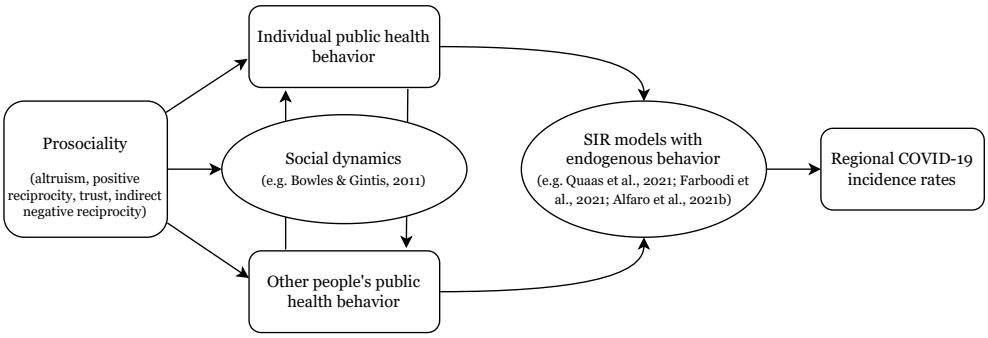


Figure 2.1. Framework

internalize the health risks that their own behavior imposes on susceptible individuals around them. This is particularly relevant for people who are uncertain about whether they are susceptible or infectious (e.g., due to asymptomatic cases and limited testing capacities), which applies to the majority of the population during our study period, since most people in Germany had not experienced a COVID-19 infection yet. To prevent that they unknowingly spread the virus, prosocial agents endogenously engage in lower levels of (risky) social activity.

While prosocial engagement in social distancing follows from an assumption on exogenously given preferences in these models, it can also be derived more explicitly from theories of human behavior that take a stance on where preferences to behave prosocially come from (e.g., Batson and Powell, 2003). For example, as an anonymous referee pointed out to us, a link between individuals' prosociality and their public health behavior can be explained by different variants of consistency theory (Festinger, 1957; Heider, 1958; Abelson et al., 1968). Specifically, individuals who hold strong prosocial values and attitudes may experience cognitive dissonance if they do not adjust their behavior in the pandemic accordingly.

In this empirical study, we consider several distinct components of prosociality that all reflect a positive disposition towards others: altruism, positive reciprocity, trust, and indirect (negative) reciprocity. Altruism constitutes a direct concern for others' well-being and links most closely to the above-mentioned models. Positive reciprocity is the tendency to return favors, which can facilitate norms of conditional cooperation (Bowles and Gintis, 2011). Trust is a composite trait reflecting preferences as well as beliefs about whether other people in general hold good intentions; higher generalized trust may encourage individuals to behave more prosocially towards friends and strangers alike. Indirect negative reciprocity describes the willingness to punish those who treat others unfairly and act detrimentally to the group. In the context of the pandemic, this could for example entail confronting others who disregard rules or norms regarding mask wearing and social distancing. This sort of third-party punishment can deter norm violation and free-riding and is therefore considered to be prosocial (Fehr and Gächter, 2002; Albrecht, Kube, and Traxler,

2018). In summary, as illustrated in Figure 2.1, individuals' prosocial attitudes can positively affect compliance with health measures both directly, out of concern for not (unintentionally) infecting others, as well as indirectly, through the social dynamics of cooperation and norm adoption. Thus, our first prediction is that more prosocial individuals are more likely to engage in preventive health measures in the pandemic.

Through the lens of a SIR model with endogenous behavior, increased compliance due to higher prosociality leads to a lower rate of disease transmission and thus fewer infections in the population, all else equal. In a dynamic setting, this positive effect is dampened, as lower incidence rates will reduce perceived infection risks and thus subsequent readjustment towards more social interactions. However, it can be shown that higher prosociality will still lead to a flatter infection curve in equilibrium (Alfaro et al., 2021b; Farboodi, Jarosch, and Shimer, 2021; Quaas et al., 2021). Thus, our second prediction is that infection rates will tend to be lower in regions with more prosocial individuals.

There are many other determinants of health behavior that are not considered in Figure 2.1. Importantly, the models highlight that behavior should adapt strongly to the perceived threat of COVID-19, which can vary based on the contemporaneous regional incidence rates and based on heterogeneity in expected health/mortality risks, e.g. due to age. Furthermore, time and risk preferences also play a role, as more patient individuals place a higher weight on future risks of infection (relative to immediate utility from social interactions) and more risk averse individuals shy away from uncertain consequences of a potential infection. Indeed, previous empirical studies have found positive associations of patience and risk aversion with better health behaviors and outcomes both in the COVID-19 pandemic (e.g., Chan et al., 2020; Alfaro et al., 2021a) and in other health-related domains such as smoking or obesity (e.g., Khwaja, Sloan, and Salm, 2006; Burks et al., 2012; Sutter et al., 2013; de Oliveira et al., 2016).

2.3 Data and Measurements

2.3.1 Survey Data

We partnered with the market research firm Dynata to recruit a target sample of 6000 German participants and conducted our web-based survey between November 11 to December 17, 2020. Participants were invited via email and sampled using demographic quotas on age, gender, and state, to achieve national-level representativeness of the population aged 18 to 65. Our final analysis sample consists of 5,843 responses that fulfilled the quality criteria for inclusion in the analysis: a minimum response duration, passing an attention check, no inconsistencies in demographic information, and no excessive straightlining.

To measure health behavior in the pandemic, we obtain responses (on a 7-point Likert scale) to ten questions about subjects' social distancing, hygiene behavior, etc. These questions were selected based on public health guidelines in Germany at that time. Using responses to these questions, we then construct an index by factor analysis. This index is our main measure of compliance to PHB. The eigenvalue of the first factor is 4.47 (0.25 for the second factor), which points towards a single underlying factor driving adherence to different PH measures. The Cronbach's α is 0.87, indicating that the different aspects of PHB are strongly interrelated.

We elicited subjects' time, risk, and social preferences using experimentally validated measures that have been employed in a large-scale representative global survey (Falk et al., 2016; Falk et al., 2018). Although the validation was conducted in a German student sample, it is plausible that the measures remain informative in our context, as language and culture are constant and there is no evidence that insights from student experiments fundamentally misrepresent behavior in the general population (Exadaktylos, Espín, and Branas-Garza, 2013; Falk, Meier, and Zehnder, 2013). To construct an individual-level measure of prosociality, we follow Falk et al. (2018) and Kosse and Tincani (2020) and combine several facets of social preferences and beliefs — altruism, trust, positive reciprocity, and indirect (negative) reciprocity — into one index variable by extracting their first principal component (eigenvalue = 1.789). This component places positive weight on all input variables and is thus congruent with the common notion of prosociality. We deviate from previous studies by also including indirect negative reciprocity, which reflects altruistic punishment and is positively correlated with our measure of altruism ($\rho = 0.257$, see Appendix Table 2.A.1).

We further collected information on demographic characteristics, education, income, political attitudes, beliefs and attitudes towards the COVID-19 pandemic, news consumption, conspiracy mentality, and Big Five personality factors. We construct the Big Five personality traits of openness, conscientiousness, neuroticism, agreeableness, and extraversion using the 15-item BFI-S scale by Gerlitz and Schupp (2005). See Appendix 2.B for a detailed description of all survey questions and variables.

2.3.2 Regional-Level Aggregation

For regional-level analyses, we aggregate our survey measures at the administrative NUTS-2 region level in Germany (38 regions; visit https://ec.europa.eu/eurostat/web/nuts/background for information on the NUTS classification system) by calculating the average of all respondents who currently live in that region. The sample size per region ranges from 46 to 427 (mean 154, median 124). We use sampling weights from a raking procedure (Battaglia, Hoaglin, and Frankel, 2009) to improve regional representativeness by age and gender (age above/below 40 × gender) as well as the share of adults with a college degree. To validate the regional represen-

tativeness of our sample, we compare vote shares of the main political parties in the 2019 election with the implied vote shares in our survey based on self-reported party preferences (Appendix, Table 2.A.7). The regional correlations are extremely high — ρ between 0.76 and 0.86 — for all parties except for the FDP, the German liberal party ($\rho = 0.29$).

We further obtain information on the official daily number of confirmed COVID-19 cases and deaths at the county-level (NUTS-3 region) reported by the Robert-Koch-Institut (RKI), the federal government agency and research institute responsible for disease control and prevention in Germany. We use data obtained from infas360 to construct a local policy stringency index by summing up a total of 23 indicator variables for whether local mandates in a certain category (e.g. curfew, school closure) were in place. We normalize this index to range between 0 (no restriction) and 100 (full restriction). Finally, we collect a host of demographic information and socio-economic indicators for each county in Germany from the joint database of the statistical offices of the German states. See Appendix 2.C for detailed descriptions of regional-level data.

2.4 Individual-Level Prosociality and Public Health Behavior

We begin by establishing a robust positive relationship between prosociality and PHB at the individual level using data from our representative online sample. To do so, we regress the PHB variable on our measures of prosociality, time and risk preferences, and a number of controls, using ordinary least squares (OLS). The statistical model is

$$PHB_{ic} = \alpha + \beta_1 \cdot Prosocial_i + \beta_2 \cdot Patience_i + \beta_3 \cdot RiskT_i + \boldsymbol{\gamma}' \boldsymbol{x}_{ic} + \varepsilon_{ic}, \qquad (2.1)$$

where PHB_{ic} is the public health behavior factor for individual i (living in county c) and $Prosocial_i$ is his or her level of prosociality. $Patience_i$ and $RiskT_i$ denote her level of patience and risk-taking, respectively, which we include as these are generally correlated with prosociality (Falk et al., 2016) and may also have an influence on individual's willingness to engage in preventive health measures. $\boldsymbol{x}_{ic}$ is a vector of control variables that differ by specifications. Standard errors are always clustered at the county level.

Table 2.1 presents the regression estimates from the baseline specification in equation 2.1 without additional control variables. Column 1 shows that prosociality strongly predicts individual behavior in the pandemic, with a one SD increase in prosociality being associated with a one third SD increase in PHB ($p < 0.001$). Additionally, we find that more patient and less risk-tolerant individuals are also more likely to adhere to social distancing and hygiene measures. These results are consistent with our theoretical predictions from Section 2.2.

People who are more prosocial also tend to differ with regard to other characteristics that may be associated with differential costs and benefits of adhering

Table 2.1. Individual-Level Association between Preferences and PHB

	Public Health Behavior (PHB)				
	(1)	(2)	(3)	(4)	(5)
Prosociality	0.3356***	0.3059***	0.3071***	0.2182***	0.1611***
	(0.0162)	(0.0165)	(0.0167)	(0.0173)	(0.0144)
Patience	0.1983***	0.1969***	0.1921***	0.1689***	0.0809***
	(0.0150)	(0.0151)	(0.0150)	(0.0149)	(0.0126)
Risk-taking	-0.2095***	-0.1710***	-0.1725***	-0.1715***	-0.0785***
	(0.0141)	(0.0144)	(0.0143)	(0.0138)	(0.0107)
Socio-demographic controls	*No*	*Yes*	*Yes*	*Yes*	*Yes*
NUTS-2 region FEs	*No*	*No*	*Yes*	*Yes*	*Yes*
Big 5 personality traits	*No*	*No*	*No*	*Yes*	*Yes*
COVID-19 perceptions	*No*	*No*	*No*	*No*	*Yes*
Observations	5843	5660	5660	5660	5660
Clusters (counties)	397	396	396	396	396
R^2	0.209	0.234	0.242	0.298	0.495

Notes: In the interest of brevity, we report only the coefficients on economic preference variables here; Appendix Table 2.A.2 reports estimates on other variables included in each specification. Socio-demographic controls include age and age-squared, gender, education, income, employment status, household size, number of children, and an indicator for having children below age 16. COVID-19 perceptions include general attitudes towards the pandemic, infection experiences, and worrying about oneself, family members, and others being infected. Standard errors (in parentheses) are clustered at the county level. See Appendix Tables 2.A.3 and 2.A.4 for detailed results using individual elements of prosociality or PHB. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

to recommended PHBs. For example, infection risk and disease severity vary with demographic factors such as age or gender, whereas economic factors such as occupation, income, or household situation could determine the costs of complying with certain preventive measures. Regional differences in current and past infection rates could further influence individual behavior, e.g., if regions hit more severely have stricter policy measures in place, or have developed stricter norms in enforcing such measures. In general, all these factors tend to be correlated with prosociality and could thus act as confounders (Falk et al., 2018). However, columns 2 and 3 of Table 2.1 show that the estimated coefficient for prosociality remains stable and highly statistically significant when controlling for demographic and socio-economic characteristics as well as region fixed effects.

Apart from economic preferences, certain psychological personality traits such as agreeableness and openness from the Big Five inventory have also been linked with stronger adherence to PH measures in the COVID-19 pandemic (Nikolov et al., 2020; Zettler et al., 2022) and are also correlated with prosociality to some degree (see e.g. Appendix Table 2.A.6). However, as the estimates in column 4 of Table 2.1

show, differences in Big Five personality traits do not drive the association between prosociality and PHB. This squares with the general observation that personality traits and economic preferences seem to be partially distinct concepts (Becker et al., 2012; Jagelka, 2020), and both retain explanatory value for individual behavior in the pandemic (see Appendix Table 2.A.2).

Finally, we also investigate to which degree the role of prosociality can be explained by individuals' perceptions and attitudes regarding the COVID-19 pandemic (Table 2.1 column 5). However, even controlling for these factors leaves a strong association between prosociality and PHB intact.

2.5 Regional-Level Prosociality and Collective Health Outcomes

In the next step, we examine how regional variation in prosociality across Germany relates to public health outcomes during the COVID-19 pandemic. For this purpose, we construct regional averages of our prosociality and PHB measures by aggregating individual survey responses at NUTS-2 level ("Regierungsbezirk") as described in Section 2.3.

2.5.1 Descriptive Overview

We document substantial variation in our measure of prosociality within Germany, as illustrated by the map in Figure 2.2a. Average prosociality ranges from −0.37 to 0.42 across NUTS-2 regions, thus spanning about 80% of an individual-level standard deviation. These regional differences are statistically significant ($p < 0.05$) and explain about 50% additional variation in individual-level prosociality compared to other socio-demographic variables alone (Appendix Table 2.A.8). Moreover, regional prosociality patterns are related to commonly used proxies for social (or civic) capital: higher average prosociality is associated with higher voter turnout in the 2019 EU election ($\varrho = 0.3098$, $p = 0.0169$) and larger density of civic associations in 2008 ($\varrho = 0.1394, p = 0.0657$), see Appendix Table 2.A.9. Thus, our measure seems to capture stable and meaningful variation.

Figure 2.2b shows that average prosociality is closely linked with average PHB in the pandemic at the regional level. In fact, the regional-level correlation ($\varrho = 0.5795$, $p < 0.001$) is substantially stronger than what would have been predicted solely based on the unconditional individual-level correlation ($\varrho = 0.3503$, $p < 0.001$), suggesting that prosocial individuals may also raise general health compliance indirectly through social influence and normative channels.

Figure 2.2c plots the evolution of COVID-19 cases per 100,000 population in Germany over the course of the pandemic, split by regions with above-median and below-median prosociality. Incidence rates in high-prosociality regions dropped persistently below those in low-prosociality regions starting from around Nov 2020, in the period of the so-called "lockdown light", which was in place at the beginning

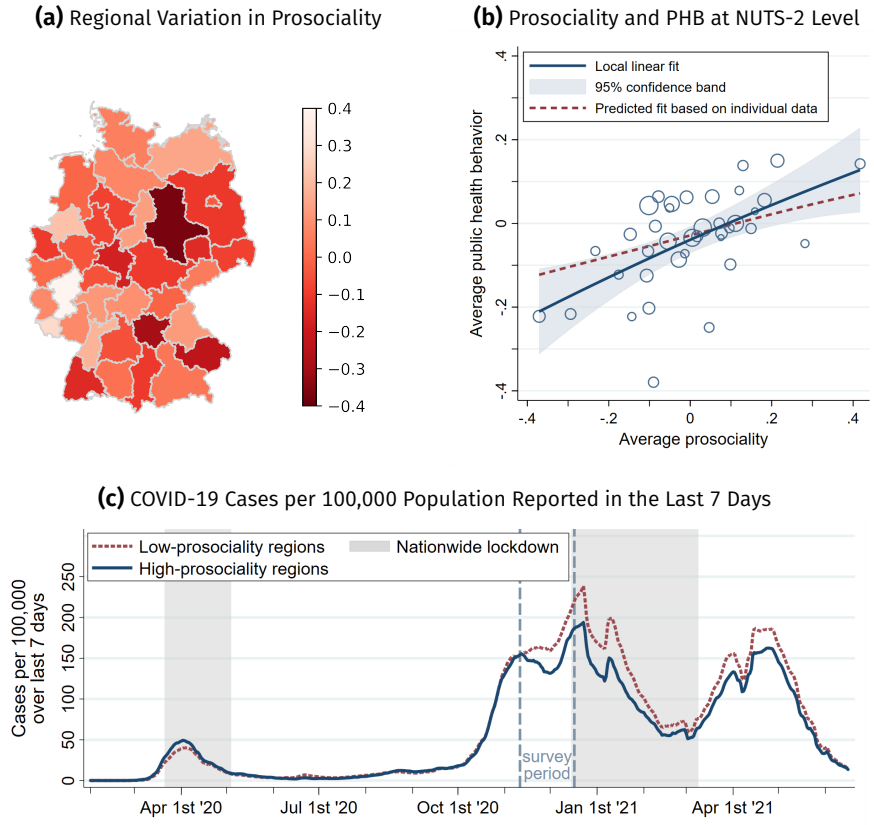


Figure 2.2. Prosociality, Public Health Behavior, and COVID-19 Incidence Rates

Notes: Panel (a): Map of the 38 NUTS-2 regions in Germany, with color intensity indicating average level of prosociality based on our survey measures. The unit is individual-level SDs. Panel (b): Relation between average prosociality and average PHB on NUTS-2 level, both expressed in terms of individual-level SDs. The solid fitted line is constructed from an unweighted local linear regression (Gaussian kernel, bandwidth = 0.3) of average PHB on average prosociality at NUTS-2 region level (N = 38). The dashed line shows the association between average prosociality and the average fitted values from an individual-level regression of PHB on prosociality and prosociality-squared. Bubbles indicate NUTS-2 regions and are proportional to population size. Panel (c): Official number of COVID-19 cases reported by RKI between Feb 1, 2020, and Jun 15, 2021. Grey shaded areas indicate time periods of strict nationwide lockdowns in Germany (as of March 8, 2021, restrictions were tied to the regional incidence rate, although the lockdown formally remained in place).

of the second wave in Germany and had the goal of reducing social contacts while avoiding a complete economic standstill. At the height of the second wave, high-prosociality regions experienced around 15-25% lower incidence rates, and 20-30% fewer COVID-19 deaths (see Appendix Figure 2.A.2, which also shows differential mobility patterns during the second wave). These descriptive observations hint at a meaningful role of prosociality in determining how well a region can slow the spread of the virus and protect vulnerable groups. However, regions with different levels of prosociality also differ by other characteristics such as population density

and socio-economic factors. Therefore, we will now move on to our formal statistical analyses.

2.5.2 Association between Prosociality and COVID-19 Incidence Rates

Our main outcome variable is the weekly COVID-19 incidence rate, i.e. the confirmed number of new cases per 100,000 population within 7 days, as reported by the RKI for each county in Germany. We additionally take the logarithm of the incidence rate to capture the exponential nature of infectious disease dynamics. Results for COVID-19 deaths are reported in the Appendix and in general very similar. As a first step in examining the relation between regional incidence rates and prosociality, we use OLS to estimate the following statistical model:

$$\log(cases_{crt}) = \alpha_t + \beta_1 \cdot \overline{Prosocial}_r + \beta_2 \cdot \overline{Patience}_r + \beta_3 \cdot \overline{RiskT}_r + \boldsymbol{\gamma}_t' \boldsymbol{x}_c + \varepsilon_{crt}\,, \tag{2.2}$$

where $log(cases_{crt})$ is the log COVID-19 incidence rate in county c (NUTS-3 level) and week t. Our main regressor of interest is $\overline{Prosocial}_r$, which is the average prosociality in NUTS-2 region r. $\overline{Patience}_r$ and $\overline{RiskT}_i$ denote the average level of patience and risk-taking, respectively. For ease of interpretation, we standardize these three preference measures to mean 0 and standard deviation 1 across regions. $\boldsymbol{x}_c$ is a vector of pre-pandemic county characteristics, which we interact with week dummies to allow the coefficient vector $\boldsymbol{\gamma}_t$ to change over time. To account for the highly dynamic nature of the pandemic, all specifications include week fixed effects α_t. We focus our analysis on the two-month period from Nov 16 to Jan 17, around the peak of the second wave in Germany, because this is when our survey measures are most applicable. Note that we include an additional month of data from the end our survey onwards, as the effects of changes in behavior or policies will only manifest themselves with a delay, which is exacerbated by reporting lags by local health authorities during Christmas and New Year. Statistical inference is robust to clustering at the NUTS-2 region level. Due to the relatively low number of clusters (38), we report confidence intervals based on a wild cluster bootstrap-t procedure (Cameron, Gelbach, and Miller, 2008; Roodman et al., 2019).

Table 2.2 presents the baseline results, which indicate a robust association between regional incidence rates and prosociality. The estimated coefficient in column 1 shows that, without controlling for any other county characteristics, a one SD higher prosociality is associated with a 13% lower weekly incidence rate in the time period we study. This effect is both statistically significant ($p < 0.001$) and quantitatively sizeable, corresponding to about 8% of the region-week SD in incidence rates(see Appendix table 2.A.16). This association remains robust to including regional-level time and risk preferences as regressors (column 2), although its precision decreases due to the covariates being correlated with each other. The estimated coefficients for patience and risk-taking are small and insignificant.

Table 2.2. Weekly Incidence at the Time of the Survey

	$y_{c,t} = \log(cases_{c,t})$ in county c and week t				
	(1)	(2)	(3)	(4)	(5)
Prosociality	-0.1391 ***	-0.1270 *	-0.1241 **	-0.1189 **	0.0183
	[-0.283, -0.061]	[-0.303, 0.010]	[-0.296, -0.021]	[-0.246, -0.033]	[-0.088, 0.106]
Patience	–	-0.0286	0.0024	-0.0054	0.0602
		[-0.211, 0.133]	[-0.117, 0.181]	[-0.111, 0.129]	[-0.019, 0.188]
Risk-taking	–	0.0106	-0.0377	-0.0454	-0.0814 *
		[-0.107, 0.126]	[-0.154, 0.092]	[-0.137, 0.072]	[-0.149, 0.005]
Public health behavior	–	–	–	–	-0.2996 ***
					[-0.443, -0.158]
Wave 1 severity	*No*	*No*	*No*	*Yes*	*Yes*
County controls × Week	*No*	*No*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38	38
R^2	0.116	0.118	0.357	0.415	0.481

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The outcome variable is the log weekly incidence rate by county, ranging from Nov 16, 2020, until Jan 17, 2021 (9 weeks). County controls include 18 variables: log population density, log GDP per capita, log average income per capita, share of college graduates, employment share, share of non-German residents, share of workers in the service sector, share of population below age 18, share of population age 65 or above, and border county dummies for each neighboring country of Germany. Controls for wave 1 severity include the log of aggregate case numbers, its square, and case fatality rate in the time period from the first confirmed infection until May 17th, 2020. See Appendix Table 2.A.10 for results with the individual elements of prosociality. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Importantly, we verify whether the association between prosociality and COVID-19 incidence rates is robust to controlling for other demographic and socio-economic county characteristics that could influence the regional spread of the virus. In column 3, we therefore add pre-pandemic county characteristics ($\boldsymbol{x}_c$) and allow their effect to vary by week. The vector of county controls consists of log population density, log GDP per capita, log average income per capita, share of college graduates, employment share, share of workers in the service sector, share of non-German residents, share of population below age 18, share of population age 65 or above, and border county dummies for each neighboring country of Germany. Another potential concern is that regional differences in severity of the pandemic experienced during the first wave may have had an impact on the level of prosociality, but simultaneously also on other factors like general attitudes or local government preparedness. To flexibly account for this, we further add control variables for counties' first wave (February-May) infection outcomes in another specification.

After including this rich set of control variables in columns 3-4 of Table 2.2, the explanatory power of the regression increases drastically by a factor of more than three. Crucially, the coefficient for prosociality remains nearly unchanged, with a one SD increase being associated with $11-12\%$ lower weekly incidence rates ($p < 0.05$).

Why is the incidence rate lower in regions with higher prosociality? Our theoretical considerations suggest that more prosocial individuals should be more willing to comply with recommended or mandatory social distancing and hygiene measures, which is confirmed empirically by our individual-level results. The models discussed in Section 2.2 would then predict that stricter engagement in preventive health behaviors leads to a lower contact and transmission rate, and thus eventually to a lower COVID-19 incidence rate in high-prosociality regions. To test this mediating role of behavior, we include our measure of average PHB as additional regressor in column 5 of Table 2.2 (Baron and Kenny, 1986). Upon doing so, the coefficient size for prosociality is reduced by 85% to almost zero, whereas we observe a remarkably strong relation between self-reported PHB and incidence rates: a one SD increase in PHB is associated with a 30% decrease in the weekly number of cases per 100,000 population. This is consistent with the hypothesis that the effect of prosociality is mediated by differences in PHB across regions. Interestingly, risk-taking has a weakly significant negative effect conditional on PHB, which could potentially be explained with a higher willingness to experiment with new strategies or to adopt new technologies.

Although we have controlled for a host of demographic and socio-economic county characteristics, there could still be other, unobserved factors that lead to generally lower levels of infections in a county, while also being positively correlated with prosociality and PHB. To circumvent this issue, we test whether regions with higher prosociality also exhibit lower growth rates of new cases, as this partials out any time-invariant differences across counties that can affect absolute levels of infection rates in the pandemic. We approximate growth rates by the weekly change in log incidence rates $\Delta \log(cases_{crt}) = \log(cases_{c,t}) - \log(cases_{c,t-1})$ in county c and week t and estimate the following statistical model:

$$\begin{aligned} \Delta \log(cases_{crt}) \;=\; & \alpha_t + \beta_1 \cdot \overline{Prosocial}_r + \beta_2 \cdot \overline{Patience}_r + \beta_3 \cdot \overline{RiskTak}_r \\ & + \boldsymbol{\gamma}_t' \boldsymbol{x}_c + \boldsymbol{\delta}' \boldsymbol{w}_c + \varepsilon_{crt}\,, \end{aligned} \tag{2.3}$$

where everything is defined as in equation 2.2. We include the full set of previously used control variables in all specifications, including the vector of controls for wave 1 severity $\boldsymbol{w}_c$.

Although high- and low-prosociality regions start from roughly similar levels of incidence at the beginning of the second wave (see Figure 2.2c), differences in the growth rate would gradually drive incidence levels apart over time, eventually resulting in large cumulative differences. Indeed, our baseline specification in Table 2.3 shows that, in the time period we study, the growth rate of new cases was about 1%p

Table 2.3. Weekly Growth Rate of Confirmed Cases at the Time of the Survey

	$y_{c,t} = \log(cases_{c,t}) - \log(cases_{c,t-1})$			
	(1)	(2)	(3)	(4)
Prosociality	-0.0091 **	-0.0097	-0.0218 ***	-0.0072
	[-0.018, -0.001]	[-0.022, 0.002]	[-0.037, -0.011]	[-0.025, 0.008]
Patience	-0.0012	-0.0015	-0.0012	0.0062
	[-0.014, 0.007]	[-0.015, 0.009]	[-0.011, 0.014]	[-0.008, 0.026]
Risk-taking	0.0002	0.0003	-0.0044	-0.0092
	[-0.012, 0.013]	[-0.012, 0.012]	[-0.016, 0.010]	[-0.026, 0.007]
Public health behavior	–	0.0012	–	-0.0340 **
		[-0.021, 0.022]		[-0.066, -0.006]
$\log(cases_{c,t-2})$	–	–	-0.1081 ***	-0.1209 ***
			[-0.126, -0.093]	[-0.146, -0.096]
Policy stringency$_{c,t-2}$	–	–	-0.2403	-0.2050
			[-0.857, 0.289]	[-0.765, 0.228]
Wave 1 severity	*Yes*	*Yes*	*Yes*	*Yes*
County controls × Week	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38
R^2	0.293	0.293	0.315	0.317

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The outcome variable is the change in log weekly incidence rate in a county, ranging from Nov 16th, 2020 until Jan 17th, 2021 (9 weeks). All control variables are defined as in Table 2.2. See Appendix Table 2.A.11 results with the individual elements of prosociality. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

lower in regions with a one SD higher prosociality ($p < 0.05$). We find no evidence for mediation through PHB in column 2 yet.

However, the estimated effects of prosociality and social distancing might be attenuated due to dynamic interactions between incidence rates, behavior, and policy responses that push towards regional convergence. For one, the share of susceptibles in the population is naturally higher in regions with fewer past infections, although this effect may have been negligible at that stage of the pandemic. Moreover, SIR models with endogenous behavior predict that in regions with lower incidence rates, people may endogenously reengage in more social contacts in response to reduced infection risks. Local governments could also feel encouraged to partially lift curtailment measures. Thus, more prosocial regions could become the victims of their own success. For this reason, we further add the 2-week lagged incidence rate $\log(cases_{c,t-2})$ as well as a 2-week lagged local policy stringency index (see Section 2.3.2) as covariates in equation 2.3. After including these lagged variables, the coef-

ficient size for prosociality more than doubles, implying a 2%p lower weekly growth rate per SD increase ($p < 0.01$) — this corresponds to about 3% of a region-week SD in incidence growth rates (see Appendix 2.A.17). This is a sizeable effect given that small differences in growth rates accumulate to large absolute differences over time. In column 4, prosociality becomes insignificant after adding average PHB, further supporting the hypothesis that better compliance with social distancing and hygiene measures mediates the effect of higher prosociality on collective health outcomes during the pandemic.

Finally, we check whether our results are influenced by comparisons between West Germany and East Germany, as previous studies document that historical institutional differences between these two regions before the German reunification still have a persistent effect on preferences, norms, and outcomes (Torgler, 2002; Alesina and Fuchs-Schündeln, 2007; Brosig-Koch et al., 2011; Becker, Mergele, and Woessmann, 2020). Therefore, we rerun our analyses adding an East-Germany dummy as control variable, and further interacting it with our measure of average prosociality (Appendix Tables 2.A.13-2.A.15). The results show that the estimated coefficients for prosociality remain robust, and that there is no evidence for a differential association between higher prosociality and lower COVID-19 incidence rates in East and West Germany, although the low number of regional units in the East precludes any conclusive statement.

2.6 Discussion

How well a group of individuals succeeds in achieving desirable collective outcomes in the face of social dilemma depends, amongst other things, on how willingly individual members engage in actions that incur personal costs but that benefit the group as a whole. We have provided suggestive evidence that, in the context of the COVID-19 pandemic, more prosocial individuals are significantly more willing to engage in public health behaviors (e.g. physical distancing and mask-wearing) aimed at slowing the spread of the virus. We further presented evidence that, in turn, regions in Germany with higher average prosociality in the population also tend to experience a lower incidence of COVID-19 cases and deaths. The estimated (conditional) correlations are quantitatively sizeable: a 1 SD higher average prosociality in a region is associated with around 11% lower COVID-19 incidence rates and 2%p lower incidence growth rates.

2.6.1 Role of the Study Context

The interpretation of our results needs to take into account the broader context in which our study is embedded, as the role of prosociality may be moderated, among others, by the stage of the pandemic, the regional severity of the outbreak, and the stringency of government-mandated restrictions and policy measures. Our sur-

vey was conducted in the late fall of 2020, before the peak of the second wave in Germany, during the so-called lockdown light. In contrast, most related studies examining determinants of PHB were conducted in the first wave of the pandemic, when more fear and uncertainty was revolving around the disease and the spread of the virus (Harper et al., 2020). Thus, we confirm previous results on the importance of prosociality (Campos-Mercade et al., 2021; Müller and Rau, 2021) also for later stages of the pandemic, when people had become more accustomed to and more weary of the situation (Petherick et al., 2021). In Table 2.A.18 of the Appendix, we compare predictors of regional incidence rates in the first and the second COVID-19 wave in Germany. We observe that the same set of demographic and socio-economic county characteristics (e.g. population density, employment share) has much higher explanatory value in the first wave ($R^2 = 0.497$) than in the second wave ($R^2 = 0.265$), possibly because behavioral responses in the population were more homogeneous early on in the pandemic.

The quickly rising case numbers at the time period of our survey might have further driven attitudes and behavioral responses apart for people in different regions and with different individual characteristics, as protecting those vulnerable to the disease becomes especially relevant when the risk of infection and transmission is high. In contrast, private gatherings may not be considered irresponsible acts of selfishness in periods of low incidence such as the summer of 2020 in Germany. Another potentially amplifying factor for the role of prosociality in our context may be that the lockdown light in Germany left plenty of wiggle room in the extent of social distancing behavior within the limits of what was allowed, thereby putting considerable weight on voluntary reduction of social contacts. Although voluntary adaptions and government-mandated restrictions can be partly substitutable (Alfaro et al., 2021b), prosociality may affect health behaviors and outcomes even under more stringent lockdown regimes, as perfect monitoring and enforcement of compliance are infeasible, and drastic government measures can also influence public perceptions of severity and social norms (Casoria, Galeotti, and Villeval, 2021; Galbiati et al., 2021).

2.6.2 Potential Endogeneity Concerns

Finally, a natural question in our context is to which extent the conditional correlations we find in our empirical analyses can be interpreted as causal. There are several potential concerns against such a causal interpretation. First, our sample may not be regionally representative due to self-selection into completing the survey. While such selection effects are hard to rule out, they could only explain our results if systematically more prosocial individuals respond to our survey in regions with lower incidence rates, which seems implausible. Second, one might worry that our measures of prosociality and economic preferences are themselves affected by the COVID-19 pandemic (Bauer et al., 2016; Branas-Garza et al., 2020; Cappelen et al., 2021;

Frondel, Osberghaus, and Sommer, 2021; Shachat, Walker, and Wei, 2021). If any influence on individuals' survey responses reflects true changes in preferences and attitudes, our measures remain internally valid for the time period around which we conducted the survey. On the other hand, we might overestimate the role of prosociality if respondents' answers to broadly framed questions overreflected their behavior during the pandemic, e.g. due to availability bias (Tversky and Kahneman, 1973). We cannot directly investigate this issue with our cross-sectional survey data, but note that regional prosociality in our data correlates with pre-pandemic outcomes such as election turnout, and that our results are robust to controlling for first-wave severity of the pandemic. Moreover, Campos-Mercade et al. (2021) provide evidence that individual health behavior during the pandemic is predicted by prosociality measured before the COVID-19 outbreak, which is consistent with the notion that individual's (social) preferences are fairly stable in general (Volk, Thöni, and Ruigrok, 2012; Carlsson, Johansson-Stenman, and Nam, 2014). A third concern is reverse causality, because regional incidence rates may also influence PHB and its relation to prosociality. However, this would presumably lead to an underestimation of the true effect since lower incidence rates allow residents and policymakers to become more lenient in their responses. Consistent with this convergence effect, we have shown in Table 2.3 that the estimated association between average prosociality and weekly incidence growth rate doubles in magnitude when controlling for lagged incidence levels.

The fourth and arguably most important concern is omitted variable bias. At the individual level, it seems unlikely that the relation between prosociality and PHB is entirely driven by some unobserved factor, as we control for a host of demographic and socio-economic characteristics, and further confirm robustness to including personality factors and political attitudes as regressors. At the regional level, we control for a variety of relevant county characteristics. However, it is difficult to rule out all potentially confounding factors, e.g., the stringency of local implementation and enforcement of containment measures, contact tracing efficiency, etc., which may themselves be a function of prosociality in the population. Most notably, the distribution of (pro-)social preferences, values, norms, and beliefs is inherently endogenous to social, cultural, political, and institutional factors. Because these factors are imperfectly observable and the underlying causal relationships highly complex and interdependent, our empirical investigation must inevitably remain correlational.

2.6.3 Concluding Remarks

Our paper is inspired by several previous studies that measure individual and geographical variation of (pro-)social behavior and preferences in order to advance our understanding of how collective societal outcomes may be shaped by the prevalent values, norms, and preferences in the population, and vice versa, how individual dispositions may vary due to ecological, cultural, or socio-economic factors (Hen-

rich et al., 2006; Nettle, Colléony, and Cockerill, 2011; Falk et al., 2018; Cohn et al., 2019; Barsbai, Lukas, and Pondorfer, 2021; Caicedo, Dohmen, and Pondorfer, 2021). Recent experimental evidence further highlights the malleability of prosociality by documenting the importance of socialization and role models (Kosse et al., 2020). Cultivating prosocial values and norms within a society may strengthen its capacity to face challenges such as pandemics or global warming that require widespread cooperation and collective action.

Appendix 2.A Additional Results and Robustness Checks

2.A.1 Supplementary Figures

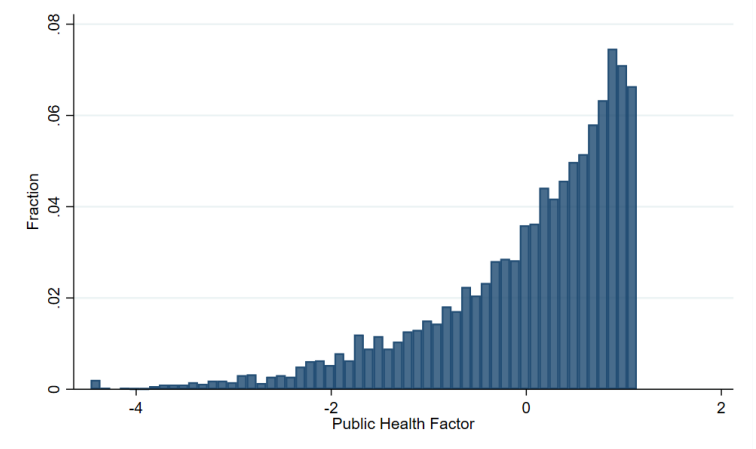
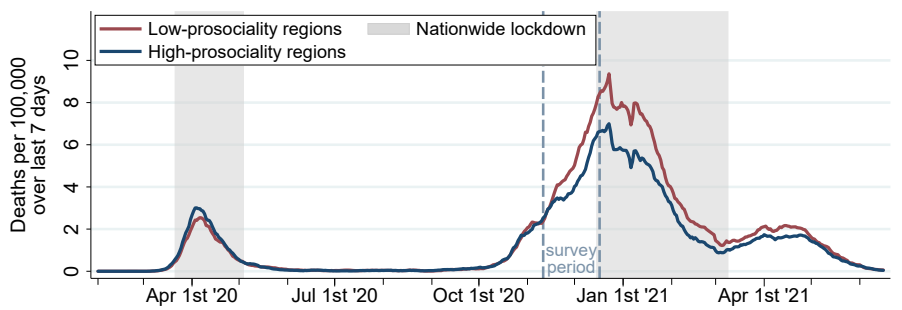


Figure 2.A.1. Histogram of PHB Values

Notes: Histogram of public health behavior, using width of 0.1.

(a) COVID-19 Deaths per 100.000 Population in 7 Days (by Date of Infection)



(b) Change in Mobility [%] Compared to 2019 (7-Day Moving Average)

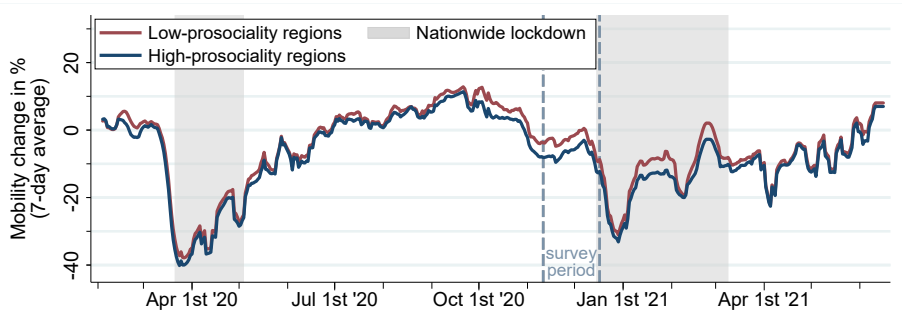


Figure 2.A.2. The COVID-19 Pandemic in Germany

Notes: The time labels in Figure 2.A.2a refer to the day the coronavirus infection of the deceased person was first reported to the RKI, not the day of death. Grey shaded areas indicate time periods of strict nationwide lockdowns in Germany (as of March 8, 2021, restrictions were tied to the regional incidence rate, although the lockdown formally remained in place).

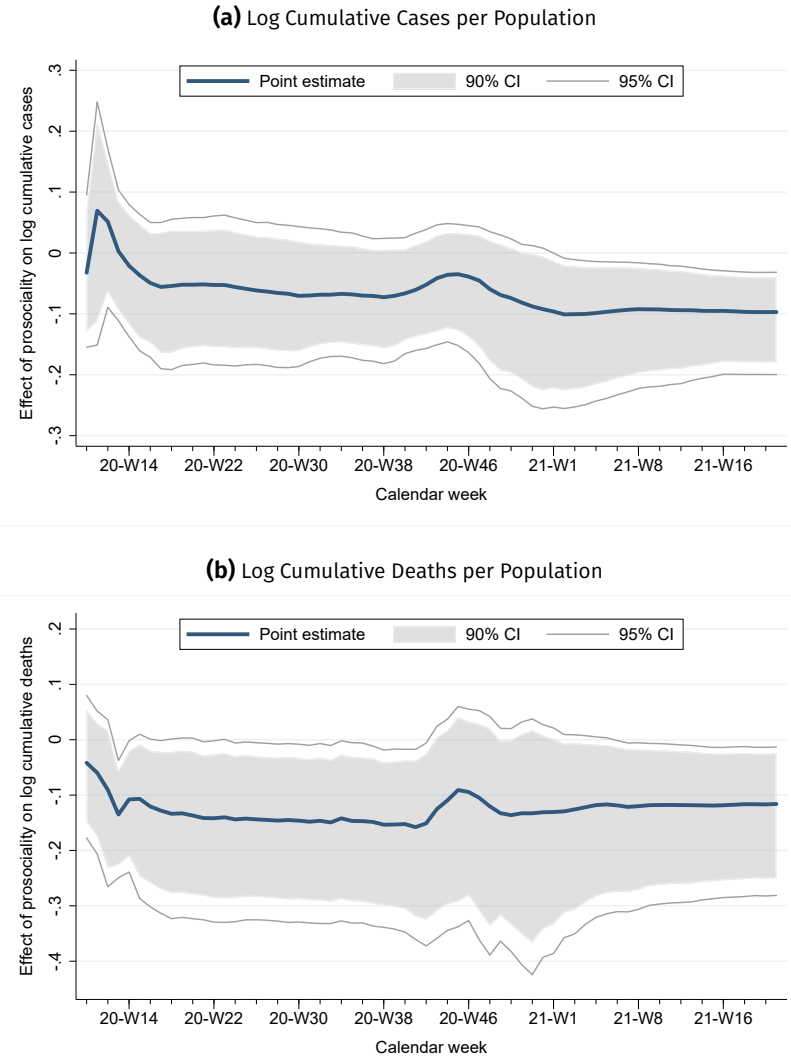


Figure 2.A.3. Estimated Effect of Prosociality on Cumulative Cases and Deaths

Notes: Confidence-intervals are obtained using the wild bootstrap (9,999 simulations) with clustering on NUTS-2 region level and Rademacher-weights. The time labels in Panel (b) refer to the day the coronavirus infection of the deceased person was first reported to the RKI, not the day of death.

2.A.2 Supplementary Tables

Table 2.A.1. Correlation Matrix of Prosociality Components

	Altruism	Positive reciprocity	Trust	Indirect neg. reciprocity
Altruism	1			
Positive reciprocity	0.3344	1		
Trust	0.2591	0.1503	1	
Indirect neg. reciprocity	0.2574	0.1705	0.1488	1
Observations		5949		

Notes: Pearson correlation coefficients of altruism, positive reciprocity, trust, and indirect (negative) reciprocity across individual survey respondents.

Table 2.A.2. Individual-Level Association between Preferences and PHB

	Public Health Behavior (PHB)				
	(1)	(2)	(3)	(4)	(5)
Prosociality	0.3356***	0.3059***	0.3115***	0.2216***	0.1625***
	(0.0162)	(0.0165)	(0.0168)	(0.0173)	(0.0148)
Patience	0.1983***	0.1969***	0.1858***	0.1633***	0.0777***
	(0.0150)	(0.0151)	(0.0155)	(0.0155)	(0.0131)
Risk-taking	-0.2095***	-0.1710***	-0.1722***	-0.1683***	-0.0790***
	(0.0141)	(0.0144)	(0.0148)	(0.0141)	(0.0110)
Negative reciprocity (Direct)	-0.1231***	-0.1078***	-0.1075***	-0.0662***	-0.0184
	(0.0141)	(0.0145)	(0.0151)	(0.0156)	(0.0127)
Female		0.1546***	0.1542***	0.0895***	0.0800***
		(0.0267)	(0.0269)	(0.0266)	(0.0225)
Age		0.0146*	0.0141*	0.0084	0.0127*
		(0.0083)	(0.0085)	(0.0081)	(0.0070)
Age^2		-0.0000	-0.0000	0.0000	-0.0001
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Big 5: Openness				0.0578***	0.0423***
				(0.0135)	(0.0116)
Big 5: Conscientiousness				0.1596***	0.1577***
				(0.0157)	(0.0129)
Big 5: Extraversion				0.0192	0.0070
				(0.0135)	(0.0114)
Big 5: Agreeableness				0.1186***	0.1055***
				(0.0162)	(0.0137)
Big 5: Neuroticism				0.0418***	-0.0121
				(0.0136)	(0.0116)
Affected by pandemic					0.0252**
					(0.0121)
Take pandemic seriously					0.2974***
					(0.0157)
Worry: Self					0.0211**
					(0.0084)
Worry: Family & Friends					0.0761***
					(0.0107)
Worry: Others					0.0557***
					(0.0101)
Socio-demographic factors	*No*	*Yes*	*Yes*	*Yes*	*Yes*
County FEs	*No*	*No*	*Yes*	*Yes*	*Yes*
Big Five	*No*	*No*	*No*	*Yes*	*Yes*
COVID-19 Perceptions	*No*	*No*	*No*	*No*	*Yes*
Observations	5843	5660	5653	5653	5653
Clusters	397	396	389	389	389
R^2	0.209	0.234	0.293	0.345	0.529

Notes: This table estimates the same specifications as in Table 2.1, but reports additional estimates that might be of interest to the reader. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.3. Individual-Level Association between Individual Preferences and PHB

	Public Health Behavior (PHB)				
	(1)	(2)	(3)	(4)	(5)
Altruism	0.1547***	0.1545***	0.1492***	0.1141***	0.0598***
	(0.0146)	(0.0148)	(0.0149)	(0.0150)	(0.0138)
Positive reciprocity	0.2383***	0.2048***	0.2125***	0.1342***	0.1361***
	(0.0170)	(0.0170)	(0.0174)	(0.0171)	(0.0149)
Negative reciprocity (Indirect)	0.0218	0.0253*	0.0258*	0.0150	0.0011
	(0.0148)	(0.0152)	(0.0152)	(0.0147)	(0.0133)
Trust	0.0708***	0.0582***	0.0663***	0.0609***	0.0512***
	(0.0130)	(0.0131)	(0.0131)	(0.0132)	(0.0113)
Patience	0.1807***	0.1813***	0.1690***	0.1561***	0.0675***
	(0.0156)	(0.0157)	(0.0159)	(0.0157)	(0.0130)
Risk-taking	-0.1949***	-0.1632***	-0.1648***	-0.1662***	-0.0772***
	(0.0140)	(0.0144)	(0.0148)	(0.0142)	(0.0110)
Negative reciprocity (Direct)	-0.0741***	-0.0665***	-0.0654***	-0.0341*	0.0086
	(0.0179)	(0.0179)	(0.0184)	(0.0187)	(0.0160)
Socio-demographic factors	—	Yes	Yes	Yes	Yes
County FEs	—	—	Yes	Yes	Yes
Big Five	—	—	—	Yes	Yes
COVID-19 Perceptions	—	—	—	—	Yes
Observations	5843	5660	5653	5653	5653
R^2	0.223	0.243	0.302	0.348	0.533
Clusters	397	396	389	389	389

Notes: This table estimates the same specifications as in Table 2.1, but with the individual social preferences of altruism, trust, positive reciprocity and indirect negative reciprocity as independent variables instead of prosociality. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.4. Individual-Level Association between Preferences and Individual PHB Survey Items

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Prosociality	0.2928***	0.3807***	0.2893***	0.3562***	0.3749***	0.4040***	0.3317***	0.3647***	0.2626***	0.3738***
	(0.0224)	(0.0268)	(0.0236)	(0.0240)	(0.0366)	(0.0301)	(0.0285)	(0.0241)	(0.0295)	(0.0244)
Patience	0.2112***	0.2257***	0.2339***	0.1562***	0.2061***	0.1473***	0.2299***	0.1531***	0.2824***	0.1608***
	(0.0223)	(0.0247)	(0.0254)	(0.0217)	(0.0353)	(0.0275)	(0.0279)	(0.0245)	(0.0277)	(0.0202)
Risk-taking	-0.1870***	-0.2419***	-0.1546***	-0.1429***	-0.1728***	-0.1473***	-0.1993***	-0.1189***	-0.3499***	-0.1206***
	(0.0208)	(0.0236)	(0.0252)	(0.0217)	(0.0350)	(0.0244)	(0.0244)	(0.0243)	(0.0248)	(0.0179)
Negative reciprocity (Direct)	-0.0835***	-0.1463***	-0.0436*	-0.1407***	-0.0766**	-0.1922***	-0.1085***	-0.1322***	-0.0569**	-0.1762***
	(0.0235)	(0.0258)	(0.0260)	(0.0224)	(0.0389)	(0.0269)	(0.0228)	(0.0248)	(0.0278)	(0.0212)
Observations	5653	5653	5653	5653	5653	5653	5653	5653	5653	5653
R^2	0.206	0.215	0.199	0.204	0.184	0.179	0.198	0.197	0.198	0.233
Clusters	389	389	389	389	389	389	389	389	389	389

Notes: This table estimates the specification (3) of Table 2.1, but using individual survey items of the PHB index as dependent variables. The columns are defined as follows: 1) Social distancing of 1.5 meters 2) Self-quarantining in the case of risky contact 3) Keeping oneself informed about the pandemic 4) Washing and disinfecting hands 5) Willingness to get vaccinated 6) Sneezing and coughing into elbow 7) Wearing mask 8) Ventilating when indoors 9) Avoiding social contacts 10) Informing others if infected. Each survey item is measured on a 7-point scale, with 1 indicating "Do not agree" and 7 indicating "Agree completely". * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.5. Economic Preferences, Personality Traits and COVID-19 Perceptions

	(1) Pandemic serious	(2) Worry: Family & Friends	(3) Worry: Others
Prosociality	0.1059***	0.1875***	0.2682***
	(0.0192)	(0.0362)	(0.0329)
Patience	0.1838***	0.2446***	0.1765***
	(0.0168)	(0.0310)	(0.0313)
Risk-taking	-0.1962***	-0.2275***	-0.1954***
	(0.0181)	(0.0349)	(0.0292)
Negative reciprocity (Direct)	-0.1429***	-0.0558	-0.0932***
	(0.0177)	(0.0349)	(0.0332)
Big 5: Openness	-0.0016	0.0814***	0.0981***
	(0.0164)	(0.0309)	(0.0279)
Big 5: Conscientiousness	-0.0116	0.0311	-0.0237
	(0.0190)	(0.0303)	(0.0294)
Big 5: Extraversion	-0.0204	0.1032***	0.1500***
	(0.0164)	(0.0325)	(0.0269)
Big 5: Agreeableness	0.0042	0.1136***	0.0518
	(0.0180)	(0.0304)	(0.0318)
Big 5: Neuroticism	-0.0144	0.3725***	0.3351***
	(0.0148)	(0.0299)	(0.0295)
Observations	5653	5653	5653
R^2	0.190	0.200	0.192
Clusters	389	389	389

Notes: Pandemic serious is a factor comprised of two survey items measuring (on a 5-point scale) how much the respondent disagrees with the statements that the media takes the pandemic too seriously, and that government measures are too strict. Worry: Family & Friends and Worry: Others measure (on a 7-point scale) how much the respondent worries about their family and friends, and others around them, respectively. All specifications include socio-demographic controls and county FEs. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.6. Correlation Matrix of Prosociality and BFI Personality Traits

	Prosociality	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Prosociality	1.0000					
Agreeableness	0.3070***	1.0000				
Conscientiousness	0.2446***	0.4353***	1.0000			
Extraversion	0.2451***	0.2347***	0.3021***	1.0000		
Neuroticism	-0.0314*	-0.0209	-0.1554***	-0.2268***	1.0000	
Openness	0.2777***	0.2399***	0.2638***	0.4142***	-0.0060	1.0000

Notes: Pearson correlation coefficients of prosociality, agreeableness, conscientiousness, extraversion, neuroticism, and openness across individual survey respondents. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.7. Regional Correlations of Vote Shares for the Major Political Parties

	Regional correlation with 2019 election outcome					
	CDU/CSU	SPD	Grüne	FDP	Die Linke	AfD
Survey vote shares	0.808***	0.854***	0.757***	0.290*	0.861***	0.784***
2017 election outcomes	0.904***	0.923***	0.844***	0.763***	0.980***	0.970***
Observations	38	38	38	38	38	38
Overall 2019 vote share [%]	22.6	15.8	20.5	5.4	5.5	11.0

Notes: The first row shows the Pearson's correlation coefficients of 2019 election vote shares with the implied vote shares from our survey on NUTS-2 region level. For comparison, the second rows shows the correlation of 2019 election outcomes with 2017 election outcomes. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.8. Variation of Prosociality across NUTS-2 Regions in Germany

	(1) *No controls*		(2) *With controls*		(3) *Only controls*
	Coefficient	SE	Coefficient	SE	
Brandenburg	-0.059	(0.102)	-0.096	(0.103)	–
Bremen	0.140	(0.124)	0.109	(0.117)	–
Direktionsbezirk Chemnitz	0.056	(0.137)	0.039	(0.136)	–
Direktionsbezirk Dresden	-0.046	(0.113)	-0.089	(0.115)	–
Direktionsbezirk Leipzig	-0.086	(0.131)	-0.105	(0.131)	–
Hamburg	0.095	(0.120)	0.079	(0.120)	–
Mecklenburg-Vorpommern	0.133	(0.118)	0.121	(0.121)	–
Reg.-Bez. Arnsberg	-0.013	(0.093)	-0.011	(0.093)	–
Reg.-Bez. Darmstadt	0.104	(0.087)	0.068	(0.088)	–
Reg.-Bez. Detmold	-0.037	(0.122)	-0.021	(0.123)	–
Reg.-Bez. Düsseldorf	-0.055	(0.083)	-0.070	(0.084)	–
Reg.-Bez. Freiburg	-0.089	(0.106)	-0.087	(0.107)	–
Reg.-Bez. Gießen	-0.017	(0.178)	-0.097	(0.193)	–
Reg.-Bez. Karlsruhe	0.157	(0.100)	0.137	(0.101)	–
Reg.-Bez. Kassel	-0.110	(0.147)	-0.171	(0.147)	–
Reg.-Bez. Köln	0.025	(0.087)	-0.003	(0.088)	–
Reg.-Bez. Mittelfranken	-0.199	(0.100)	-0.239	(0.099)	–
Reg.-Bez. Münster	0.181	(0.101)	0.171	(0.103)	–
Reg.-Bez. Niederbayern	-0.153	(0.137)	-0.177	(0.142)	–
Reg.-Bez. Oberbayern	0.044	(0.089)	0.023	(0.090)	–
Reg.-Bez. Oberfranken	0.011	(0.114)	-0.019	(0.117)	–
Reg.-Bez. Oberpfalz	0.111	(0.124)	0.100	(0.128)	–
Reg.-Bez. Schwaben	-0.056	(0.116)	-0.080	(0.117)	–
Reg.-Bez. Stuttgart	-0.020	(0.089)	-0.048	(0.090)	–
Reg.-Bez. Tübingen	0.034	(0.111)	0.002	(0.113)	–
Reg.-Bez. Unterfranken	0.093	(0.128)	0.077	(0.131)	–
Saarland	0.232	(0.140)	0.245	(0.144)	–
Sachsen-Anhalt	-0.256	(0.106)	-0.278	(0.105)	–
Schleswig-Holstein	0.061	(0.095)	0.011	(0.097)	–
Statistische Region Braunschweig	0.118	(0.117)	0.096	(0.120)	–
Statistische Region Hannover	-0.043	(0.098)	-0.016	(0.099)	–
Statistische Region Lüneburg	0.074	(0.117)	0.025	(0.119)	–
Statistische Region Weser-Ems	0.014	(0.104)	-0.012	(0.105)	–
Thüringen	-0.055	(0.101)	-0.074	(0.104)	–
früher: Reg.-Bez. Koblenz	0.333	(0.186)	0.299	(0.203)	–
früher: Reg.-Bez. Rheinhessen-Pfalz	0.078	(0.104)	0.057	(0.106)	–
früher: Reg.-Bez. Trier	0.078	(0.168)	0.053	(0.169)	–
Constant	-0.021	(0.064)	0.023	(0.205)	–
Socio-demographic controls	–		*yes*		*yes*
Observations	5843		5660		5660
F-statistic (NUTS-2 dummies)	1.426		1.480		–
p-value (NUTS-2 dummies)	.0455		.0307		–
R^2	0.011		0.034		0.023

Notes: The baseline region is Berlin. Socio-demographic controls include age and age-squared, gender, education, income, employment status, household size, number of children, and an indicator for having children below age 16. Robust standard errors in parentheses.

Table 2.A.9. Prosociality and Measures of Social Capital

	Turnout in 2019 election [%]			Civic associations per 100k pop. in 2008		
	(1)	(2)	(3)	(4)	(5)	(6)
Prosociality	1.52 **	1.57 **	1.51 ***	14.63 *	10.79 *	10.92
	[0.37, 2.51]	[0.36, 2.93]	[0.56, 2.55]	[-1.06, 23.85]	[-1.91, 19.16]	[-7.49, 24.83]
Patience	–	-0.46	-0.26	–	12.62 *	12.78
		[-1.93, 0.58]	[-1.40, 0.79]		[-1.57, 30.41]	[-12.39, 40.19]
Risk-taking	–	0.74	0.36	–	-10.75	-16.82 **
		[-0.50, 1.74]	[-1.15, 1.75]		[-24.60, 3.87]	[-30.38, -1.64]
County controls	*No*	*No*	*Yes*	*No*	*No*	*Yes*
Population mean	61.37	61.37	61.37	280.82	280.82	280.82
Observations	401	401	401	401	401	401
Clusters	38	38	38	38	38	38
R^2	0.096	0.117	0.542	0.019	0.035	0.415

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. Control variables include log GDP per capita, log average income per capita, share of college graduates, share of non-German residents, share of population below age 18, share of population age 65 or above, and indicators for the degree of urbanization. Under civic associations, we include (non-profit) organizations focused on social and economic welfare, political asssociations, and interest groups, following a classification by Franzen and Botzen (2011). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.10. Weekly Incidence at the Time of the Survey

	$y_{c,t} = \log(cases_{c,t})$ in county c and week t				
	(1)	(2)	(3)	(4)	(5)
Altruism	-0.1041 *	-0.1042 *	-0.0745	-0.0742	0.0229
	[-0.227, 0.011]	[-0.238, 0.018]	[-0.211, 0.047]	[-0.188, 0.024]	[-0.122, 0.163]
Trust	0.1055	0.1095	0.0655	0.0493	0.0649
	[-0.048, 0.293]	[-0.065, 0.304]	[-0.079, 0.217]	[-0.072, 0.182]	[-0.023, 0.177]
Positive Reciprocity	-0.0640	-0.0657	-0.0640	-0.0527	-0.0296
	[-0.179, 0.079]	[-0.178, 0.086]	[-0.209, 0.089]	[-0.183, 0.071]	[-0.133, 0.067]
Negative Reciprocity (ind.)	-0.1018	-0.1017	-0.1032	-0.0910 *	-0.0428
	[-0.239, 0.028]	[-0.258, 0.056]	[-0.240, 0.041]	[-0.198, 0.020]	[-0.135, 0.055]
Patience	–	0.0051	0.0323	0.0215	0.0718 *
		[-0.168, 0.175]	[-0.102, 0.210]	[-0.081, 0.158]	[-0.021, 0.230]
Risk-taking	–	-0.0157	-0.0484	-0.0550	-0.0837 **
		[-0.092, 0.066]	[-0.139, 0.058]	[-0.128, 0.044]	[-0.158, -0.020]
Public health behavior	–	–	–	–	-0.2878 ***
					[-0.431, -0.144]
Wave 1 severity	*No*	*No*	*No*	*Yes*	*Yes*
County controls × Week	*No*	*No*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401	401
Clusters	38	38	38	38	38
R^2	0.170	0.171	0.385	0.433	0.492

Notes: This table estimates the same specifications as in Table 2.2, but with the individual social preferences of altruism, trust, positive reciprocity and indirect negative reciprocity as independent variables instead of prosociality. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.11. Weekly Growth Rate of Confirmed Cases at the Time of the Survey

	$y_{c,t} = \log(cases_{c,t}) - \log(cases_{c,t-1})$			
	(1)	(2)	(3)	(4)
Altruism	-0.0028	-0.0025	-0.0119	-0.0010
	[-0.015, 0.009]	[-0.020, 0.016]	[-0.031, 0.005]	[-0.027, 0.028]
Trust	-0.0037	-0.0037	0.0037	0.0062
	[-0.016, 0.012]	[-0.017, 0.014]	[-0.010, 0.023]	[-0.010, 0.026]
Positive Reciprocity	-0.0073	-0.0072	-0.0114	-0.0092
	[-0.022, 0.006]	[-0.021, 0.006]	[-0.031, 0.007]	[-0.026, 0.006]
Negative Reciprocity (ind.)	0.0016	0.0017	-0.0116	-0.0066
	[-0.010, 0.015]	[-0.010, 0.016]	[-0.030, 0.006]	[-0.024, 0.011]
Patience	-0.0029	-0.0028	0.0013	0.0074
	[-0.013, 0.006]	[-0.014, 0.007]	[-0.015, 0.022]	[-0.010, 0.032]
Risk-taking	0.0005	0.0004	-0.0055	-0.0097
	[-0.012, 0.014]	[-0.014, 0.014]	[-0.020, 0.009]	[-0.028, 0.007]
Public health behavior	–	-0.0010	–	-0.0341 **
		[-0.027, 0.023]		[-0.069, -0.005]
$\log(cases_{c,t-2})$	–	–	-0.1112 ***	-0.1234 ***
			[-0.129, -0.096]	[-0.148, -0.101]
Policy stringency$_{c,t-2}$	–	–	-0.2759	-0.2260
			[-1.009, 0.298]	[-0.901, 0.250]
Wave 1 severity	*Yes*	*Yes*	*Yes*	*Yes*
County controls × Week	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609
R^2	0.294	0.294	0.316	0.318

Notes: This table estimates the same specifications as in Table 2.3, but with the individual social preferences of altruism, trust, positive reciprocity and indirect negative reciprocity as independent variables instead of prosociality. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.12. Effect of Preferences and Behavior on Weekly Deaths

	$y = \log deaths_t$			$y = \log deaths_t - \log deaths_{t-1}$	
	(1)	(2)	(3)	(4)	(5)
Prosociality	-0.1272 *	-0.1241 *	0.0488	-0.0134 *	-0.0051
	[-0.315, 0.009]	[-0.288, 0.007]	[-0.089, 0.176]	[-0.033, 0.000]	[-0.035, 0.019]
Patience	-0.0095	-0.0163	0.0678	-0.0010	0.0032
	[-0.174, 0.207]	[-0.180, 0.180]	[-0.051, 0.222]	[-0.015, 0.020]	[-0.013, 0.024]
Risk-taking	-0.0271	-0.0307	-0.0852	-0.0147	-0.0181
	[-0.139, 0.110]	[-0.134, 0.107]	[-0.196, 0.022]	[-0.048, 0.013]	[-0.053, 0.016]
Public health behavior	–	–	-0.3851 ***	–	-0.0197
			[-0.520, -0.240]		[-0.056, 0.022]
$\log cases_{t-2}$	–	–	–	-0.1476 ***	-0.1549 ***
				[-0.195, -0.103]	[-0.211, -0.101]
Policy measures$_{t-2}$	–	–	–	-0.2032	-0.1806
				[-1.079, 0.321]	[-0.948, 0.299]
Wave 1 severity		*Yes*	*Yes*	*Yes*	*Yes*
County controls × Week	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3395	3395	3395	3213	3213
Spatial units (counties)	401	401	401	401	401
Clusters	38	38	38	38	38
R^2	0.249	0.257	0.299	0.090	0.090

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The outcome variable is the (change in the) log of weekly deaths per 100000 population in a county, ranging from Nov 11th 2020 until Jan 17th 2021. Controls for wave 1 severity include the log of aggregate case numbers, its square, and case fatality rate in the time period from the first confirmed case until May 17th, 2020. County controls include log population density, log GDP per capita, log average income per capita, share of college graduates, employment share, share of non-German residents, share of workers in the service sector, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.13. Individual-Level Association with PHB — East and West Germany

	Public Health Behavior (PHB)				
	(1)	(2)	(3)	(4)	(5)
Prosociality	0.3269***	0.2971***	0.2974***	0.2073***	0.1529***
	(0.0177)	(0.0180)	(0.0181)	(0.0184)	(0.0155)
Prosociality × East Germany	0.0573	0.0569	0.0589	0.0647*	0.0485
	(0.0392)	(0.0385)	(0.0391)	(0.0375)	(0.0302)
Patience	0.1939***	0.1924***	0.1913***	0.1680***	0.0804***
	(0.0149)	(0.0149)	(0.0149)	(0.0149)	(0.0126)
Risk-taking	-0.2100***	-0.1710***	-0.1718***	-0.1708***	-0.0780***
	(0.0140)	(0.0143)	(0.0143)	(0.0138)	(0.0108)
Negative reciprocity (Direct)	-0.1228***	-0.1075***	-0.1070***	-0.0671***	-0.0178
	(0.0141)	(0.0145)	(0.0145)	(0.0150)	(0.0124)
Socio-demographic factors	*No*	*Yes*	*Yes*	*Yes*	*Yes*
East Germany dummy	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
NUTS-2 region FEs	*No*	*No*	*Yes*	*Yes*	*Yes*
Big 5 personality traits	*No*	*No*	*No*	*Yes*	*Yes*
COVID-19 perceptions	*No*	*No*	*No*	*No*	*Yes*
Observations	5843	5660	5660	5660	5660
Clusters	397	396	396	396	396
R^2	0.213	0.239	0.243	0.299	0.495

Notes: Socio-demographic controls include age and age-squared, gender, education, income, employment status, household size, number of children, and an indicator for having children below age 16. COVID-19 perceptions include general attitudes towards the pandemic, infection experiences, and worrying about oneself, family members, and others being infected. SEs (in parentheses) are clustered at the county level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.14. Weekly Incidence at the Time of the Survey — East and West Germany

	$y_{c,t} = \log(cases_{c,t})$ in county c and week t				
	(1)	(2)	(3)	(4)	(5)
Prosociality	-0.1108 **	-0.1286 **	-0.0943 **	-0.0927 **	0.0021
	[-0.246, -0.024]	[-0.331, -0.016]	[-0.238, -0.003]	[-0.200, -0.024]	[-0.116, 0.091]
Prosociality × East Germany	0.0312	0.0375	-0.0178	0.0005	-0.0056
	[-1.577, 0.992]	[-1.422, 1.100]	[-1.885, 0.793]	[-1.557, 0.821]	[-1.352, 0.787]
Patience	–	0.0256	0.0447	0.0386	0.0741 **
		[-0.092, 0.208]	[-0.052, 0.206]	[-0.052, 0.162]	[0.007, 0.194]
Risk-taking	–	0.0386	-0.0299	-0.0377	-0.0661
		[-0.110, 0.188]	[-0.144, 0.095]	[-0.128, 0.082]	[-0.120, 0.038]
Public health behavior	–	–	–	–	-0.2194 ***
					[-0.354, -0.071]
Wave 1 severity	*No*	*No*	*No*	*Yes*	*Yes*
County controls × Week	*No*	*No*	*Yes*	*Yes*	*Yes*
East Germany × Week	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38	38
R^2	0.189	0.194	0.424	0.483	0.513

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The time period of analysis ranges from Nov 16, 2020, until Jan 17, 2021. County controls include log population density, log GDP per capita, log average income per capita, share of college graduates, employment share, share of non-German residents, share of workers in the service sector, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany. Controls for wave 1 severity include the log of aggregate case numbers, its square, and case fatality rate in the time period from the first confirmed case until May 17th, 2020. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.15. Weekly Growth Rate of Confirmed Cases — East and West Germany

	$y_{c,t} = \log(cases_{c,t}) - \log(cases_{c,t-1})$			
	(1)	(2)	(3)	(4)
Prosociality	-0.0053	-0.0102 *	-0.0163 ***	-0.0086
	[-0.016, 0.004]	[-0.024, 0.001]	[-0.034, -0.006]	[-0.027, 0.004]
Prosociality × East Germany	-0.0074	-0.0071	-0.0068	-0.0072
	[-0.064, 0.009]	[-0.068, 0.016]	[-0.220, 0.129]	[-0.210, 0.123]
Patience	0.0015	-0.0003	0.0058	0.0090 *
	[-0.012, 0.010]	[-0.014, 0.008]	[-0.003, 0.019]	[-0.002, 0.025]
Risk-taking	0.0009	0.0024	-0.0038	-0.0065
	[-0.010, 0.013]	[-0.008, 0.012]	[-0.013, 0.011]	[-0.018, 0.008]
Public health behavior	–	0.0113	–	-0.0186
		[-0.010, 0.030]		[-0.043, 0.005]
$\log(cases_{c,t-2})$	–	–	-0.1232 ***	-0.1285 ***
			[-0.147, -0.103]	[-0.155, -0.105]
Policy stringency$_{c,t-2}$)	–	–	-0.1672	-0.1548
			[-0.632, 0.181]	[-0.593, 0.153]
Wave 1 severity	*Yes*	*Yes*	*Yes*	*Yes*
County controls × Week	*Yes*	*Yes*	*Yes*	*Yes*
East Germany × Week	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38
R^2	0.302	0.302	0.327	0.328

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The outcome variable is the change in the log of weekly cases per capita in a county, ranging from Nov 16th 2020 until Jan 17th 2021. County controls include log population density, log GDP per capita, log average income per capita, share of college graduates, employment share, share of non-German residents, share of workers in the service sector, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany. Controls for wave 1 severity include the log of aggregate case numbers, its square, and case fatality rate in the time period from the first confirmed case until May 17th, 2020. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.16. Weekly Incidence at the Time of the Survey — Standardized

	$y_{c,t} = \log(cases_{c,t})$ in county c and week t				
	(1)	(2)	(3)	(4)	(5)
Prosociality	-0.0820 ***	-0.0749 *	-0.0732 **	-0.0701 **	0.0108
	[-0.167, -0.036]	[-0.178, 0.006]	[-0.174, -0.013]	[-0.145, -0.019]	[-0.052, 0.063]
Patience	–	-0.0168	0.0014	-0.0032	0.0355
		[-0.125, 0.079]	[-0.069, 0.106]	[-0.065, 0.076]	[-0.011, 0.111]
Risk taking	–	0.0062	-0.0222	-0.0268	-0.0480 *
		[-0.063, 0.074]	[-0.091, 0.054]	[-0.081, 0.042]	[-0.088, 0.003]
Public health behavior	–	–	–	–	-0.1767 ***
					[-0.262, -0.093]
Wave 1 severity	*No*	*No*	*No*	*Yes*	*Yes*
County controls × Week	*No*	*No*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38	38
R^2	0.116	0.118	0.357	0.415	0.481

Notes: This table estimates the same specifications as Table 2.2, but with the dependent variable standardized. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.17. Weekly Growth Rate of Confirmed Cases — Standardized

	$y_{c,t} = \log(cases_{c,t}) - \log(cases_{c,t-1})$			
	(1)	(2)	(3)	(4)
Prosociality	-0.0133 **	-0.0141	-0.0318 ***	-0.0105
	[-0.027, -0.002]	[-0.033, 0.003]	[-0.055, -0.016]	[-0.037, 0.011]
Patience	-0.0018	-0.0021	-0.0017	0.0090
	[-0.020, 0.010]	[-0.022, 0.013]	[-0.016, 0.020]	[-0.012, 0.038]
Risk taking	0.0002	0.0004	-0.0064	-0.0134
	[-0.017, 0.019]	[-0.018, 0.018]	[-0.023, 0.014]	[-0.038, 0.011]
Public health behavior	–	0.0018	–	-0.0496 **
		[-0.031, 0.032]	–	[-0.096, -0.009]
$\log(cases_{c,t-2})$	–	–	-0.1578 ***	-0.1765 ***
			[-0.183, -0.135]	[-0.214, -0.141]
Policy stringency$_{c,t-2}$	–	–	-0.3508	-0.2993
			[-1.251, 0.422]	[-1.117, 0.332]
Wave 1 severity	*Yes*	*Yes*	*Yes*	*Yes*
County controls × Week	*Yes*	*Yes*	*Yes*	*Yes*
Week fixed effects	*Yes*	*Yes*	*Yes*	*Yes*
Observations	3609	3609	3609	3609
Spatial units (counties)	401	401	401	401
Clusters (NUTS-2 regions)	38	38	38	38
R^2	0.293	0.293	0.315	0.317

Notes: This table estimates the same specifications as Table 2.3, but with the dependent variable standardized. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.18. Overall Number of Confirmed Cases in First and Second Wave

	y_i = log overall confirmed cases per 100000 population in county i			
	"first wave"		"second wave"	
	(1)	(2)	(3)	(4)
Prosociality	–	-0.0546 [-0.186, 0.053]	–	-0.0913 ** [-0.231, -0.011]
Patience	–	0.0113 [-0.110, 0.182]	–	0.0025 [-0.092, 0.146]
Risk-taking	–	0.0938 [-0.017, 0.212]	–	-0.0238 [-0.124, 0.089]
log population density	0.4055 ** [0.045, 0.738]	0.4142 ** [0.047, 0.757]	0.0634 [-0.192, 0.347]	0.0847 [-0.151, 0.341]
Employed / population	3.5720 *** [2.072, 5.091]	3.6969 *** [2.150, 5.276]	1.5709 * [-0.056, 3.428]	1.4675 * [-0.156, 3.458]
Share of jobs in service sector	-3.1460 *** [-4.694, -1.551]	-3.0334 *** [-4.559, -1.429]	-1.4531 * [-3.077, 0.086]	-1.4196 * [-3.052, 0.078]
Further county characteristics	*Yes*	*Yes*	*Yes*	*Yes*
Observations	401	401	401	401
Clusters	38	38	38	38
R^2	0.497	0.509	0.265	0.323

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The "first wave" is defined as the time period until May 17th, 2020; the "second wave" is defined as time period between Sep 28th 2020 and Feb 28th 2021. Further regressors include log GDP per capita, log average income per capita, share of college graduates, share of non-German residents, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.19. Aggregate Number of Deaths in First and Second Wave

	y_i = *log COVID-19 deaths per 100000 population in county i*			
	"First wave"		"Second wave"	
	(1)	(2)	(3)	(4)
Prosociality	–	-0.1835 **	–	-0.1157 *
		[-0.373, -0.043]		[-0.312, 0.003]
Patience	–	0.0571	–	-0.0345
		[-0.106, 0.261]		[-0.157, 0.185]
Risk-taking	–	0.2022 ***	–	-0.0254
		[0.066, 0.376]		[-0.144, 0.101]
log population density	0.2898	0.3214	0.0433	0.0686
	[-0.175, 0.786]	[-0.147, 0.789]	[-0.256, 0.378]	[-0.200, 0.353]
Employed / population	5.1239 ***	5.4715 ***	1.1927	0.9480
	[2.655, 7.635]	[2.984, 7.960]	[-0.743, 3.316]	[-1.224, 3.444]
Share of jobs in service sector	-4.1070 ***	-3.8792 ***	-1.1468	-1.0467
	[-6.428, -1.782]	[-6.176, -1.563]	[-3.186, 0.791]	[-3.141, 0.904]
Further county controls	*Yes*	*Yes*	*Yes*	*Yes*
Observations	381	381	401	401
Clusters	38	38	38	38
R^2	0.288	0.322	0.272	0.321

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. The "first wave" is defined as the time period until May 17th, 2020; the "second wave" is defined as time period between Sep 28th 2020 and Feb 28th 2021. Further controls include log average income per capita, share of college graduates, share of non-German residents, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.A.20. Aggregate Number of Cases and Deaths in Third Wave

	"Third wave": starting from March 1st, 2021			
	log cumulative cases		log cumulative deaths	
	(1)	(2)	(3)	(4)
Prosociality	–	-0.1020 ***	–	-0.0947 **
		[-0.186, -0.049]		[-0.240, -0.004]
Patience	–	0.0106	–	-0.0257
		[-0.064, 0.121]		[-0.134, 0.163]
Risk-taking	–	0.0220	–	-0.0097
		[-0.036, 0.106]		[-0.110, 0.113]
log population density	0.0973	0.1196	0.0773	0.0977
	[-0.072, 0.258]	[-0.030, 0.266]	[-0.259, 0.389]	[-0.230, 0.410]
log GDP per capita	-0.0826 **	-0.1105 **	-0.4362 ***	-0.4030 ***
	[0.489, 9.790]	[0.237, 9.521]	[3.679, 12.716]	[3.037, 12.165]
Employed / population	2.3504 ***	2.3284 ***	2.0686	1.8922
	[0.731, 4.251]	[0.676, 4.329]	[-0.514, 4.643]	[-0.962, 4.837]
Share of jobs in service sector	-2.3614 ***	-2.2861 ***	-2.3779 *	-2.2885 *
	[-4.179, -0.668]	[-4.092, -0.604]	[-4.840, 0.099]	[-4.825, 0.192]
Population share age 65 or above	5.3880	5.0146	11.4636	10.9036
	[-0.412, 0.301]	[-0.415, 0.236]	[-0.516, 0.549]	[-0.517, 0.579]
Further county controls	*Yes*	*Yes*	*Yes*	*Yes*
Observations	401	401	401	401
Clusters	38	38	38	38
R^2	0.305	0.365	0.319	0.346

Notes: Bootstrapped 95%-confidence-intervals in brackets (clustered at NUTS-2 level), obtained using wild bootstrapping with Rademacher-weights and 9,999 simulations. Dependent variables are log cumulative cases (deaths) per 100000 population. The time period of analysis goes until July 8, 2021. Further controls include log average income per capita, share of college graduates, share of non-German residents, share of population below age 18, share of population age 65 or above, and border country dummies for each neighboring country of Germany. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix 2.B Survey Questions and Data

In this section, we describe all the survey questions that respondents were asked to complete as part of the questionnaire (subsections 2.B.1- 2.B.6), including those that we use to construct major dependent or independent variables for the main paper, i.e. pandemic-related behavior and prosocial preferences. We translated all questions into English for this Section. For the complete original questionnaire in German language, see Section 2.D. In subsections 2.B.7, we describe our sample recruiting and data cleaning procedures, and in subsection 2.B.8, we describe how we construct our individual-level variables based on the survey items.

2.B.1 Public Health Behavior

To what extent do the following statements apply to your own behavior? *Please rate on a scale from 1 to 7. The value 1 means: does not apply at all. The value 7 means: fully applies.*

- I keep a distance of at least 1.5 meters from other people.
- I will socially isolate myself if I have had contact with an infected person.
- I always keep up to date on news about the pandemic.
- I wash and disinfect my hands regularly.
- I am going to get vaccinated against the coronavirus when a vaccine becomes available.
- I cough and sneeze into the crook of my elbow.
- I wear a face mask in public.
- I ventilate regularly when several people are using a room.
- I avoid social contacts as much as possible.
- I will inform other people if I am infected with the coronavirus.

2.B.2 Questions from the Peference Survey Module

To elicit time, risk, and social preferences, we included some questions from experimentally-validated preference survey module by Falk et al. (2016) and Falk et al. (2018) in our questionnaire. All qualitative questions were rated on an 11-point Likert scale from 0 to 10, where the value of 0 indicates complete disagreement or unwillingness, and the value 10 indicates complete agreement or willingness.

Altruism was elicited using one qualitative question and a quantitative decision involving a hypothetical donation. Positive reciprocity, indirect negative reciprocity, and trust were elicited using one qualitative question each. Direct negative reciprocity was elicited using two qualitative questions, and patience and risk taking were elicited using one qualitative item each.

Altruism, Reciprocity, and Trust

How willing are you to give to good causes without expecting anything in return.

not willing at all □—□—□—□—□—□—□—□—□—□—□ very willing to do it

Imagine the following situation:

Today you unexpectedly received 1,000 euros. How much of this amount would you donate to a good cause? _____

If someone does me a favor, I am willing to return it.

does not describe me at all □—□—□—□—□—□—□—□—□—□—□ describes me perfectly

How willing are you to punish someone who treats you unfairly, even if there may be costs for you?

not at all willing to do it □—□—□—□—□—□—□—□—□—□—□ very willing to do it

If I am treated very unfairly, I will take revenge at the first occasion, even if there is a cost to do so.

does not describe me at all □—□—□—□—□—□—□—□—□—□—□ describes me perfectly

How willing are you to punish someone who treats others unfairly, even if there may be costs for you?

not at all willing to do it □—□—□—□—□—□—□—□—□—□—□ very willing to do it

I assume that people have only the best intentions.

does not describe me at all □—□—□—□—□—□—□—□—□—□—□ describes me perfectly

Risk and time preferences

In general, how willing are you to take risks?

completely unwilling to take risks □—□—□—□—□—□—□—□—□—□—□ very willing to take risks

How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?

completely unwilling to do so □—□—□—□—□—□—□—□—□—□—□ very willing to do so

Control questions

I am good at math.

does not describe me at all □—□—□—□—□—□—□—□—□—□ describes me perfectly

I tend to put off tasks even when I know it would be better to do them now.

does not describe me at all □—□—□—□—□—□—□—□—□—□ describes me perfectly

2.B.3 Demographic and socio-economic questions

Please enter your year of birth. _____

Please select your gender.

- □ Female
- □ Male
- □ Others

Which state do you live in? _____

Please enter your zip code? _____

How long have you been living at your current place of residence? _____

What is your highest educational qualification?

- □ No degree
- □ Elementary/secondary school certificate (GDR: 8th grade)
- □ Secondary school leaving certificate (GDR: 10th grade)
- □ Fachhochschulreife (qualification from a technical college)
- □ Abitur/university entrance qualification
- □ Fachhochschule (formerly: engineering school, teacher training, GDR: engineer and technical college degree)
- □ University, college degree
- □ Doctorate
- □ Other educational qualification

How many people live in your household (ie unit living and working together)? _____

How many children do you have? _____

Which of the following best describes your current employment status?

- □ full-time employed

- ☐ part-time employed
- ☐ self-employed
- ☐ in educational/vocational training
- ☐ non-employed

What is approximately your net monthly household income in Euro?

- ☐ under €900
- ☐ €900 to under €1,300
- ☐ €1,300 to less than €1,500
- ☐ €1,500 to less than €2,000
- ☐ €2,000 to less than €2,600
- ☐ €2,600 to less than €3,200
- ☐ €3,200 to less than €4,500
- ☐ €4,500 to less than €6,000
- ☐ €6,000 and more
- ☐ not specified

2.B.4 Big Five personality index

How well does each of the following statements describe you as a person?
Please answer as honestly and spontaneously as possible on a scale from 1 to 7. The value 1 means: Not at all applicable. The value 7 means: Completely applies.

I am someone who...

- ... works thoroughly.
- ... is communicative, talkative.
- ... is sometimes rude to others.
- ... is original, brings in new ideas.
- ... worries a lot.
- ... can forgive.
- ... is rather lazy.
- ... is outgoing and sociable.
- ... values artistic experiences.
- ... gets nervous easily.
- ... gets tasks done effectively and efficiently.
- ... is reserved.
- ... treats others with respect and kindness.
- ... has a vivid imagination.
- ... is relaxed, can handle stress well.

2.B.5 Other pandemic-related questions

How much do you agree with the following statements? *Please rate on a scale from 1 to 5. The value 1 means: completely disagree. The value 5 means: completely agree.*

- The pandemic has a negative effect on my financial situation.
- The pandemic has a negative effect on my personal life.
- The government's measures against the pandemic are way too strict.
- Overall, Germany has managed the pandemic well so far.
- The media takes COVID-19 way too seriously.

Have you contracted COVID-19 before?

- ☐ Yes
- ☐ No
- ☐ prefer not to say

Do you personally know someone who has contracted COVID-19?

- ☐ Yes
- ☐ No
- ☐ prefer not to say

Do you personally know someone who has died from Covid-19?

- ☐ Yes
- ☐ No
- ☐ prefer not to say

When was the last time you had the flu? _____

The 7-day incidence rate indicates the number of new COVID-19 cases (i.e., people who tested positive for the coronavirus) per 100,000 inhabitants within the past seven days. It is considered an important indicator for assessing the current pandemic situation.

Please estimate the 7-day incidence rate in your city (note: as of December 17, the value for all of Germany is 179). _____

Please rate on a scale from 1 to 7. The value 1 means: not worried at all. The value 7 means: extremely worried.
How worried are you about ...

- ... contracting COVID-19 yourself.
- ... friends or relatives contracting COVID-19.
- ... other people in general contracting COVID-19.

How high do you rate the risk of contracting COVID-19 within the next 3 months?

Very low □—□—□—□—□—□—□ Very high

Have you installed the Corona-Warn-App on your current mobile phone?

- □ Yes
- □ No

Which type of smartphone do you use most of the time?

- □ Android smartphone (e.g. Samsung, Huawei, ...)
- □ iPhone (Apple)
- □ other smartphone (e.g. Windows-Phone, Blackberry, ...)
- □ I don't use a smartphone

If not: How likely is it that you would install the Corona-Warn-App within the next few weeks?

Very unlikely □—□—□—□—□—□—□ Very likely

If yes: How likely is it that you would report your infection status to the Corona-Warn-App in case you would be tested positiv?

Very unlikely □—□—□—□—□—□—□ Very likely

How much do you agree with the following statements about the Corona-Warn-App?
Please answer on a scale from 1 to 7. The value 1 means: I do not agree at all. The value 7 means: I completely agree.

The Corona-Warn-App ...

- ... helps to slow down the spread of the coronavirus in Germany.
- ... helps to slow down the spread of the coronavirus in my city.
- ... is of no real use to me personally.
- ... is a good way to trace infection chains.
- ... is not used by enough people yet.

Donation option: The following scenario has a 25% probability of actually being implemented. So you should think carefully about what you want to do. It may involve real money.

You have 1 Euro at your disposal. You are free to decide how much of this you donate and what share you keep for yourself. Your donation will be used for an online advertising campaign on social media, which encourages more people (including in your region) to use the Corona-Warn-App. Past data has shown that 50 cents of advertising expenditure correspond to 1 additional Corona-Warn-App installation on average. You will get to keep the rest of the amount that you don't donate.

At the end of the survey, a random number generator determines whether this donation and the additional remuneration will actually be paid out. Please move the slider to decide on your allocation:

How much would you like to donate? Your donation is _____ Euro.

2.B.6 News consumption, political attitudes, and values

Where do you inform yourself about the news?
People use different news sources to learn about what is happening around them and in the world. For each of the following sources, please indicate how often you use them:

	Daily	Weekly	Monthly	Less than monthly	Never
Newspaper	☐	☐	☐	☐	☐
TV	☐	☐	☐	☐	☐
Radio	☐	☐	☐	☐	☐
News sites on the internet	☐	☐	☐	☐	☐
Mobile phone (WhatsApp, Telegram, etc.)	☐	☐	☐	☐	☐
Social media (Facebook, Twitter, etc.)	☐	☐	☐	☐	☐
Conversations with friends, colleagues and acquaintances	☐	☐	☐	☐	☐

About how much time do you spend on social media (e.g. Facebook, Instagram)?: _____

If there were general elections tomorrow, which party would you vote for?

- ☐ CDU/CSU
- ☐ Buendnis'90/Die Gruenen
- ☐ SPD
- ☐ AfD
- ☐ Die Linke
- ☐ FDP
- ☐ Other

Would you actually vote?

- ☐ Yes
- ☐ No
- ☐ Undecided

How satisfied are you with how the political system in Germany works today?
Please rate on a scale of 0 to 10, where 0 is "not at all satisfied" and 10 is "completely satisfied".

not at all satisfied ☐—☐—☐—☐—☐—☐—☐—☐—☐—☐—☐ completely satisfied

	totally agree not to				totally agree and quite to
	1	2	3	4	5
There are many very important things happening in the world which the public is never informed about.	☐	☐	☐	☐	☐
Government agencies monitor all citizens.	☐	☐	☐	☐	☐
There are secret powers that control the world.	☐	☐	☐	☐	☐

How much do you agree with the following statements?

People have different views about themselves and how strong they feel connected to their environment and the rest of the world.
How strongly do you feel connected to...

	Not at all connected	A little connected	Somewhat connected	Quite connected	Very connected
The town or city you live in	☐	☐	☐	☐	☐
The region you live in	☐	☐	☐	☐	☐
Germany	☐	☐	☐	☐	☐
Europe	☐	☐	☐	☐	☐
The whole world	☐	☐	☐	☐	☐

2.B.7 Data cleaning

The survey was administered to a sample of individuals in Germany through the market research company Dynata. Participants between 18 and 65 years old were recruited via email-invitation, with quotas on age, gender, and state to achieve national-level representativeness along these dimensions for the relevant age group of our sample. The questionnaire was web-based could be completed online on PC, laptop, tablet, or smartphone. It consisted of 20 pages in total and the median response time was about 13 minutes. A total 7,052 individuals responded to our survey, and 6,826 respondents completed every survey question on preferences and public health behavior. In accordance with Dynata policy, we used several different criteria to check response quality and to exclude bad responses: speeding (i.e. unreasonable quick response time), inconsistencies or conflicting answers, excessive straightlining (e.g. always ticking the same box in Likert scales), and an attention check question.

To check for speeding, we recorded the duration spent on answering questions on each page of the survey, as well as for completing the entire survey. We immediately

excluded all responses where the survey-taker spent less than 2 seconds per question on average on at least 3 pages. Then, we flagged responses as potentially bad if the survey was completed in less than one-third of the median completion time. With regard to inconsistencies, we flagged responses as potentially bad if they would imply that the respondent became a parent at the age of 12 or younger, that the respondent lived at the current place of residence since before he or she was born, or if the zip code did not match the state of residence. With regard to straightlining, we flagged responses as potentially bad if they included at least 2 modules of Likert-scale-type sequences (e.g. preferences survey module, public health behavior) in which always the same value was selected. Finally, we flagged responses as potentially bad if the survey-taker failed an attention check question at the beginning of the survey. The attention check consisted of an absurd question ("[...] How interested are you in learning about the impact of traffic noise on the singing bird population in German cities?") for which the description prescribed a particular response in order to "demonstrate that you answer this survey carefully". We excluded all responses which were flagged as potentially bad in at least 2 out of 4 criteria. In total, 992 responses (i.e. below 15%) were removed for our analyses, thus giving us our main sample size of 5,843. In some analyses that include socioeconomic variables as controls, an additional 183 responses drop out due to missing information about education or income.

2.B.8 Variable Construction

Public Health Behavior

To construct the factor variable on public health behavior (PHB), we assume that compliance to public health behavior is driven by one underlying factor, and conduct factor analysis on the ten survey items on PHB (see Section 2.B.1). The results of our factor analysis support this notion. From Figure 2.B.1, we see that the eigenvalue on the first factor is 4.47, whereas those on the remaining factors are below 1. Table 2.B.1, which shows the factor loadings on each survey item, indicates that all survey items are highly correlated with the underlying factor. Furthermore, Cronbach's alpha is 0.87, indicating that all the PHB items are highly interrelated.

Prosociality

We construct the prosociality variable via principal component analysis on the five (standardized) survey items for altruism, positive reciprocity, trust, and indirect negative reciprocity. See Section 2.B.2 for the wording and scale of the questions. Note that we do not include the two questions on direct negative reciprocity ("If I am

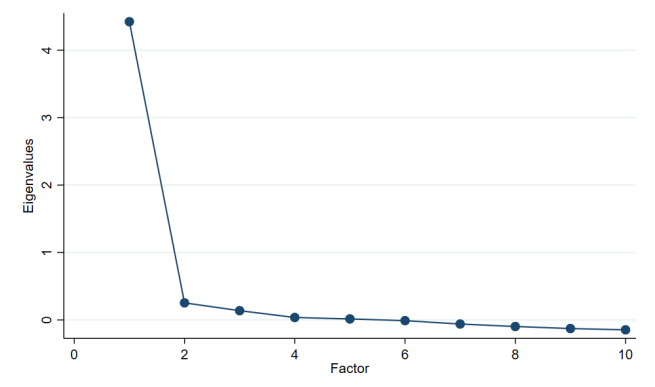


Figure 2.B.1. Scree Plot PHB

Notes: Eigenvalues on factors obtained from a factor analysis of PHB survey items.

Table 2.B.1. Factor Loadings PHB

	Factor loadings
Social distancing 1.5 meters	0.769
Self-quarantine if risky contact	0.746
Keep informed about pandemic	0.618
Wash and disinfect hands	0.686
Get vaccinated when vaccine available	0.434
Sneeze and cough in elbow	0.589
Wear mask	0.690
Regular ventilation when indoors	0.707
Avoid social contacts	0.713
Would inform others if infected	0.633

Notes: Factor loadings on survey items used to construct PHB.

treated very unfairly, I will take revenge at the first occasion, even if there is a cost to do so.", and "How willing are you to punish someone who treats you unfairly, even if there may be costs for you?") as these do not square with our notion of prosociality. From Table 2.B.2, we see that the first principal component for prosociality explains approximately 36% of the total variance. The subsequent components explain 20%, 17%, 17%, and 10% of the variance respectively, which suggest that there are several distinct aspects to social preferences. Though these components could also explain adherence to PHB, this is not the aim of our study. Rather, our analysis is guided by theoretical considerations— We are interested in how a particular aspect of social preferences, i.e. prosociality, predicts adherence to PHB. In this regard, we see from Table 2.B.3 that the first principal component assigns weights to the underlying variables that are congruent with our notion of prosociality: 0.2 and 0.6 for the

two altruism survey items, 0.49 for positive reciprocity, 0.4 for trust, and 0.4 for indirect negative reciprocity.

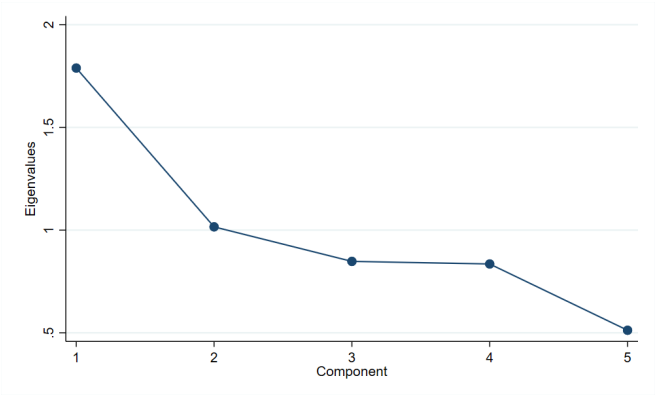


Figure 2.B.2. Scree Plot Prosociality

Notes: Eigenvalues on components obtained from a principal component analysis of prosocial preference survey items.

Table 2.B.2. Eigenvalues and Proportion of Total Variance, Prosocial Preferences Components

	Eigenvalues	Proportion
Component 1	1.789	0.358
Component 2	1.016	0.203
Component 3	0.848	0.170
Component 4	0.835	0.167
Component 5	0.512	0.102

Notes: Eigenvalues and proportion of total variance on components of principal component analysis on standardized prosocial preferences survey items.

Table 2.B.3. Weights on Prosociality Survey Items, Prosocial Preferences Components

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5
Willingness to give for a good cause	0.602	-0.0638	0.00970	-0.269	-0.749
Donation amount out of 1000 Euro	0.257	0.835	0.129	-0.381	0.274
Postive reciprocity	0.485	-0.514	-0.0608	-0.403	0.578
Negative reciprocity (indirect)	0.415	-0.00562	0.679	0.592	0.130
General trust towards people	0.405	0.187	-0.720	0.519	0.113

Notes: Weights on prosociality survey items for each component obtained by principal component analysis of prosocial preferences survey items.

Other Variables

Table 2.B.4. Overview of All Individual-Level Control Variables Used in the Paper

Variable	Question(s)	Value formatting
Age	Please state your year of birth.	Age in years and age^2
Gender	Please select your gender.	Categorical
County	Please enter your zip code	Categorical
Education	What is your highest educational attainment?	4 categories: No degree, Secondary degree, Abitur, University degree
Income	What is approximately your net monthly household income in Euro?	10 categories: <900, 900-1300, 1300-1500, 1500-2000, 2000-2600, 1.6k-3.2k, 3200-4500, 4500-6000, >6000, prefer not to say
Employment status	Which of the following best describes your current employment status?	5 categories: full-time employed, part-time employed, self-employed, educational/vocational training, non-employed
Household size	How many people live in your household?	5 categories: 1, 2, 3, 4, 5 or more
Number of children	How many children do you have?	4 categories: none, 1, 2, 3 or more
Children below 16	What is the age of your youngest child?	Indicator for age $\leq$ 16 years
Pandemic skeptical	1) The government's measures against the pandemic are way too strict. 2) The media takes COVID-19 way too seriously.	Mean of two 5-point Likert scales (standardized)
Pandemic affected	1) The pandemic has a negative effect on my financial situation. 1) The pandemic has a negative effect on my personal life.	Mean of two 5-point Likert scales (standardized)
Worry self	How worried are you about contracting COVID-19 yourself?	7-point Likert scales
Worry family	How worried are you about friends or family contracting COVID-19?	7-point Likert scales
Worry others	How worried are you about people in general contracting COVID-19?	7-point Likert scales
Infected	Have you contracted COVID-19 before?	Categorical: yes, no, prefer not to say
Know infected	Do you personally know someone who has contracted COVID-19?	Categorical: yes, no, prefer not to say
Know died	Do you personally know someone who has died from COVID-19?	Categorical: yes, no, prefer not to say
Patience	How willing are you to give up sth that is beneficial for you today in order to benefit more from that in the future?	11-point Likert scale
Risk taking	In general, how willing are you to take risks?	11-point Likert scale
Big Five personality	15 item BFI-S (Gerlitz and Schupp, 2005)	5 standardized variables: openness, conscientiousness, extraversion, agreeableness, neuroticism

Table 2.B.5. Survey Items Used to Construct Big Five Personality Factors

Personality Trait	Definitions (Becker et al., 2012, p.466)	Survey items
Openness	Individual differences in the tendency to be open to new aesthetic, cultural, and intellectual experiences	... is original, brings in new ideas. ... values artistic experiences. ... has a vivid imagination.
Conscientiousness	The tendency to be organized, responsible, and hardworking; located at one end of a dimension of individual differences (conscientiousness versus lack of direction)	... works thoroughly. ... is rather lazy. ... tasks done effectively and efficiently.
Extraversion	An orientation of one's interests and energies toward the outerworld of people and things rather than the inner world of subjective experience; includes the qualities of being outgoing, gregarious, sociable, and openly expressive	... is reserved. ... is communicative, talkative. ... can be outgoing, is sociable.
Agreeableness	The tendency to act in a cooperative, unselfish manner; located at one end of a dimension of individual differences (agreeableness versus disagreeableness)	... sometimes being rude to others. ... can forgive. ... treat others with respect and kindness.
Neuroticism	A chronic level of emotional instability and proneness to psychological distress	... often worries. ... is relaxed, can handle stress well. ... gets nervous easily.

Notes: We construct each Big Five personality factor by conducting factor analysis on the relevant survey items, and then standardizing the resultant factor variable. Definitions are taken from Becker et al. (2012, p.466).

Appendix 2.C Data Sources for Regional Data

2.C.1 Aggregation of Survey Measures

As the sample for our online survey was recruited to be representative only at the national-level, we weight observations to improve representativeness of at the NUTS-2 level. Specifically, we obtain official data on age, gender, and education by region (see 2.C.3) and calculate sampling weights to match the regional population with regard to age-by-gender (2×2 matrix of age above/below 40 with female/male) and the share of adults with a university degree. We do so using a simple stepwise raking procedure (Battaglia, Hoaglin, and Frankel, 2009), in which we first calculate initial weights so that our sample matches the population age-gender distribution, then readjust these weights to match the share of adults with a university degree, then readjust to match age-gender again, and so on, until the weights converge. Using the final sampling weights, we then calculate the NUTS-2 region-level average of the PHB, prosociality, patience, and risk taking measures described in Section 2.B.

2.C.2 COVID-19 Incidences and Deaths

We obtained official data on the daily number of confirmed COVID-19 cases and deaths as reported by the Robert-Koch-Institut (RKI), the the federal government agency and research institute responsible for disease control and prevention in Germany. It can be publicly accessed via the Corona data hub (https://npgeo-corona-npgeo-de.hub.arcgis.com). The information is updated daily at the county level, although there can be delays in reporting by local health authorities, especially on weekends and on holidays. We therefore aggregate all numbers to the weekly level, with each week beginning on Monday and ending on Sunday. Furthermore, we adjust the number of cases and deaths by each county's population size to obtain the incidence rates, defined as number of confirmed cases/deaths per 100,000 population in a period of 7 days.

2.C.3 Demographic and Socio-Economic Information

We collect data on pre-pandemic county characteristics from the publicly accessible official database of the German federal statistical office and the state statistical offices (Regionaldatenbank, https://www.regionalstatistik.de/genesis/online). This includes information on population and demographics, education, economic indicators, employment statistics, etc. We complement this with data collected by infas360 in an effort to synthesize databases that can be relevant with regard to COVID-19 and make them available to researchers (Corona-Datenplattform, https:

//www.corona-datenplattform.de). In Table 2.C.1, we provide a complete list of all variables that we use in the paper, the data source, and from which year it is.

Table 2.C.1. Overview of All County-Level Control Variables Used in the Paper

Variable	Year	Source(s)
Population density (settlement area only)	2018	Corona-Datenplattform
GDP per capita	2017	Regionaldatenbank
Average disposable income per capita	2017	Regionaldatenbank
Share of population with college degree	2018	Regionaldatenbank
Employment share	2019	Regionaldatenbank
Share of employees in service sector	2017	Corona-Datenplattform
Share of non-German residents	2019	Regionaldatenbank
Share of population below age 18	2019	Regionaldatenbank
Share of population aged 65 or above	2019	Regionaldatenbank
Border country indicators	2021	Any map of choice
Local policy restrictions	2021	Corona-Datenplattform
2019 EU parliament election turnout & vote shares	2019	Regionaldatenbank
2017 general election turnout & vote shares	2017	Regionaldatenbank
Civic associations per 100,000 population	2008	Franzen and Botzen (2011)

2.C.4 Local Policy Stringency

Finally, to evaluate the role of county-level stringency of policy restrictions aimed to combat the pandemic, we obtain data from the infas360 Corona-Datenplattform (https://www.corona-datenplattform.de/dataset/massnahmen_oberkategorien_kreise) which indicates for 23 categories of possible restrictions (e.g. curfew, school closure, ...) whether they were in place in a certain county at a particular point in time. To construct a local policy stringency index, we sum up all 23 indicator variables and then normalize this index to range between 0 and 100, where 0 means that not a single restriction was in place, and 100 means that every single restriction was mandated by the local government. The 23 categories entail restrictions regarding: private gatherings, public gatherings, secondary schools, primary schools, daycare centers, indoor public events, outdoor public events, cultural

institutions (museums, theaters, ...), retail and wholesale, gastronomy, services and craft, nightclubs and bars, hotels, indoor sports, outdoor sports, domestic travel, international travel, mask wearing, workplace, curfews, public transport, physical distancing, testing.

Appendix 2.D Full Original Questionnaire (in German)

Einführung

COVID-19 (ugs. auch Corona) ist eine Infektionskrankheit, die von einem neu entdeckten Coronavirus ausgelöst wird. Die aktuelle Coronavirus-Pandemie hat sicher auch Ihr Leben stark verändert. Bei dieser Umfrage geht es um Ihre Einstellungen zur Corona-Pandemie sowie Ihre allgemeinen Einstellungen im Leben.
Die Umfragedauer beträgt in etwa 20 Minuten. Für die mobile Version empfehlen wir Ihnen das Smartphone im Querformat zu nutzen.

Datenschutz- und Einwillungserklärung
In der folgenden Umfrage werden wir Ihnen Fragen stellen zu möglicherweise sensiblen Themen bezüglich Ihrer Gesundheit und Ihren politischen Einstellungen. Ziel der Studie ist es, die Einstellungen von Menschen in Deutschland zur aktuellen Corona- Pandemie besser zu verstehen.
Die Studie wird duchgeführt von Forschern der Universität Bonn. Verantwortlicher für die Erhebung und Auswertung der Daten ist Ximeng Fang. Die erhobenen Daten werden auf sicheren Servern der Universität Bonn aufgezeichnet und ausschließlich für Forschungszwecke verwendet. Dabei sind keine Rückschlüsse auf Ihre Person möglich, Sie bleiben also jederzeit vollständig anonym.
Bitte geben Sie an, ob Sie mit der Erhebung und Auswertung Ihrer Daten aus dieser Umfrage einverstanden sind. Sie können Ihre Einwilligungserklärung jederzeit widerrufen.

Sind Sie mit der Erhebung und Auswertung Ihrer Daten aus dieser Umfrage einverstanden?

- ☐ Ich bin einverstanden und möchte an dieser Umfrage teilnehmen.
- ☐ Ich bin nicht einverstanden.

Manchmal lesen Umfrageteilnehmer nicht sorgfältig die Fragen und klicken sich einfach nur schnell durch. Das kann die Ergebnisse von wissenschaftlichen Studien verfälschen. Geben Sie darum bei der folgenden Frage bitte den Wert 2 als Antwort an, um zu zeigen, dass Sie diese Umfrage sorgfältig beantworten.

In Anbetracht dieses Problems: Wie stark sind Sie interessiert an den Auswirkungen von Verkehrslärm auf die Vogelpopulation in deutschen Städten?

Überhaupt nicht interessiert ☐—☐—☐—☐—☐—☐—☐ Sehr stark interessiert

Allgemeine Angaben zu Ihrer Person

Vielen Dank, dass Sie sich die Zeit nehmen für unsere Umfrage! Wir möchten Ihnen zunächst einige allgemeine Fragen zu Ihrer Person stellen.

Bitte geben Sie ihr Geburtsjahr an. _____

Bitte geben Sie Ihr Geschlecht an.

- ☐ Weiblich
- ☐ Männlich
- ☐ Divers

Wie viel Zeit verbringen Sie in etwa auf sozialen Medien (z.B. Facebook, Instagram)?: _____

Was für ein Smartphone benutzen Sie im Alltag?

- ☐ Android-Smartphone (z.B. Samsung, Huawei, …)
- ☐ iPhone (Apple)
- ☐ anderes Smartphone (z.B. Windows-Phone, Blackberry, …)
- ☐ Ich benutze kein Smartphone

In welchem Bundesland leben Sie? _____

Was ist Ihre Postleitzahl? _____

Seit welchem Jahr leben Sie an Ihrem aktuellen Wohnort?: _____

Menschen haben verschiedene Ansichten über sich selbst und wie stark Sie sich mit ihrem Umfeld und dem Rest der Welt verbunden fühlen.
Wenn Sie sich einmal diese Liste ansehen, **wie stark fühlen Sie sich verbunden mit…**

	Überhaupt nicht verbunden	Nicht sehr verbunden	Ein wenig verbunden	Ziemlich verbunden	Sehr verbunden
Dem Ort oder der Stadt, in der Sie leben	☐	☐	☐	☐	☐
Der Region, in der Sie leben	☐	☐	☐	☐	☐
Deutschland	☐	☐	☐	☐	☐
Europa	☐	☐	☐	☐	☐
Der ganzen Welt	☐	☐	☐	☐	☐

Wie sehr stimmen Sie den folgenden Aussagen zu?
Bitte bewerten Sie auf einer Skala von 1 bis 5. Der Wert 1 bedeutet: Stimme überhaupt nicht zu. Der Wert 5 bedeutet: Stimme voll und ganz zu.

	stimme überhaupt nicht zu				stimme voll und ganz zu
	1	2	3	4	5
Ich bin finanziell negativ betroffen von der Corona-Pandemie.	☐	☐	☐	☐	☐
Ich bin in meinem persönlichen Leben stark eingeschränkt durch die Pandemie.	☐	☐	☐	☐	☐
Ich finde die Regierungsmaßnahmen gegen Corona überzogen.	☐	☐	☐	☐	☐
Insgesamt betrachtet hat Deutschland die Corona-Krise bisher gut bewältigt.	☐	☐	☐	☐	☐
Die Medien nehmen das Coronavirus viel zu ernst.	☐	☐	☐	☐	☐

Haben Sie sich in der Vergangenheit mit dem Coronavirus infiziert?

- ☐ Ja
- ☐ Nein
- ☐ keine Angabe

Kennen Sie persönlich jemanden, der sich mit dem Coronavirus infiziert hat?

- ☐ Ja
- ☐ Nein
- ☐ weiß nicht

Kennen Sie persönlich jemanden, der an Covid-19 gestorben ist?

- ☐ Ja
- ☐ Nein
- ☐ weiß nicht

Wann sind Sie das letzte Mal an Grippe erkrankt? _____

Nun etwas ganz anderes. Unsere alltäglichen Handlungen werden davon beeinflusst, welche Grundüberzeugungen wir haben. Darüber ist in der Wissenschaft wenig bekannt. In den folgenden Seiten zeigen wir Ihnen einige unterschiedliche Eigenschaften, die eine Person haben kann. Wahrscheinlich werden manche Eigenschaften auf Sie persönlich mehr zutreffen als andere.

Bei allen Fragen geht es darum, wie Sie sich tatsächlich einschätzen, und nicht darum, wie Sie gerne sein würden. Bitte antworten Sie deshalb so ehrlich und spontan wie möglich. Es gibt keine richtigen oder falschen Antworten.

Versuchen Sie im Allgemeinen, Risiken zu vermeiden, oder sind Sie im Allgemeinen ein risikobereiter Mensch? *Bitte schätzen Sie sich persönlich ein, auf einer Skala von 0 bis 10. Der Wert 0 bedeutet: Überhaupt nicht bereit, Risiken einzugehen. Der Wert 10 bedeutet: Sehr bereit, Risiken einzugehen.*

überhaupt riskobereit □—□—□—□—□—□—□—□—□—□—□ sehr risikobereit

Wir fragen Sie nun nach Ihrer Bereitschaft sich in einer bestimmten Art zu verhalten.
Bitte verwenden Sie wieder eine Skala von 0 bis 10. Der Wert 0 bedeutet: Überhaupt nicht bereit es zu tun. Der Wert 10 bedeutet: Sehr bereit es zu tun.

Wie sehr wären Sie bereit auf etwas zu verzichten, das für Sie heute Nutzen bringt, um dadurch in Zukunft mehr zu profitieren?

überhaupt nicht bereit es zu tun □—□—□—□—□—□—□—□—□—□—□ sehr bereit es zu tun

Wie sehr wären Sie bereit jemanden zu bestrafen, der Sie unfair behandelt, selbst wenn dies für Sie negative Konsequenzen haben würde?

überhaupt nicht bereit es zu tun □—□—□—□—□—□—□—□—□—□—□ sehr bereit es zu tun

Wie sehr wären Sie bereit jemanden zu bestrafen, der andere unfair behandelt, selbst wenn dies für Sie Kosten verursachen würde?

überhaupt nicht bereit es zu tun □—□—□—□—□—□—□—□—□—□—□ sehr bereit es zu tun

Wie sehr wären Sie bereit für einen guten Zweck zu geben, ohne etwas als Gegenleistung zu erwarten.

überhaupt nicht bereit es zu tun □—□—□—□—□—□—□—□—□—□—□ sehr bereit es zu tun

Wie gut beschreibt jede der nachfolgenden Aussagen Sie als Person?
Bitte verwenden Sie erneut eine Skala von 0 bis 10. Der Wert 0 bedeutet: Beschreibt mich überhaupt nicht. Der Wert 10 bedeutet: Beschreibt mich perfekt.

Wenn mir jemanden einen Gefallen tut, bin ich bereit ihn zu erwidern.

beschreibt mich überhaupt nicht □—□—□—□—□—□—□—□—□—□ beschreibt mich perfekt

Wenn ich sehr ungerecht behandelt werde, räche ich mich bei der ersten Gelegenheit, selbst wenn Kosten entstehen um das zu tun.

beschreibt mich überhaupt nicht □—□—□—□—□—□—□—□—□—□ beschreibt mich perfekt

Ich vermute, dass Leute nur die besten Absichten haben.

beschreibt mich überhaupt nicht □—□—□—□—□—□—□—□—□—□ beschreibt mich perfekt

Ich bin gut in Mathematik.

beschreibt mich überhaupt nicht □—□—□—□—□—□—□—□—□—□ beschreibt mich perfekt

Ich neige dazu Aufgaben zu verschieben, auch wenn ich weiß, dass es besser wäre sie gleich zu tun.

beschreibt mich überhaupt nicht □—□—□—□—□—□—□—□—□—□ beschreibt mich perfekt

Stellen Sie sich die folgende Situation vor:

Heute haben Sie unerwartet 1000 Euro erhalten. Wie viel von dem Geld würden Sie einem guten Zweck spenden? _____

Woher beziehen Sie Ihre Nachrichten?
Menschen nutzen unterschiedliche Quellen, um zu erfahren, was um sie herum und in der Welt passiert. Geben Sie bitte für jede der folgenden Quellen an, wie oft Sie diese nutzen:

	Täglich	Wöchentlich	Monatlich	Seltener als monatlich	Niemals
Zeitung	□	□	□	□	□
Fernsehsendungen	□	□	□	□	□
Radiosendungen	□	□	□	□	□
Nachrichtenseiten im Internet	□	□	□	□	□
Mobiltelefon (WhatsApp, Telegram, etc.)	□	□	□	□	□
Social media (Facebook, Twitter, etc.)	□	□	□	□	□
Gespräche mit Freunden, Kollegen und Bekannten	□	□	□	□	□

Ihre politischen Einstellungen

Wenn morgen Bundestagswahl wäre, welche Partei würden Sie dann wählen?

- ☐ CDU/CSU
- ☐ Bündnis '90/Die Grünen
- ☐ SPD
- ☐ AfD
- ☐ Die Linke
- ☐ FDP
- ☐ Sonstige

Würden Sie tatsächlich wählen gehen?

- ☐ Ja
- ☐ Nein
- ☐ Unentschlossen

Wie zufrieden sind Sie damit, wie das politische System in Deutschland heutzutage funktioniert?
Bewerten Sie bitte auf einer Skala von 0 bis 10, auf der 0 für „überhaupt nicht zufrieden" und 10 für „voll und ganz zufrieden" steht.

überhaupt nicht zufrieden ☐—☐—☐—☐—☐—☐—☐—☐—☐—☐—☐ voll und ganz zufrieden

Wie sehr stimmen Sie den folgenden Aussagen zu?

	stimme überhaupt nicht zu				stimme voll und ganz zu
	1	2	3	4	5
Es geschehen viele sehr wichtige Dinge in der Welt, über die die Öffentlichkeit nie informiert wird.	☐	☐	☐	☐	☐
Regierungsbehörden überwachen alle Bürger genau.	☐	☐	☐	☐	☐
Es gibt geheime Mächte, die die Welt steuern.	☐	☐	☐	☐	☐

Wie gut beschreibt jede der nachfolgenden Aussagen Sie als Person? *Bitte antworten Sie so ehrlich und spontan wie möglich, auf einer Skala von 1 bis 7. Der Wert 1 bedeutet: Trifft überhaupt nicht zu. Der Wert 7 bedeutet: Trifft voll und ganz zu.*

Ich bin jemand, der …

	Trifft überhaupt nicht zu						Trifft voll und ganz zu
	1	2	3	4	5	6	7
… gründlich arbeitet.	☐	☐	☐	☐	☐	☐	☐
… kommunikativ, gesprächig ist.	☐	☐	☐	☐	☐	☐	☐
… manchmal etwas grob zu anderen ist.	☐	☐	☐	☐	☐	☐	☐
… originell ist, neue Ideen einbringt.	☐	☐	☐	☐	☐	☐	☐
… sich oft Sorgen macht.	☐	☐	☐	☐	☐	☐	☐

Ich bin jemand, der …

	Trifft überhaupt nicht zu						Trifft voll und ganz zu
	1	2	3	4	5	6	7
… verzeihen kann.	☐	☐	☐	☐	☐	☐	☐
… eher faul ist.	☐	☐	☐	☐	☐	☐	☐
… aus sich herausgehen kann, gesellig ist.	☐	☐	☐	☐	☐	☐	☐
… künstlerische Erfahrungen schätzt.	☐	☐	☐	☐	☐	☐	☐
… leicht nervös wird.	☐	☐	☐	☐	☐	☐	☐

Ich bin jemand, der …

	Trifft überhaupt nicht zu						Trifft voll und ganz zu
	1	2	3	4	5	6	7
… Aufgaben wirksam und effizient erledigt.	☐	☐	☐	☐	☐	☐	☐
… zurückhaltend ist.	☐	☐	☐	☐	☐	☐	☐
… rücksichtsvoll und freundlich mit anderen umgeht.	☐	☐	☐	☐	☐	☐	☐
… eine lebhafte Phantasie, Vorstellungen hat.	☐	☐	☐	☐	☐	☐	☐
… entspannt ist, mit Stress gut umgehen kann.	☐	☐	☐	☐	☐	☐	☐

Wie sehr treffen die folgenden Aussagen auf Ihr eigenes Verhalten zu?
Bitte bewerten Sie erneut auf einer Skala von 1 bis 7. Der Wert 1 bedeutet: Trifft überhaupt nicht zu. Der Wert 7 bedeutet: Trifft voll und ganz zu.

Ich halte mindestens 1,5m Abstand zu Mitmenschen.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich werde mich sozial isolieren, wenn ich Kontakt hatte mit einer infizierten Person.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich halte mich stets auf dem Laufenden über Neuigkeiten zur Corona-Pandemie.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich wasche bzw. desinfiziere regelmäßig meine Hände.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich werde mich gegen das Coronavirus impfen lassen, wenn ein Impfstoff verfügbar ist.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich huste und niese in die Ellbogenbeuge.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich trage in der Öffentlichkeit einen Mund-Nasen-Schutz.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich lüfte regelmäßig durch, wenn mehrere Personen einen Raum benutzen.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich vermeide soziale Kontakte soweit es geht.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Ich werde Mitmenschen darüber informieren, wenn ich mich mit Corona infiziert habe.

Trifft überhaupt nicht zu □—□—□—□—□—□—□ Trifft voll und ganz zu

Die Corona-Warn-App ist eine Smartphone-App, die Nutzer informieren soll, ob sie in Kontakt mit einer infizierten Person geraten sind und daraus ein erhöhtes Ansteckungsrisiko anzunehmen ist.

Haben Sie die Corona-Warn-App auf Ihrem aktuellen Mobiltelefon installiert?

- ☐ Ja
- ☐ Nein

Für den Fall, dass Sie positiv auf Corona getestet werden würden: Wie wahrscheinlich ist es, dass Sie dies über die Corona-Warn-App melden?

Sehr unwahrscheinlich ☐—☐—☐—☐—☐—☐—☐ Sehr wahrscheinlich

Wie sehr stimmen Sie den folgenden Aussagen zur Corona-Warn-App zu? *Bitten antworten Sie auf einer Skala von 1 bis 7. Der Wert 1 bedeutet: Stimme überhaupt nicht zu. Der Wert 7 bedeutet: Stimme voll und ganz zu.*

Die Corona-Warn-App …

… hilft dabei, die Ausbreitung von Corona in Deutschland zu verlangsamen.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

… hilft dabei, die Ausbreitung von Corona in meiner Stadt zu verlangsamen.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

… hat für mich persönlich keinen großen Nutzen.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

… ist datenschutzrechtlich bedenklich.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

… ist ein gutes Mittel um Infektionsketten nachzuverfolgen.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

… wird noch nicht von genügend Menschen genutzt.

Stimme überhaupt nicht zu ☐—☐—☐—☐—☐—☐—☐ Stimme voll und ganz zu

Die sogenannte 7-Tage-Inzidenz gibt die Zahl der Corona-Neuinfektionen (d.h. positiv auf Corona getestete Personen) pro 100.000 Einwohnern innerhalb der vergangenen sieben Tage an. Sie gilt als wichtige Kennziffer zur Einschätzung der aktuellen Corona-Lage (Hinweis: Stand 17.12. liegt der Wert für Gesamtdeutschland bei 179).

Bitte schätzen Sie die 7-Tage Inzidenzrate in Ihrer Stadt. _____

Wie besorgt sind Sie über die Möglichkeit, dass …

	Überhaupt nicht besorgt						Sehr besorgt
	1	2	3	4	5	6	7
... Sie selbst an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐
... Freunde oder Verwandte an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐
... andere Menschen in Ihrer Umgebung an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐

Wie hoch schätzen Sie das Risiko ein, dass Sie sich innerhalb der nächsten 3 Monate mit COVID-19 anstecken?

Sehr unwahrscheinlich ☐—☐—☐—☐—☐—☐—☐ Sehr wahrscheinlich

Spendenmöglichkeit

Wichtig: Das folgende Szenario wird mit 25% Wahrscheinlichkeit tatsächlich umgesetzt. Sie sollten also sorgfältig überlegen, was Sie tun wollen. Es handelt sich womöglich um reale Geldbeträge.

Ihnen steht ein Geldbetrag in Höhe von 1 Euro zur Verfügung. Sie können frei entscheiden, welchen Anteil davon Sie spenden wollen, und welchen Anteil Sie für sich selbst behalten. Ihre Spende wird für eine Online-Werbekampagne auf sozialen Medien eingesetzt, die mehr Menschen (u.a. in Ihrer Region) zur Nutzung der Corona-Warn-App ermutigt. In der Vergangenheit hat sich gezeigt, dass 1 Corona-Warn-App-Installation durchschnittlich knapp 50 Cent Werbeausgaben entspricht. Den Teil des Geldbetrags, den Sie nicht spenden, erhalten Sie als zusätzliche Entlohnung in Form von Panelpunkten.

Am Ende der Umfrage lost ein Zufallsgenerator aus, ob die Spende und die zusätzliche Entlohnung tatsächlich ausgezahlt werden. Bitte bewegen Sie den Schieberegler, um über Ihr Budget zu entscheiden:

Welchen Betrag möchten Sie spenden? Ihre Spende beträgt _____ Euro.

Bitte schauen Sie sich das folgende Video an.

In Kiel wurden zuletzt 117,1 Corona-Neuinfektionen pro 100.000 Einwohnern in 7 Tagen gemeldet, das ist 27% höher als der landesweite Durchschnitt. (Quelle: Robert-Koch-Institut, Stand 17.12.)

Sie nähern sich nun dem Ende des Fragebogens. Einige der folgenden Fragen werden Ihnen bekannt vorkommen. Wundern Sie sich nicht, das ist ein ganz normaler Teil der Umfrage. Bitte beantworten Sie diese Fragen genauso sorgfältig und gewissenhaft wie die vorherigen.

Für den Fall, dass Sie positiv auf Corona getestet werden würden: Wie wahrscheinlich ist es, dass Sie dies über die Corona-Warn-App melden?

Sehr unwahrscheinlich □—□—□—□—□—□—□ Sehr wahrscheinlich

Wie sehr stimmen Sie den folgenden Aussagen zur Corona-Warn-App zu? *Bitten antworten Sie auf einer Skala von 1 bis 7. Der Wert 1 bedeutet: Stimme überhaupt nicht zu. Der Wert 7 bedeutet: Stimme voll und ganz zu.*

Die Corona-Warn-App …

… hilft dabei, die Ausbreitung von Corona in Deutschland zu verlangsamen.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

… hilft dabei, die Ausbreitung von Corona in meiner Stadt zu verlangsamen.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

… hat für mich persönlich keinen großen Nutzen.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

… ist datenschutzrechtlich bedenklich.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

… ist ein gutes Mittel um Infektionsketten nachzuverfolgen.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

… wird noch nicht von genügend Menschen genutzt.

Stimme überhaupt nicht zu □—□—□—□—□—□—□ Stimme voll und ganz zu

Die sogenannte 7-Tage-Inzidenz gibt die Zahl der Corona-Neuinfektionen (d.h. positiv auf Corona getestete Personen) pro 100.000 Einwohnern innerhalb der vergangenen sieben Tage an (Hinweis: Stand 17.12. liegt der Wert für Gesamtdeutschland bei 179).

Bitte schätzen Sie die 7-Tage Inzidenzrate in Ihrer Stadt. _____

Wie besorgt sind Sie über die Möglichkeit, dass …

	Überhaupt nicht besorgt						Sehr besorgt
	1	2	3	4	5	6	7
... Sie selbst an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐
... Freunde oder Verwandte an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐
... andere Menschen in Ihrer Umgebung an COVID-19 erkranken.	☐	☐	☐	☐	☐	☐	☐

Wie hoch schätzen Sie das Risiko ein, dass Sie sich innerhalb der nächsten 3 Monate mit COVID-19 anstecken?

Sehr unwahrscheinlich ☐—☐—☐—☐—☐—☐—☐ Sehr wahrscheinlich

Erneute Spendenmöglichkeit

Sie stehen erneut der gleichen Spendenentscheidung gegenüber wie zuvor. Das Szenario auf dieser Seite wird wieder mit 25% Wahrscheinlichkeit tatsächlich umgesetzt. Maximal eine Ihrer beiden Spendenentscheidungen wird zufällig ausgewählt, nie jedoch beide gleichzeitig.

Zur Erinnerung: Ihnen steht ein Geldbetrag in Höhe von 1 Euro zur Verfügung. Sie können frei entscheiden, welchen Anteil davon Sie spenden wollen für eine Online-Werbekampagne zur Corona-Warn-App. In der Vergangenheit hat sich gezeigt, dass 1 Corona- Warn-App-Installation durchschnittlich knapp 50 Cent Werbeausgaben entspricht. Den Teil des Geldbetrags, den Sie nicht spenden, erhalten Sie als zusätzliche Entlohnung in Form von Panelpunkten.

Am Ende der Umfrage lost ein Zufallsgenerator aus, ob die Spende und die zusätzliche Entlohnung tatsächlich ausgezahlt werden. Bitte bewegen Sie den Schieberegler, um über Ihr Budget zu entscheiden:

Welchen Betrag möchten Sie spenden? Ihre Spende beträgt _____ Euro.

Finale Angaben zu Ihrer Person

Fast geschafft! Zum Abschluss der Umfrage möchten wir Sie gerne noch um einige letzten Angaben zu Ihrer Person bitten.

Welches ist Ihr höchster Bildungsabschluss?

- ☐ Schule ohne Abschluss verlassen
- ☐ Volks-/Hauptschulabschluss (DDR: 8. Klasse)
- ☐ Realschulabschluss/Mittlere Reife (DDR: 10. Klasse)
- ☐ Fachhochschulreife (Abschluss einer Fachoberschule)
- ☐ Abitur/Hochschulreife
- ☐ Fachhochschule (früher: Ingenieurschule, Lehrerbildung, DDR: Ingenieur und Fachschulabschluss)
- ☐ Universitäts-, Hochschulabschluss
- ☐ Promotion
- ☐ Sonstiger Bildungsabschluss

Wie viele Personen leben in Ihrem Haushalt (d.h. zusammen wohnende und wirtschaftende Einheit)? _____

Wie viele Kinder haben Sie? _____

Was beschreibt Ihren aktuellen Erwerbsstatus am besten?

- ☐ Vollzeit angestellt
- ☐ Teilzeit angestellt
- ☐ Selbstständig
- ☐ im Studium/in Ausbildung
- ☐ Nicht erwerbstätig, nicht in Ausbildung

Welches ist Ihr höchster Bildungsabschluss?

- ☐ unter 900 €
- ☐ 900 € bis unter 1.300 €
- ☐ 1.300 € bis unter 1.500 €
- ☐ 1.500 € bis unter 2.000 €
- ☐ 2.000 € bis unter 2.600 €
- ☐ 2.600 € bis unter 3.200 €
- ☐ 3.200 € bis unter 4.500 €
- ☐ 4.500 € bis unter 6.000 €
- ☐ 6.000 € und mehr
- ☐ keine Angabe

Vielen Dank für Ihre Teilnahme an dieser Umfrage!

Sie haben nun das Ende des Fragebogens erreicht. Sie konnten im Laufe der Umfrage zwei Mal über einen Geldbetrag von jeweils 1 Euro entscheiden. Von einem Zufallsgenerator wurde ausgelost, ob eines dieser Szenarien tatsächlich umgesetzt wird. Folgende Auszahlung wurde für Sie bestimmt:

Sie spenden _____ Euro an eine Online-Werbekampagne für die Corona-Warn-App.

Sie erhalten _____ Euro als zusätzliche Entlohnung in Form von Panelpunkten. Bitte beachten Sie, dass es 4 bis 6 Wochen dauern kann, bis diese Ihrem Konto gutgeschrieben werden..

Diese Umfrage wurde durchgeführt von Forschern der Universität Bonn. Ziel der Studie ist es, mehr über die Einstellungen zur Corona-Pandemie in Deutschland zu erfahren. Dabei ging es unter anderem auch um die Bereitschaft zur Nutzung der Corona- Warn-App. Für weitere Informationen zur App haben wir für Sie im Folgenden einige Antworten auf häufig gestellte Fragen (FAQs) zusammengestellt.

Sobald Sie fertig sind, klicken Sie bitte auf Umfrage abschließen.

References

Abelson, Robert P, Elliot Aronson, William J McGuire, Theodore M Newcomb, Milton J Rosenberg, and Percy H Tannenbaum. 1968. *Theories of Cognitive Consistency: A Sourcebook.* Edited by Robert P Abelson, Elliot Aronson, William J McGuire, Theodore M Newcomb, Milton J Rosenberg, and Percy H Tannenbaum. Chicago: Rand-McNally. [72]

Albrecht, Felix, Sebastian Kube, and Christian Traxler. 2018. "Cooperation and Norm Enforcement - The Individual-Level Perspective." *Journal of Public Economics* 165: 1–16. https://doi.org/10.1016/j.jpubeco.2018.06.010. [70, 72]

Alesina, Alberto, and Nicola Fuchs-Schündeln. 2007. "Good-Bye Lenin (or Not?): The Effect of Communism on People's Preferences." *American Economic Review* 97 (4): 1507–28. [83]

Alfaro, Laura, Ester Faia, Nora Lamersdorf, and Farzad Saidi. 2021a. "Health Externalities and Policy: The Role of Social Preferences." *ECONtribute Discussion Paper No. 109.* [70, 73]

Alfaro, Laura, Ester Faia, Nora Lamersdorf, and Farzad Saidi. 2021b. "Social Interactions in Pandemic." *ECONtribute Discussion Paper No. 110,* https://doi.org/10.3386/w27134. [70, 71, 73, 84]

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2021. "Fighting Climate Change: The Role of Norms, Preferences, and Moral Values." *IZA Discussion Paper 14518.* [70]

Baron, Reuben M, and David A Kenny. 1986. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51 (6): 1173–82. https://doi.org/https://psycnet.apa.org/doi/10.1037/0022-3514.51.6.1173. [81]

Barrios, John M., Efraim Benmelech, Yael V. Hochberg, Paola Sapienza, and Luigi Zingales. 2021. "Civic Capital and Social Distancing during the Covid-19 Pandemic." *Journal of Public Economics* 193: 104310. https://doi.org/10.1016/j.jpubeco.2020.104310. [70]

Barsbai, Toman, Dieter Lukas, and Andreas Pondorfer. 2021. "Local Convergence of Behavior across Species." *Science* 371 (6526): 292–95. https://doi.org/10.5281/ZENODO.4159697. [86]

Bartscher, Alina, Sebastian Seitz, Michaele Slotwinski, Nils Wehrhöfer, and Sebastian Siegloch. 2020. "Social Capital and the Spread of Covid-19: Insights from European Countries." *ZEW Discussion Paper 20-023.* [70]

Batson, Daniel C., and Adam A Powell. 2003. "Altruism and Prosocial Behavior." In *Handbook of Psychology: Personality and Social Psychology,* edited by Theodore Millon and Melvin J. Lerner, 463–84. Wiley. https://doi.org/10.1002/0471264385.wei0519. [72]

Battaglia, Michael P, David C Hoaglin, and Martin R Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice* 2 (5): 2953. [74, 120]

Bauch, Chris T, and David JD Earn. 2004. "Vaccination and the Theory of Games." *Proceedings of the National Academy of Sciences* 101 (36): 13391–94. [71]

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts. 2016. "Can War Foster Cooperation?" *Journal of Economic Perspectives* 30 (3): 249–74. [84]

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. 2012. "The Relationship between Economic Preferences and Psychological Personality Measures." *Annual Review of Economics* 4 (1): 453–78. https://doi.org/10.1146/annurev-economics-080511-110922. [77, 119]

Becker, Sascha O, Lukas Mergele, and Ludger Woessmann. 2020. "The Separation and Reunification of Germany: Rethinking a Natural Experiment Interpretation of the Enduring Effects of Communism." *Journal of Economic Perspectives* 34 (2): 143–71. [83]

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution.* Princeton University Press. https://doi.org/doi:10.1515/9781400838837. [72]

Branas-Garza, Pablo, Diego Andrés Jorrat, Antonio Alfonso, Antonio M. Espin, Teresa García, and Jaromir Kovarik. 2020. "Exposure to the Covid-19 Pandemic and Generosity." *MPRA Paper No. 103389,* https://doi.org/10.31234/osf.io/6ktuz. [84]

Brosig-Koch, Jeannette, Christoph Helbach, Axel Ockenfels, and Joachim Weimann. 2011. "Still Different after All These Years: Solidarity Behavior in East and West Germany." *Journal of Public Economics* 95 (11-12): 1373–76. [83]

Burks, Stephen, Jeffrey Carpenter, and Lorenz Goette. 2009. "Performance Pay and Worker Cooperation: Evidence from an Artefactual Field Experiment." *Journal of Economic Behavior & Organization* 70 (3): 458–69. [70]

Burks, Stephen, Jeffrey Carpenter, Lorenz Götte, and Aldo Rustichini. 2012. "Which Measures of Time Preference Best Predict Outcomes: Evidence from a Large-Scale Field Experiment." *Journal of Economic Behavior & Organization* 84 (1): 308–20. [73]

Caicedo, Felipe Valencia, Thomas Dohmen, and Andreas Pondorfer. 2021. "Religion and Prosociality across the Globe." *Working Paper.* [86]

Cameron, A. Colin, Jonah Gelbach, and Douglas Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–27. https://doi.org/10.3386/t0344. [79]

Campos-Mercade, Pol, Armando N. Meier, Florian H. Schneider, and Erik Wengström. 2021. "Prosociality Predicts Health Behaviors during the COVID-19 Pandemic." *Journal of Public Economics* 195: 104367. https://doi.org/10.1016/j.jpubeco.2021.104367. [70, 84, 85]

Cappelen, Alexander W., Ranveig Falch, Erik Ø. Sørensen, and Bertil Tungodden. 2021. "Solidarity and Fairness in Times of Crisis." *Journal of Economic Behavior & Organization* 186: 1–11. https://doi.org/10.1016/j.jebo.2021.03.017. [84]

Carlsson, Fredrik, Olof Johansson-Stenman, and Pham Khanh Nam. 2014. "Social Preferences Are Stable over Long Periods of Time." *Journal of Public Economics* 117: 104–14. https://doi.org/10.1016/j.jpubeco.2014.05.009. [85]

Casoria, Fortuna, Fabio Galeotti, and Marie Claire Villeval. 2021. "Perceived Social Norm and Behavior Quickly Adjusted to Legal Changes during the COVID-19 Pandemic." *Journal of Economic Behavior & Organization* 190: 54–65. [84]

Chan, Ho Fai, Ahmed Skali, David A. Savage, David Stadelmann, and Benno Torgler. 2020. "Risk Attitudes and Human Mobility during the COVID-19 Pandemic." *Scientific Reports* 10 (1): 19931. https://doi.org/10.1038/s41598-020-76763-2. [73]

Cohn, Alain, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd. 2019. "Civic Honesty around the Globe." *Science* 365 (6448): 70–73. https://doi.org/10.1126/science.aau8712. [86]

de Oliveira, Angela C.M., Tammy C.M. Leonard, Kerem Shuval, Celette Sugg Skinner, Catherine Eckel, and James C. Murdoch. 2016. "Economic Preferences and Obesity among a Low-Income African American Community." *Journal of Economic Behavior & Organization* 131: 196–208. https://doi.org/10.1016/j.jebo.2015.11.002. [73]

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2008. "Homo Reciprocans: Survey Evidence on Behavioural Outcomes." *Economic Journal* 119 (536): 592–612. https://doi.org/10.26481/umaror.2008007. [70]

Durante, Ruben, Luigi Guiso, and Giorgio Gulino. 2021. "Asocial Capital: Civic Culture and Social Distancing during COVID-19." *Journal of Public Economics* 194: 104342. https://doi.org/10.2139/ssrn.3611606. [70]

Epper, Thomas, Ernst Fehr, and Julien Senn. 2020. "Other-Regarding Preferences and Redistributive Politics." *University of Zurich Working Paper No. 339,* https://doi.org/10.2139/ssrn.3526809. [70]

Exadaktylos, Filippos, Antonio M Espín, and Pablo Branas-Garza. 2013. "Experimental Subjects Are Not Different." *Scientific Reports* 3 (1): 1–6. [74]

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4): 1645–92. https://doi.org/10.1093/qje/qjy013. [70, 74, 76, 86, 107]

Falk, Armin, Anke Becker, Thomas J. Dohmen, David Huffman, and Uwe Sunde. 2016. "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences." *IZA Discussion Paper No. 9674,* https://doi.org/10.2139/ssrn.2725874. [71, 74, 75, 107]

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples." *Journal of the European Economic Association* 11 (4): 839–52. https://doi.org/10.1111/jeea.12019. eprint: https://academic.oup.com/jeea/article-pdf/11/4/839/10461629/jeea0839.pdf. [74]

Farboodi, Maryam, Gregor Jarosch, and Robert Shimer. 2021. "Internal and External Effects of Social Distancing in a Pandemic." *Journal of Economic Theory* 196: 105293. [70, 71, 73]

Fehr, Ernst, and Urs Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425: 785–91. https://doi.org/10.1046/j.0013-0427.2003.00027.x. [70]

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415 (6868): 137–40. https://doi.org/10.1038/415137a. [70, 72]

Fehr, Ernst, and Ivo Schurtenberger. 2018. "Normative Foundations of Human Cooperation." *Nature Human Behaviour* 2 (7): 458–68. https://doi.org/10.1038/s41562-018-0385-5. [70]

Fenichel, Eli P, Carlos Castillo-Chavez, M Graziano Ceddia, Gerardo Chowell, Paula A Gonzalez Parra, Graham J Hickling, Garth Holloway, Richard Horan, Benjamin Morin, Charles Perrings, et al. 2011. "Adaptive Human Behavior in Epidemiological Models." *Proceedings of the National Academy of Sciences* 108 (15): 6306–11. [71]

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance.* Stanford University Press. [72]

Fischbacher, Urs, and Simon Gächter. 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." *American Economic Review* 100 (1): 541–56. [70]

Franzen, Axel, and Katrin Botzen. 2011. "Vereine in Deutschland Und Ihr Beitrag Zum Wohlstand Der Regionen." *Soziale Welt* 62: 391–413. [96, 121]

Frondel, Manuel, Daniel Osberghaus, and Stephan Sommer. 2021. "Corona and the Stability of Personal Traits and Preferences: Evidence from Germany." *ZEW Discussion Paper 21-029,* https://doi.org/10.2139/ssrn.3820484. [84]

Fuhrmann-Riebel, Hanna, Ben D'Exelle, and Arjan Verschoor. 2021. "The Role of Preferences for Pro-Environmental Behaviour among Urban Middle Class Households in Peru." *Ecological Economics* 180: 106850. https://doi.org/10.1016/j.ecolecon.2020.106850. [70]

Galbiati, Roberto, Emeric Henry, Nicolas Jacquemet, and Max Lobeck. 2021. "How Laws Affect the Perception of Norms: Empirical Evidence from the Lockdown." *PloS ONE* 16 (9): e0256624. [84]

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. "Zur Erhebung Der Big-Five-Basierten Persoenlichkeitsmerkmale Im SOEP." *DIW Research Notes* 4: 2005. [74, 118]

Gollwitzer, Anton, Killian Mcloughlin, Martel, Cameron: Marshall, Julia, Johanna M. Höhs, and John A. Bargh. 2021. "Linking Self-Reported Social Distancing to Real-World Behavior during the COVID-19 Pandemic." *Preprint.* [71]

Harper, Craig A., Liam P. Satchell, Dean Fido, and Robert D. Latzman. 2020. "Functional Fear Predicts Public Health Compliance in the COVID-19 Pandemic." *International Journal of Mental Health and Addiction,* 1–14. https://doi.org/10.1007/s11469-020-00281-5. [84]

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations.* Wiley. [72]

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, et al. 2006. "Costly Punishment across Human Societies." *Science* 312 (5781): 1767–70. https://doi.org/10.1126/science.1127333. [85]

Jagelka, Tomas. 2020. "Are Economists' Preferences Psychologists' Personality Traits? A Structural Approach." *Working Paper.* [77]

Jensen, Ulrich Thy. 2020. "Is Self-Reported Social Distancing Susceptible to Social Desirability Bias? Using the Crosswise Model to Elicit Sensitive Behaviors." *Journal of Behavioral Public Administration* 3 (2): 1–11. https://doi.org/10.30636/jbpa.32.182. [71]

Jones, Callum, Thomas Philippon, and Venky Venkateswaran. 2021. "Optimal Mitigation Policies in a Pandemic: Social Distancing and Working from Home." *Review of Financial Studies* 34 (11): 5188–223. [71]

Keeling, Matt J, and Pejman Rohani. 2011. *Modeling Infectious Diseases in Humans and Animals.* Princeton university press. [71]

Kermack, William Ogilvy, and Anderson G McKendrick. 1927. "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 115 (772): 700–721. [71]

Khwaja, Ahmed, Frank Sloan, and Martin Salm. 2006. "Evidence on Preferences and Subjective Beliefs of Risk Takers: The Case of Smokers." *International Journal of Industrial Organization* 24 (4): 667–82. https://doi.org/10.1016/j.ijindorg.2005.10.001. [73]

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk. 2020. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment." *Journal of Political Economy* 128 (2): 434–67. https://doi.org/10.1086/704386. [86]

Kosse, Fabian, and Michela M. Tincani. 2020. "Prosociality Predicts Labor Market Success around the World." *Nature Communications* 11 (1): 5298. https://doi.org/10.1038/s41467-020-19007-1. [70, 74]

Lades, Leonhard K., Kate Laffan, and Till O. Weber. 2021. "Do Economic Preferences Predict Pro-Environmental Behaviour?" *Ecological Economics* 183: 106977. https://doi.org/10.1016/j.ecolecon.2021.106977. [70]

Makridis, Christos A., and Cary Wu. 2021. "How Social Capital Helps Communities Weather the COVID-19 Pandemic." *PloS ONE* 16 (1): e0245135. https://doi.org/10.1371/journal.pone.0245135. [70]

Müller, Stephan, and Holger A. Rau. 2021. "Economic Preferences and Compliance in the Social Stress Test of the COVID-19 Crisis." *Journal of Public Economics* 194: 104322. https://doi.org/10.1016/j.jpubeco.2020.104322. [70, 84]

Nettle, Daniel, Agathe Colléony, and Maria Cockerill. 2011. "Variation in Cooperative Behaviour within a Single City." *PloS ONE* 6 (10): e26922. https://doi.org/10.1371/journal.pone.0026922.t001. [86]

Nikolov, Plamen, Andreas Pape, Ozlem Tonguc, and Charlotte Williams. 2020. "Predictors of Social Distancing and Mask-Wearing Behavior: Panel Survey in Seven U.S. States." *IZA Discussion Paper 13745.* [76]

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997." *American Political Science Review* 92 (1): 1–22. [69]

Petherick, Anna, Rafael Goldszmidt, Eduardo B. Andrade, Rodrigo Furst, Thomas Hale, Annalena Pott, and Andrew Wood. 2021. "A Worldwide Assessment of Changes in Adherence to COVID-19 Protective Behaviours and Hypothesized Pandemic Fatigue." *Nature Human Behaviour,* https://doi.org/10.1038/s41562-021-01181-x. [84]

Quaas, Martin F., Japer N. Meya, Hanna Schenk, Björn Bos, Moritz A. Drupp, and Till Requate. 2021. "The Social Cost of Contacts: Theory and Evidence for the First Wave of the COVID-19 Pandemic in Germany." *PLoS ONE* 16 (3): e0248288. [70, 71, 73]

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using Boottest." *Stata Journal* 19 (1): 4–60. https://doi.org/10.1177/1536867X19830877. [79]

Shachat, Jason, Matthew J Walker, and Lijia Wei. 2021. "How the Onset of the Covid-19 Pandemic Impacted pro-Social Behaviour and Individual Preferences: Experimental Evidence from China." *Journal of Economic Behavior & Organization* 190: 480–94. [85]

Sutter, Matthias, Martin G. Kocher, Daniela Glätzle-Rützler, and Stefan T. Trautmann. 2013. "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior." *American Economic Review* 103 (1): 510–31. https://doi.org/10.1257/aer.103.1.510. [73]

Torgler, Benno. 2002. "Does Culture Matter? Tax Morale in an East-West-German Comparison." *FinanzArchiv / Public Finance Analysis* 59 (4): 504–28. [83]

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–32. [85]

van Hulsen, Merel, Kirsten I. M. Rohde, and Job van Exel. 2020. "Inter-Temporal and Social Preferences Predict Compliance in a Social Dilemma: An Application in the Context of COVID-19." *Working Paper,* https://doi.org/10.2139/ssrn.3665978. [70]

Volk, Stefan, Christian Thöni, and Winfried Ruigrok. 2012. "Temporal Stability and Psychological Foundations of Cooperation Preferences." *Journal of Economic Behavior & Organization* 81 (2): 664–76. [85]

Zettler, Ingo, Christoph Schild, Lau Lilleholt, Lara Kroencke, Till Utesch, Morten Moshagen, Robert Böhm, Mitja D Back, and Katharina Geukes. 2022. "The Role of Personality in COVID-19-related Perceptions, Evaluations, and Behaviors: Findings across Five Samples, Nine Traits, and 17 Criteria." *Social Psychological and Personality Science* 13 (1): 299–310. [76]

Chapter 3

The effect of workweek reforms on labor supply preferences: Evidence from the German public sector

Joint with Thomas Dohmen

3.1 Introduction

Labor supply is a main determinant of an economy's output. However, many developed economies currently face a scarcity of this resource. In standard labor supply models, wages play a prominent role in explaining individuals' labor supply choices. This has also been the main focus of research in labor supply. Though less emphasis is placed on non-monetary incentives or an individual's underlying taste for work, these factors may also impact the number of hours an individual is willing to supply at a given wage— different non-monetary incentives and tastes imply different choices at the same wage.

In this paper, we study how changes in institutional rules influence individuals' preferred working hours. More specifically, we study the effect of changes to the length of the standard workweek on employees' work hour choices in the context of the German public sector. The German public sector is divided into two categories: public sector employees (*Angestellte im öffentlichen Dienst*) and civil servants (*Beamte*). Work conditions are governed by collective agreements and are periodically re-negotiated between employer and employee unions.[1] The standard workweek refers to the contractual weekly number of hours in these collective agreements. Between 1989 and 1991, the standard workweek was shortened from 40

1. Employers are represented by employer unions at the federal, state, or municipal level, while public sector employees are represented by the United Services Union (*Vereinigte Dienstleistungsgewerkschaft*). Civil servants are not allowed to unionize and are instead represented by an interest group (*Deutscher Beamtenbund*).

to 38.5 hours for civil servants and public sector employees. For most civil servants and all public sector employees, this workweek reduction was implemented in two stages— from 40 hours to 39 hours in 1989 and from 39 hours to 38.5 hours in 1990. This decrease in work hours was not accompanied by a decrease in nominal monthly income and essentially translated to an increase in hourly wages. However, between 1994 and 2004, the standard workweeks for state- and municipal-level civil servants were eventually increased by 1.5 to 3.5 hours, at different points in time across states. As with the workweek decreases, increases to the standard workweek were not compensated by increases in income.

To see how these changes affect individuals' workweek preferences, we use data from the German Socio-Economic Panel (SOEP), waves 1985 to 2017 and study responses to the question: "If you could choose your own number of working hours, taking into account that your income would change according to the number of hours: How many hours would you want to work?". We bin these responses into 1-hour categories and find that following changes to the standard workweek, individuals are more likely to prefer the new standard workweek, and less likely to prefer the old standard workweek. Comparing the binned responses of all civil servant and public sector employee observations in the five years before and the four years after a decrease in the standard workweek from 40 to 38.5 hours between 1990 and 1991, we find that the fraction of individuals preferring a 38-hour workweek category increases by 13 pp ($p<0.01$) post-reform, from a pre-reform average of 5.7%. At the same time, the fraction of individuals preferring a 40-hour workweek decreases by 21 pp ($p<0.01$), from a pre-reform average of 41%. Our event-study model indicates that these changes are immediate. We also find evidence of these changes at the individual level. Amongst individuals preferring a 40-hour workweek category in a given year, the probability of switching to a preferred workweek category of 39 hours increases by 11 pp in the year prior to the implementation of the 39-hour workweek, while the probability of switching to a preferred workweek category of 38 hours two years later increases by 17 pp in the year prior to the introduction of the 38.5-hour workweek.

In contrast, the change in the composition of preferred workweeks following increases to the standard workweek is milder and more gradual. The increase of the standard workweek amongst civil servants is initially only followed by an 8.5 pp increase in the fraction of individuals preferring the new standard workweek category, from a pre-reform average of 13.3%. This fraction gradually increases in magnitude between the fifth and seventh year post-reform, to 12.2 pp ($p<0.01$). Similarly, the proportion of individuals preferring a 38-hour workweek category decreases by an average 6.1 pp in the first four years post-reform, and by 8.2 pp ($p<0.01$) in the fifth to seventh year following the introduction of the longer standard workweek, relative to a pre-reform average of 23.4%. Furthermore, in contrast to workweek reductions, we do not find an increase in individuals switching directly from preferring the old standard workweek to the new standard workweek. The effect we

observe instead stems from individuals switching from preferring other workweeks to the new standard workweek.

These changes in preferred workweeks could potentially be explained by several mechanisms. Firstly, because changes in the standard workweek are not accompanied by offsetting changes in income, it could be that individuals choose the optimal number of work hours in response to these hourly wages changes. However, several points suggest that wage changes may not be the sole reason underpinning the effects we find. Although substantial income increases were negotiated for a minority of civil servants close to the workweek reductions, the effect of the workweek reforms do not significantly differ for this group of individuals. Furthermore, although decreases in hourly wages, induced by increases to the standard workweek, are partly reversed in later years, we do not see a reversal of the effects described above.

Indeed, there are many alternative explanations for the effects we find. Because the standard workweek also influences the actual number of hours individuals are required to work, changes in preferences could occur through habit formation, where longer (shorter) working hours in the past lead to stronger preferences for longer (shorter) workweeks in subsequent periods. Alternatively, given the salience and wide coverage of the standard workweek, preferences could change in response to changes in others' preferences or choices, or in response to altered social norms. These considerations can be incorporated into a model of reference-dependent preferences.

There is a literature showing that reference-dependent preferences can explain heterogeneity in labor supply. Camerer et al. (1997) find that taxi drivers tend to work fewer hours on days where the hourly wage is high, which is interpreted as evidence of daily income targeting; although this was subsequently refuted by Farber (2005) and Farber (2015).[2,3] Meanwhile, Fehr and Goette (2007) conduct a field experiment that increased the compensation of bicycles messengers for a month and find lower effort provision per shift, but higher overall labor supply, in the month with increased compensation. More importantly, they are able to link the negative effort elasticity to loss averse individuals, which is consistent with Kahneman

2. Standard life-cycle models of labor supply predict that a temporary increase in wages should be met with increased willingness to work more in that period. On the other hand, if an individual has reference-dependent preferences in the income domain, higher wages helps them meet their target more quickly, so that they opt to stop working after fewer hours.

3. Farber (2005) questions Camerer et al. (1997)'s assumption that the hourly wage is constant throughout the day, and the exogeneity of other drivers' wages as an instrument for a driver's own hourly wage. In light of these econometric issues, Farber (2005) estimates a probability stopping model and find that while cumulative hours predict the probability of stopping, cumulative income does not. Using newer, more detailed data on taxi drivers from 2009 to 2013, Farber (2015) attempts to replicate Camerer et al. (1997)'s analysis, but fail to find the large negative elasticities of Camerer et al. (1997).

and Tversky (1979)'s prospect theory. Following Kőszegi and Rabin (2006)'s seminal work on rational expectations as reference points, Crawford and Meng (2011) define expected income and hours as reference points and estimate a reference-dependent labor supply model using the same data as Farber (2005). In line with Kőszegi and Rabin (2006)'s idea that anticipated wage increases are incorporated into income targets, several papers find evidence that the reference point can change. Farber (2015) finds more negative elasticities when controlling for anticipated wage changes, which suggests that only unanticipated wage changes leads to income targeting behavior. Further evidence consistent with reference point adaptation is provided by Thakral and Tô (2021) — using data of trips that end 10 minutes around the median number of hours worked, they find that more recent increases in earnings have a higher probability of inducing shift-quitting, which is consistent with indivduals incorporating older earnings levels into the reference point as the day proceeds.

Meanwhile, Behaghel and Blau (2012) and Seibold (2021) suggest that external quantities, such as statutory retirement ages, influence individuals' reference points, so that changes in these retirement ages alter the ages individuals choose to retire at. Using administrative data from Germany, Seibold (2021) finds bunching of individuals retiring at statutory retirement ages that is inconsistent changes in financial incentives alone. He also finds that a change in the statutory retirement age from 60 to 65 in one-month increments by cohort is accompanied by a bunching of individuals retiring at the cohort-specific statutory retirement age. Similarly, Behaghel and Blau (2012) study a change in the official retirement age from 65 to 66 in two-month increments in the United States and find that the benefit claiming hazard rate at retirement age moves in lockstep with cohort. As with Seibold (2021) and Behaghel and Blau (2012), we provide evidence supporting this phenomenon by documenting increased masses in the distribution of preferred hours at various standard workweeks. We add to the findings of these two papers by documenting within-individual shifts in preferred hours from the old to the new standard workweek in response to changes in standard workweeks.

We next contribute to the literature on the effects of workweek changes on various outcomes. Several papers study the effects of workweek changes on employee satisfaction and health outcomes. Lepinteur (2019) documents higher job and leisure satisfaction amongst workers following workweek reductions in Portugal and France. Similarly, Hamermesh, Kawaguchi, and Lee (2017) finds increased life satisfaction amongst individuals likely facing shorter workweeks due to the implementation of an overtime penalty. Ahn (2016) reports lower incidences of smoking and a higher probability of regular exercise amongst employees in response to a decrease in the standard workweek from 44 to 40 hours in Korea, while Berniell and Bietenbeck (2020) also finds that individuals are less likely to smoke following a workweek reduction in France. Cygan-Rehm and Wunder (2018) study the same civil servant workweek reforms as we do and find that longer work hours lead to

a decline in self-assessed health and increased frequency of medical appointments. We extend the scope of this literature to another employee outcome, and thus contribute to a more comprehensive view of the consequences of workweek changes on employees. Secondly, with the exception of Cygan-Rehm and Wunder (2018), the studies mentioned above either study workweek extensions or reductions. We study workweek reforms in both directions within the same context and document asymmetric effects.

We proceed as follows: in Section 3.2, we briefly describe the German public sector and the workweek changes. In section 3.3, we describe our dataset and the merging of standard workweek information to the dataset. Next, we outline our empirical strategy in section 3.4, before presenting our main results in section 3.5. Lastly, we discuss potential mechanisms in Section 3.6, before concluding.

3.2 Context

Work conditions are regulated by law for civil servants, and by collective agreements for public sector employees, and potentially differ by administrative level (federal, state, and municipal level). In contrast to the private sector, where more individualized work agreements are commonplace, work conditions in the public sector typically apply to all individuals within an employment group and administrative level. The standard workweek thus refers to the number of contractually required work hours per week that is common to all indivduals under a particular collective agreement or law. Tables 3.A.1 and 3.A.2 outline changes in standard workweeks in the time period we study (1985 to 2017).

Reductions in the standard workweek, from 40 to 38.5 hours, were implemented for both public sector employees and civil servants in 1989 and 1990, with no reduction in income. This reduction was likely part of an overall trend towards shorter working hours during this period— labor unions in the metal and printing industries had been pushing for a 35-hour workweek in the 1980s as part of an effort to counteract high unemployment rates. Thus, by 1989, the 38.5-hour workweek had already been implemented in various industries in the private sector, such as the metal, electrical and car-manufacturing industries.

However, this workweek reduction was soon reversed for state- and municipal-level civil servants in Schleswig-Holstein, Bavaria, Baden-Württemberg, Bremen, Lower Saxony, and Rhineland-Palatinate in the 1990s, and for civil servants of these administrative levels in Hamburg and Saarland in the early 2000s. In North Rhine-Westphalia and Hessia, the standard workweek for state and municipal civil servants was extended beyond 40 hours in 2004, to 41 and 42 hours respectively. Between 2002 and 2006, further extensions were implemented in Schleswig-Holstein, Bavaria, and Baden-Württemberg, so that the standard workweek for state and municipal civil servants in these states were also extended beyond 40 hours. As for

civil servants employed at the federal level, the 40-hour workweek was reinstated in 2004 and further increased to 41 hours in 2006. Thus, by the mid-2000s, civil servants at all administrative levels and in all states had a standard workweek of 40 hours or more. Lastly, we note a reversal of the trend towards longer standard workweeks in Bavaria, where the standard workweek was reduced from 42 to 40 hours in two stages in 2012 and 2013.

As for public sector employees, the standard workweek was also increased in the mid-2000s in conjunction with the implementation of new collective agreements. However, unlike for civil servants, where the workweek was extended at least to 40 hours, work hour increases were relatively mild for a large proportion of public sector employees, likely due to the right of public sector employees to strike. Demands of employers to return to the 40-hour workweek were met by two month-long strikes, so that for federal and most municipal employees, the workweek was extended at most by 0.5 hours, whereas for state-level public sector employees, the workweek was extended between 0.2 and 1.6 hours.[4] Furthermore, employees at medical institutions were excluded from any work hour increases. Thus, by the mid-2000s, the standard workweek for public sector employees differed between state-level employees across states, and between employees of different administrative levels within-state.

3.3 Data

To study how these changes in the standard workweek affect work hour preferences, we use data from the German Socio-Economic Panel Study (SOEP), waves 1985 to 2017. The SOEP surveys a nationally representative sample of households on a number of topics, including employment and income. It thus contains self-stated information on whether an individual is a civil servant, public or private sector employee, the state they live in, their employment status (full- or part-time), the industry they work in, as well as weekly actual (*ha*) and contractual (*hc*) hours. More importantly, the SOEP also surveys individuals on their preferred weekly work hours (*hp*). Specifically, respondents are asked the following question:"If you could choose your own number of working hours, taking into account that your income would change according to the number of hours: How many hours would you want to work?"[5] Except

4. To prevent the automatic implementation of any workweek extension amongst federal and municipal employees under the "most-favored treatment" clause (*Meistbegünstigungklausel*), the length of the extension was determined using a convoluted calculation based on the difference between the standard and actual work hours of employees. As a result, the standard workweek was extended by unconventional magnitudes across states, with larger extensions implemented in states where employees were already working longer hours.

5. Note that individuals are explicitly asked to consider changes in income when answering this question, so that reporting a higher (lower) number of work hours is accompanied by an income increase (decrease).

for 1996, the SOEP contains data on this variable in all waves. This variable forms the basis for our main outcome variables. Prior to 2000, no decimal points were allowed in answering this question, whereas from 2000 onwards, one decimal point was allowed. We round down responses to this question to the nearest integer value (hp'), bin these responses into categories, and use binary variables based on these categories as dependent variables.

We restrict our sample to individuals between the ages of 15 and 49 (inclusive) who are regularly employed as a civil servant or white-collar public sector employee, with a contractual workweek. We exclude employees aged 50 and older from the sample, because public sector employees and civil servants above this age are eligible for special arrangements with respect to weekly hours. Furthermore, we exclude all individuals in full-time education, as well as individuals with an additional job. We also note that prior to 2000, the SOEP groups Saarland and Rhineland-Palatinate together in reporting the state respondents live in. Since the inhabitants of Saarland constitute only a minor fraction of the sample, we neglect this issue.

Table 3.1 summarizes several characteristics of individuals in our dataset. Several differences are apparent— public-sector employment is more female-dominated (70% female as opposed to 34% female in the civil service). Part-time employment, at 37%, is also more common amongst public sector employees. The proportion of individuals with a college degree is higher amongst civil servants, at 36%. In comparison, only 23% of public-sector employees have a college degree. This likely explains the higher average labor income of civil servants (1,834 EUR) compared to those of public-sector employees (1,217 EUR). We note also that individuals switch between the civil service, public sector and private sector employment. Amongst individuals who are observed in at least once as civil servants, 15% have switched or will switch to public- or private sector employment. Amongst public-sector employees, this fraction is 32%. The actual hours of the full-time employed, i.e., hours including overtime work, are higher than contractual hours, for both civil servants and public sector employees. Furthermore, desired hours are lower then contractual and actual hours, indicating that on average, full-time individuals would prefer to work fewer hours than they currently do.

3.3.1 Assignment to standard hours regime

As the SOEP does not survey individuals on their standard workweek, we merge this information to the dataset. For individuals in the public sector, we do so based on the state that they reside in, and whether they report being employed as a civil servant or public sector employee. We first discard occupational groups that are not subject to a standard workweek. For civil servants, these are soldiers, judges and firefighters. We also exclude teachers, because although the state or municipal standard workweek apply to them, they are also subject to a mandatory number of teaching hours per week. These mandatory teaching hours might be more salient— indeed,

Table 3.1. Descriptive statistics of civil servants and public sector employees

	Civil servants	Public sector employees
Female	0.34	0.70
Full-time employed	0.84	0.63
Lives with spouse	0.77	0.72
Child in household	0.53	0.51
Age	38.05	37.40
Net labor income	1,834.66	1,217.56
Has vocational degree	0.77	0.76
Has college degree	0.36	0.23
Observations	5,335	21,219
Individuals	1,036	5,856
Mean panel length	5.15	3.62
Proportion switch status	0.15	0.32
Net labor income (full-time)	1,957.83	1,506.71
Actual hours (full-time)	42.67	41.71
Contractual hours (full-time)	39.77	38.68
Desired hours (full-time)	38.31	36.43

Notes: Net labor income refers to net labor income earned through a main job in the previous month, converted to EURs and inflation-adjusted to 1995 terms. Hours variables are based on hours per week.

the majority of full-time employed teachers in our dataset report exceptionally low workweeks as a response to the survey item on contractual hours. Because we do not have complete information on the changes in mandatory teaching hours throughout the years, we remove teachers from our analysis. Lastly, as the post, telecommunications and railway transport sectors were privatized in the 1990s, we also discard individuals employed in these sectors.

As mentioned in Section 3.2, the standard workweek may differ within-state, depending on the administrative level that a civil servant or public sector employee is employed at. Because the SOEP lacks information pertaining to an individual's administrative level, the merging of standard hours is not perfect. As the standard workweek differs only between civil servants employed at the federal and non-federal (state and municipal) levels, and the former group are a small minority (10%), we assign all civil servants the state and municipal standard workweek, according to Panel A of Table 3.A.1.

As for public sector employees, the standard workweek is the same for all administrative levels until October 2005, so that incorrect assignment of the standard workweek is not an issue for the workweek reductions of 1989 and 1990. After 2005, the standard workweek differs between federal, state and municipal employees. Because both state and municipal employees form sizeable proportions of the public sector, we are unable to accurately assign standard hours to public sector employees

in this period and therefore do not study the workweek extensions of public sector employees.[6]

3.3.1.1 Changes in stated contractual hours

We check the validity of the merged standard hours variable (*hs*) using the surveyed contractual hours question.[7] As with responses to the survey question to desired hours, we round down responses to these questions to the nearest integer value and bin the rounded-down responses (hc' and hs') into the following categories: strictly less than 38 hours, 38 hours, 39 hours, 40 hours, 41 and 42 hours, and 43 hours and above. Figure 3.1 plots the fractions of responses within these categories, separately for full-time civil servants and full-time public sector employees. Prior to the first workweek reduction in 1989, almost all civil servants and public sector employees report having a contractual workweek of 40 hours. Following the implementation of the 38.5-hour workweek, a majority of (but not all) individuals report a contractual workweek category of 38 hours. However, 10% of civil servants and 20% of public sector employees report a contractual workweek category of less than 38 hours. This could be because these individuals are governed by industry contracts where the standard workweek is less than 38 hours, or by individual contracts.

In conjunction with the civil servant standard workweek increases, the compositions of contractual workweek categories of civil servants and public sector employees diverge from 1994 onwards. The fraction of civil servants with a contractual workweek category of 40 hours increases to 40% by 1997, and close to 60% of civil servants report contractual workweek categories of 41 or 42 hours after 2004. Correspondingly, the fraction of individuals with a contractual workweek category of 38 hours decreases to close to 0% by 2004. In contrast, the fraction of public sector employees with a contractual workweek category of 38 hours remains at 70% until 2005. Thereafter, the fractions of public sector employees with stated contractual workweek categories of 39 or 40 hours increase, which correspond to the public sector employee workweek extensions during this period.

As the workweek extensions of state- and municipal-level civil servants differ by state, we also plot in Figure 3.A.2 the binned stated contractual hours of full-time employed civil servants separately by region.[8] Following the implementation of the workweek increases, only 70 to 80% of respondents report having a contractual workweek category with the same number of hours as the standard workweek

6. The proportions of federal, state, and municipal employees in 2005 in direct public employment are 9%, 35%, and 56% respectively. Note that public service also includes indirect employment, which comprised 23% of public sector employment in 2005.

7. In English: "How many hours per week are stipulated in your contract (excluding overtime)?"; in German: "Wie viele Wochenstunden beträgt Ihre vereinbarte Arbeitszeit ohne Überstunden?"

8. Due to the low number of observations in some states, we group states together by reform timing, pre-reform standard workweek, and magnitude of change in the standard workweek.

category. This is likely due to the presence of federal-level civil servants in our sample, whose standard workweek differs from that of state- and municipal-level civil servants. Firstly, following the introduction of the 40-hour standard workweek in Baden-Württemberg, Bavaria, Lower Saxony, Bremen, and Rhineland-Palatinate in the 1990s, 20 to 30% of respondents in these states still report 38 hours, the federal standard workweek category, as their weekly contractual workweek category (see Figures 3.A.2a, 3.A.2b, 3.A.2c). Secondly, in conjunction with the implementation of workweek extensions at the federal level in 2004 from 38.5 to 40 hours, the fraction of individuals with a contractual workweek category of 38 hours decreases to 0% in all states, while the fraction of individuals with a 40-hour contractual workweek category increases by 10-20% in Lower Saxony, Bremen, and Rhineland-Palatinate, North Rhine-Westphalia and Hessia.[9] Note that due to the change in state and municipal standard workweeks from 40 to 41 or 42 hours in Baden-Württemberg and Bavaria in 2004, the fraction of civil servants with a 40-hour contractual workweek category decreases, but does not decrease to 0, which is consistent with federal-level civil servants having a contractual workweek of 40 hours.

3.3.1.2 Changes in stated actual hours

The SOEP also surveys individuals on their actual weekly work hours, which includes overtime hours.[10] Given the changes in stated contractual hours, it might be expected that actual hours also change proportionally in response to changes in the standard workweek.

Because stated actual hours vary more than contractual hours, we plot in Figure 3.2 the average, median and interquartile range of actual hours of full-time employed civil servants and public sector employees. As some civil servants and public sector employees report contractual workweek categories that do not correspond to the standard workweek category, Figure 3.2 also plots these statistics excluding these responses. As opposed to stated contractual hours, the workweek reductions in 1989 and 1990 did not lead to a large decrease in either the average or median actual hours, though the 25th and 75th percentiles decrease by 1.5 hours.

We next plot the average, median and interquartile range of actual hours of full-time employed civil servants separately by region in Figure 3.A.3. In contrast to the workweek reductions, workweek extensions generally lead to full increases in actual work hours— average actual hours of full-time Bavarian civil servants increase by two to four hours, and by two hours, following the workweek extensions

9. For federal-level civil servants, the standard workweek from 2006 onwards is 41 hours; however, individuals with at least one child under age of 12 are entitled to a 40-hour contractual workweek without a reduction in income.

10. In English: " And how many hours do you generally work, including any overtime?"; In German: "Und wieviel beträgt im Durchschnitt Ihre tatsächliche Arbeitszeit einschließlich eventueller Überstunden?"

in 1994 and 2004 respectively. Similarly, average actual hours also increase by two hours following the the workweek extensions in Baden-Württemberg, Lower Saxony, Bremen and Rhineland-Palatinate, North Rhine-Westphalia, and Hessia.

3.4 Empirical strategy

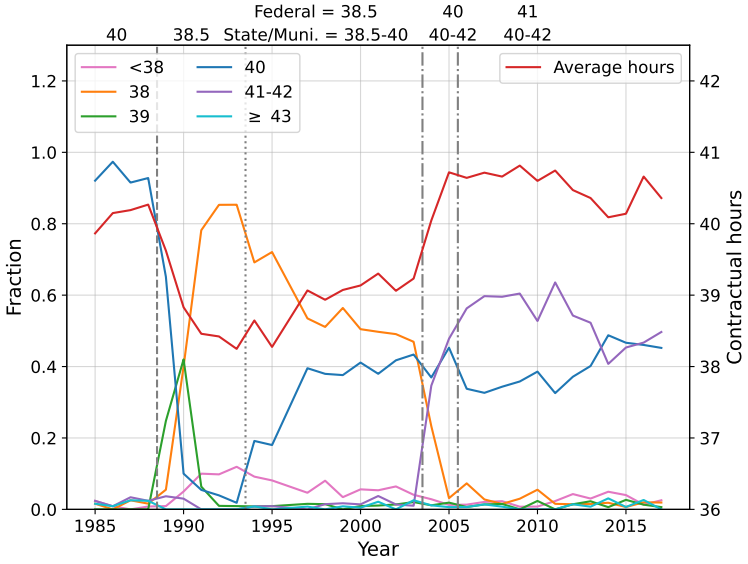
Our main empirical strategy involves comparing the desired hour categories of civil servants or public sector employees in the years before and after a change in the standard workweek. We estimate both dynamic and (partly) static specifications, as the former allows for tracking of the evolution of desired hours over time, while the latter allows for a sufficient number of observations to study the effect of changes separately for each state.

More specifically, we estimate the following dynamic specification:

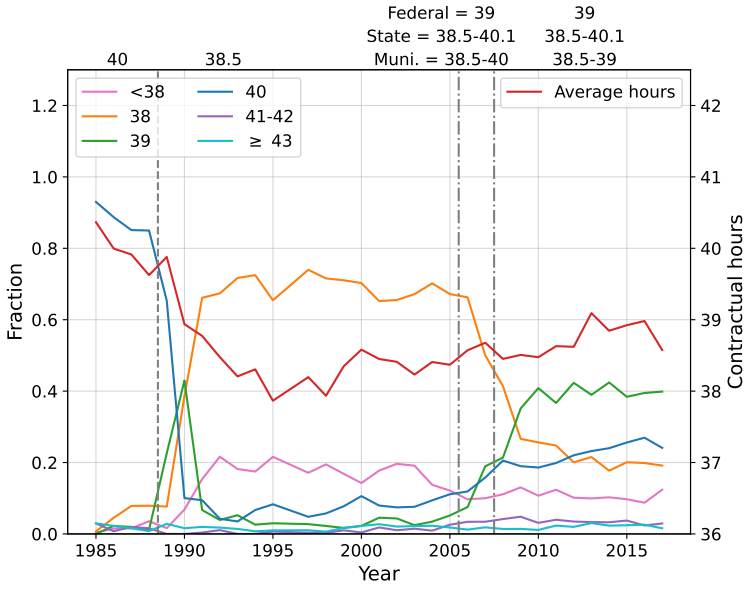
$$\mathbb{1}\{hp' \in k\}_{it} = \alpha + \sum_{l \neq b} \delta_l \cdot D_{it}^l + \varepsilon_{it} \tag{3.1}$$

where i, t, and l denote individual, year, and relative year respectively. The indicator $\mathbb{1}\{hp' \in k\}_{it}$ equals one if individual i's (rounded-down) preferred workweek (hp') is in category k; b refers to the base year, which is typically either -1 or -2, that is, one or two years preceding the change in the standard workweek respectively. D_{it}^l are event-time dummies for the years preceding and following the standard workweek reform. δ_l consequently refers to the difference in the fractions of individuals with a desired workweek category of k in period l and the base year. Due to adjustment costs, we expect workweek reductions and extensions to have different effects. We therefore estimate equation (3.1) separately for the workweek reductions between 1989 and 1991 using all civil servant and public sector employee observations, and for the civil servant workweek extensions between 1994 and 2004. For the workweek reductions, we define l as years relative to 1988, the year immediately preceding the introduction of the 39-hour workweek in most states.[11] We bin the rounded-down responses into the following categories: strictly less than 38 hours, 38 hours, 39 hours, 40 hours, 41 hours and above, and estimate equation (3.1) using binary variables based on these categories. As for the civil servant workweek extensions, l is defined as years relative to year of the first workweek extension. We set $b = -2$ because the question on preferred hours is not asked in 1996, and this coincides with $l = -1$ for states with a workweek extension in 1997. We estimate equation (3.1) with binary variables based on the following categories— strictly less than 38 hours,

11. This corresponds to $b = -1$ for all public sector employees and civil servants in most states. Note that the civil servant standard workweeks in Schlewsig-Holstein and Hessia were decreased directly from 40 to 38.5 hours in 1990 and 1991 respectively.



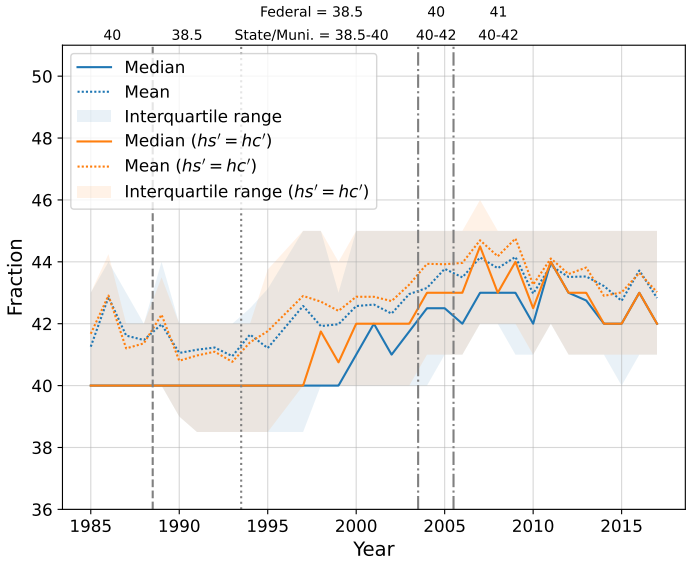
(a) Full-time civil servants



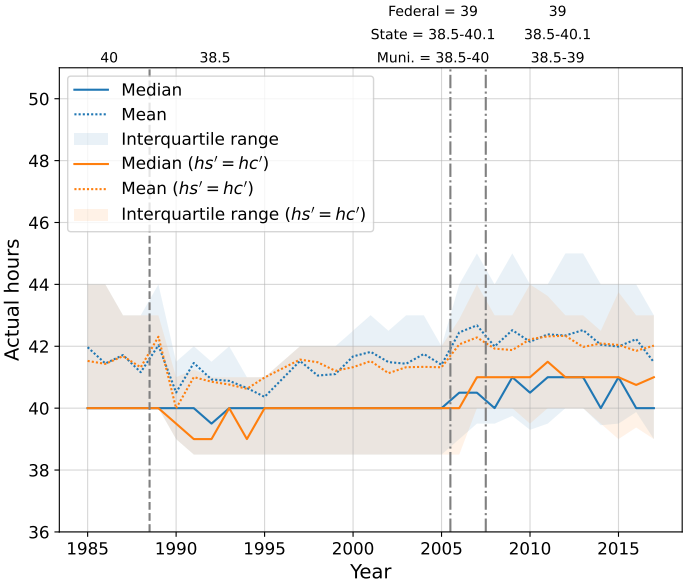
(b) Full-time public sector employees

Figure 3.1. Contractual workweek categories of full-time civil servants and public sector employees

Notes: Each line corresponding to the left axis plots the fraction of individuals whose rounded-down contractual workweek (hc') lies in a particular category, for the following categories: strictly less than 38 hours, 38 hours, 39 hours, 40 hours, 41 or 42 hours, and 43 hour or more. The line corresponding to the right axis plots average stated contractual hours. The text above the figure refers to the standard workweek in place for different administrative levels. See Tables 3.A.1 and 3.A.2 for information on standard workweeks.



(a) Full-time civil servants



(b) Full-time public sector employees

Figure 3.2. Actual hours statistics of full-time civil servants and public sector employees

Notes: Panels (a) and (b) plot various statistics of stated actual hours of full-time civil servants and public sector employees respectively. The bold line, dashed line, and shaded area plot the average, median, and interquartile range respectively. The blue lines and shaded area refer to statistics computed using all full-time individuals, while the orange lines and shaded area refer to statistics computed using only full-time individuals whose stated contractual workweek category equals the standard workweek category.

38 hours, 39 hours, the same number of hours as the new standard workweek, and any number of hours strictly more than the new standard workweek.[12]

12. Due to differences in the new standard workweeks implemented, responses are binned differently by state. For states where the new standard workweek is 40 hours, the categories are defined

We also estimate partially static specifications by grouping years into intervals. For the workweek reductions, we first estimate:

$$\mathbb{1}\{hp' \in k\}_{it} = \alpha + \delta_1 Post_{0\leq l\leq 4,it} + \delta_2 Post_{l\geq 5,it} + \varepsilon_{it} \tag{3.2}$$

where $k = [40, 41)$, and l is defined as years from the introduction of the first workweek reduction. The indicator variable $Post_{0\leq l\leq 4,it}$ is 1 for observations in the first five years after the removal of the 40-hour workweek, and 0 otherwise. Because the first civil servant workweek extensions were implemented in 1994, we include an indicator $Post_{l\geq 5,it}$ for observations from the sixth year after the removal of the 40-hour workweek. δ_1 thus measures the change in the fraction of individuals preferring a 40-hour workweek category in the four to five years after the introduction of either a 39-hour or 38.5-hour standard workweek, relative to the four years prior. Because the introduction of the 38.5-hour workweek does not coincide with the removal of the 40-hour workweek for all public sector employees and civil servants in most states, we estimate a modified version of equation (3.2) when using $\mathbb{1}\{hp' \in [38, 39)\}_{it}$ and $\mathbb{1}\{hp' \in [0, 38)\}_{it}$ as dependent variables. In this case, l indexes years since the introduction of the 38.5-hour workweek. As the 38.5-hour workweek was introduced in 1990 in most states, and the first civil servant workweek extensions were implemented in 1994, we replace $Post_{0\leq l\leq 4,it}$ with $Post_{0\leq l\leq 3,it}$, and $Post_{l\geq 5,it}$ with $Post_{l\geq 4,it}$. For the workweek extensions, we estimate:

$$\mathbb{1}\{hp' \in k\}_{it} = \alpha + \beta Pre_{l\leq -4,it} + \delta_1 Post_{0\leq l\leq 4,it} + \delta_2 Post_{5\leq l\leq 7,it} + \delta_3 Post_{l\geq 8,it} + \varepsilon_{it} \tag{3.3}$$

where $Post_{0\leq l\leq 4,it}$ is defined as in equation (3.2). l indexes years from introduction of first workweek extension. Because of the timing of the workweek reductions and first workweek extension in 1994 in Bavaria and Schleswig-Holstein, we include the indicator variable $Pre_{l\leq -4,it}$ in equation (3.3). Similarly, due to the implementation of a second workweek extension seven years after the first workweek extensions in Schleswig-Holstein and Baden-Württemberg, we include the indicator $Post_{l\geq 8,it}$ for observations in the eight years after the implementation of the first workweek extension. δ_1 thus measures the change in the fraction of individuals preferring a k-hour workweek category in the four to five years after the introduction of the new standard workweek, relative to the three years prior, while δ_2 measures the same relative change from the fifth to seventh year.

---

as follows: strictly less than 38 hours, 38 hours, 39 hours, 40 hours, and 41 hours and above. As for North Rhine-Westphalia, where the new standard workweek is 41 hours, the last two categories are defined as 41 hours, and 42 hours and above. Similarly, for Hessia, where the new standard workweek is 42 hours, the last two categories are defined as 42 hours, and 43 hours and above.

3.5 Effect of workweek changes on desired hours

3.5.1 Workweek reductions

We first present results on the workweek reductions of civil servants and public sector employees in 1989 and 1990. Since the workweek reductions involved two changes for most individuals— from 40 to 39 hours, and subsequently from 39 to 38.5 hours, we expect to see a decrease in the fraction of individuals preferring a 40-hour workweek category, a temporary increase in the fraction of individuals preferring a 39-hours workweek category, and a longer-lasting increase in the fraction of individuals preferring a 38-hour workweek category. Figure 3.3 plots the event-study estimates of equation (3.1), using with binary variables of various hour categories as dependent variables.

Figure 3.3a estimates equation (3.1) using $\mathbb{1}\{hp' = 39\}$ and $\mathbb{1}\{hp' = 38\}$ as dependent variables. Pre-reduction trends are constant, indicating no major changes in the fractions of individuals preferring either a 38- or 39-hour workweek category. The fraction of individuals preferring a 39-hour workweek category increases by 10 pp in the first two years after the introduction of the 39-hour workweek. At the same time, the fraction of individuals preferring the old standard workweek category of 40 hours decreases by 10 pp in $l = 0$ and by 20 pp in $l = 1$, relative to $l = -1$. We find a similar pattern with the introduction of the 38-hour workweek— the fraction of individuals preferring this workweek category increases by 10 pp while the fraction of individuals preferring a 39-hour workweek category decreases. Figure 3.3b, which plots estimates using $\mathbb{1}\{hp' < 38\}$ and $\mathbb{1}\{hp' > 41\}$ as dependent variables, also indicates flat trends pre-reform. Though there is also a slight increase in the fraction of individuals preferring a lower than 38-hour workweek category, this is likely due to some individuals facing different standard workweeks post-reform.

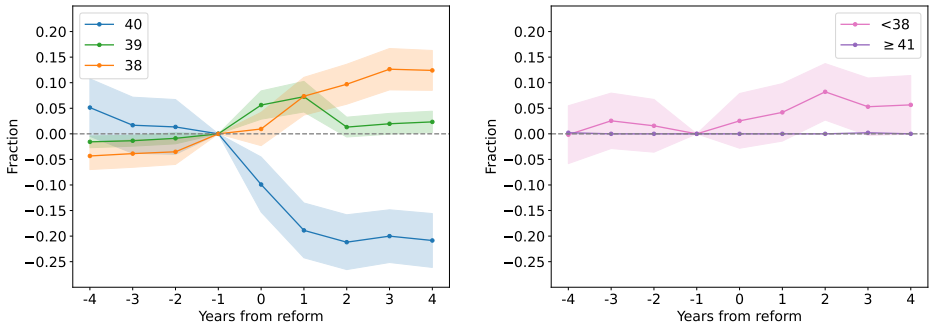
Table 3.2 presents estimates of equation (3.2) using as dependent variables $\mathbb{1}\{hp' = 38\}$, $\mathbb{1}\{hp' = 40\}$, and $\mathbb{1}\{hp' < 38\}$ in Panel A, B and C respectively. Column (1) of Panel A indicates an average 12.9 pp increase in the fraction of individuals preferring to work 38 hours in the four to five years after the introduction of the 38.5-hour standard workweek, relative to the four years prior. Including individual FEs in column (2), as well as controls in column (3) yield smaller but nonetheless significant estimates of 11.1 pp and 11.3 pp respectively. As mentioned in Section 3.3.1, not all civil servants and public sector employees are full-time employed; furthermore only a subset of these individuals report having a contractual workweek that corresponds to the standard workweek. Columns (4) and (5), which present estimates using only full-time employed individuals and individuals whose stated contractual hours equal the standard workweek respectively, indicate larger increases, at 15.9 and at 21.9 pp respectively.

Columns (1) to (3) of Panel B in turn indicate decreases between 17.8 and 20.5 pp in the fraction of individuals preferring the old standard 40-hour workweek post-

Table 3.2. Effect of standard workweek decrease on probability of preferring various workweek categories

	(1)	(2)	(3)	(4)	(5)
Panel A: 38 hours					
$\text{Post}_{0\leq l\leq 3}$	0.1291***	0.1113***	0.1134***	0.1592***	0.2187***
	(0.0114)	(0.0131)	(0.0132)	(0.0133)	(0.0172)
Pre-reform avg.	0.057	0.058	0.058	0.064	0.047
Panel B: 40 hours					
$\text{Post}_{0\leq l\leq 4}$	-0.2053***	-0.1793***	-0.1784***	-0.2285***	-0.2442***
	(0.0165)	(0.0185)	(0.0184)	(0.0182)	(0.0203)
Pre-reform avg.	0.411	0.409	0.409	0.466	0.489
Panel C: < 38 hours					
$\text{Post}_{0\leq l\leq 3}$	0.0488***	0.0425**	0.0378**	0.0317*	0.0005
	(0.0175)	(0.0177)	(0.0174)	(0.0185)	(0.0214)
Pre-reform avg.	0.502	0.501	0.501	0.436	0.431
Observations	22,928	20,855	20,836	15,771	8,502
Clusters	5,814	3,741	3,736	4,050	2,278
Sample	All	All	All	Full-time	$hc' = hs'$
Individual FEs	—	*Yes*	*Yes*	—	—
Controls	—	—	*Yes*	—	—

Notes: OLS estimates of equation (3.2), with $\mathbb{1}\{hp' = 38\}$, $\mathbb{1}\{hp' = 40\}$, and $\mathbb{1}\{hp' < 38\}$ as dependent variables in Panels A, B, and C respectively. Column (1) uses the full sample of civil servants and public sector employees. Columns (2) and (3) include individual fixed effects. Column (3) also includes the following controls: whether there is a child below the age of 16 in the individual's household, whether the individual lives with a spouse, and whether the individual has a college or vocational degree. Column (4) uses only full-time observations, whereas column (5) uses observations whose stated contractual workweek category equals the standard workweek category. The decrease in observations in column (5) is due to our inability to distinguish between state and municipal public sector employees, and the workweeks for these two groups differ from 2005 onwards. Coefficient estimates on $Post_{l\geq 5}$ and $Post_{l\geq 4}$ are omitted from this table. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



(a) Standard categories **(b)** Other categories

Figure 3.3. Effect of standard workweek decrease on probability of preferring various workweek categories

Notes: OLS estimates of equation (3.1) with $\mathbb{1}\{hp' = 40\}$, $\mathbb{1}\{hp' = 39\}$ and $\mathbb{1}\{hp' = 38\}$ as dependent variables in Panel (a), $\mathbb{1}\{hp' < 38\}$ and $\mathbb{1}\{hp' > 41\}$ as dependent variables in Panel (b), using all civil servant and public sector employee observations. Only estimates of event-times corresponding to years before the first civil servant workweek extensions in 1994 are plotted. Bands indicate 95% confidence intervals. Standard errors clustered at the individual level.

reform. Columns (4) and (5), which are based on full-time observations and observations whose stated contractual workweek equals the standard workweek respectively, indicate decreases of 22.9 and 24.4 pp respectively, and thus suggest that the smaller increases in the fraction of individuals preferring the new standard workweek is partly due to the presence individuals whose contractual workweek do not correspond to the standard workweek. Estimates from Panel C also support this: columns (1) to (4), which either use the full sample or the sample of full-tme employed observations, indicate a 3.2 to 4.9 pp increase in the fraction of individuals preferring a workweek with less than 38 hours. However, column (5), which uses only observations where stated contractual hours equals the standard workweek, yields an estimate that is close to 0.

3.5.2 Workweek extensions

Having established that a decrease in the standard workweek leads to a decrease (increase) in the fraction of individuals preferring the old (new) standard workweek, we now turn to the standard workweek extensions. As in the previous section, we first estimate equation (3.1) using indicator variables of various desired hours categories as dependent variables in Figure 3.4.

Due to the workweek reductions implemented between 1989 and 1990, Figure 3.4a indicates an increase in the fraction of individuals preferring a 38-hour workweek category, as well as a decrease in the fraction of individuals preferring the new

standard workweek category, nine to four years pre-reform.[13] Following the introduction of the new, longer standard workweek, the fraction of individuals preferring the new standard workweek gradually increases and plateaus at 15 pp four to five years post-reform. The fraction of individuals preferring the old standard workweek also decreases in the four to five years post-reform, before stabilizing at -10 pp from $l = 5$ onwards. These findings contrast with those for the workweek reductions of Figure 3.3a, which show an immediate and larger change in desired hours categories. Figure 3.4b, which plots estimates of equation (3.1) using as dependent variables binary variables based on desired hours categories that equal neither the old nor new standard workweek categories, indicates no changes either before or after the workweek extension.

Table 3.3 reports estimates of equation (3.3) with $\mathbb{1}\{hp' = hs'_{new}\}$ and $\mathbb{1}\{hp' = 38\}$ as dependent variables in Panels A and B respectively. Estimates in column (1) of Panel A indicate a 8.5 pp increase in the fraction of civil servants preferring the new standard workweek in the first five years after the standard workweek increase, from a pre-reform average of 13.3 pp.[14] This fraction increases to 12.2 pp ($p<0.01$) in the subsequent three years. The inclusion of individual fixed effects, in column (2), leads to slightly smaller estimates— 7.1 pp and 9.8 pp in the first five years and subsequent three years respectively. Including controls (alongside individual FEs) in column (3) does not change the estimates much. Using only full-time employed observations, in column (4), yields similar estimates to those of the full sample in column (1). Lastly, limiting the sample only to observations where the contractual workweek category equals the standard workweek category yields larger estimates of 15.6 pp and 23.4 pp. Estimates from Panel B indicate a parallel but smaller decrease in the fraction of individuals preferring the 38-hour workweek category after the implementation of the new (non-38.5-hour) workweek. Column (1) of Panel B points to a 6.1 pp ($p<0.05$) decrease in the first five years after the workweek extension, and a significant 8.2 pp ($p<0.01$) decrease in the sixth to eight years post-reform. Columns (2) and (3) and (4) of Panel B yield similar estimates to column (1), while column (5), which uses only observations where the contractual workweek category equals the standard workweek category, yields larger estimates of -12.6 pp and -17.5 pp.

Overall, this section suggests that workweek extensions induce smaller and more gradual changes in preferred hours compared to workweek reductions.

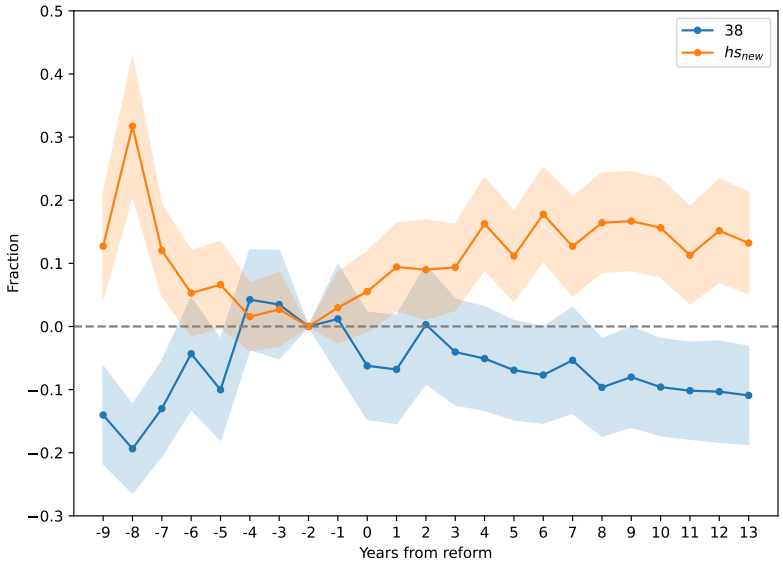
13. The old standard workweek is 38.5 hours in all states, while the new standard workweek is 40 hours in most states.

14. The pre-reform average varies largely across states depending on the new standard workweek implemented. In states where the new standard workweek is 40 hours, a sizeable fraction of individuals (32%) prefer a 40-hour workweek even while the 38.5-hour standard workweek was in place. This is not the case for 39-, 41-, or 42-hour standard workweeks— the fractions of individuals preferring these workweeks under a standard workweek of 38.5 hours is close to 0.

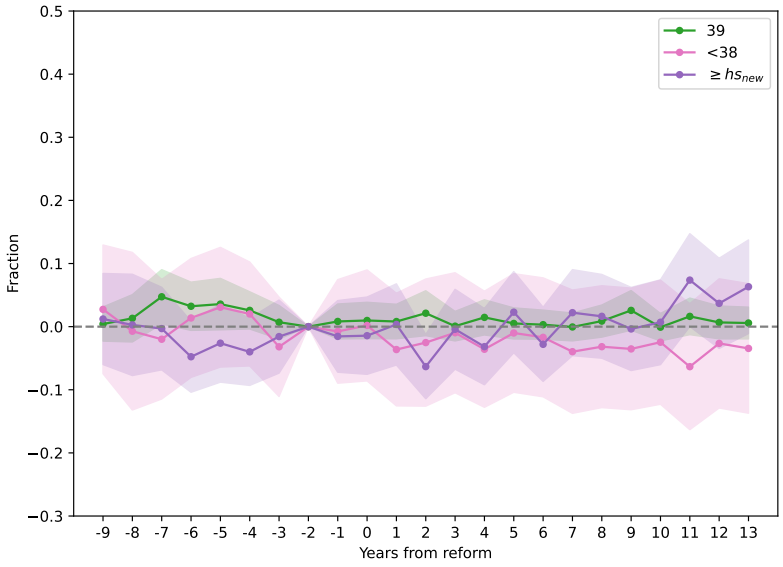
Table 3.3. Effect of standard workweek increase on probability of preferring various workweek categories

	(1)	(2)	(3)	(4)	(5)
Panel A: New standard workweek					
$\text{Post}_{0\leq l\leq 4}$	0.0850***	0.0712***	0.0759***	0.0957***	0.1566***
	(0.0217)	(0.0244)	(0.0246)	(0.0253)	(0.0307)
$\text{Post}_{5\leq l\leq 7}$	0.1221***	0.0980***	0.1048***	0.1408***	0.2344***
	(0.0279)	(0.0347)	(0.0355)	(0.0319)	(0.0391)
Pre-reform avg.	0.133	0.129	0.129	0.156	0.155
Panel B: 38 hours					
$\text{Post}_{0\leq l\leq 4}$	-0.0609**	-0.0499*	-0.0528*	-0.0584**	-0.1262***
	(0.0250)	(0.0272)	(0.0270)	(0.0286)	(0.0346)
$\text{Post}_{5\leq l\leq 7}$	-0.0822***	-0.0683*	-0.0726**	-0.0835**	-0.1751***
	(0.0291)	(0.0353)	(0.0348)	(0.0330)	(0.0384)
Pre-reform avg.	0.234	0.229	0.229	0.258	0.302
Observations	5,335	5,018	5,015	4,486	3,368
Clusters	1,036	719	719	903	740
Sample	All	All	All	Full-time	$hc' = hs'$
Individual FEs	—	*Yes*	*Yes*	—	—
Controls	—	—	*Yes*	—	—

Notes: OLS estimates of equation (3.3), with $\mathbb{1}\{hp' = hs'_{new}\}$ and $\mathbb{1}\{hp' = 38\}$ as the dependent variables in Panels A and B respectively. Column (1) uses the full sample of civil servant observations. Columns (2) and (3) include individual fixed effects. Column (3) also includes the following controls: whether there is a child below the age of 16 in the individual's household, whether the individual lives with a spouse, and whether the individual has a college or vocational degree. Column (4) uses only full-time observations, whereas column (5) uses observations where the stated contractual workweek category equals the standard workweek category. Estimated coefficients on $Pre_{l\leq -4}$ and $Post_{l\geq 8}$ omitted from this table. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



(a) Standard categories



(b) Other categories

Figure 3.4. Effect of standard workweek increase on probability of preferring various workweek categories

Notes: OLS estimates of equation (3.1) with $\mathbb{1}\{hp' = hs'_{new}\}$ and $\mathbb{1}\{hp' = 38\}$ as dependent variables in Panel (a), $\mathbb{1}\{hp' < 38\}$, $\mathbb{1}\{hp' = 39\}$, and $\mathbb{1}\{hp' > hs'_{new}\}$ as dependent variables in Panel (b), using all civil servant observations. Only event-time indicators where all states should be represented are plotted; note however that $l = -8$, $l = -1$, and $l = 2$ are missing observations from some states because the question on desired hours was not asked in 1996. Bands indicate 95% confidence intervals. Standard errors clustered at individual level.

3.5.2.1 Heterogeneous effects by state

Sections 3.5.1 and 3.5.2 reported the overall effect of standard workweek changes, averaged across states. Due to differences in the magnitude of standard workweek extensions (ranging from 1 to 2.5 hours) and differences in sample sizes across states, this section examines the effect of workweek changes by state.

For the workweek reductions, we estimate a variant of equation (3.2) by interacting the *Post* indicators with state dummies. More specifically, we estimate:

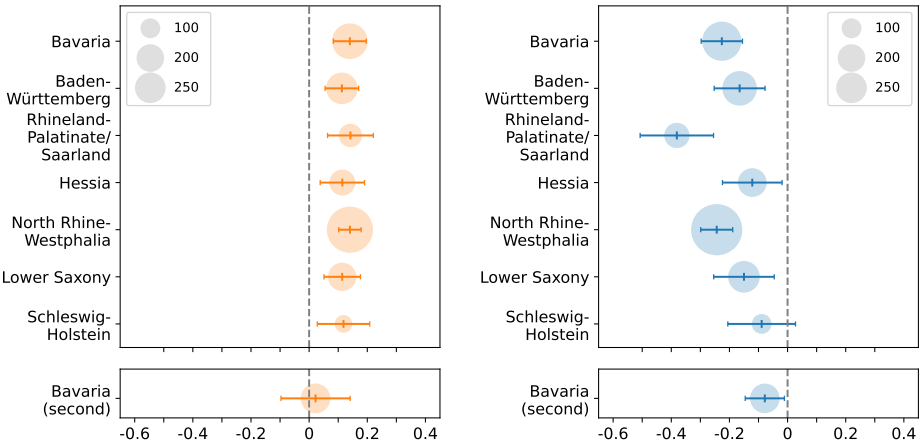
$$\begin{aligned}\mathbb{1}\{hp' \in k\}_{it} = \alpha_i + \delta_1 Post_{0\leq l\leq 4,it} + \delta_2 Post_{l\geq 5,it} \\ + \sum_{s\neq NRW} \mathbb{1}\{state = s\} \cdot \left(\gamma_s + \delta_{1s} Post_{0\leq l\leq 4,it} + \delta_{2s} Post_{l\geq 5,it}\right) + \varepsilon_{it}\end{aligned} \tag{3.4}$$

where *NRW* denotes North Rhine-Westphalia. The upper panels of Figure 3.5 plot the average difference in the fraction of individuals preferring the old (new) standard workweeks post-reform for each state. For North Rhine-Westphalia, this is the estimate of δ_1; while for other states, this is the estimate of $\delta_1 + \delta_{1s}$. Because we omitted the civil servant workweek reductions of 2012 and 2013 in Bavaria (from 42 to 41 hours, and subsequently from 41 to 40 hours) from the analysis in section 3.5.1, we estimate equation (3.2) using the subsample of civil servants in Bavaria and plot the coefficients on the *Post* indicators in the lower panels of each subfigure in Figure 3.5. Figure 3.5a shows that the increase in the fraction of individuals preferring a workweek category of 38 hours is fairly uniform, ranging from 2.2 to 14.0 pp, while Figure 3.5b suggests more variation in the decrease in the fraction of individuals preferring the old standard workweek, with estimates ranging from -7.8 to -38.1 pp.

Similarly, for the workweek extensions, we interact the grouped time indicators in equation (3.3) with state dummies:

$$\begin{aligned}\mathbb{1}\{hp' \in k\}_{it} = \alpha_i + \beta Pre_{l\leq -4,it} + \delta_1 Post_{0\leq l\leq 4,it} + \delta_2 Post_{5\leq l\leq 7,it} + \delta_3 Post_{l\geq 8,it} \\ + \sum_{s\neq NRW} \mathbb{1}\{state = s\} \cdot \Big(\gamma_s + \beta_s Pre_{l\leq -4,it} + \delta_{1s} Post_{0\leq l\leq 4,it} \\ + \delta_{2s} Post_{5\leq l\leq 7,it} + \delta_{3s} Post_{l\geq 8,it}\Big) + \varepsilon_{it}\end{aligned} \tag{3.5}$$

and plot the state-specific post-reform changes in the fraction of individuals preferring the old (new) standard workweek categories in the upper panels of each subfigure in Figure 3.6. For North Rhine Westphalia, these are the estimated δ_1 and δ_2, while for other states these are the estimated $\delta_1 + \delta_{1s}$ and $\delta_2 + \delta_{2s}$. Figure 3.6 indicates that the effect of the workweek extensions vary more across states, ranging between -8.5 and 14.0 pp for the new standard workweek, and between -15.9 and 6.6 pp for the old standard workweek. Indeed, the average effect in Section



(a) New standard workweek

(b) Old standard workweek

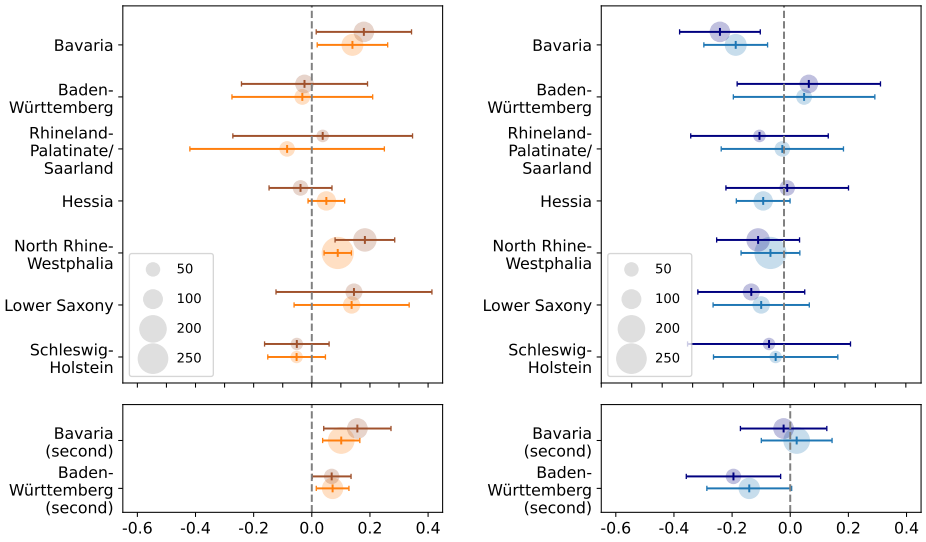
Figure 3.5. Effect of standard workweek decrease on change in probability of preferring new and old standard workweek categories, by state

Notes: Upper panels plot OLS estimates of equation (3.4) using all civil servant and public sector employee observations. For North Rhine-Westphalia, this is the coeffiecient on $Post_{0\leq l\leq 4,it}$ (δ_1), while for other states, this is the sum of coefficient on $Post_{0\leq l\leq 4,it}$ and $\mathbb{1}\{state = s\} \cdot Post_{0\leq l\leq 4,it}$ ($\delta_1 + \delta_{1s}$). Lower panels plot OLS estimates of the coefficient on $Post_{0\leq l\leq 4,it}$ (δ_1) from equation (3.2) using only civil servant observations from Bavaria. Panels (a) and (b) plot estimates from specifications using as dependent variables $\mathbb{1}\{hp' = hs'_{new}\}$ and $\mathbb{1}\{hp' = hs'_{old}\}$ respectively. States with fewer than 20 observations are omitted from the figure. Notches denote estimates, while circles denote the number of observations in a particular state post-reform. Bars represent 95% confidence intervals. Standard errors clustered at individual level.

3.5.2 is largely driven by the more populous states of Bavaria and North Rhine-Westphalia, and to a smaller extent, Hessia and Lower Saxony. Because we omitted the second extensions of Bavaria and Baden-Württemberg from the main analysis, we also estimate a variant of equation (3.5) using only civil servants from Bavaria and Baden-Württemberg in the lower panels of Figure 3.6, where relative time l is defined according to the second workweek extension in each state, and the grouped time indicators are interacted with a dummy variable for observations from Bavaria. The lower panel of Figure 3.6a shows that effect sizes for these two reforms are similar to that of North Rhine-Westphalia, Lower Saxony, and the first extension in Bavaria. Interestingly, for the second workweek extension in Bavaria, although the fraction of individuals preferring the new standard workweek increases, there is no corresponding decrease in the fraction preferring the old standard workweek. Overall, it appears that the workweek extensions implemented in the 2000s have larger effects compared to the workweek extensions implemented in the 1990s.

3.5.2.2 Individual changes in desired hours

The previous sections show that changes in the standard workweek, in particularly decreases in the standard workweek, lead to an increase (decrease) in the propor-



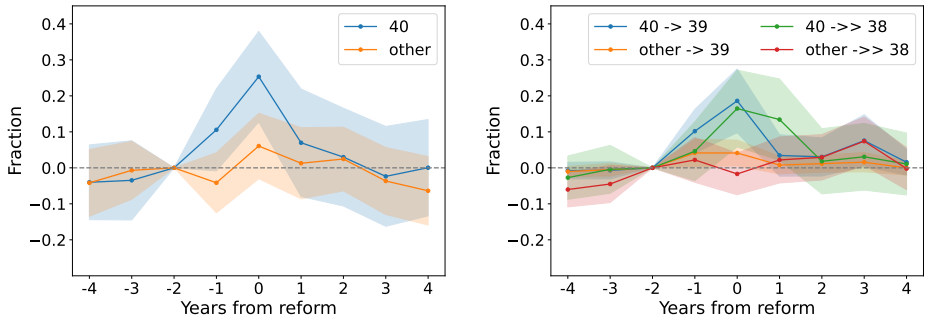
(a) New standard workweek **(b)** Old standard workweek

Figure 3.6. Effect of standard workweek increase on change in probability of preferring new and old standard workweek categories, by state

Notes: Upper panels plot OLS estimates of equation (3.5) using all civil servant observations. For North Rhine-Westphalia, these are the coeffiecients on $Post_{0\leq l\leq 4,it}$ (δ_1) and $Post_{5\leq l\leq 7,it}$ (δ_2), while for other states, these are the sum of coefficients on $Post_{0\leq l\leq 4,it}$ and $\mathbb{1}\{state = s\} \cdot Post_{0\leq l\leq 4,it}$ ($\delta_1 + \delta_{1s}$), and the sum of coefficients on $Post_{5\leq l\leq 7,it}$ and $\mathbb{1}\{state = s\} \cdot Post_{5\leq l\leq 7,it}$ ($\delta_2 + \delta_{2s}$). Lower panels plot analogous OLS estimates from a variant of equation (3.5), where the grouped time indicators are interacted with a dummy variable for observations from Bavaria, using only civil servant observations from Baden-Württemberg and Bavaria. Panels (a) and (b) plot estimates from specifications using as dependent variables $\mathbb{1}\{hp' = hs'_{new}\}$ and $\mathbb{1}\{hp' = hs'_{old}\}$ respectively. States with fewer than 20 observations are omitted from the figure. Notches denote estimates, while circles denote the number of observations in a particular state post-reform. Bars represent 95% confidence intervals. Standard errors clustered at individual level.

tion of individuals preferring the new (old) standard workweek. In this section, we exploit the panel structure of our dataset and link the changes in aggregate fractions to changes in individual choices.

Figure 3.7a plots the change in the fraction of individuals preferring a different workweek in the next year (switchers) for years around the reform, relative to two years prior to the first workweek reduction. We do so separately for individuals preferring the old standard workweek category (40 hours), and those preferring neither the old nor new standard workweek categories of 38 or 39 hours. Relative to two years prior to the first workweek extension, the fraction of individuals who prefer a 40-hour workweek category and who switch to a different preferred workweek in the next year increases by 10.6 pp in the year prior to the implementation of the 39-hour standard workweek, and by 25.3 pp a year later. In contrast, we do not see large increases in switching amongst individuals preferring workweeks below 38 hours or 41 hours and above.



(a) Switch out **(b)** Switching out to new standard workweeks

Figure 3.7. Effect of introduction of 39- and 38.5-hour standard workweeks on within-individual changes in desired workweek categories

Notes: OLS estimates of equation (3.1). Panel (a) plots estimates from specifications using $\mathbb{1}\{hp'_{t+1} \neq hp'_t\}$ as the dependent variable. Panel (b) plots estimates from specifications using $\mathbb{1}\{hp'_{t+1} = 39\}$ (blue and orange lines) and $\mathbb{1}\{hp'_{t+2} = 38\}$ (green and red lines) as dependent variables. The blue and green lines (in either panel) plot estimates using observations where the stated desired workweek category equals the old standard workweek category, while the orange and red lines plot estimates using observations with preferred workweek categories that do not equal the old or new standard workweeks. We have 144 individuals with a preferred workweek category of 40 hours, and 196 individuals with a preferred workweek of less than 38 hours or 41 hours or more in $l = -2$. Bands represent 95% confidence intervals. Standard errors clustered at individual level.

Figure 3.7b plots the change in the fraction of individuals switching to a preferred workweek category of 39 hours in the next year, and the change in the fraction of individuals switching a preferred workweek category of 38 hours after two years, relative to two years prior to the first workweek decrease. The fraction of individuals preferring a 40-hour workweek and switching to a preferred workweek of 39 hours in the next year increases by 10.2 and 18.6 pp in $l = -1$ and $l = 0$. Consistent with the implementation of workweek reductions, the fraction of individuals preferring a 40-hour workweek and switching to a preferred 38-hour workweek category two years later also increases by a similar magnitude. This indicates that individuals preferring a 40-hour workweek category prefer a 39-hour workweek category when the 39-hour workweek is implemented, and in turn prefer a 38-hour workweek category a year later, in conjunction with the implementation of the 38.5-hour workweek. On the other hand, we find no increases in switching to a preferred workweek category of 39 or 38 hours amongst individuals preferring a workweek of less than 38 hours or more than 41 hours. Overall, Figure 3.7 suggests that the effects from Section 3.5.1 are largely due to individuals switching from preferring the old standard workweek to preferring the new standard workweek.

As for the workweek extensions, we bin together observations from several years due to the smaller number of observations and estimate a specification similar to equation (3.3), but with finer grouped time indicators, as we expect higher rates of switching only in the first few years post-reform. More specifically, we estimate the

following equation:

$$\mathbb{1}\{hp'_{t+1} \in k\} = \alpha + \beta Pre_{l\leq -5,it} + \delta_1 Post_{-1\leq l\leq 2,it} + \delta_2 Post_{3\leq l\leq 6,it} + \delta_3 Post_{l\geq 7,it} + \varepsilon_{it} \tag{3.6}$$

where l denotes years to the reform. Note that because the dependent variable is a future response, $Post_{-1\leq l\leq 2,it}$ also includes the year immediately preceding the reform.

Column (1) of Panel A of Table 3.4 indicates that individuals preferring the old standard workweek are not significantly more likely to switch in the following year to a different preferred workweek post-reform; and are in fact significantly less likely to switch to a different workweek category three to six years post-reform. At the same time, column (3) suggests that the effect we find is not due to individuals switching directly from preferring the old standard workweek to preferring the new standard workweek. Columns (4) and (5) instead point to changes in switching behavior of individuals preferring neither the new nor old workweek as the main driver underlying the increased popularity of the new standard workweek post-reform—these individuals are more (less) likely to switch to prefer the new (old) standard workweek post-reform. For the second workweek extensions in Baden-Württemberg and Bavaria, we estimate equation (3.6) using the subsample of civil servants in these states, with relative time l defined as years from the second workweek extension in each respective state. We find only insignificant increases in switching out amongst individuals preferring a 40-hour workweek category, and amongst individuals preferring neither the old nor new standard workweeks. In contrast to Panel A, individuals preferring the old standard workweek are more likely to switch to prefer the new standard workweek.

Given the large increase in standard hours in states such as North Rhine-Westphalia and Hessia, it might be expected that individuals initially switch to a preferred workweek between the old and new standard workweeks, and only prefer the new standard workweek in later years. We find some evidence consistent with this idea. We estimate in column (1) of Table 3.A.4 equation (3.6) using as a sample individuals preferring the old standard workweek, but with a binary variable indicating if the individual prefers a workweek between the old and new standard workweeks as a dependent variable. Column (1) indicates an insignificant 7.3 pp increase in the fraction of individuals preferring the old standard workweek and switching to prefer an intermediate workweek in the next year.[15] Column (2) shows a significant 19.9 pp increase in switching to a preferred workweek corresponding to the new standard workweek category three to six years post-reform amongst individuals preferring an intermediate workweek, while column (3) indicates that this

15. In unpublished specifications, we do find a significant increase in the fraction of individuals preferring the old standard workweek who switch to prefer either an intermediate workweek or the new workweek post-reform.

group of individuals are not significantly less likely to switch to prefer the old standard workweek post-reform. However, this could also potentially reflect the baseline tendency of indviduals switching from preferring non-standard workweek to standard workweek categories in the next year. To see if this is the case, we define as a dependent variable an indicator that equals one if an individual's preferred workweek category corresponds to the standard workweek category in a given year.[16] Column (4) of Table 3.A.4 indicates no significant increase in the rate of switching into preferring the standard workweek amongst individuals preferring an intermediate workweek three to six years post-reform.

We next estimate in Table 3.A.5 equation (3.6) using as a sample individuals preferring workweek categories above the new standard workweek or below the old standard workweek. Column (1) indicates a significant 8.6 pp increase in the fraction of individuals switching to prefer the new standard workweek in years close to the reform, as well as a 10.3 pp increase in the third to sixth year post-reform. At the same time, column (2) indicates a corresponding decrease in the fraction of individuals switching to prefer the old standard workweek. Column (3) of Table 3.A.5 indicates that individuals preferring workweek categories above the new standard workweek or below the old standard workweek not more likely to switch into preferring a workweek corresponding to the standard workweek category post-reform. Overall, Table 3.A.5 suggests that the changes in desired hours amongst individuals preferring workweeks above the new standard workweek or below the old standard workweek likely reflect baseline switching from preferring non-standard to standard workweeks from one year to the next.

3.6 Discussion

The previous sections show that changes in the standard workweek are accompanied by decreases (increases) in the fraction of individuals preferring the old (new) standard workweek. In this section, we discuss several mechanisms that are consistent with these empirical findings.

3.6.1 Changes in wages

As with the standard workweek, compensation in the public sector is regulated by collective agreements. Individuals are assigned to compensation groups (*Entgeltgruppe*) based on their education level and their job position, and to levels (*Stufe*) based on experience in the public sector. Compensation within each group×level typically consists of a base pay (*Grundgehalt*), extra allowance (*Sonderzuwendung*),

16. Note that the standard workweek differs pre- and post-reform, so that this variable equals 1 if an individual's preferred workweek category corresponds to the old (new) standard workweek pre-reform (post-reform).

Table 3.4. Effect of standard workweek increase on within-individual changes in desired workweek categories

	$hp'_{t+1} \neq hp'_t$		$hp'_{t+1} = hs'_{new}$		$hp'_{t+1} = hs'_{old}$
	(1)	(2)	(3)	(4)	(5)
Panel A: First extensions					
$\text{Post}_{-1\leq l\leq 2}$	0.0387	-0.0046	0.0539	0.0636**	-0.0572*
	(0.0714)	(0.0502)	(0.0552)	(0.0253)	(0.0328)
$\text{Post}_{3\leq l\leq 6}$	-0.1691**	-0.0561	0.0279	0.0998***	-0.1004***
	(0.0737)	(0.0453)	(0.0517)	(0.0289)	(0.0304)
Pre-reform avg.	0.598	0.619	0.115	0.065	0.170
Observations	544	2,466	544	2,466	2,466
Clusters	231	579	231	579	579
Panel B: Second extensions					
$\text{Post}_{-1\leq l\leq 2}$	-0.0122	0.0547	0.0764**	0.0615**	-0.0385
	(0.0779)	(0.0750)	(0.0332)	(0.0241)	(0.0492)
$\text{Post}_{3\leq l\leq 6}$	0.0769	0.0864	0.1420**	0.0619***	-0.0850*
	(0.1133)	(0.0692)	(0.0678)	(0.0228)	(0.0480)
Pre-reform avg.	0.397	0.568	0.016	0.000	0.200
Observations	410	769	410	769	769
Clusters	125	204	125	204	204
Sample (hp'_t)	$= hs'_{old}$	$\neq hs'_{old\|new}$	$= hs'_{old}$	$\neq hs'_{old\|new}$	$\neq hs'_{old\|new}$

Notes: OLS estimates of equation (3.6), with $\mathbb{1}\{hp'_{t+1} \neq hp'_t\}$, $\mathbb{1}\{hp'_{t+1} = hs'_{new}\}$, and $\mathbb{1}\{hp'_{t+1} = hs'_{old}\}$ as dependent variables in columns (1) and (2), (3) and (4), and (5) respectively. Columns (1) and (3) use as a sample observations where the desired workweek category equals the old standard workweek category (38 hours). Columns (2), (4) and (5) use as a sample observations where the desired hours category equals neither the old nor new standard workweek categories. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

possibly holiday allowance (*Urlaubsgeld*) and parental allowance (*Familienzuschlag*). The base pay is negotiated periodically to keep up with inflation, with changes typically applied as a fixed percentage increase across all group×levels, occasionally with absolute minimums so that the total percentage increases of lower-paid groups are higher in a given negotiation round. As this information is publicly available, we are able to check for changes in contractual income, and thus wages, of individuals around the time of the reforms.

Figure 3.8 plots the real hourly wage (base pay divided by monthly standard hours) relative to the month preceding the workweek reduction for a representative group×level.[17] Although decreases in the standard workweek, which were implemented for most civil servants and all public sector employees on 01.04.1989 and 01.04.1990, do not coincide exactly with a negotiation round, large wage increases were implemented in the 01.01.1990 negotiation round. Changes in base pay from the 01.01.1990 negotiation round differ across groups, with larger increases for civil servants of the simple (*einfach*) and middle (*mittler*) services on the one hand, and smaller increases for civil servants in elevated (*gehoben*) and higher (*höher*) services, and public sector employees on the other hand. Following this negotiation round, real hourly wages increase by 15 to 20%, 7.5 to 10%, and 2 to 5% relative to the month before the first workweek reduction for civil servants of the simple service, civil servants of the middle service, and civil servants of elevated and higher service and all public sector employees respectively.

In contrast to the workweek reductions, no large changes in the base pay were implemented close to the workweek extensions. Figures 3.9 and 3.10 plot the percentage change in real hourly wages for a representative civil servant group×level relative to the month before the first workweek extensions in North Rhine-Westphalia, Rhineland Palatinate and Lower Saxony, and Baden Württemberg, Bavaria and Hessia respectively. The increase in the standard workweek is not compensated by increased monthly base pay, so that the hourly wage decreases immediately thereafter and only reverts to pre-extension levels after several years.

How plausible is it that the effect we see is solely due to changes in hourly wages? Table 3.5 estimates equations (3.2) and (3.3) with desired hours as a dependent variable, using only full-time observations. Column (1) indicates a significant 0.56-hour decrease in average desired hours in the first four to five years following the workweek reductions. As for the civil servant workweek extensions, columns (3) and (4) show that average desired hours increase insignificantly five to seven years post-reform, by 0.39 and 0.26 hours respectively.[18] The wage changes implied by

17. Due to small variations in negotiated income increases across payscale groups within the service level, we plot only wage changes of a representative payscale group×level for the simple and middle service levels.

18. Due to the large variation in desired hours, it is not possible to detect with sufficient power the changes that would be implied by the partial shift in preferred hours from the old to the new

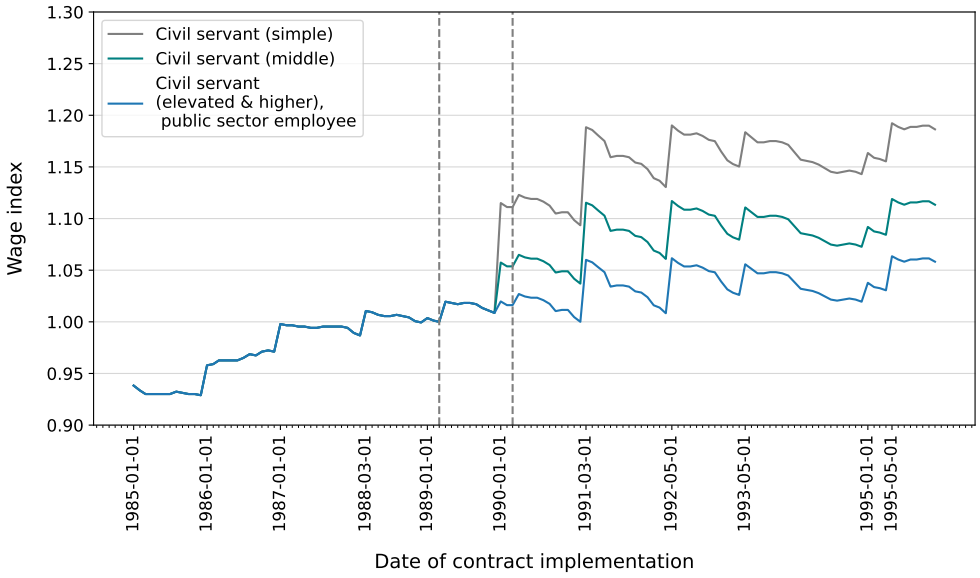


Figure 3.8. Wage changes relative to month preceding workweek decrease

Notes: Each line plots the inflation-adjusted hourly wage of a reference payscale group×level in a negotiation round divided by the respective inflation-adjusted hourly wage of 01.03.1989. The reference payscale group×levels are A2×5 for the simple service, and A7×5 for the middle service. For civil servants of the elevated and higher service, as well as public sector employees, there was no variation in wage changes across payscale group×levels in this time frame. Hourly wage is defined as monthly base pay (*Grundgehalt*) divided by monthly standard hours.

the payscale changes above translate to a negative elasticity of -0.24 in the case of workweek reductions, and an imprecisely estimated negative elasticity of -0.18 in the case of workweek extension.[19]

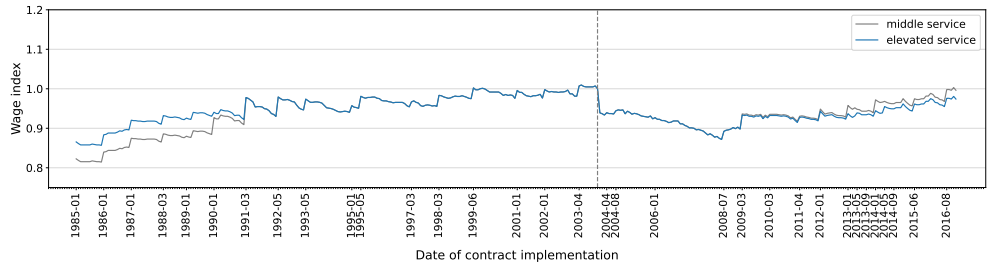
Estimated static or steady-state uncompensated intensive-margin labor supply elasticities tend to be positive and small for men, and positive and large for women.[20]

---

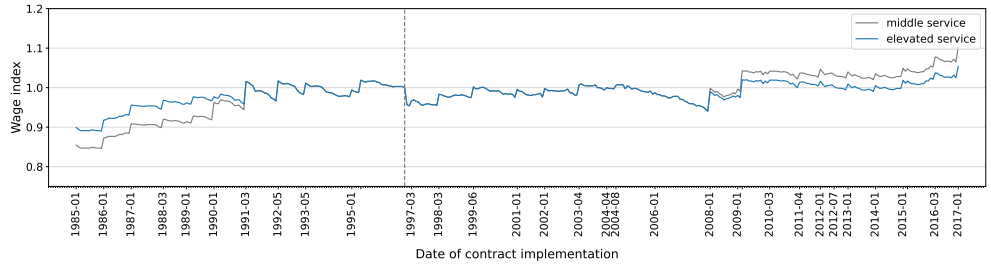
standard workweek we document. The minimum detectable effect for the coefficient on $\text{Post}_{0\leq l\leq 4}$ in the case of workweek reductions with power of 0.8 is -(1.96 + 0.84)*0.27 = -0.76, and that for workweek extensions is (1.96 + 0.84)*0.36 = 1.01. Barring other changes, this would require at least 38% of full-time individuals to switch from a preferred workweek of 40 hours to 38 hours in the case of workweek reductions, and 50% of full-time individuals to switch from a preferred workweek of 38 hours to 40 hours (in states where the workweek was increased from 38.5 to 40 hours).

19. This is computed as the percentage change in average desired hours in the four to five years after the workweek reform, relative to average desired hours in the four to five years pre-reform, divided by the average percentage change in real hourly wage (as implied by the change in negotiated income and standard hours). For workweek reductions, this is calculated as (-0.56/36.10)/(0.065) = -0.24. As for workweek extensions, this is calculated as (0.25/37.85)/(-0.036) =-0.18

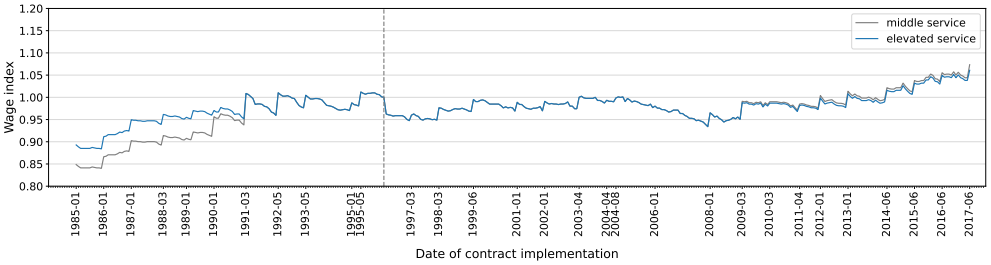
20. See Bargain and Peichl (2016) for an overview. More recent papers suggest larger elasticities: Chetty (2012) shows that frictions such as adjustment costs and inattention can explain the small elasticities obtained by previous studies and estimates a steady-state Hicksian elasticity of 0.33 correcting for these frictions. Keane (2015) estimates a life cycle model with human capital accumulation and obtain larger Hicksian elasticities in response to a permanent wage change.



(a) North Rhine-Westphalia



(b) Rhineland-Palatinate

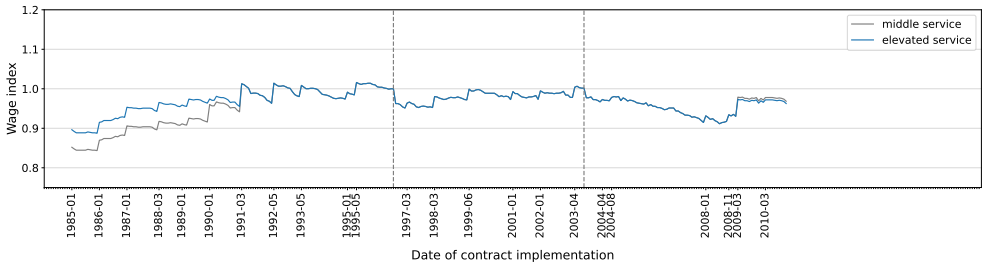


(c) Lower Saxony

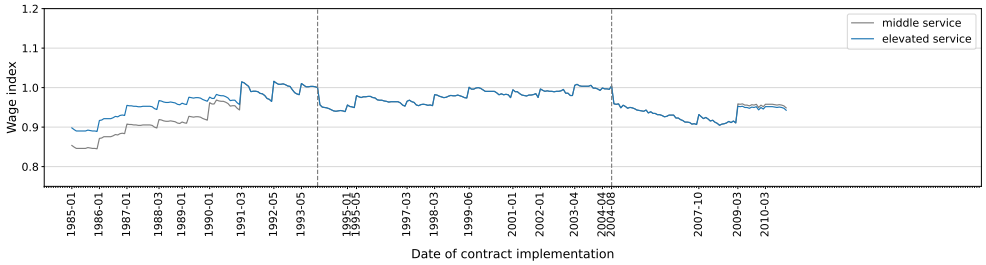
Figure 3.9. Wage changes relative to month preceding workweek increase

Notes: Each line plots the inflation-adjusted hourly wage of a reference payscale group×level in a negotiation round divided by the inflation-adjusted hourly wage of the month before the reform. Contracts are always implemented on the first day of the month. Note that due to space constraints, only the date of the first contract is printed in the figure for contracts two months apart or less. The reference payscale group×level is A2×5 for civil servants from the simple service, A7×5 for civil servants from the middle service, and A12×5 for civil servants from the elevated and higher services. Note that due to a reform implemented in 2017 in North Rhine-Westphalia that re-defines groups and levels, the wage index is only plotted until the end of 2016. Hourly wage is defined as monthly base pay (*Grundgehalt*) divided by monthly standard hours.

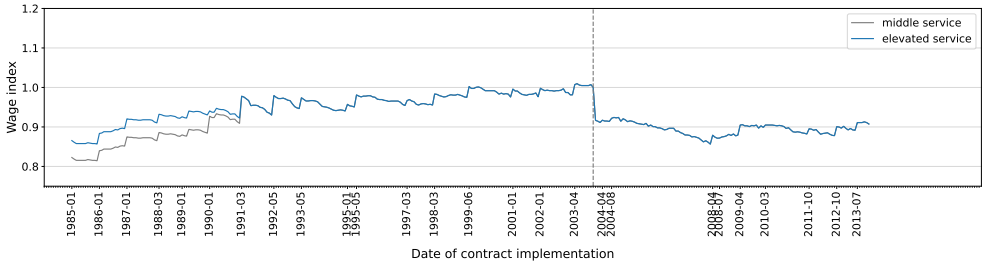
However, a more mixed picture emerges when individuals are allowed to freely choose work hours (as is the case with the survey question on desired hours). Using permanent wage increases in 1996 and 2004 as sources of exogeneous wage changes amongst New York taxi drivers, Ashenfelter, Doran, and Schaller (2010) estimate an uncompensated elasticity of -0.2. Similarly, Motghare (2021) uses a permanent wage increase in 2012 amongst taxi drivers and estimates an elasticity of -0.5. Pencavel (2015) use changes in input and output prices, log and plywood respectively, as sources of exogeneous variation in wages of workers at plywood co-



(a) Baden-Württemberg



(b) Bavaria



(c) Hessia

Figure 3.10. Wage changes relative to month preceding workweek increase

Notes: Each line plots the inflation-adjusted hourly wage of a reference payscale group×level in a negotiation round divided by the inflation-adjusted hourly wage of the month before the reform. The reference payscale group×level is A7×5 for civil servants from the middle service, and A12×5 for civil servants from the elevated service. Note that due to reforms implemented in 2011 in Baden-Württemberg and Bavaria, and in 2014 in Hessia, that re-defined groups and levels, the wage index is only plotted until the end of 2010 in the former states and until the end of 2013 in the latter state. Hourly wage is defined as monthly base pay (*Grundgehalt*) divided by monthly standard hours.

ops and estimate (imprecisely) an uncompensated elasticity ranging between -0.006 and -0.37.

Though our elasticity estimates are in line with those that are purely the result of financial incentives alone, several points suggest that wage changes may not be the sole driver of the preferred hour changes we observe. We note firstly that civil servants and public sector employees are compensated on an monthly basis, so that the hourly wage is less salient for these individuals. Furthermore, although civil servants from different service levels face different wage increases during the workweek reductions, the effect of the reform does not significantly differ between

Table 3.5. Effect of standard workweek changes on average desired hours of full-time employed individuals

	Reductions		Extensions	
	(1)	(2)	(3)	(4)
$Post_{0\leq l\leq 4}$	-0.5625**	-0.5237*	0.2471	0.2055
	(0.2738)	(0.2742)	(0.3599)	(0.3646)
$Post_{5\leq l\leq 7}$			0.3931	0.2643
			(0.4418)	(0.4714)
Pre-reform avg.	36.101	36.101	37.845	37.845
Observations	14,356	14,341	4,204	4,202
Clusters	2,635	2,631	621	621
Sample	All	All	Civil	Civil
Individual FEs	*Yes*	*Yes*	*Yes*	*Yes*
Controls	—	*Yes*	—	*Yes*

Notes: OLS estimates of equations (3.2) and (3.3) with desired hours (hp_t) as a dependent variable. Columns (1) and (2) estimate equation (3.2) using full-time employed civil servants and public sector employees, while columns (3) and (4) estimate equation (3.3) using full-time employed civil servants. Controls are as defined in Table 3.2. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

civil servants of simple and middle service on one hand, and other civil servants and public sector employees on the other hand.[21] Lastly, we note that the changes in the composition of desired hours do not seem to move in line with wage changes. Despite the 5.3% increase in hourly wages from 1985 to 1987 (see Figure 3.8), we do not find changes in the composition of desired workweek categories during this period, as seen from the constant pre-trends in Figure 3.3. Similarly, although base pay increases between 5 and 8% in 2008 and 2009 for civil servants, so that the previous decreases in hourly wages are compensated, Figure 3.6 does not suggest a reversal in desired workweek changes in the later years following the workweek extensions.

3.6.2 Changes in preferences

Other than through wage changes, the changes in desired hours may also be due to altered preferences. Because the standard workweek reforms were accompanied by changes in the actual workweek, work hour preferences may change through habit formation. Evidence of habit formation has been found in various contexts, such

21. We estimate in Table 3.A.3 a variant of (3.2), where $Post_{0\leq l\leq 4,it}$ and $Post_{l\geq 5,it}$ are interacted with binary variables indicating if an individual is a civil servants of simple and middle service, and do not find significant estimates on the interacted variables.

as excercising (Charness and Gneezy, 2009; Royer, Stehr, and Sydnor, 2015; Harris and Kessler, 2019; Carrera et al., 2020), consumption of healthy foods (Loewenstein, Price, and Volpp, 2016), weight loss (Augurzky et al., 2018), showering (Byrne et al., 2022), hygiene (Hussam et al., 2022), voting (Fujiwara, Meng, and Vogl, 2016), and blood donation (Bruhin et al., 2021). Direct evidence of habit formation in the context of labor supply is scarce. However, past studies have incorporated general notions of preference inseparability across periods in estimating life-cycle models. Johnson and Pencavel (1984) use data and wage variation from negative income tax experiments conducted in Seattle and Denver (SIME/DIME) to estimate a life-cycle model where consumption and work hours from the previous period are allowed to influence current consumption and work hours, and find significant estimates on coefficients on past hours and consumption. Using data from the Michigan Panel of Income Dynamics (PSID), Hotz, Kydland, and Sedlacek (1988) estimate the consumption Euler equation of a life-cycle model where current leisure is allowed to depend on past leisure and finds better fit to compared to a standard model that assumes full separability of preferences across time. Lastly, also using PSID data, Bover (1991) estimates labor supply elasticities based on a life cycle model with habit formation in work hours.

Alternatively, because changes in the standard workweek also apply to one's peers, changes in preferences may arise in response to changes in the behavior or preferences of one's peers. This could be due to complementarities in leisure, where leisure spent with friends or spouses yields higher utility than time spent alone. Thus, having longer work hours yields less disutility because one's colleagues or spouse also work longer hours. Georges-Kot, Goux, and Maurin (2017) find that individuals without children time their paid leave to coincide with school holidays in France, while Georges-Kot, Goux, and Maurin (2022) find that self-employed individuals tend to take a day off when their spouse has a paid day off during public holidays. Similarly, Goux, Maurin, and Petrongolo (2014) find that husbands work 0.5 hours less in response to an average two-hour reduction in their wives' working hours, induced by a 4-hour workweek reduction in France. Collewet, de Grip, and de Koning (2017) find a positive correlation between men's working hours and those of their self-reported peers. Lastly, Lalive and Parrotta (2011) find that the retirement of older employees increases the probability of employment exit of younger employees.

At the same time, the standard workweek might form the basis for a social norm, so that changes in the standard workweek alters the perceived socially acceptable number of hours to work. In contrast to leisure complementarities, social norms are typically modelled as disutility from deviating from the actions of others (Grodner and Kniesner, 2006). Evidence on (perceived) social norms affecting labor market behaviour is found mostly in the contexts of female labor supply decision and fairness in compensation. Bertrand, Kamenica, and Pan (2015) and Codazzi, Pero, and Albuquerque Sant'Anna (2018) find that wives in houesholds where they are potentially likely to earn more than their husbands are less likely to participate in the

labor force, and tend work work fewer hours if they do participate in the labor force. Bursztyn, González, and Yanagizawa-Drott (2020) find evidence of misperceived social norms amongst husbands in Saudi Arabia regarding the level of support for women working outside the home, and that correcting these beliefs results in husbands supporting their wife's search for outside work. Boneva, Kaufmann, and Rauh (2021) conduct a survey and find evidence of a social norm in Germany regarding maternal labor supply— most individuals think that their friends and family would prefer that mothers work part-time or not at all. Furthermore, absent constraints on finding suitable full-time chidcare, the perceived opinions of friends and family predict maternal labor supply decisions.

With regards to perceptions of fair wages, Kahneman, Knetsch, and Thaler (1986) explore different circumstances under which an employer's decision to cut wages is considered fair by survey participants. Similarly, Bewley (1999) surveys managers and find that firms are reluctant to cut pay during recessions because this would damage employee morale, which is derived in part from employees' beliefs in the fairness of the firm's actions. Breza, Kaur, and Krishnaswamy (2019) conduct a field experiment in India and find that although workers would like to accept wages below the socially accepted level, they do not do so if this decision is observable by their peers. The above studies mainly find evidence of existing norms— an intriguing question is whether norms can be shaped by external forces. Galbiati et al. (2021) compares the responses of individuals surveyed before the implementation of a lockdown in the UK during the COVID-19 pandemic to those of individuals surveyed afterwards, and find that individuals in the latter group are more likely to believe that others support behaviors aimed at decreasing the spread of the disease, such as staying at home, store closures, and refraining from social gatherings. Similarly, Casoria, Galeotti, and Villeval (2021) find that survey participants rate the behavior of inviting friends over as more socially inappropriate (appropriate) following the implementation (lifting) of COVID-19 rules. Lastly, Lane, Nosenzo, and Sonderegger (2023) find discontinuities in the social appropriateness of various activities such as alcohol consumption, drunk driving, or speeding, in line with cutoffs in legal regulations.

3.7 Robustness Checks

3.7.1 Workweek changes amongst public and private sector employees

The results of the previous sections are limited to civil servants and public sector employees. In this section, we provide further descriptive evidence of similar shifts in preferred hours from the workweek extensions amongst public sector employees, and workweek reductions amongst private sector employees.

As mentioned in Section 3.3.1, the standard workweek of public sector employees differs from October 2005 onwards between the state, municipal, and federal

levels. The SOEP does not contain information on the administrative level at which public sector employees are employed. However, regardless of the administrative level or sector, the standard workweeks of public sector employees from 2006 onwards are between 38.5 and 39.8 hours in most states.[22] Figure 3.A.1 plots the fractions of public sector employees preferring particular workweek categories over time. In line with the workweek extensions, we see that the fraction of individuals preferring a 39-hour workweek category increases gradually by 8 pp between 2005 and 2010, while the fraction of individuals preferring a 38-hour workweek category decreases by 12 pp. On the other hand, the fractions of public sector employees preferring a 40 or 41 to 42 hour workweek category remains constant, which supports the idea that the changes in preferred hours amongst civil servants post-extension are indeed attributable to standard workweek changes amongst this group of individuals.

Collective agreements and standard workweeks also exist in various industries of the private sector, and are typically negotiated between labor and employer unions. However, these agreements do not necessarily apply to all employees within an industry, and the SOEP does not have information on whether an individual is covered by a collective agreement.[23] Additionally, information on the industry an individual is employed in is limited, so that incorrect assignment of standard workweek to individual is likely. Nonetheless, we obtain from Bispinck and Schulten (2017) information on standard workweek changes for certain private sector industries, and assign these to individuals based on NACE Rev 1.1 industry codes.

As in Section 3.3.1, we assess the accuracy of the standard workweek assignment by plotting in Figure 3.A.5 the proportions of individuals with various stated contractual work hour categories in the metal and electrical industry, the chemical industry, retail sector and construction sector. Likely due to the aforementioned issues, contractual hours do not strictly follow standard hours. For the metal and electrical industries, as well as the retail trade sector, the fraction of individuals with a contractual workweek within the new (old) standard workweek category increases (decreases) after the implementation of a new standard workweek. As for the chemical industry, the increase in the fraction of individuals with a contractual workweek category of 37 hours appears to precede the introduction of a 37-hour workweek. Lastly, for the construction sector, the fractions of individuals with contractual workweeks of 38 and 37 hours increase from 1985 and 1990 onwards respectively, despite the 39-hour standard workweek being in place during this time.

22. From Figure 3.1b, we see that the fraction of individuals with a 39-hour contractual workweek category increases by 40 pp between 2005 and 2010, while the fraction of individuals with a 38-hour contractual workweek category decreases by 45 to 50 pp.

23. Whether a collective agreement (between employer and employee unions) applies to an employee depends several factors— whether the firm is part of the employer union, whether the employee is a member of the labor union, and whether a particular collective agreement is declared by the state to be generally applicable.

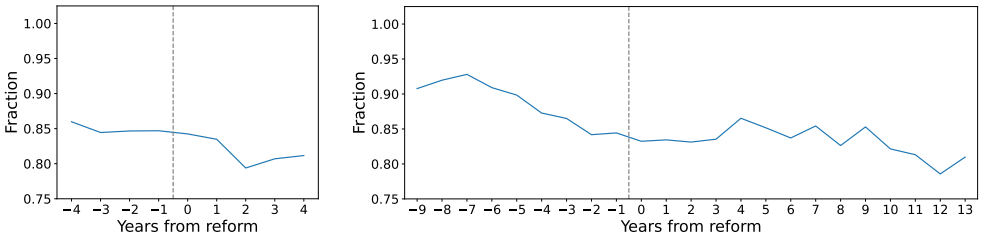
With these caveats in mind, Figure 3.A.6 plots the proportions of full-time employed individuals in the metal and electrical industry, the chemical industry, retail sector and construction sector with various preferred work hour categories over time. Across all industries, the fraction of individuals preferring a 40-hour workweek decreases following the introduction of a non-40-hour workweek—by 25 pp from 1985 to 1990 in the metal and electrical industry, 15 pp from 1986 to 1987 in the retail trade sector, and approximately 20 pp in the chemical and construction industries from 1989 to 1991. However, the decrease in the chemical industry seems to be part of a pre-existing trend.

Evidence of emergence of preferences for the new standard workweek is perhaps the clearest for metal and electrical industries. Between 1986 and 1995, the standard workweek was gradually decreased from 40 hours to 35 hours in five stages—first with the introduction of a 38.5-hour standard workweek in 1986, followed by a 37.5-hour workweek in 1988, a 37-hour workweek in 1989, and 36- and 35-hour workweeks in 1993 and 1995 respectively. With each workweek decrease, the fraction of individuals preferring the new (old) standard workweek increases (decreases) by 10 to 20 pp. As for the retail sector, the fractions of individuals preferring 38- and 37-hour workweek categories following the introduction of 38.5- and 37-hour workweeks increase by 14 pp and 10 pp respectively. We also find a similar increase in the fraction of individuals preferring a 39-hour workweek after the implementation of this workweek in the chemical and construction industries in 1989 and 1990 respectively.

3.7.2 Selection in or out of treatment

In section 3.5, we find larger estimates when the sample is limited to full-time employed individuals and individuals whose stated contractual workweek categories equal the standard workweek category. This could firstly be because only full-time employees are fully subject to the change in standard hours, or that the standard workweek does not apply to certain individuals, such as federal civil servants. However, it is also possible that an individual selects into their preferred contractual workweek by changing their employment status after the implementation of a new standard workweek. For example, it could be that the individuals averse to longer workweeks switch from full- to part-time employment following a workweek extension. We explore this by looking at changes in employment status in the years before and after the reforms.

Figure 3.11a plots the full-time employment rate of civil servants and public sector employees five years before and five years after the first workweek reduction. The decrease in full-time employment, from 85% to 80%, suggests that the slightly larger estimates obtained using the full-time employed sample in Table 3.2 is unlikely to be due to part-time individuals switching into full-time employment. Figure 3.11b plots the full-time employment rate of civil servants nine years before



(a) Workweek decrease **(b)** Workweek increase

Figure 3.11. Full-time employment rates before and after workweek reforms

Notes: Panel (a) plots the fraction of civil servants and public sector employees in full-time employment before and after the introduction of the first workweek reductions (between 1989 and 1991), while panel (b) plots the fraction of civil servants in full-time employment before and after the first workweek increase.

and thirteen years after the introduction of the workweek extension. Though the full-time employment rate decreases from 90% to 85% in the years preceding the workweek extension, it remains relatively stable in the years around the workweek extensions, from $l = -2$ to $l = 9$, before dropping to 80% from $l = 10$ onwards. Thus, it does not appear that the workweek extensions led to an increase in individuals switching out of full-time employment.

It is also possible that individuals selecting out of the new standard workweek are replaced by individuals selecting into the new standard workweek, so that average full-time employment rates remain constant. In this case, switching of employment status (in either direction) should be higher after a reform. Table 3.6 estimates equation (3.6) using a binary variable indicating if an individual changes employment status in the next year, for workweek reductions in columns (1), (2) and (3), and for workweek extensions in columns (4), (5) and (6). Columns (1) and (4), which use the full sample of observations, indicate no significant increases in the fractions of individuals switching employment status following workweek reductions and workweek extensions respectively. Based on columns (2) and (5), we also do not find significant increases in the fractions of full-time individuals switching into part-time employment after either a workweek reduction or extension. Similarly, columns (3) and (6) do not indicate significant increases in the fraction of part-time individuals switching into full-time employment post-reform.

3.8 Conclusion

In this paper, we study the preferred workweek choices of individuals in response to changes in the standard workweek. Following the introduction of a new standard workweek, the fraction of individuals preferring the new (old) standard workweek increases (decreases). This effect is immediate for workweek reductions, and gradual and smaller for workweek extensions. The change in the composition of desired

Table 3.6. Effect of workweek reforms on probability of switching employment status

	Workweek decrease			Workweek increase		
	(1)	(2)	(3)	(4)	(5)	(6)
$Post_{-1\leq l\leq 2}$	0.0133	0.0048	0.0497	0.0106	0.0051	0.0266
	(0.0083)	(0.0060)	(0.0377)	(0.0098)	(0.0075)	(0.0481)
$Post_{3\leq l\leq 6}$	0.0034	0.0021	-0.0156	-0.0058	-0.0027	-0.0167
	(0.0083)	(0.0061)	(0.0338)	(0.0088)	(0.0061)	(0.0460)
Pre-reform avg.	0.033	0.019	0.115	0.020	0.009	0.083
Sample	All	Full-time	Part-time	All civil	Full-time civil	Part-time civil
Observations	17,204	12,066	5,138	4,356	3,667	689
Clusters	3,853	2,782	1,560	738	639	183

Notes: OLS estimates of equation (3.6), with a binary variable indicating if an individual changes employment status in the next year as dependent variable. Column (1) contains estimates using all civil servant and public sector employee observations, while columns (2) and (3) use only full-time employed and part-time employed civil servant and public sector employee observations respectively. Column (4) contains estimates using all civil servant observations, while columns (5) and (6) use only full-time employed and part-time employed civil servant observations respectively. Estimated coefficients on $Pre_{l\leq -4}$ and $Post_{l\geq 8}$ are omitted. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

hours following the workweek decrease is driven by individuals directly switching from preferring the old to the new standard workweek. As for workweek extensions, the effect mainly stems from individuals preferring workweeks than the old standard workweek switching to prefer the new standard workweek post-reform. Because the workweek reforms were not compensated by offsetting changes in income, these findings could potentially be due to individuals optimizing preferred work hours in response to changes in hourly wages. However, we do not find strong evidence of wage changes as the sole driver of the effect we observe. Alternatively, it could also be that the workweek reforms induce a change in preferences through habit formation, or by inducing a change in behavior or preferences of one's peers, or by altering social norms. We are unable to distinguish between these channels and leave this to future research.

Appendix 3.A Supplementary Figures and Tables

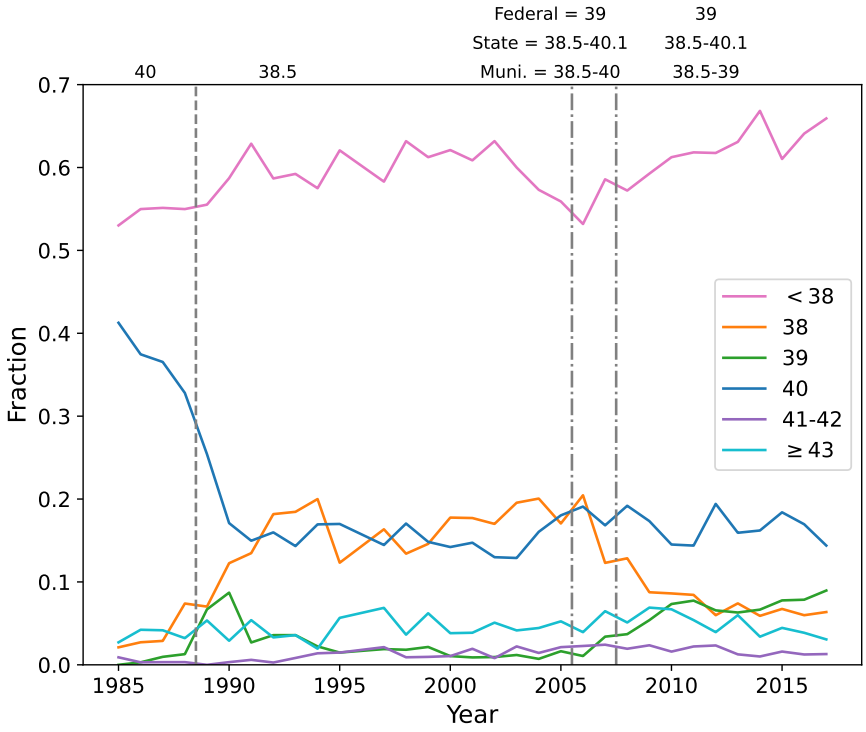
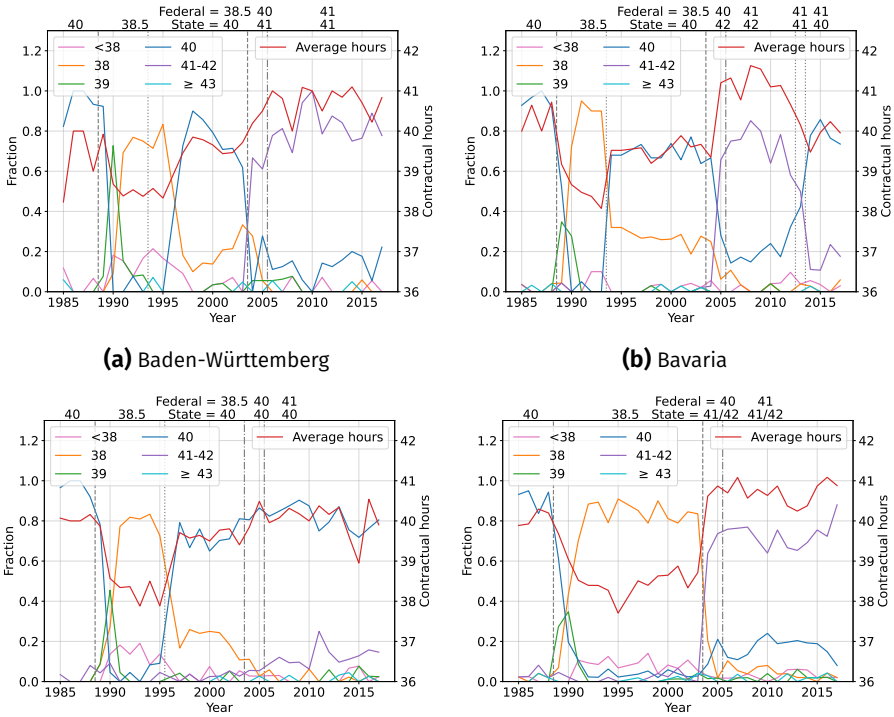


Figure 3.A.1. Desired workweek categories of public sector employees

Notes: Each line plots the fraction of public sector employees preferring a particular workweek category, for the following categories: strictly less than 38 hours, 38 hours, 39 hours, 40 hours, 41 to 42 hours, and 43 hours and above. Dashed vertical lines indicate timing of workweek reforms. Text above the figure denotes the standard workweek in place.



(a) Baden-Württemberg

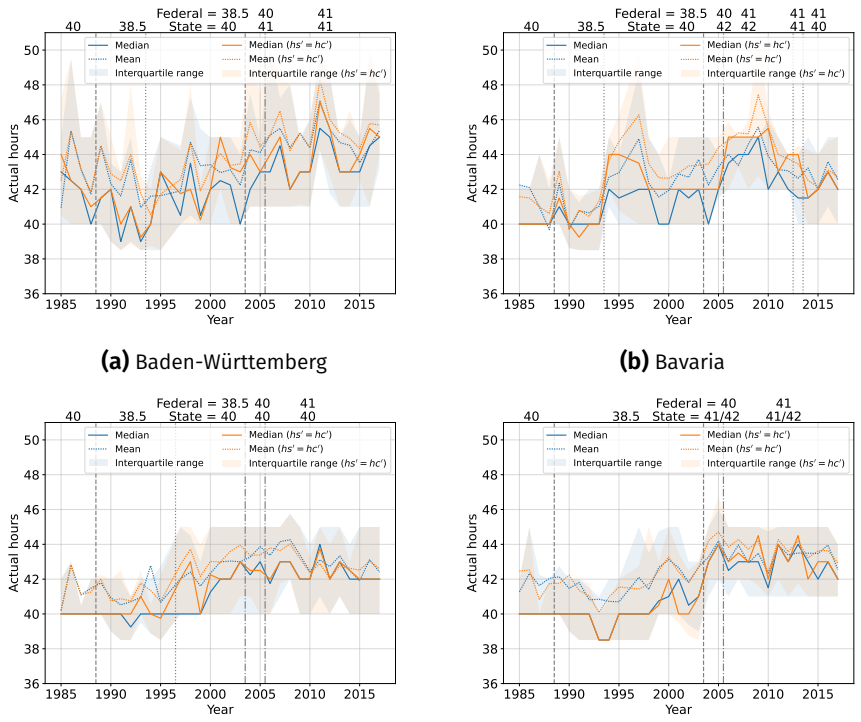
(b) Bavaria

(c) Lower Saxony, Bremen and Rhineland-Palatinate

(d) North Rhine-Westphalia and Hessia

Figure 3.A.2. Contractual workweek categories of full-time civil servants by state

Notes: Each line corresponding to the left axis plots the fractions of individuals whose stated contractual workweek corresponds to a particular workweek category, for the following categories: strictly less than 38 hours, 38 hours, 40 hours, and 41 to 42 hours. The line corresponding to the right axis plots average stated contractual hours. Text above the figure refers to the standard workweek in place for different administrative levels. See Tables 3.A.1 and 3.A.2 for information on standard workweeks.



(a) Baden-Württemberg

(b) Bavaria

(c) Lower Saxony, Bremen and Rhineland-Palatinate

(d) North Rhine-Westphalia and Hessia

Figure 3.A.3. Actual hours statistics of civil servants by state

Notes: Blue lines and shading refers to statistics of stated actual hours of full-time employed civil servants, orange lines and shading refers to statistics of stated actual hours of civil servants whose stated contractual workweek category equals the standard workweek category. Stated actual hours refers to responses to the question "And how many hours do you generally work, including any overtime?". Bold lines refer to the average, dashed lines to the median, and the interval refers to the interquartile range.

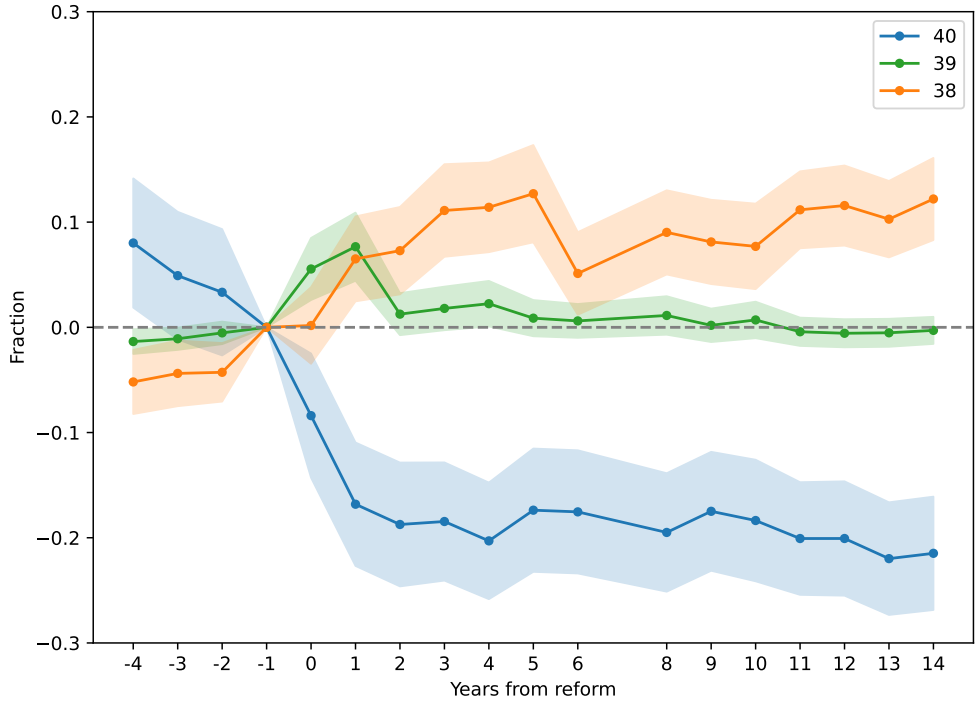
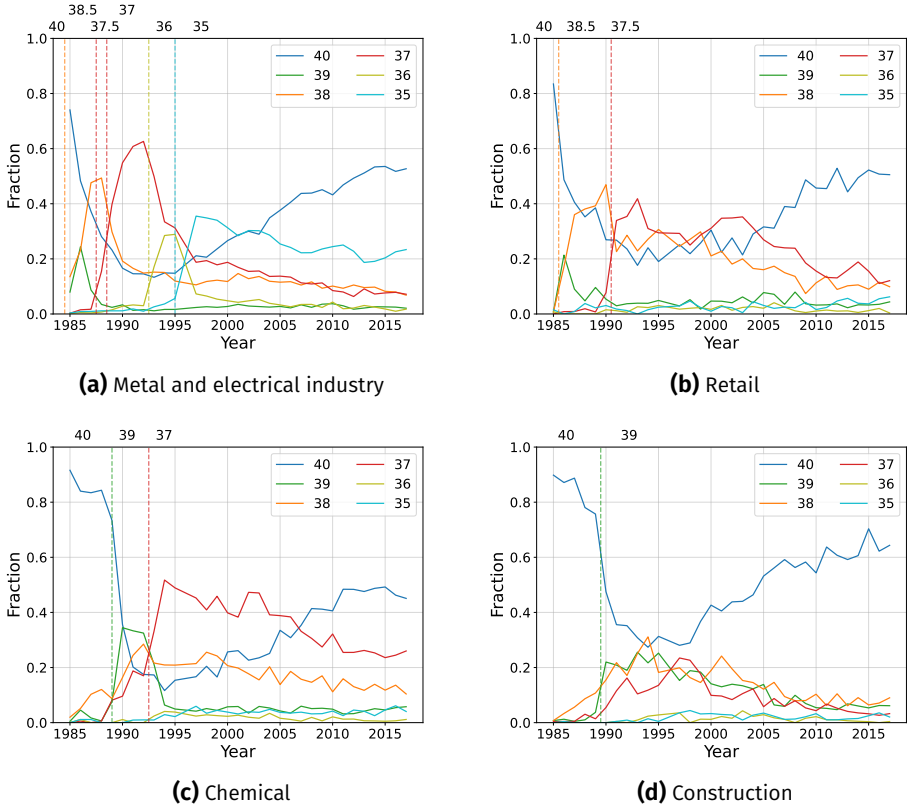


Figure 3.A.4. Effect of standard workweek decrease on probability of preferring various workweek categories

Notes: Estimates of equation (3.1), with $\mathbb{1}\{hp' = 40\}$ (blue line), $\mathbb{1}\{hp' = 39\}$ (green line), and $\mathbb{1}\{hp' = 38\}$ (orange line) as dependent variables, using all public sector employee observations and civil servant observations in North Rhine-Westphalia and Hessia. Bands indicate 95% confidence intervals. Standard errors clustered at individual level.



(a) Metal and electrical industry

(b) Retail

(c) Chemical

(d) Construction

Figure 3.A.5. Contractual workweek categories of full-time private sector employees by industry

Notes: Each line corresponding to the left axis plots the fractions of individuals whose stated contractual workweek corresponds to a particular workweek category, for the following categories: 35 hours, 36 hours, 37 hours, 38 hours, 39 hours, and 40 hours.

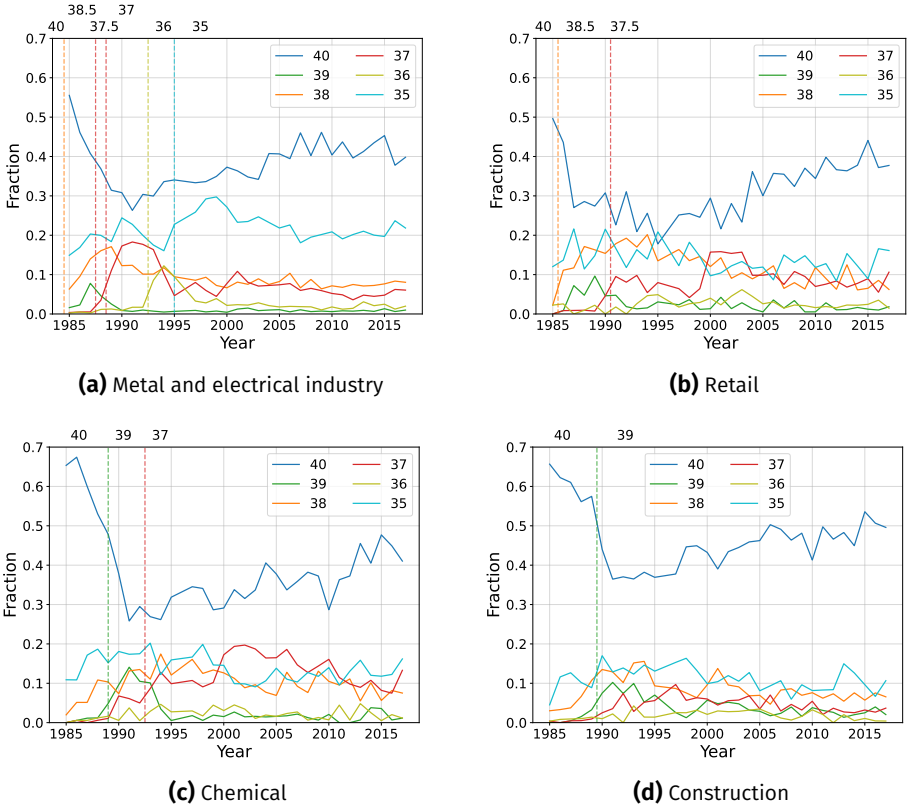


Figure 3.A.6. Preferred workweek categories of full-time private sector employees by industry

Notes: Each line plots the fractions of individuals whose stated preferred workweek category corresponds to a particular workweek category, for the following categories: 35 hours, 36 hours, 37 hours, 38 hours, 39 hours, and 40 hours.

Table 3.A.1. Length of the standard workweek of civil servants

	’85 -’88	’89	’90	’91	’92 -’93	’94	’95	’96	’97	’98 -’00	’01	’02	’03	’04	’05	’06 -’11	’12	’13 -’16	’17
Panel A: Civil servants employed at municipalities and states																			
Baden-W.	40	**39**	**38.5**	38.5	38.5	38.5	38.5	**40**[10]	40	40	40	40	**41**[9]	41	41	41	41	41	41
Bavaria	40	**39**	**38.5**	38.5	38.5	**40**[1]	40	40	40	40	40	40	40	**42**[9]	42	42	**41**[8]	**40**[8]	40
Bremen	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	**40**[4]	40	40	40	40	40	40	40	40	40	40
Hamburg	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	**40**[8]	40	40	40	40	40	40	40
Hesse	40	40	40	**38.5**	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	**42**[1]	42	42	42	42	**41**[8]
Lower Sax.	40	**39**	**38.5**	38.5	38.5	38.5	38.5	**40**[4]	40	40	40	40	40	40	40	40	40	40	40
North Rhine-W.	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	**41**[1]	41	41	41	41	41
Rhineland-P.	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	**40**[1]	40	40	40	40	40	40	40	40	40	40
Saarland	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	38.5	38.5	**40**[1]	40	40	40	40	40	40	40	40
Schleswig-H.	40	40	**38.5**	38.5	38.5	**39.5**[1]	39.5	39.5	39.5	39.5	39.5	**40**[1]	40	40	40	**41**[8]	41	41	41
Panel B: Civil servants employed at federal bodies																			
Federal bodies	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	**40**[10]	40	**41**[3]	41	41	41

Notes: Changes in the standard workweek are depicted in bold font, and are always effective on the first day of the month. The workweek decreases between 1989 and 1991 became effective on April first of the respective year. The numerical superscripts indicate the month in which changes in standard workweeks after 1991 became effective. *Source: Federal Ministry of the Interior*

Table 3.A.2. Length of the standard workweek of public sector employees

	'85-'88	'89	'90-'04	'05	'06	'07	'08	'09	'10	'11-'17
Panel A: Public sector employees employed at states										
Baden-W.	40	**39**	**38.5**	38.5	**39.5**[11]	39.5	39.5	39.5	39.5	39.5
Bavaria	40	**39**	**38.5**	38.5	**40.1**[11]	40.1	40.1	40.1	40.1	40.1
Bremen	40	**39**	**38.5**	38.5	**39.2**[11]	39.2	39.2	39.2	39.2	39.2
Hamburg	40	**39**	**38.5**	38.5	**39**[11]	39	39	39	39	39
Hesse	40	**39**	**38.5**	38.5	38.5	38.5	38.5	38.5	**40**[1]	40
Lower Sax.	40	**39**	**38.5**	38.5	**39.8**[11]	39.8	39.8	39.8	39.8	39.8
North Rhine-W.	40	**39**	**38.5**	38.5	**39.84**[11]	39.84	39.84	39.84	39.84	39.84
Rhineland-P.	40	**39**	**38.5**	38.5	**39**[11]	39	39	39	39	39
Saarland	40	**39**	**38.5**	38.5	**39.5**[11]	39.5	39.5	39.5	39.5	39.5
Schleswig-H.	40	**39**	**38.5**	38.5	**38.7**[11]	38.7	38.7	38.7	38.7	38.7
Panel B: Public sector employees employed at municipalities										
Baden-W.	40	**39**	**38.5**	38.5	**39**[5]	39	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Bavaria	40	**39**	**38.5**	38.5	38.5	38.5	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Bremen	40	**39**	**38.5**	38.5	38.5	38.5	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Hamburg	40	**39**	**38.5**	38.5	**38-40**[4]	40	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Hesse	40	**39**	**38.5**	38.5	**39/38.5**[1]	39/38.5	39/38.5	39/38.5	39/38.5	39/38.5
Lower Sax.	40	**39**	**38.5**	38.5	**38.5-39**[4]	38.5-39	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
North Rhine-W.	40	**39**	**38.5**	38.5	38.5	38.5	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Rhineland-P.	40	**39**	**38.5**	38.5	38.5	38.5	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Saarland	40	**39**	**38.5**	38.5	38.5	38.5	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Schleswig-H.	40	**39**	**38.5**	38.5	**38-39**[1]	38-39	**39/38.5**[7]	39/38.5	39/38.5	39/38.5
Panel C: Public sector employees employed at federal bodies										
Federal bodies	40	**39**	**38.5**	**39**[10]	39	39	39	39	39	39

Notes: Changes in the standard workweek are depicted in bold font, and are always effective on the first day of the month. The workweek decreases in 1989 and 1990 became effective on April first of the respective year. The numerical superscripts indicate the month in which changes in standard workweeks after 1990 became effective. "-" indicates that standard hours may differ by occupational groups, or age, or family status. "/" indicates that the standard workweek differs for hospital staff. *Sources: Ver.di*

Table 3.A.3. Effect of standard workweek decrease on probability of preferring various workweek categories, by civil servant job level.

	$hp' = 38$		$hp' = 40$	
	(1)	(2)	(3)	(4)
Post	0.1611***	0.1393***	-0.1636***	-0.2062***
	(0.0327)	(0.0377)	(0.0510)	(0.0519)
Lower	-0.0015	-0.0288	0.0853	-0.1098
	(0.0165)	(0.0481)	(0.0607)	(0.0748)
Post × Lower	0.0215	-0.0050	-0.0857	0.0427
	(0.0482)	(0.0504)	(0.0665)	(0.0668)
Pre-reform avg.	0.038	0.038	0.509	0.509
Observations	5,335	5,015	5,335	5,015
Clusters	1,036	719	1,036	719
Sample	Civil	Civil	Civil	Civil
Individual FEs	—	*Yes*	—	*Yes*
Controls	—	*Yes*	—	*Yes*

Notes: OLS estimates of a variation of equation (3.2), where the grouped time indicators are interacted with a binary variable (Lower) indicating if an individual from the middle or simple service, using all civil servant observations. Columns (1) and (2) contain estimates from specifications using $\mathbb{1}\{hp' = 38\}$ as the dependent variable, while columns (3) and (4) contain estimates from specifications using $\mathbb{1}\{hp' = 40\}$ as the dependent variable. See section 3.4 for differences in the definitions of l and *Post* when using $\mathbb{1}\{hp' = 38\}$ and $\mathbb{1}\{hp' = 40\}$ as dependent variables. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.A.4. Effect of standard workweek increase on within-individual changes in preferred workweek categories

	$hp'_{t+1} \in [39, hs'_{new})$	$hp'_{t+1} = hs'_{new}$	$hp'_{t+1} = 38$	$hp'_{t+1} = hs'_{t+1}$
	(1)	(2)	(3)	(4)
$\text{Post}_{-1\leq l\leq 2}$	0.0734	0.0096	-0.0321	-0.0847
	(0.0621)	(0.0592)	(0.0706)	(0.0792)
$\text{Post}_{3\leq l\leq 6}$	-0.0380	0.1993**	-0.0409	0.1050
	(0.0371)	(0.0826)	(0.0761)	(0.0860)
Pre-reform avg.	0.105	0.113	0.208	0.208
Sample (hp_t)	$= 38$	$\in [39, hs'_{new})$	$\in [39, hs'_{new})$	$\in [39, hs'_{new})$
Observations	512	464	464	464
Clusters	217	163	163	163

Notes: OLS estimates of equation (3.6), with $\mathbb{1}\{hp'_{t+1} \in [39, hs'_{new})\}$, $\mathbb{1}\{hp'_{t+1} = hs'_{new}\}$, $\mathbb{1}\{hp'_{t+1} = 38\}$, $\mathbb{1}\{hp'_{t+1} = hs'_{t+1}\}$ as dependent variables in columns (1), (2), (3) and (4) respectively. Column (1) uses observations where the preferred workweek category equals the old standard workweek category of 38 hours. Columns (2), (3) and (4) use observations where the preferred workweek category is between 39 hours and the new standard workweek category. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.A.5. Effect of standard workweek increase on within-individual changes in preferred workweek categories

	$hp'_{t+1} = hs'_{new}$	$hp'_{t+1} = 38$	$hp'_{t+1} = hs'_{t+1}$
	(1)	(2)	(3)
$\text{Post}_{-1\leq l\leq 2}$	0.0862***	-0.0702*	-0.0285
	(0.0319)	(0.0379)	(0.0363)
$\text{Post}_{3\leq l\leq 6}$	0.1031***	-0.1187***	-0.0117
	(0.0313)	(0.0346)	(0.0375)
Pre-reform avg.	0.049	0.164	0.164
Sample (hp_t)	$\notin [38, hs'_{new}]$	$\notin [38, hs'_{new}]$	$\notin [38, hs'_{new}]$
Observations	1,797	1,797	1,797
Clusters	487	487	487

Notes: OLS estimates of equation (3.6), with $\mathbb{1}\{hp'_{t+1} = hs'_{new}\}$, $\mathbb{1}\{hp'_{t+1} = hs'_{old}\}$ and $\mathbb{1}\{hp'_{t+1} = hs'_{t+1}\}$ as dependent variables in columns (1), (2) and (3) respectively, using observations where the preferred workweek category is either strictly less than the old standard workweek category of 38 hours, or strictly more than the new standard workweek category. Standard errors in parantheses, clustered at individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

References

Ahn, Taehyun. 2016. "Reduction of Working Time: Does It Lead to a Healthy Lifestyle?" *Health Economics* 25 (8): 969–83. https://doi.org/https://doi.org/10.1002/hec.3198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.3198. [146]

Ashenfelter, Orley, Kirk Doran, and Bruce Schaller. 2010. "A Shred of Credible Evidence on the Long-run Elasticity of Labour Supply." *Economica* 77 (308): 637–50. https://doi.org/10.1111/j.1468-0335.2010.00858.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0335.2010.00858.x. [172]

Augurzky, Boris, Thomas K. Bauer, Arndt R. Reichert, Christoph M. Schmidt, and Harald Tauchmann. 2018. "Habit formation, obesity, and cash rewards" [in eng]. Ruhr Economic Papers 750. Essen. https://doi.org/10.4419/86788871. [175]

Bargain, Olivier, and Andreas Peichl. 2016. "Own-wage labor supply elasticities: variation across time and estimation methods." *IZA Journal of Labor Economics* 5 (1): 1–31. [171]

Behaghel, Luc, and David M. Blau. 2012. "Framing Social Security Reform: Behavioral Responses to Changes in the Full Retirement Age." *American Economic Journal: Economic Policy* 4 (4): 41–67. https://doi.org/10.1257/pol.4.4.41. [146]

Berniell, Ines, and Jan Bietenbeck. 2020. "The effect of working hours on health." *Economics & Human Biology* 39: 100901. https://doi.org/https://doi.org/10.1016/j.ehb.2020.100901. [146]

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender Identity and Relative Income within Households." *Quarterly Journal of Economics* 130 (2): 571–614. https://doi.org/10.1093/qje/qjv001. eprint: https://academic.oup.com/qje/article-pdf/130/2/571/30631743/qjv001.pdf. [175]

Bewley, Truman F. 1999. *Why Wages Don't Fall during a Recession.* Harvard University Press. Accessed August 13, 2023. http://www.jstor.org/stable/j.ctv1pncnkx. [176]

Bispinck, Reinhard, and Thorsten Schulten. 2017. "WSI Arbeitszeitkalender 2017: Tarifdaten aus 25 Wirtschaftszweigen." Working paper. Wirtschafts- und Sozialwissenschaftliches Institut. [177]

Boneva, Teodora, Katja Kaufmann, and Christopher Rauh. 2021. "Maternal Labor Supply: Perceived Returns, Constraints, and Social Norms," https://repec.iza.org/dp14348.pdf. [176]

Bover, Olympia. 1991. "Relaxing Intertemporal Separability: A Rational Habits Model of Labor Supply Estimated from Panel Data." *Journal of Labor Economics* 9 (1): 85–100. http://www.jstor.org/stable/2535115. [175]

Breza, Emily, Supreet Kaur, and Nandita Krishnaswamy. 2019. "Propping up the wage floor: Collective labor supply without unions." Working paper. CEPR Discussion Papers. [176]

Bruhin, Adrian, Lorenz Goette, Simon Haenni, Lingqing Jiang, et al. 2021. "Oops!... I Did It Again: Understanding Mechanisms of Persistence in Prosocial Behavior." Working paper. CEPR Discussion Papers. [175]

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110 (10): 2997–3029. https://doi.org/10.1257/aer.20180975. [176]

Byrne, David P, Lorenz Goette, Leslie A Martin, Amy Miles, Alana Jones, Samuel Schob, Thorsten Staake, and Verena Tiefenbeck. 2022. "The habit forming effects of feedback: Evidence from a large-scale field experiment." *Available at SSRN 3974371.* [175]

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics* 112 (2):

407–41. https://doi.org/10.1162/003355397555244. eprint: https://academic.oup.com/qje/article-pdf/112/2/407/5291730/112-2-407.pdf. [145]

Carrera, Mariana, Heather Royer, Mark Stehr, and Justin Sydnor. 2020. "The Structure of Health Incentives: Evidence from a Field Experiment." *Management Science* 66 (5): 1890–908. https://doi.org/10.1287/mnsc.2018.3271. eprint: https://doi.org/10.1287/mnsc.2018.3271. [175]

Casoria, Fortuna, Fabio Galeotti, and Marie Claire Villeval. 2021. "Perceived social norm and behavior quickly adjusted to legal changes during the COVID-19 pandemic." *Journal of Economic Behavior & Organization* 190: 54–65. https://doi.org/https://doi.org/10.1016/j.jebo.2021.07.030. [176]

Charness, Gary, and Uri Gneezy. 2009. "Incentives to Exercise." *Econometrica* 77 (3): 909–31. Accessed July 4, 2023. http://www.jstor.org/stable/40263846. [175]

Chetty, Raj. 2012. "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* 80 (3): 969–1018. https://doi.org/https://doi.org/10.3982/ECTA9043. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9043. [171]

Codazzi, Karen, Valéria Pero, and André Albuquerque Sant'Anna. 2018. "Social norms and female labor participation in Brazil." *Review of Development Economics* 22 (4): 1513–35. https://doi.org/https://doi.org/10.1111/rode.12515. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rode.12515. [175]

Collewet, Marion, Andries de Grip, and Jaap de Koning. 2017. "Conspicuous work: Peer working time, labour supply, and happiness." *Journal of Behavioral and Experimental Economics* 68: 79–90. https://doi.org/https://doi.org/10.1016/j.socec.2017.04.002. [175]

Crawford, Vincent P., and Juanjuan Meng. 2011. "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income." *American Economic Review* 101 (5): 1912–32. https://doi.org/10.1257/aer.101.5.1912. [146]

Cygan-Rehm, Kamila, and Christoph Wunder. 2018. "Do working hours affect health? Evidence from statutory workweek regulations in Germany." *Labour Economics* 53: 162–71. European Association of Labour Economists 29th annual conference, St.Gallen, Switzerland, 21-23 September 2017, https://doi.org/https://doi.org/10.1016/j.labeco.2018.05.003. [146, 147]

Farber, Henry S. 2005. "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers." *Journal of Political Economy* 113 (1): 46–82. https://doi.org/10.1086/426040. eprint: https://doi.org/10.1086/426040. [145, 146]

Farber, Henry S. 2015. "Why you Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers." *Quarterly Journal of Economics* 130 (4): 1975–2026. https://doi.org/10.1093/qje/qjv026. eprint: https://academic.oup.com/qje/article-pdf/130/4/1975/30637404/qjv026.pdf. [145, 146]

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97 (1): 298–317. https://doi.org/10.1257/aer.97.1.298. [145]

Fujiwara, Thomas, Kyle Meng, and Tom Vogl. 2016. "Habit Formation in Voting: Evidence from Rainy Elections." *American Economic Journal: Applied Economics* 8 (4): 160–88. https://doi.org/10.1257/app.20140533. [175]

Galbiati, Roberto, Emeric Henry, Nicolas Jacquemet, and Max Lobeck. 2021. "How laws affect the perception of norms: Empirical evidence from the lockdown." *PLOS ONE* 16 (9): 1–14. https://doi.org/10.1371/journal.pone.0256624. [176]

Georges-Kot, Simon, Dominique Goux, and Eric Maurin. 2017. "Following the Crowd: Leisure Complementarities beyond the Household." *Journal of Labor Economics* 35 (4): 1061–88. https://doi.org/10.1086/692511. eprint: https://doi.org/10.1086/692511. [175]

Georges-Kot, Simon, Dominique Goux, and Eric Maurin. 2022. "The value of leisure synchronization." [175]

Goux, Dominique, Eric Maurin, and Barbara Petrongolo. 2014. "Worktime Regulations and Spousal Labor Supply." *American Economic Review* 104 (1): 252–76. Accessed July 4, 2023. http://www.jstor.org/stable/42920694. [175]

Grodner, Andrew, and Thomas J. Kniesner. 2006. "Social Interactions in Labor Supply." *Journal of the European Economic Association* 4 (6): 1226–48. https://doi.org/10.1162/JEEA.2006.4.6.1226. eprint: https://academic.oup.com/jeea/article-pdf/4/6/1226/10317538/jeea1226.pdf. [175]

Hamermesh, Daniel S., Daiji Kawaguchi, and Jungmin Lee. 2017. "Does labor legislation benefit workers? Well-being after an hours reduction." *Journal of the Japanese and International Economies* 44: 1–12. https://doi.org/https://doi.org/10.1016/j.jjie.2017.02.003. [146]

Harris, Matthew C., and Lawrence M. Kessler. 2019. "Habit formation and activity persistence: Evidence from gym equipment." *Journal of Economic Behavior & Organization* 166: 688–708. https://doi.org/https://doi.org/10.1016/j.jebo.2019.08.010. [175]

Hotz, V. Joseph, Finn E. Kydland, and Guilherme L. Sedlacek. 1988. "Intertemporal Preferences and Labor Supply." *Econometrica* 56 (2): 335–60. Accessed November 8, 2023. http://www.jstor.org/stable/1911075. [175]

Hussam, Reshmaan, Atonu Rabbani, Giovanni Reggiani, and Natalia Rigol. 2022. "Rational Habit Formation: Experimental Evidence from Handwashing in India." *American Economic Journal: Applied Economics* 14 (1): 1–41. https://doi.org/10.1257/app.20190568. [175]

Johnson, T. R., and J. H. Pencavel. 1984. "Dynamic Hours of Work Functions for Husbands, Wives, and Single Females." *Econometrica* 52 (2): 363–89. http://www.jstor.org/stable/1911494. [175]

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* 76 (4): 728–41. Accessed July 10, 2023. http://www.jstor.org/stable/1806070. [176]

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–91. Accessed January 8, 2024. http://www.jstor.org/stable/1914185. [145]

Keane, Michael P. 2015. "Effects of permanent and transitory tax changes in a life-cycle labor supply model with human capital." *International Economic Review* 56 (2): 485–503. https://doi.org/https://doi.org/10.1111/iere.12112. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12112. [171]

Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–65. Accessed August 27, 2023. http://www.jstor.org/stable/25098823. [146]

Lalive, Rafael, and Pierpaolo Parrotta. 2011. "Coworker interactions in labor supply." Working paper. https://www.researchgate.net/profile/Rafael-Lalive/publication/228535182_Coworker_Interactions_in_Labor_Supply. [175]

Lane, Tom, Daniele Nosenzo, and Silvia Sonderegger. 2023. "Law and Norms: Empirical Evidence." *American Economic Review* 113 (5): 1255–93. https://doi.org/10.1257/aer.20210970. [176]

Lepinteur, Anthony. 2019. "The shorter workweek and worker wellbeing: Evidence from Portugal and France." *Labour Economics* 58: 204–20. https://doi.org/https://doi.org/10.1016/j.labeco.2018.05.010. [146]

Loewenstein, George, Joseph Price, and Kevin Volpp. 2016. "Habit formation in children: Evidence from incentives for healthy eating." *Journal of Health Economics* 45: 47–54. https://doi.org/https://doi.org/10.1016/j.jhealeco.2015.11.004. [175]

Motghare, Swapnil. 2021. "The long-run elasticity of labor supply: New evidence for New York City taxicab drivers." *Labour Economics* 71: 102025. https://doi.org/https://doi.org/10.1016/j.labeco.2021.102025. [172]

Pencavel, John. 2015. "The labor supply of self-employed workers: The choice of working hours in worker co-ops." *Journal of Comparative Economics* 43 (3): 677–89. https://doi.org/https://doi.org/10.1016/j.jce.2014.10.001. [172]

Royer, Heather, Mark Stehr, and Justin Sydnor. 2015. "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company." *American Economic Journal: Applied Economics* 7 (3): 51–84. https://doi.org/10.1257/app.20130327. [175]

Seibold, Arthur. 2021. "Reference Points for Retirement Behavior: Evidence from German Pension Discontinuities." *American Economic Review* 111 (4): 1126–65. https://doi.org/10.1257/aer.20191136. [146]

Thakral, Neil, and Linh T. Tô. 2021. "Daily Labor Supply and Adaptive Reference Points." *American Economic Review* 111 (8): 2417–43. https://doi.org/10.1257/aer.20170768. [146]