Anti-corruption and compliance policies: A behavioural approach

Nikita Grabher-Meyer
nikita.grabhermeyer@gmail.com
Registration number: 100314589

A thesis presented for the degree of
Doctor of Philosophy

School of Economics
University of East Anglia

November 2025

Abstract

Fraud, corruption, and regulatory noncompliance impose significant costs on organizations and society. Across three chapters, this thesis explores how different governance instruments — a legal directive, an integrity training program, and a supplier code of conduct — shape perceptions and, through them, influence ethical conduct and compliance behaviour in public, professional, and private organizations.

The first chapter evaluates the short-term effects of the 2019 EU Whistleblower Protection Directive's transposition into national law on corruption perceptions. Using panel data for 27 EU Member States and their regions from 2010 to 2023, the analysis exploits the gradual timing of transposition in a Difference-in-Differences design to identify early impacts. Results show a modest deterioration in expert-assessed corruption among early adopters, with no clear changes in reporting or actual corruption, but with rising scepticism about enforcement, consistent with greater expert scrutiny.

The second chapter tests whether an integrity training course for Ukrainian law students reduces corrupt behaviour. In a field experiment, students were randomly assigned to receive the training or not. They later participated in a bribery game as intermediaries in a potential corrupt transaction. Some (randomly selected) students were also told that most peers had completed the training. While the training alone had little effect on corrupt behaviour, those receiving the information treatment overestimated peers' integrity and behaved more ethically, aligning with a misperceived social norm.

The third chapter uses a contextualized online experiment simulating a multi-tier supply chain to test whether framing and incentivization of buyers' compliance requests affect first-tier suppliers' behaviour through fairness perceptions. Deterrence sustains high monitoring of sub-suppliers, while collaborative framing reduces strict monitoring, and incentives further erode fairness perceptions and compliance.

Collectively, these chapters show how perceptions shape the translation of rules and expected standards of behaviour into action.

Contents

List of Figures

List of Tables

Acknowledgements

I have never liked running. In fact, I've always tried to stay far away from running of any kind — let alone marathons — until I started this PhD "marathon". This has been a long, exhausting run filled with unexpected detours, unmarked trails, and more than a few steep uphills. Yet, I haven't regretted a single moment. Every step brought a lesson, and for that, I feel deeply grateful to the people I met along the way.

First, I want to sincerely thank my main supervisor, Oana. Thank you for giving me the chance to embark on this journey with you. Thank you for your trust and for believing in me at a moment in my life when I barely believed in myself. You have been one of the most important mentors I've had so far, and if I'm ever lucky enough to mentor someone one day, I hope I can do it with the same care and encouragement you've shown me.

A big thank you also goes to my other "coaches" who supported me along this route. Many thanks to Amrish, my co-supervisor, for introducing me to the strategic thinking of game theory, and to our co-authors Stephanie and Ted for generously sharing your expertise in experimental methods. I'm particularly indebted to Ted for all the time, patience, and incredible expertise invested in helping me through this race. I would also like to thank my other co-authors, Theresa and Azam, for taking me on such an important detour along the way. From applying for (and winning!) a small grant to designing a survey field experiment, I learned more than I ever expected, and it was a joy to share that path with you.

I want to thank all the fellow runners and new friends I met along the way. Thank you Alex, Andrea, Chip, Vincent, Jack, Kensley, and Jan for the moments of laughter and shared struggle. You have each shaped the runner — and the person — I am today, and I am immensely grateful.

My heartfelt thanks also go to the other runners and cheerleaders who supported us all this time: Grace, Valentina, Sabria, Bianca, Keila, Aayushi, and Deanna. Thank you for the chats, the venting, the encouragement, and the moral support during the darkest stretches. I miss our calorie breaks at N33 and Old Post Office Court. You were my energy source when I needed it most.

I'm equally thankful to my old friends cheering from different corners of the world. Thank you Cristina, Claudia, Alessandro, Jo, and João, for your patience and for regularly checking my pulse throughout this long run. And thank you to my university friends in Italy, who may still be wondering what on earth I've been doing all this time!

I am also grateful to the colleagues at the ILO who inspired me so deeply that I eventually found the courage to take on this marathon. First, my sincere thanks go to my first supervisor and mentor, Moazam. Thank you for introducing me to the world of research and policymaking, and for trusting that I could find my place in it. A special thank you to Arianna, whose passion, determination, and career choices inspired my own ten years ago. Thank you, Pawel, for your impressive determination in running your own PhD marathon alongside a full-time job. You made it look so easy that I thought I might try as well. It was a steep uphill, but worth every step. And thank you to Kidist, Roopa, and all my colleagues at Better Work for your encouragement and patience, even when my response to your Monday morning "Anything fun this weekend?" stayed the same for two years: "Still running my marathon...".

Finally, and most importantly, I want to thank my parents, to whom I dedicate this work. You have always been — and will always be — my greatest inspiration, even if you may not quite believe it. Your hard work and perseverance, especially during the most difficult times, have continually motivated me to keep going, even when the path felt steep. Thank you for encouraging me to follow my dreams, even when doing so may have caused you a few worries along the way. Thank you for being the most wonderful parents I could have wished for.

Introduction

Fraud, corruption, and regulatory noncompliance impose significant costs on organisations and society. From large-scale corporate tax evasion, through bribery facilitated by legal intermediaries, to factory disasters where social audits failed to detect risks, unethical and noncompliant practices take many forms yet carry severe economic and social costs. Governance instruments such as legal directives, integrity training programmes, and corporate codes of conduct are key to maintaining social order, but their existence alone rarely ensures ethical conduct or compliance more broadly.

Van Rooij and Sokol (2021) conceptualise compliance as the interaction between rules (R) and behaviour (B). Adopting a broad perspective, they define rules as formal laws as well as private and informal norms that guide conduct in professional and social settings. Likewise, their framework encompasses a wide variety of deviant behaviours ranging from fraud and bribery, to street crime, traffic violations, and human rights abuses.

In a similar vein, our view of compliance is broad and ex ante — that is, it focuses on how rules shape future behaviour rather than, as traditionally done, on how laws respond to past misconduct (Van Rooij and Sokol, 2021). In other words, we are interested in understanding how individuals, organisations, and societies interpret rules and adjust their behaviour accordingly. However, our work expands Van Rooij and Sokol (2021)'s behavioural approach to compliance by introducing perceptions (P) as a key link between rules and behaviour ($R \leftrightarrow P \leftrightarrow B$). In this work, we understand compliance as a belief-driven process in which perceptions of enforcement credibility, collective norms, and fairness mediate how rules are translated into action. Across three chapters, we investigate how different governance instruments — a legal directive, an integrity training programme, and a

supplier code of conduct — shape these perceptions and, through them, influence behavioural responses in public, professional, and private organizations.

The first chapter evaluates the short-term effects of the 2019 EU Whistleblower Protection Directive's transposition into national law on corruption perceptions. These perceptions can provide an early barometer of whether such reforms are viewed as credible and capable of encouraging reporting and deterring wrongdoing within societies. Using panel data for 27 EU Member States and 210 regions from 2010 to 2023, the analysis exploits the gradual timing of transposition in a Difference-in-Differences design to identify early impacts. Results show no significant changes in citizen-based perceptions of corruption, or in composite measures of corruption perceptions aggregating household, firm, and expert-based sources, but a modest deterioration in expert-assessed corruption among early adopters. Exploratory evidence indicates no change in reporting behaviour or actual corruption, but a rise in scepticism about enforcement, consistent with heightened expert scrutiny.

The second chapter investigates whether integrity training can reduce corrupt behaviour among intermediaries operating in a culture of widespread rule-breaking. In collaboration with the USAID New Justice Program, we delivered an integrity training course to undergraduate law students in Ukraine. In a field experiment, students were randomly assigned either to receive the training or to a control group. Afterwards, all participants took part in a bribery game, acting as intermediaries in a potential corrupt transaction. Some (randomly selected) students also received information indicating that most of their peers had participated in the training. Results show that the training alone did not significantly reduce corrupt behaviour. However, students informed that most peers had completed the training — regardless of whether they themselves received it — overestimated its effectiveness and thus behaved more ethically to conform to a misperceived social norm.

The third chapter employs a contextualised online experiment simulating a multi-tier supply chain to examine how the framing and incentivization of buyers' social compliance requests shape first-tier suppliers' effort allocation across production, their own compliance, and monitoring of sub-suppliers. We compare a baseline deterrence approach with two treatments: deterrence combined with a collaborative "leading by example" framing, and deterrence combined with collaboration plus a

conditional financial incentive. Results show that deterrence sustained high monitoring of sub-suppliers, whereas collaborative framing reduced strict monitoring, and incentives further eroded fairness perceptions and compliance, highlighting potential risks associated with hybrid governance approaches.

Taken together, the three chapters contribute to the anti-corruption and compliance literature in two main ways. First, they contribute to the growing integration of behavioural public policy and compliance theory by showing how perceptions mediate the translation of rules and expected standards of behaviour into action. Although perceptions have long been recognised as important precursors of compliant behaviour (Cialdini and Goldstein, 2004; Rosenson, 2009; Tankard and Paluck, 2016; Banuri, 2021), this work strengthens the case for their strategic consideration in the design of governance tools. It highlights the need to assess the signals that rules and standards send — particularly whether they are perceived as credibly enforced, collectively shared, and fair — before and during implementation. Communication, information, and framing are therefore not peripheral to compliance but core instruments of governance that policymakers and regulatory actors should harness to enhance the effectiveness of compliance systems.

Second, this work advances a methodological contribution. It demonstrates how experimental and quasi-experimental methods can enrich compliance policy learning at different stages of institutional and organisational reform. From assessing institutional credibility through quasi-experimental evaluation using secondary data, to observing behavioural change as interventions unfold in the field, and testing compliance incentives ex ante through contextualised online experiments, this work shows how a behavioural toolkit comprising a range of quantitative methodologies can help policymakers evaluate and possibly anticipate perceptual and behavioural responses to anti-corruption and compliance initiatives.

Chapter 1

Signalling integrity? Early effects of the EU Whistleblower Protection Directive on corruption perceptions

We conduct the first evaluation of the short-term impacts of the 2019 EU Whistleblower Protection Directive transposition into national law on corruption perceptions. Using panel data for 27 EU Member States and 210 regions from 2010 to 2023, we exploit the gradual timing of the Directive's transposition in a Difference-in-Differences framework to identify potential early impacts. Results show no significant changes in citizen-based perceptions of corruption (European Quality of Government Index), or in composite measures of corruption perceptions aggregating household, firm, and expert-based sources (World Governance Indicators), but a slight deterioration in expert-assessed corruption (V-Dem) among early adopters. Exploratory evidence shows no changes in reporting behaviour or actual corruption but points to rising scepticism about enforcement, consistent with heightened expert scrutiny.

[0]This chapter is joint work with Oana Borcan.

1.1 Introduction

Corruption imposes significant costs on organizations and society. It is commonly understood as the abuse of entrusted power for private gain (Transparency International, 2025). Such abuses can occur in both public and private organizations and may take many forms, including bribery, misuse of public resources, procurement irregularities, fraud, and other forms of unlawful activities. Employees working for or closely with organizations are often the first to detect wrongdoing and are uniquely positioned to report it. According to the Association of Certified Fraud Examiners (2024), 43% of fraud cases are uncovered through whistleblower tips, which is more than three times the share detected by internal audits. Despite their critical role in exposing misconduct, few employees come forward due to fear of retaliation from co-workers or management (Dyck et al., 2010; Transparency International Ireland, 2025). In response, governments backed by civil society have adopted whistleblower protection laws to encourage reporting. However, empirical evidence on their effectiveness in increasing detection and reducing corruption remains limited.

To address this gap, this study investigates the short-term effects of transposing the EU Whistleblower Protection Directive into national law on corruption perceptions across EU Member States. Adopted in December 2019, the Directive required all 27 countries to establish confidential reporting channels, ensure proper investigation of reported misconduct, and protect whistleblowers from retaliation (European Parliament and Council of the European Union, 2019). While members had until December 2021 to transpose the Directive, transposition and subsequent legal entry into force were staggered between 2021 and 2024 (European Commission, 2024).[1] We exploit this variation in timing to assess whether countries that transposed the Directive earlier (early adopters) experienced changes in corruption perceptions compared to those that transposed it later (late adopters).

Understanding how such reforms may influence corruption perceptions — and, in turn, behaviour — requires unpacking the mechanisms through which whistleblower protection laws operate. By protecting whistleblowers from retaliation, these poli-

[1] As of June 2024, all EU members had adopted the Directive, with 5 countries transposing it in 2021, 9 in 2022, 11 in 2023, and the last 2 in 2024. Official transposition dates did not always coincide with actual enforcement: in several cases, legal provisions entered into force between 0-6 months after formal transposition. Our analysis relies on the dates of transposition.

cies aim to encourage individuals to report wrongdoing, thereby strengthening the detection, investigation, and prosecution of legal breaches. As the expected probability of detection and punishment increases, potential wrongdoers are deterred from engaging in corruption (Krügel and Uhl, 2023). However, such behavioural effects are unlikely to materialize shortly after legal transposition. Moreover, corruption is inherently difficult to observe – becoming visible only when detected or reported. As a result, the initial effects of such reforms are expected to operate through perceptions. The introduction and early implementation of measures such as confidential reporting channels and investigative procedures generate visible institutional signals that shape individuals' expectations about the reporting environment, including the ease and safety of reporting and the likelihood that misconduct will be investigated and sanctioned. These expectations about enforcement effectiveness in turn translate into broader perceptions of corruption control. When perceived as credible, such signals can enhance confidence in corruption control; when seen as symbolic or weakly implemented, they may instead generate scepticism. Hence, perceptions of control of corruption can provide an early barometer of whether institutional reforms are interpreted as credible signals of greater enforcement effectiveness and government commitment to integrity and accountability (Reinders Folmer, 2021).

We use country- and regional-level panel data from 2010 to 2023 covering 27 EU Member States and 210 regions. We focus on perceptions of public sector corruption, as this is one of the key domains targeted by the Directive and is also the dimension most consistently captured by comparable cross-country indicators. Corruption perceptions are measured through three complementary indicators that differ in their level of aggregation and sources of information, capturing perspectives from citizens and experts and allowing us to account for how these groups access, interpret, and respond to institutional and legal changes. At the regional level, the corruption pillar of the European Quality of Government Index (EQI) captures citizens' perceived and experienced corruption in public service delivery, providing granular subnational variation. Higher standardized scores indicate lower corruption (or higher corruption control) (Quality of Government Institute, 2024). To complement this measure with broader temporal and country-level coverage, we use the control of corruption indicator from the World Bank's Worldwide Governance Indicators (WGI), a composite measure of corruption perceptions aggregating

household surveys, firm surveys, and expert assessments of the extent to which public power is used for private gain. Higher scores indicate better corruption control (World Bank, 2024b). Finally, we draw on the public sector corruption index from the Varieties of Democracy project (V-Dem Institute) as an expert-based measure at the country level, capturing assessments of the extent to which public officials engage in bribery or theft. Higher scores reflect higher levels of corruption (or lower corruption control) (V-Dem Institute, 2024).

We estimate a standard two-by-two Difference-in-Differences (DiD) model comparing early adopters (i.e., countries that transposed the Directive in 2021 or early 2022) to late adopters transposing it between late 2022 and 2024. We define treatment this way to allow for a one- to two-year window after transposition for any changes in perceptions to take place. Results show no significant changes in citizen-based perceptions of corruption (EQI) or in composite measures of corruption perceptions aggregating household, firm, and expert-based sources (WGI), but a slight deterioration in expert-only assessments (V-Dem) among early adopters. This apparent deterioration likely reflects increased scrutiny following transposition rather than an actual decline in integrity, consistent with the absence of short-term effects on corruption or economic crime rates in complementary objective data. Effects on expert perceptions remain stable across dynamic specifications and do not extend to other governance indicators. Visual DiD diagnostics confirm parallel pre-trends between treatment and a control group that includes EU candidate countries to increase power and comparability.

Additional exploratory evidence shows no change in actual reporting behaviour. While respondents in early-adopting countries perceived the burden of proof as a lesser barrier to reporting, they were also more likely to view the lack of punishment as a key obstacle. This rise in scepticism about enforcement may help explain why reporting did not increase in the short term. Heterogeneity tests confirm that the observed transposition effects were primarily captured through expert perceptions, which likely adjust faster than citizens' views. Deterioration was stronger in countries with lower reporting activities, where concerns about impunity were higher, corruption more pervasive and socially tolerated, and trust in government weaker. The decline was also more marked in EU-engaged countries, suggesting that experts in such contexts may have reacted more critically to early implementation

gaps. Taken together, these patterns point to two plausible interpretations: weak enforcement credibility in low-trust settings and a potential "expectations gap" in pro-EU contexts.

There is limited empirical evidence on whether institutional reforms aimed at curbing corruption translate into measurable changes in perceptions of corruption. Dynamic panel analysis of 82 countries using WGI and ICRG indicators shows that greater fiscal transparency improves perceived control of corruption (Montes and Luna, 2020), while quasi-experimental evidence from Ukraine finds that traffic police reforms improved citizens' perceptions of integrity in reformed institutions (Pop-Eleches and Robertson, 2024). Yet other studies point to a paradox: reforms designed to strengthen integrity may initially heighten perceptions of corruption by increasing scrutiny and media exposure. In the United States, stronger campaign finance regulations were associated with higher perceived corruption by journalists due to intensified coverage of unethical behaviour (Rosenson, 2009). Survey experiments in China similarly show that scandal-oriented anti-corruption reporting can worsen public perceptions by exposing more corruption (Sun et al., 2022).

This study extends this literature to a supranationally mandated reform to protect whistleblowers in advanced democracies, providing the first quasi-experimental evaluation of the short-term impacts of the 2019 EU Whistleblower Protection Directive on perceptions of corruption. Our findings are consistent with the paradox identified in previous studies, showing short-term deterioration in expert-assessed corruption but stable citizen perceptions. In addition, the paper contributes to the debate on perception-based indicators (Arndt and Oman, 2006; Olken, 2009; Charron, 2015) by highlighting the analytical value of expert perceptions as early diagnostics of institutional credibility, especially where observable behavioural change lags behind legal reform.

This paper also advances the literature on whistleblower protection. While theoretical and experimental research has examined the role and design of mechanisms that encourage whistleblowing (Abbink et al., 2014; Choo et al., 2019; Butler et al., 2019; Mechtenberg et al., 2020; Banuri, 2021; Krügel and Uhl, 2023), empirical evidence on their effectiveness in increasing reporting or reducing misconduct remains limited. This is largely due to measurement and identification challenges, including the inherent invisibility of undetected wrongdoing. Recent studies in sectoral

or single-country settings — such as corporate governance and tax enforcement in the United States and Israel — provide rare evidence of deterrence effects, showing that whistleblower programs can reduce tax evasion and misreporting (Dyck et al., 2010; Wilde, 2017; Amir et al., 2018; Johannesen and Stolper, 2021; Lee et al., 2024). This paper extends this literature by providing the first multi-country, multi-region quasi-experimental evaluation of the short-term effects of a multi-sectoral reform to protect whistleblowers on corruption perceptions — an important precursor of individuals' willingness to report wrongdoing or refrain from misconduct.

Finally, this study further contributes to understanding how institutional credibility and trust condition the effects of whistleblower protection. Experimental evidence highlights that credible enforcement is crucial: when sanctions are uncertain or symbolic, whistleblowing systems fail to deter wrongdoing and can even backfire by crowding out intrinsic compliance (Krügel and Uhl, 2023). Credible protection also increases reporting, but improvements in detection and deterrence may not materialize if prosecutors become less inclined to investigate when protection is in place (Mechtenberg et al., 2020). Broader institutional trust has also been shown to shape attitudes toward reporting and perceptions of corruption. Survey and experimental studies in the United States and Armenia show that citizens with greater trust in government express stronger support for whistleblowing (Antinyan et al., 2020), while distrust amplifies perceptions of corruption (Wroe et al., 2013; Li and Meng, 2020). Consistent with this evidence, our heterogeneity analysis suggests that the Directive's perceptual effects depend on pre-existing perceptions of enforcement and institutional trust. When these are weak, reforms may increase scepticism rather than reassurance. In more EU-engaged contexts, high expectations can instead amplify disappointment when implementation lags behind legal commitments.

The rest of the paper is structured as follows. Section 1.2 provides background on the EU Whistleblower Protection Directive. Section 1.3 outlines the conceptual framework, the data sources, and empirical strategy. Section 1.4 presents the main results and robustness checks. Section 1.5 discusses the broader implications of the findings and directions for future research.

1.2 The EU Whistleblower Protection Directive

Until recently, whistleblower protection in the European Union was either absent or uneven across Member States and policy areas. In the early 2010s, only about 40% of Member States had dedicated whistleblower protection regulations. Existing laws differed widely in scope, application, and the level of protection offered. In most cases, legal frameworks were sectoral rather than horizontal — that is, they applied only to specific policy areas or categories of workers, rather than providing comprehensive protection across both the public and private sectors (Andreis, 2019).[2]

In response to this fragmented legal landscape, advocacy from civil society and repeated calls from the European Parliament for a horizontal whistleblower protection law, the EU adopted the Whistleblower Protection Directive in 2019. The Directive sets minimum standards for reporting breaches of Union law and protecting those who report them. Key provisions require the establishment of secure reporting channels, confidentiality safeguards, follow-up procedures, and protection against retaliation. It applies broadly across areas such as public procurement, financial services, anti-money laundering, food safety, transport safety, consumer protection, environmental protection, and public health (European Parliament and Council of the European Union, 2019).

Member States were required to transpose the Directive into national law by December 2021. To support implementation, the European Commission established an Expert Group to assist national authorities. Transposition, however, occurred unevenly. Only 5 of the 27 Member States met the 2021 deadline, prompting the European Commission to initiate infringement procedures (EU Whistleblowing Monitor, 2024; European Commission, 2024). We use information from the EU Whistleblowing Monitor – cross-checked with complementary web searches and official national sources – to document the timing of transposition for our empirical analysis.[3]

[2]Only a few Member States had adopted horizontal whistleblower protection laws (i.e., comprehensive frameworks covering both the public and private sectors) before the Directive's adoption. These include Malta (2013), Hungary (2013), Ireland (2014), Slovakia (2014), Sweden (2016), France (2016), Netherlands (2016), Italy (2017), Lithuania (2019), Latvia (2019), Croatia (2019).

[3]The EU Whistleblowing Monitor, established by the Whistleblowing International Network (WIN) and Transparency International (TI) Ireland, tracks each Member State's progress in trans-

In July 2024, the European Commission released its first assessment of the transposition. While all countries had formally adopted the Directive by then, the report identified major delays and inconsistencies. Key shortcomings were found in the areas of protection conditions, liability exemptions, and penalties, which risk undermining the Directive's objectives (European Commission, 2024). Independent assessments by civil society organisations reached similar conclusions. Transparency International's 2023 review found that 19 out of 20 Member States did not fully comply with the Directive's requirements in at least one of four key areas: the right to report directly to the authorities, access to remedies and full compensation for damage suffered, free and easily accessible advice, and effective penalties for those violating whistleblower protections (Transparency International, 2023). Together, these findings highlight a persistent gap between formal transposition and effective protection in practice.

1.3 Methodology

1.3.1 Conceptual framework

Whistleblower protection laws are designed to reduce the personal and professional risks faced by individuals who report misconduct, such as job loss, blacklisting, workplace ostracism, and psychological harm (Schmolke and Utikal, 2025). By protecting whistleblowers from retaliation, these policies aim to encourage individuals to report wrongdoing, thereby strengthening the detection, investigation, and prosecution of legal breaches. In turn, more credible detection and enforcement mechanisms are expected to deter potential wrongdoing, as misconduct becomes less likely to go undetected (Krügel and Uhl, 2023).

While these mechanisms ultimately operate through behavioural change, such effects are unlikely to materialize shortly after legal transposition. Moreover, corruption is inherently difficult to observe — becoming visible only when detected or reported. As a result, the initial effects of whistleblower protection laws are expected to operate through perceptions. Specifically, we conceptualize the trans-

posing the Directive through its various legislative stages. Today, it continues to provide updates as countries refine and implement their laws: https://whistleblowingmonitor.eu/.

position of the EU Whistleblower Protection Directive as triggering a chain of perceptual updates. First, legal transposition, together with early implementation of key requirements (e.g., confidential reporting channels in the workplace, designated investigative procedures, etc.), generates visible institutional signals. Second, individuals observe these signals and update their expectations about the reporting environment, including the ease and safety of reporting wrongdoing and the likelihood that reports will be followed by investigation and sanction. Third, these updated assessments of enforcement effectiveness in turn translate into broader perceptions of corruption control.

The direction of this effect depends on the perceived credibility of the signal. When reforms are seen as credible and effectively implemented, they can improve perceptions of corruption control by increasing confidence that misconduct will be detected and sanctioned. However, when reforms are perceived as symbolic, weakly enforced, or inconsistent with prior institutional performance, they may fail to shift expectations or may even generate scepticism. In such cases, increased attention to corruption — arising both from the reform itself, which may signal a previously under-addressed problem, and from greater media exposure and public scrutiny following its introduction — can increase the visibility of misconduct. Combined with unmet expectations regarding enforcement, this may lead to a deterioration in perceived corruption control in the short term.

The credibility of these signals is likely to depend on pre-existing institutional conditions, including baseline perceptions of corruption, perceived enforcement capacity, and trust in government. Where institutional quality is higher, signals of reform may be more readily interpreted as credible; where it is lower, they may be discounted or interpreted as "toothless".

Perceptions of corruption may also differ across types of observers, particularly between citizens and experts. On one hand, citizens likely rely more on indirect signals, personal experiences, and broader societal narratives, which may lead to slower or more heterogeneous adjustments in perceptions. On the other hand, experts with professional or academic expertise in political and institutional processes are likely to have greater access to information about legal reforms and their implementation. As a result, experts may update their perceptions more rapidly in response to institutional changes, as they can more readily interpret the credibility

and implications of new legal frameworks.

Based on this, we derive our main testable hypotheses:

- H1: The transposition of the EU Whistleblower Protection Directive affects perceptions of corruption control in the short term by signaling changes in the reporting environment and expected enforcement, with stronger or earlier effects expected in expert-based assessments, given their greater exposure to information about legal reforms and their implementation.
- H2: The direction and magnitude of this effect depend on baseline institutional conditions such as pre-existing perceptions of corruption, reporting environment, levels of trust in government, and perceived institutional capacity.

To better understand the mechanisms through which perceptions might adjust, we conduct exploratory analysis of perception-based indicators related to reporting, such as perceived barriers (e.g., lack of reporting channels, burden of proof, lack of punishment, lack of protection) and awareness of reporting channels. These indicators help identify whether the Directive has begun to shift expectations about the feasibility and safety of reporting wrongdoing.

Finally, we examine objective behavioural outcomes, including corruption reporting and corruption incidence. Given the short post-transposition period, no measurable behavioural effects are expected at this stage; rather, these outcomes provide a reference point to assess whether perceptual shifts are accompanied by early signs of behavioural change.

1.3.2 Data

Our analysis combines several complementary datasets, summarized in Table 1.A.1 in the Appendix. We begin with regional data from the corruption pillar of the European Quality of Government Index (EQI), which captures citizens' perceptions and experiences of corruption in public service delivery (Quality of Government Institute, 2024). The EQI, developed by the Quality of Government Institute at the University of Gothenburg, is based on citizen surveys conducted in five waves (2010, 2013, 2017, 2021, and 2024) at the NUTS2 regional level across the European

Union.[4] This measure provides granular subnational variation and allows us to capture how citizen-based corruption perceptions evolve within countries. Other EQI dimensions — quality of public services such as education, healthcare, and police, and impartiality in their access — are used in robustness checks. Scores are standardized as z-scores and centred on national governance values (from the WGI) to preserve subnational variation. Higher scores indicate greater government quality and lower perceived or experienced corruption (Charron et al., 2024).[5]

To complement this measure with annual data and broader country-level comparability, we use the control of corruption indicator from the World Bank's Worldwide Governance Indicators (WGI), which measures the extent to which public power is used for private gain, including both petty and grand corruption as well as state capture by elites and private interests (World Bank, 2024b). The WGI covers 214 economies from 1996 to 2023 and aggregates information from over 30 data sources reflecting the views and experiences of citizens, entrepreneurs, and experts across the public, private, and NGO sectors. Measurements from different data sources are first rescaled to a 0-1 range and then combined using an unobserved components model (UCM), which adjusts for differences in scale and measurement error across sources to produce a weighted average (Kaufmann and Kraay, 2024). By aggregating information across many sources and assigning greater weight to those that are more strongly correlated, this approach reduces noise and improves precision, but also tends to smooth short-term variation, as changes are only reflected when they are consistently observed across multiple inputs. Other WGI dimensions — government effectiveness, political stability, rule of law, regulatory quality, and voice and accountability — are used in our robustness checks. Scores are standardized in units of a standard normal distribution (mean = 0, SD = 1, approximately -2.5 to 2.5), with higher values indicating stronger governance quality and control of corruption.[6]

Finally, to capture expert-only assessments of corruption, we use the public sec-

[4]Fieldwork for the 2024 round took place between September 2023 and March 2024. For consistency across country- and regional-level analyses, we treat this wave as 2023.

[5]More details on the methodology can be found in Charron et al. (2024). The data can be downloaded here: https://www.gu.se/en/quality-government/qog-data/data-downloads/european-quality-of-government-index.

[6]More details on the methodology can be found in Kaufmann and Kraay (2024). The WGI dataset can be downloaded here: https://www.worldbank.org/en/publication/worldwide-governance-indicators.

tor corruption index from the Varieties of Democracy dataset (V-Dem Institute), which assesses the extent to which public officials engage in bribery, theft, and misuse of public office for private gain (V-Dem Institute, 2024). The V-Dem dataset provides annual expert-based assessments of governance and democracy for over 200 countries from 1789 to the present. These assessments are produced by a global network of country experts, including academics and practitioners with specialized knowledge of national political and institutional contexts, who provide independent ratings for each country-year. For subjective indicators such as corruption, V-Dem typically aggregates ratings from multiple experts per country using a Bayesian measurement model that corrects for coder bias and enhances cross-country comparability (Coppedge et al., 2019). Compared to the WGI, which aggregates information from multiple sources and tends to smooth year-on-year variation, V-Dem relies solely on expert judgments and may therefore be more sensitive to recent institutional and legal changes. Other V-Dem dimensions — Electoral, Liberal, Participatory, Deliberative, and Egalitarian Democracy — are used in robustness checks. Unlike these other V-Dem indicators, which typically range from normatively worse (0) to better (1), the public sector corruption index is scaled such that higher values indicate greater perceived corruption.[7]

To capture more objective and behavioural dimensions of corruption and reporting — alongside the perceptions that shape them — we incorporate additional data sources described in Table 1.A.1. Data from the United Nations Office on Drugs and Crime (UNODC) provide annual national statistics on reported corruption and economic offences, including bribery, fraud, and other corruption-related crimes (United Nations Office on Drugs and Crime, 2024).[8] To examine reporting behaviour and related perceptions, we use five waves of the Eurobarometer public opinion surveys (2013, 2017, 2019, 2022, 2023), aggregated at the country level (Eurobarometer, 2024).[9] These provide measures of corruption reporting, knowledge

[7]The V-Dem indicators report median estimates from a Bayesian measurement model on a continuous latent scale (approximately -5 to +5, centred around 0). While the scale resembles a standardized z-score, it is not normalized (mean = 0, SD = 1). More details on the methodology can be found in Coppedge et al. (2019). The V-Dem dataset can be downloaded here: https://v-dem.net/data/the-v-dem-dataset/.

[8]This data is compiled from member state submissions, national surveys, and scientific studies, and standardized for comparability. The dataset and more details on the methodology can be found here: https://dataunodc.un.org/dp-crime-corruption-offences.

[9]Eurobarometer is a series of multi-topic, pan-European surveys undertaken for the European

of reporting channels, and perceived barriers such as unclear reporting channels, difficulty proving wrongdoing, lack of punishment, and lack of protection. The Eurobarometer data also include measures of perceived and experienced corruption, corruption tolerance, and the perceived effectiveness of national anti-corruption and prosecution efforts.

For the exploratory heterogeneity analysis, we use baseline indicators of institutional trust in government from the OECD (2018 and the 2019-21 average) and trust in the judiciary, public administration, and the written press from the Eurobarometer (2018) (OECD, 2024; Eurobarometer, 2024).[10] To capture differences in European integration and political engagement, we include Eurobarometer (2018) baseline indicators of trust in the EU, perceptions of the Union's direction, understanding of EU decision-making, and political interest, complemented by author-compiled measures of OECD membership and EU membership maturity (Eurobarometer, 2024).

Country-level time-variant (yearly) control variables were drawn primarily from the World Bank's World Development Indicators (WDI) (World Bank, 2024a). These include standard measures of economic structure, labour market composition, and political participation. Specifically, we control for economic development and structural characteristics (GDP per capita, total population, rural population share, agricultural value added), labour market conditions (employment, unemployment, and female employment rates, including part-time employment), and human capital (share of labour force with intermediate education and population with upper secondary education, disaggregated by gender). To account for institutional and political factors potentially correlated with both transposition timing and corruption perceptions, we include women's representation in national parliaments and the Participatory Democracy Index from V-Dem. Comparable regional-level controls were obtained from Eurostat to ensure consistency across specifications (Eurostat,

Commission since 1970, covering attitudes towards European integration, policies, institutions, social conditions, health, culture, the economy, citizenship, security, information technology, the environment and other topics. Standard and Special Eurobarometer surveys consist of regular face-to-face interviews with approximately 1,000 subjects in the EU member states. The dataset and more details on the methodology can be found here: https://europa.eu/eurobarometer/screen/home.

[10]OECD data, and particularly on trust in government, can be found here: https://www.oecd.org/en/data/indicators/trust-in-government.html.

2024).[11]

Finally, to construct the treatment variable, we compiled information on the timing of each Member State's transposition of the Directive, distinguishing between early and late adopters. Transposition dates were obtained from the EU Whistleblowing Monitor (2024) and cross-checked with multiple public sources, including the European Commission (2024), Transparency International (2023), and media reports.

1.3.3 Empirical strategy

While all EU Member States were expected to transpose the Directive by December 2021, the actual timing of transposition varied substantially across countries. Table 1.1 reports the dates of transposition into national law, which we use to construct our treatment variable, as well as information on whether countries had already enacted horizontal whistleblower protection laws before 2021.

We classify as early adopters the ten countries that transposed the Directive in 2021 or during the first half of 2022 (up to July), allowing for roughly one to two years of exposure before the 2023 data round.[12] Countries that transposed the Directive after July 2022 (late 2022-2024) constitute the main control group.

For robustness and to increase power, we also include eight EU candidate countries that were not legally required to transpose the Directive but were expected to align their anti-corruption frameworks as part of the EU accession process. As shown in Table 1.1, many of these EU candidates had already introduced some form of horizontal whistleblower protection laws prior to 2021.

We exploit this variation in transposition timing and estimate treatment effects using a canonical 2x2 Difference-in-Differences (DiD) framework. A DiD approach is appropriate in this context because variation in the timing of transposition generates differential exposure to the reform across countries over time, allowing us to compare changes in outcomes between early adopters and countries not yet exposed (Angrist

[11] WDI data can be found here: https://datatopics.worldbank.org/world-development-indicators/. Eurostat data can be found here: https://ec.europa.eu/eurostat/data/database.

[12] Because the underlying corruption perception indicators combine citizen-based surveys and expert assessments gathered at different points throughout each year, countries that transposed the Directive in the second half of 2022 (mainly after July) were likely not fully exposed to the reform for an entire year by the time 2023 data were compiled.

Table 1.1: Timing of transposition of the EU Whistleblower Protection Directive

Country	N. of NUTS2 regions	Date of transposition	Early adopter	Year of adoption of pre-existing horizontal whistleblower protection law
Denmark	5	24/06/21	1	
Sweden	8	29/09/21	1	2016
Malta	1	15/12/21	1	2013
Lithuania	2	16/12/21	1	2019
Portugal	7	20/12/21	1	
Cyprus	1	20/01/22	1	
Latvia	1	20/01/22	1	2019
France	27	21/03/22	1	2016
Croatia	4	15/04/22	1	2019
Ireland	3	21/07/22	1	2014
Greece	13	11/11/22	0	
Belgium	3	16/12/22	0	
Romania	8	16/12/22	0	
Finland	5	20/12/22	0	
Netherlands	12	24/01/23	0	2016
Slovenia	2	27/01/23	0	
Austria	9	01/02/23	0	
Bulgaria	6	02/02/23	0	
Spain	17	21/02/23	0	
Italy	21	30/03/23	0	2017
Hungary	8	11/04/23	0	2013
Germany	16	12/05/23	0	
Luxembourg	1	16/05/23	0	
Slovakia	4	01/06/23	0	2014
Czech Republic	8	07/06/23	0	
Estonia	1	15/05/24	0	
Poland	17	14/06/24	0	
Albania	n/a	n/a	0	2017
Bosnia and Herzegovina	n/a	n/a	0	
Georgia	n/a	n/a	0	
Moldova	n/a	n/a	0	2018
North Macedonia	n/a	n/a	0	2016
Montenegro	n/a	n/a	0	2016
Serbia	n/a	n/a	0	2014
Türkiye	n/a	n/a	0	

Note: The table reports, for each EU Member State (27) and its NUTS-2 regions (210), the transposition date, treatment status (early adopters = 1; late adopters = 0), and the year of adoption of any pre-existing horizontal whistleblower protection law. The EU candidate countries (8) in the control group are also listed to illustrate prior adoption of horizontal whistleblower protection laws.

and Pischke, 2009). Given that our last post-treatment year is 2023 and that changes in perceptions are likely to emerge with a lag, we consider late adopters (who transposed the Directive after July 2022) as effectively untreated in 2023. This setting gives us just enough time to capture any short-term or transitional effects for early adopters.

While recent advances in the literature have highlighted the advantages of staggered Difference-in-Differences estimators in settings with heterogeneous treatment timing (Baker et al., 2025), their application in our context is constrained by the

limited number of countries and the uneven distribution of treatment cohorts. In particular, most cohorts contain a relatively small number of countries, and our post-treatment period extends only to 2023, providing limited post-treatment observations for several groups. These features result in imprecise and potentially noisy cohort-specific estimates. For this reason, we adopt a parsimonious 2×2 DiD framework that compares early and late adopters, allowing us to maximize statistical power and focus on short-term average treatment effect for countries with meaningful exposure to the Directive by 2023.

We estimate the average treatment effect on the treated at time t (ATT_t) using the following specification:

$$\text{Corruption Perception}_{it} = \alpha_i + \gamma_t + \beta\,\text{DiD}_{it} + \delta\,\text{post21}_t + \lambda\,\text{early adopters}_i + \theta\boldsymbol{X}_{it} + \varepsilon_{it} \tag{1.1}$$

$\text{Corruption Perception}_{it}$ denotes the outcome (extracted from the WGI, EQI, or V-Dem) for country or region i at time t, observed from 2010 to 2023 (with yearly frequency for countries, and approximately 3-yearly frequency for regions).[13] The same specification is applied to our additional outcomes: reported corruption and economic offences (UNODC) as well as reporting behaviour and related perceptions (Eurobarometer).

α_i represents country (or region) fixed effects which account for time-invariant unobserved heterogeneity across jurisdictions. γ_t represents time fixed effects which control for common shocks affecting all countries. DiD_{it} is the treatment variable defined as the interaction between a post-2021 indicator and a dummy for early adopter countries. It equals 1 for years after 2021 in countries that implemented the Directive early, and 0 otherwise. The coefficient β captures the DiD estimate of the Directive's average treatment effect.

The vector $\boldsymbol{X}_{it}$ includes time-varying control variables at the country and regional level (described in more detail in Table 1.A.1). At the country level, controls include: a time-varying dummy that switches to one in the year a horizontal whistleblower protection law was adopted (before 2021), capturing prior institu-

[13]The 2010-2023 timeline was selected to ensure consistency between country- and regional-level analyses. Regional data from the European Quality of Government Index (EQI) are available only from 2010 onward, constraining the common observation period across datasets.

tional preparedness for the Directive; the logarithm of GDP per capita (PPP, constant 2021 international dollars); the share of agriculture, forestry, and fishing in GDP; the rural population share; total and female employment rates (ages 15+); the share of female part-time employment; the proportion of parliamentary seats held by women; the share of the labour force with intermediate education; and the V-Dem degree of participatory democracy. At the regional level, controls include: the logarithm of regional population; regional unemployment rates (total and female); the share of the regional population (total and female) with at least upper secondary education; and the logarithm of regional GDP per capita (PPP, constant 2020 international dollars). The inclusion of these variables accounts for differences in economic structure, labour market composition, educational attainment, and institutional inclusiveness, which may jointly influence both corruption perceptions and the pace of institutional adoption. The error terms ε_{it} are clustered at country level, where variation in treatment timing occurs.

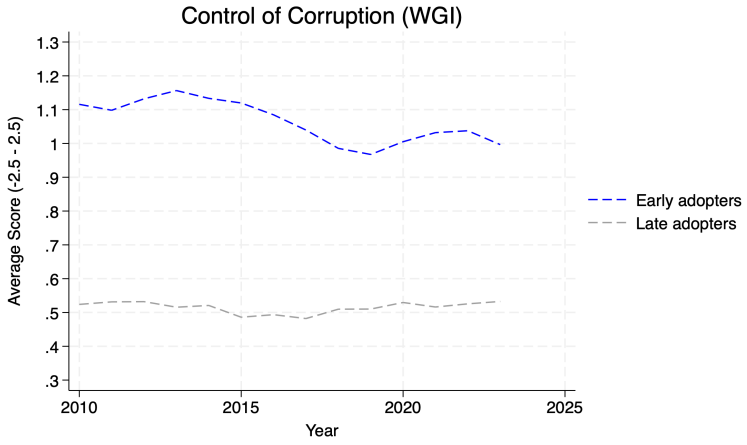
We assume no systematic anticipation effects at the national level prior to the Directive's enforcement in December 2021. Although the Directive was formally adopted in December 2019, it became legally binding only upon transposition into national law. While preparatory discussions may have occurred earlier, particularly among experts, the practical implications for enforcement and institutional behaviour were minimal before transposition. Moreover, public awareness of EU legislative initiatives is generally low. According to Eurobarometer baseline data (2018), only 13.7% of respondents across countries in our sample frequently discussed European political matters with friends, suggesting limited diffusion of information that could have influenced perceptions or behaviour prior to national implementation.[14]

In addition, our identification strategy relies on the parallel trend assumption, which requires that, absent treatment, treated and comparison groups would have followed similar trends in corruption perceptions. Visual inspection of pre-treatment trends suggests broadly parallel trajectories between early and late adopters when EU candidates are included in the control group (Figure 1.1). This pattern weakens

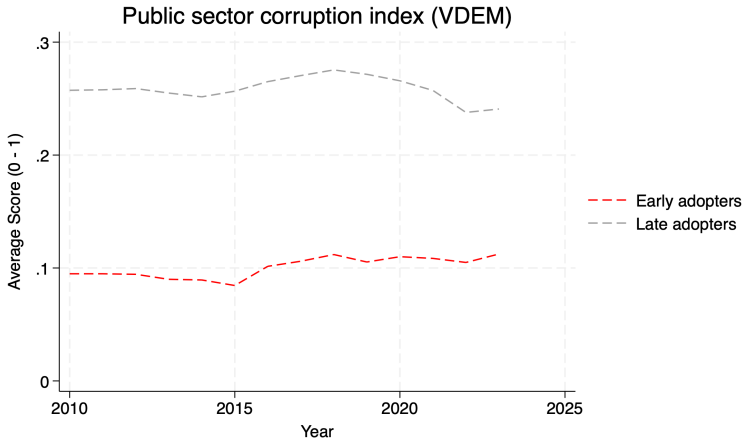
[14]The Eurobarometer sample covers 32 countries: 27 EU Member States and 5 EU candidates. Excluding the candidates, 13.2% of respondents reported frequently discussing EU politics with friends.

when restricting the sample to EU Member States only (Figure 1.A.1).

Figure 1.1: Outcome trends in control of corruption (WGI) and public sector corruption index (V-Dem) when EU members and candidates are included in the control group



(a) WGI



(b) V-Dem

Covariate balance tests (Table 1.A.2) show no systematic pre-treatment differences between early and late adopters when restricting the sample to EU Member States. However, early adopters were substantially more likely to have pre-existing national horizontal whistleblower protection laws (70% versus 23.5%, $p < 0.05$), suggesting underlying differences in institutional capacity, legal preparedness, or political commitment to transparency and accountability. These patterns indicate

that early transposition was not random, but likely reflects pre-existing institutional and normative conditions that made compliance with the Directive less costly or more politically salient.

When EU candidate countries are included in the control group, this gap narrows but remains weakly significant ($p < 0.1$), while a significant difference emerges in participatory democracy ($p < 0.05$). At the regional level, treated and control groups also differ in the share of the population with upper secondary education.

While such differences raise concerns about selection into early adoption, they are likely to reflect relatively stable country characteristics that affect baseline levels of corruption perceptions rather than their short-term evolution. Our empirical strategy accounts for these factors by including country fixed effects, which absorb time-invariant institutional and political differences, and time-varying covariates capturing observable changes in economic and institutional conditions. Moreover, early adopters do not appear to be driven by a common shock or synchronized reform dynamic that could generate systematically different pre-treatment trends. As such, it is plausible that these differences do not induce differential trends in corruption perceptions prior to treatment, an assumption that is supported by the similarity of observed pre-treatment trajectories. Under this assumption, the DiD framework remains valid. To support this, we include all covariates in the main specifications and conduct additional statistical and graphical tests of the parallel trends assumption.

1.4 Main results

We begin by examining regional-level evidence based on the corruption pillar of the European Quality of Government Index (EQI), which provides granular measures of citizens' perceived and experienced corruption in public service delivery. Table 1.2 reports DiD estimates for the EQI corruption dimension and its sub-indices. All models include country and year fixed effects, the full set of time-varying covariates listed in Table 1.A.2, and cluster-robust standard errors at the country level. Models (2) and (3) include time-varying regional controls too. Given the limited number of clusters, we also report p-values based on wild cluster bootstrap procedures with 20,000 replications. Control means correspond to the average outcome for control

Table 1.2: Standard DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on regional-level corruption perceptions (EQI)

Variable	**EQI Corruption**			**EQI Corruption Perceptions**		**EQI Corruption Experience**	
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(1)**	**(2)**
Early transposition (DiD)	-0.042	0.024	-0.033	0.088	0.013	0.226	0.109
	[0.093]	[0.093]	[0.113]	[0.111]	[0.098]	[0.230]	[0.211]
Post transposition of EU	0.333*	0.225	0.385*	0.549	0.486	1.062**	0.991*
Directive	[0.188]	[0.190]	[0.217]	[0.431]	[0.452]	[0.407]	[0.507]
Observations	1,050	933	571	617	559	617	559
R-squared	0.109	0.143	0.213	0.278	0.335	0.481	0.538
Control mean (2021)	-3.3e-5	-3.3e-5	-3.3e-5	7.1e-8	7.1e-8	8.1e-8	8.1e-8
Number of regions	210	202	198	210	198	210	198
EU candidates included	NO	NO	NO	NO	NO	NO	NO
Country controls	YES	YES	YES	YES	YES	YES	YES
Regional controls	NO	YES	YES	NO	YES	NO	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	NO	NO
Clust. P-value	0.654	0.801	0.770	0.432	0.900	0.335	0.610
Boot. Std. Error	0.0933	0.0933	0.113	0.111	0.0982	0.230	0.211
Boot. P-value (DiD)	0.718	0.826	0.847	0.422	0.891	0.418	0.664

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on regional-level perceptions of corruption, using the EQI indices. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) and (3) include time-varying regional controls too. Standard errors (in brackets) are clustered by country. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

countries in 2021, the last pre-treatment year. Across all specifications, we find no statistically significant effects of early transposition on regional-level corruption perceptions, suggesting no detectable short-term changes in citizens' perceptions or experiences of corruption.

To complement this analysis with annual country-level data and broader temporal coverage, Table 1.3 presents corresponding DiD estimates using the control of corruption indicator from the World Bank's Worldwide Governance Indicators (WGI) and the public sector corruption index from the Varieties of Democracy (V-Dem) dataset. Model (1) restricts the sample to EU Member States, while Models (2) and (3) extend the control group to include EU candidate countries. Model (3) further adds two lags of the dependent variable to account for potential serial correlation. Consistent with the EQI results, we find no statistically significant effects on corruption perceptions as captured by the WGI, a composite indicator combining household, firm, and expert-based sources. In contrast, estimates based on V-Dem indicate a small, positive, and statistically significant effect on public sector corruption, suggesting a modest deterioration in expert-assessed corruption

Table 1.3: Standard DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level corruption perceptions (WGI & V-Dem)

Variable	**Control of corruption index (WGI)**			**Public sector corruption index (V-Dem)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
Early transposition (DiD)	-0.038	-0.020	-0.006	0.029**	0.026**	0.016*
	[0.063]	[0.058]	[0.020]	[0.012]	[0.013]	[0.008]
Post transposition of EU Directive	0.183	0.198*	0.116*	-0.051	-0.103***	-0.040***
	[0.112]	[0.109]	[0.062]	[0.032]	[0.033]	[0.013]
Observations	378	468	403	378	468	403
R-squared	0.362	0.300	0.722	0.397	0.416	0.694
Control mean (2021)	0.947	0.664	0.664	0.125	0.215	0.215
Number of countries	27	35	35	27	35	35
EU candidates included	NO	YES	YES	NO	YES	YES
Country controls	YES	YES	YES	YES	YES	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	YES
Clust. P-value	0.546	0.736	0.765	0.0302**	0.0498**	0.0680*
Boot. Std. Error	0.0626	0.0580	0.0204	0.0125	0.0128	0.00821
Boot. P-value (DiD)	0.623	0.783	0.764	0.0466**	0.0578*	0.0739*

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption, using the WGI and V-Dem indices. All models include country and year fixed effects and the full list of time-varying country controls. Standard errors (in brackets) are clustered by country. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

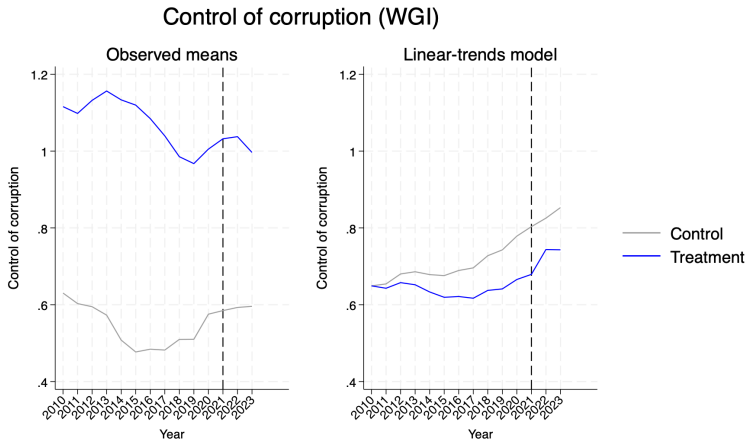
perceptions among early adopters.

Robustness checks using alternative governance indicators from EQI, WGI, and V-Dem (Tables 1.A.4, 1.A.5, 1.A.3) as placebo outcomes confirm the absence of treatment effects. Additional visual DiD diagnostics indicate that pre-treatment trends are broadly parallel when EU candidate countries are included in the control group (Figure 1.2), while there is some heterogeneity in pre-trends among EU members (Figure 1.A.2).

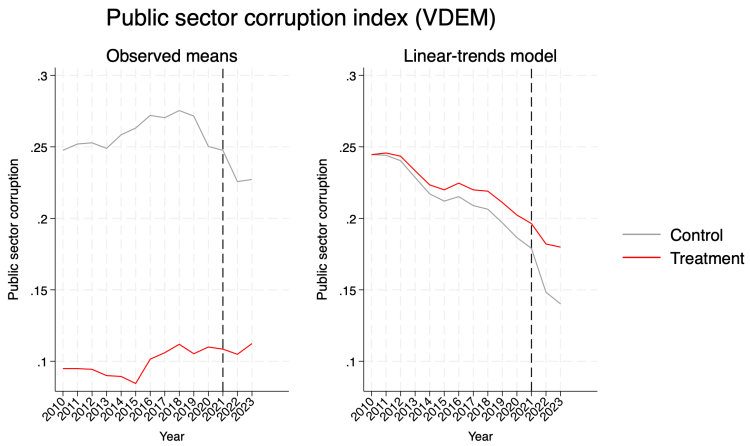
Exploratory analyses using UNODC country-level data on corruption and other economic crime (Table 1.4) show no evidence of behavioural change at this stage. This suggests that the modest deterioration in expert-assessed corruption perceptions among early adopters likely reflects heightened scrutiny following transposition rather than an actual decline in integrity. Further exploratory analysis using Eurobarometer data on corruption reporting and perceived reporting barriers (Table 1.5) also shows no change in actual reporting behaviour. However, respondents in early-adopting countries were less likely to view the difficulty of providing proof as a barrier to reporting, yet more likely to cite the lack of punishment as an obsta-

cle. This growing scepticism about enforcement may help explain the absence of a short-term behavioural response.

Figure 1.2: Visual DiD diagnostics of pre-treatment trends in control of corruption (WGI) and public sector corruption index (V-Dem) when EU members and candidates are included in the control group



(a) WGI
H0: Linear trends are parallel
F(1,34)=1.10
Prob > F = 0.3018



(b) V-Dem
H0: Linear trends are parallel
F(1,34)=0.48
Prob > F = 0.4923

Table 1.4: Standard DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level crime rates (UNODC)

Variable	**Corruption**			**Other forms of corruption**			**Bribery**			**Fraud**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
Early transposition (DiD)	0.863	-0.040	3.564	-0.826	-2.120	3.078	1.159	1.090	0.877	-90.742	-98.693	-82.364
	[6.868]	[6.813]	[2.781]	[8.338]	[8.031]	[2.466]	[1.581]	[1.405]	[1.114]	[63.601]	[65.730]	[52.963]
Post transposition of EU Directive	-1.651	-1.704	-3.000	8.363	5.235	-8.119	-7.228*	-6.372*	-5.439	341.990*	365.307*	87.913
	[13.286]	[12.364]	[6.701]	[20.110]	[17.067]	[11.066]	[3.944]	[3.638]	[3.495]	[194.954]	[186.606]	[69.461]
Observations	262	302	231	218	256	195	245	293	224	237	262	195
R-squared	0.110	0.094	0.263	0.171	0.133	0.382	0.144	0.121	0.192	0.453	0.425	0.696
Control mean (2021)	25.78	25.04	25.04	24.07	23.81	23.81	5.577	5.075	5.075	449.5	405.7	405.7
Number of countries	27	35	33	24	30	28	25	33	31	27	32	30
EU candidates included	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES
Country controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES
Clust. P-value	0.901	0.995	0.209	0.922	0.794	0.223	0.470	0.443	0.437	0.166	0.143	0.131
Boot. Std. Error	6.854	6.801	2.774	8.316	8.013	2.459	1.577	1.402	1.111	63.45	65.59	52.81
Boot. P-value (DiD)	0.904	0.995	0.239	0.928	0.825	0.331	0.476	0.450	0.451	0.192	0.167	0.191

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level crime rates, using UNODC data. All models include country and year fixed effects and the full list of time-varying country controls. Models (1) include EU Member States only, whereas models (2) and (3) include also EU candidate countries in the control group. Models (3) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.5: Standard DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level corruption reporting and perceived barriers (Eurobarometer)

Variable	**Corruption reporting**		**Perceived barriers to reporting corruption**			
	Knows where to report	**Has reported corruption**	**No clear reporting channels**	**Hard to prove**	**No punishment expected**	**No protection for reporters**
Early transposition (DiD)	-0.005	-0.021	-0.009	-0.029**	0.024**	-0.005
	[0.017]	[0.024]	[0.012]	[0.012]	[0.011]	[0.018]
Post transposition of EU Directive	-0.069*	0.021	0.077***	0.001	-0.025	-0.055**
	[0.035]	[0.067]	[0.025]	[0.028]	[0.033]	[0.022]
Observations	135	133	135	135	135	135
R-squared	0.441	0.160	0.347	0.354	0.278	0.208
Control mean (2019)	0.426	0.185	0.205	0.447	0.334	0.297
Number of countries	27	27	27	27	27	27
EU candidates included	NO	NO	NO	NO	NO	NO
Country controls	YES	YES	YES	YES	YES	YES
Two lags of dependent variable included	NO	NO	NO	NO	NO	NO
Clust. P-value	0.788	0.391	0.463	0.0284**	0.0442**	0.772
Boot. Std. Error	0.0168	0.0242	0.0118	0.0124	0.0114	0.0178
Boot. P-value (DiD)	0.794	0.423	0.502	0.0377**	0.0613*	0.806

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level beliefs on corruption reporting, using Eurobarometer data. Models include only EU Member States, country and year fixed effects, and the full list of time-varying country controls. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Heterogeneity tests (Tables 1.6 and 1.7) confirm that the effects of transposition were concentrated in expert-based corruption perceptions, which tend to adjust more rapidly than citizen views and are more sensitive to awareness of the Directive's adoption. We examined several potential mediating factors using baseline data from the Eurobarometer (2018 or 2019) and the OECD (2018 or 2019-2021 average), grouped into three domains: (i) reporting behaviour and related corruption perceptions, (ii) institutional readiness and trust in national institutions, and (iii) interest and trust in EU institutions. For each indicator, countries were classified as having high or low baseline values based on whether they fell above or below the sample median.

In terms of reporting behaviour and related corruption perceptions, results in Table 1.6 show that expert-assessed corruption perceptions deteriorated more in countries where reporting activity was relatively lower, where concerns about impunity were stronger, and where there were fewer concerns about protections available to whistleblowers. Perceptions also deteriorated more in countries with greater tolerance of corruption. Additional heterogeneity tests are included in Table 1.A.6. Taken together, these patterns suggest that in contexts where corruption is viewed as pervasive yet socially tolerated — and where enforcement is expected to be weak — the Directive's introduction may have increased scepticism rather than reassurance among experts.

Regarding institutional trust and preparedness, results in Table 1.7 show that deterioration in expert-assessed perceptions was stronger in countries with weaker institutional readiness (i.e. no prior horizontal whistleblower protection law), lower perceived effectiveness of government anti-corruption efforts and prosecutions, and lower trust in government, in public administration, and in the media. Additional heterogeneity tests are included in Table 1.A.7. Overall, greater deterioration remained concentrated in countries with weak perceived government effectiveness.

Finally, perceptions declined more in countries with stronger support for the EU (Table 1.A.8), suggesting that more EU-engaged contexts may have reacted more critically to early implementation gaps, interpreting them as a sign that the Directive's promises were not yet matched by national practice. Additional heterogeneity tests are included in Table 1.A.8.

Table 1.6: Heterogeneity analysis of reporting behaviour and corruption-related perceptions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
High corruption reporting (Eurobarometer 2019)				
Early transposition	0.094	0	0.051**	0.033**
	[0.058]	[0.027]	[0.020]	[0.015]
Early transposition * High corruption reporting	-0.205**	-0.026	-0.034*	-0.023*
	[0.082]	[0.037]	[0.018]	[0.011]
Post transposition of EU Directive	0.211*	0.087	-0.047	-0.026
	[0.104]	[0.065]	[0.031]	[0.016]
Observations	378	324	378	324
R-squared	0.382	0.711	0.407	0.632
Control mean (2019)	0.526	0.526	0.177	0.177
Coeff: T + T*Mediator = 0	-0.111	-0.0257	0.0166	0.00958
P-value: T + T*Mediator = 0	0.174	0.421	0.135	0.245
Lack of punishment highly perceived as a reporting barrier (Eurobarometer 2019)				
Early transposition	-0.038	-0.009	0.014	0.008
	[0.071]	[0.039]	[0.013]	[0.009]
Early transposition * Lack of punishment barrier	-0.001	-0.012	0.024*	0.016*
	[0.116]	[0.040]	[0.014]	[0.009]
Post transposition of EU Directive	0.183	0.082	-0.050	-0.028
	[0.115]	[0.065]	[0.031]	[0.017]
Observations	378	324	378	324
R-squared	0.362	0.711	0.402	0.629
Control mean (2019)	1.249	1.249	0.071	0.071
Coeff: T + T*Mediator = 0	-0.0386	-0.0212	0.0385	0.024
P-value: T + T*Mediator = 0	0.682	0.350	0.0135**	0.0389**
Lack of protection highly perceived as a reporting barrier (Eurobarometer 2019)				
Early transposition	0.108**	-0.013	0.036**	0.028**
	[0.051]	[0.037]	[0.013]	[0.010]
Early transposition * Lack of protection barrier	-0.244**	-0.007	-0.012	-0.018*
	[0.098]	[0.045]	[0.013]	[0.009]
Post transposition of EU Directive	0.153	0.082	-0.053	-0.032*
	[0.111]	[0.065]	[0.033]	[0.018]
Observations	378	324	378	324
R-squared	0.389	0.711	0.398	0.630
Control mean (2019)	1.13	1.13	0.0872	0.0872
Coeff: T + T*Mediator = 0	-0.136	-0.0191	0.0239	0.0105
P-value: T + T*Mediator = 0	0.0995	0.502	0.0996*	0.306
Low tolerance of corruption (Eurobarometer 2019)				
Early transposition	0.048	-0.030	0.052***	0.036**
	[0.076]	[0.029]	[0.017]	[0.013]
Early transposition * Low tolerance of corruption	-0.121	0.020	-0.033**	-0.025**
	[0.096]	[0.042]	[0.015]	[0.010]
Post transposition of EU Directive	0.184	0.083	-0.051	-0.029
	[0.113]	[0.065]	[0.032]	[0.017]
Observations	378	324	378	324
R-squared	0.368	0.711	0.405	0.632
Control mean (2019)	0.808	0.808	0.148	0.148
Coeff: T + T*Mediator = 0	-0.0737	-0.0103	0.0191	0.0102
P-value: T + T*Mediator = 0	0.334	0.741	0.102	0.232
Number of countries	27	27	27	27
EU candidates included	NO	NO	NO	NO
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline reporting behaviour and corruption-related perceptions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.7: Heterogeneity analysis of readiness of and trust in national institutions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
High perceived effectiveness of government anti-corruption efforts (Eurobarometer 2019)				
Early transposition	-0.045	-0.028	0.032**	0.020*
	[0.081]	[0.022]	[0.014]	[0.011]
Early transp. * High perceived effectiveness (govt. efforts)	0.023	0.041	-0.010	-0.007
	[0.118]	[0.045]	[0.013]	[0.008]
Post transposition of EU Directive	0.182	0.082	-0.051	-0.029
	[0.113]	[0.064]	[0.032]	[0.017]
Observations	378	324	378	324
R-squared	0.363	0.711	0.397	0.627
Control mean (2019)	0.845	0.845	0.141	0.141
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	-0.022	0.0122	0.0218	0.0129
P-value: T + T*Mediator = 0	0.798	0.778	0.116	0.171
High perceived effectiveness of corruption prosecutions (Eurobarometer 2019)				
Early transposition	-0.047	-0.028	0.029**	0.019*
	[0.072]	[0.025]	[0.012]	[0.009]
Early transp. * High perceived effectiveness (prosecutions)	0.061	0.087*	-0.005	-0.007
	[0.150]	[0.046]	[0.016]	[0.010]
Post transposition of EU Directive	0.192*	0.098	-0.052	-0.030
	[0.100]	[0.064]	[0.034]	[0.018]
Observations	378	324	378	324
R-squared	0.363	0.713	0.397	0.627
Control mean (2019)	0.986	0.986	0.138	0.138
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	0.0137	0.0592	0.0245	0.0119
P-value: T + T*Mediator = 0	0.913	0.188	0.218	0.386
Trust in the national public administration (Eurobarometer 2018)				
Early transposition	-0.082	-0.038*	0.037***	0.024**
	[0.100]	[0.021]	[0.013]	[0.011]
Early transp. * High trust in public administration	0.093	0.047	-0.018	-0.013
	[0.127]	[0.039]	[0.012]	[0.008]
Post transposition of EU Directive	0.188*	0.086	-0.052	-0.030*
	[0.101]	[0.062]	[0.033]	[0.017]
Observations	378	324	378	324
R-squared	0.367	0.712	0.400	0.629
Control mean (2018)	0.418	0.418	0.193	0.193
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	0.0108	0.00903	0.0193	0.0105
P-value: T + T*Mediator = 0	0.876	0.814	0.159	0.283
Trust in the written press (Eurobarometer 2018)				
Early transposition	-0.030	-0.005	0.036**	0.024**
	[0.085]	[0.024]	[0.015]	[0.010]
Early transp. * High trust in written press	0.028	-0.013	-0.026*	-0.014*
	[0.098]	[0.038]	[0.013]	[0.008]
Post transposition of EU Directive	0.196*	0.097	-0.100**	-0.033**
	[0.106]	[0.064]	[0.042]	[0.015]
Observations	436	375	436	375
R-squared	0.307	0.707	0.381	0.634
Control mean (2018)	0.259	0.259	0.270	0.270
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	-0.00258	-0.0174	0.0108	0.0103
P-value: T + T*Mediator = 0	0.968	0.610	0.391	0.184
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline readiness of and trust in national institutions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

To investigate potential mechanisms of change, we plan to complement this analysis with organisation-level data on integrity incidents, share of employees who reported misconduct, and general perceptions about the whistleblowing environment from the EY Global Integrity Report (Ernst & Young, 2024). The dataset is based on a survey of 5,464 board members, senior managers, managers and employees in a sample of large organizations and public bodies in 53 countries and territories across the Americas, Asia Pacific and Europe, the Middle East, India and Africa.

1.5 Conclusion

This paper offers the first impact evaluation of the short-term effects of transposing the 2019 EU Whistleblower Protection Directive into national law on perceptions of corruption. Using panel data for EU countries from 2010 to 2023, we exploit the gradual timing of the Directive's transposition to identify early effects on corruption perceptions.

Perceptions matter because they serve as an early barometer of how new anti-corruption reforms are received and whether they are viewed as credible. As discussed, whistleblower protection laws are expected to operate initially through perceptions rather than behaviour: by introducing reporting channels, protection mechanisms, and investigative procedures, they generate institutional signals that shape expectations about the ease and safety of reporting and the likelihood that misconduct will be detected and sanctioned. These expectations in turn translate into broader perceptions of corruption control. Whether such signals improve or worsen perceptions depends on their perceived credibility and on pre-existing institutional conditions.

Our findings indicate no measurable short-term improvements in citizen-based perceptions of control of corruption (EQI) or in composite indicators combining citizen, firm, and expert assessments (WGI), and no changes in reporting behaviour (Eurobarometer). By contrast, we find a modest deterioration in expert-assessed corruption (V-Dem) among early adopters. This pattern is consistent with the interpretation that transposition triggered increased scrutiny and a reassessment of enforcement expectations, rather than a deterioration in underlying integrity.

This interpretation is further supported by the absence of short-term effects on corruption and economic crime in objective data (UNODC).

Heterogeneity analyses reinforce this interpretation. Perceptual shifts are concentrated among experts, who are more likely to be aware of legal reforms and better positioned to assess their credibility. Deterioration was stronger in countries with lower reporting activities, where concerns about impunity were higher, corruption more pervasive and socially tolerated, and trust in government weaker. The decline was also more marked in EU-engaged countries — suggesting that experts in such contexts may have reacted more critically to early gaps between formal transposition and effective implementation. Taken together, these patterns point to two plausible interpretations: weak enforcement credibility in low-trust settings and a potential "expectations gap" in pro-EU contexts where reforms raise standards that are not immediately met.

These results underline that the initial impact of whistleblower protection reforms is perceptual rather than structural. In the short-term, transposition appears to trigger a reassessment of institutional credibility rather than immediate improvements in behaviour. This interpretation is consistent with the Eu-ropean Commission's 2024 assessment and civil society reviews, which document delays, inconsistencies, and gaps in protection and enforcement (Transparency International, 2023; European Commission, 2024). Over time, however, consistent and credible implementation could narrow this credibility gap, strengthening trust in reporting systems and improving perceptions of corruption control as enforcement mechanisms become more visible and effective.

Several limitations warrant caution. First, the short post-treatment window limits our ability to capture medium- or long-term behavioural effects. Future studies with extended panels could employ dynamic or event-study DiD models to track temporal adjustments more fully. Second, variation in transposition timing may be endogenous to institutional capacity or political will. While we control for observable differences, account for time-invariant heterogeneity through fixed effects, and include EU candidate countries to improve comparability, other unobserved time-varying factors may still bias estimates. Third, although graphical and statistical diagnostics support the plausibility of parallel trends in the full sample, this assumption is weaker among EU Member States alone. With more data in the com-

ing years, future work could apply group-time average treatment effect estimators that explicitly account for staggered adoption and allow for treatment effect heterogeneity (Baker et al., 2025). Finally, perception-based indicators — although useful for early detection — are inherently subjective and may be only partly correlated with actual institutional progress. Future research should revisit these relationships as more post-treatment data become available and integrate more objective indicators of reporting behaviour, enforcement outcomes, and case resolution rates.

A key policy implication is the importance of credible and consistent implementation. The effectiveness of whistleblower protection laws depends not only on their formal adoption but also on how they are perceived in practice. Strengthening reporting systems, ensuring follow-up and enforcement, and improving transparency around outcomes are essential to reinforcing credibility. The European Commission and Member States could support this process by systematically collecting and harmonizing data on implementation quality, enforcement capacity, protection practices, and public awareness. Such efforts would enable more rigorous evaluation of governance reforms over time.

Despite its short-term scope, this study provides an off-the-shelf framework for evaluating how institutional reforms operate through perceptions before translating into behavioural change, offering a foundation for future research on the medium- and long-term effects of whistleblower protection policies.

References

Abbink, K., Dasgupta, U., Gangadharan, L. and Jain, T.: 2014, Letting the briber go free: An experiment on mitigating harassment bribes, *Journal of Public Economics* **111**, 17–28.

Amir, E., Lazar, A. and Levi, S.: 2018, The deterrent effect of whistleblowing on tax collections, *European Accounting Review* **27**.

Andreis, E.: 2019, Towards common minimum standards for whistleblower protection across the eu, *European Papers* **4**, 575–588.

Angrist, J. D. and Pischke, J.-S.: 2009, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press.

Antinyan, A., Corazzini, L. and Pavesi, F.: 2020, Does trust in the government matter for whistleblowing on tax evaders? survey and experimental evidence, *Journal of Economic Behavior & Organization* **171**, 77–95.

Arndt, C. and Oman, C.: 2006, *Uses and abuses of governance indicators*, OECD Development Centre Studies.

Association of Certified Fraud Examiners: 2024, Occupational fraud 2024: A report to the nations, *Technical report*, Association of Certified Fraud Examiners.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A. and Sant'Anna, P. H. C.: 2025, Difference-in-differences designs: A practitioner's guide, *Technical report*, arXiv.

Banuri, S.: 2021, A behavioural economics perspective on compliance, *in* A. Riley, A. Stephan and A. Tubbs (eds), *Perspectives on Antitrust Compliance*, Concurrences.

Butler, J., Serra, D. and Spagnolo, G.: 2019, Motivating whistleblowers, *Management Science* **66**.

Charron, N.: 2015, Do corruption measures have a perception problem? assessing the relationship between experiences and perceptions of corruption among citizens and experts, *Technical report*, European Political Science Review.

Charron, N., Lapuente, V. and Bauhr, M.: 2024, The geography of quality of government in europe: Subnational variations in the 2024 european quality of government index and comparisons with previous rounds, *Technical report*, Department of Political Science, University of Gothenburg.

Choo, L., Grimm, V., Horváth, G. and Nitta, K.: 2019, Whistleblowing and diffusion of responsibility: An experiment, *European Economic Review* **119**, 287–301.

Coppedge, M., Gerring, J. et al.: 2019, The methodology of "varieties of democracy" (v-dem), *Technical report*, V-Dem Institute.

Dyck, A., Morse, A. and Zingales, L.: 2010, Who blows the whistle on corporate fraud?, *The Journal of Finance* **65**, 2213–2253.

Ernst & Young: 2024, Ey global integrity report 2024: How can trust survive without integrity? why taking the human-centered approach empowers an ethical culture, *Technical report*, Ernst & Young.

EU Whistleblowing Monitor: 2024, Eu whistleblowing monitor.

Eurobarometer: 2024, Eurobarometer data.

European Commission: 2024, Report on the transposition of the whistleblower protection directive (directive 2019/1937) on the protection of persons who report breaches of union law, *Technical report*, European Commission.

European Parliament and Council of the European Union: 2019, Directive (eu) 2019/1937 on the protection of persons who report breaches of union law.

Eurostat: 2024, Eurostat database.

Johannesen, N. and Stolper, T.: 2021, The deterrence effect of whistleblowing, *The Journal of Law and Economics* **64**.

Kaufmann, D. and Kraay, A.: 2024, The worldwide governance indicators: Methodology and 2024 update, *Technical report*, World Bank.

Krügel, S. and Uhl, M.: 2023, Internal whistleblowing systems without proper sanctions may backfire, *Journal of Business Economics* **93**, 1355–1383.

Lee, Y., Ng, S., Shevlin, T. and Venkat, A.: 2024, The deterrence effects of tax whistleblower laws: Evidence from new york's false claims acts, *Management Science* **71**.

Li, H. and Meng, T.: 2020, Corruption experience and public perceptions of anti-corruption crackdowns: Experimental evidence from china, *Journal of Chinese Political Science* **25**, 431–456.

Mechtenberg, L., Muehlheusser, G. and Roider, A.: 2020, Whistleblower protection: Theory and experimental evidence, *European Economic Review* **126**.

Montes, G. C. and Luna, P. E.: 2020, Fiscal transparency, legal system and perception of the control on corruption: empirical evidence from panel data, *Empirical Economics* **60**, 2500–2537.

OECD: 2024, Oecd data: Trust in government.

Olken, B.: 2009, Corruption perceptions vs. corruption reality, *Journal of Public Economics* **93**(7), 950–964.

Pop-Eleches, G. and Robertson, G.: 2024, Do reforms reduce corruption perceptions? evidence from police reform in ukraine, *Post-Soviet Affairs* **40**, 345–361.

Quality of Government Institute: 2024, European quality of government index (eqi) data.

Reinders Folmer, C.: 2021, Crowding-out effects of laws, policies and incentives on compliant behaviour, *in* B. van Rooij and D. D. Sokol (eds), *The Cambridge Handbook of Compliance*, Cambridge University Press, pp. 326–340.

Rosenson, B. A.: 2009, The effect of political reform measures on perceptions of corruption, *Election Law Journal* **8**, 31–46.

Schmolke, K. U. and Utikal, V.: 2025, Whistleblowing: Incentives and situational determinants, *Journal of Business Economics* **95**, 725–748.

Sun, Z., Zhu, L. and Ni, X.: 2022, How does anti-corruption information affect public perceptions of corruption in china?, *China Review* **22**(2), 113–143.

Transparency International: 2023, How well do eu countries protect whistleblowers? assessing the transposition of the eu whistleblower protection directive, *Technical report*, Transparency International.

Transparency International: 2025, What is corruption?

Transparency International Ireland: 2025, Speak up report 2025, *Technical report*, Transparency International Ireland. See PDF for full list of contributors.

United Nations Office on Drugs and Crime: 2024, Unodc data: Crime and corruption offences.

V-Dem Institute: 2024, V-dem dataset.

Wilde, J.: 2017, The deterrent effect of employee whistleblowing on firms' financial misreporting and tax aggressiveness, *The Accounting Review* **92**(5), 247–280.

World Bank: 2024a, World development indicators.

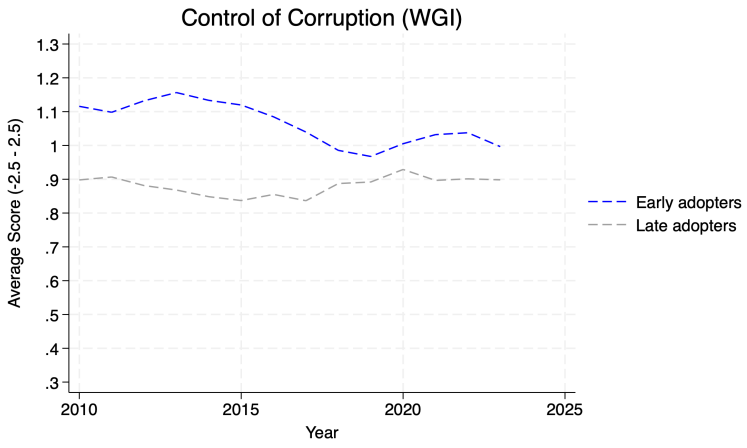
World Bank: 2024b, Worldwide governance indicators: 2024 data update.

Wroe, A., Allen, N. and Birch, S.: 2013, The role of political trust in conditioning perceptions of corruption, *European Political Science Review* **5**(2), 175–195.

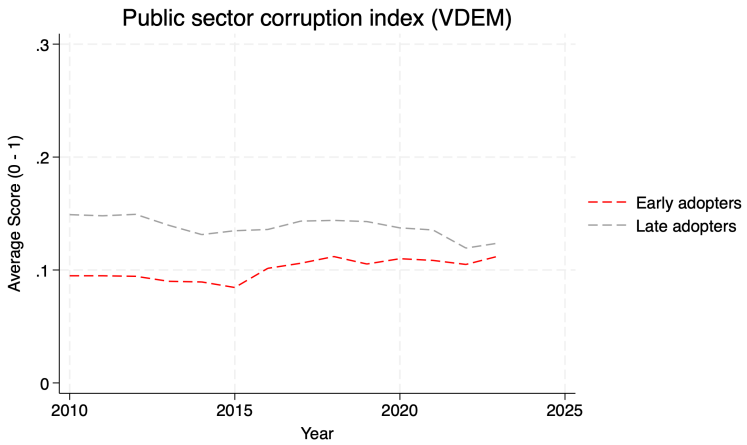
Appendix

1.A Additional tables and figures

Figure 1.A.1: Outcome trends in control of corruption (WGI) and public sector corruption index (V-Dem) when only EU members are included in the control group



(a) WGI



(b) V-Dem

Table 1.A.1: Summary data and statistics

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Outcome variables (WGI – yearly / V-Dem – yearly / EQI – 2010, 2013, 2017, 2021, 2024 / UNODC – yearly)							
Control of corruption	WGI	Extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.664	0.878	-0.666	2.333
Government effectiveness	WGI	Quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.726	0.733	-1.073	1.960
Political stability and absence of violence	WGI	Likelihood of political instability and/or politically motivated violence, including terrorism. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.491	0.503	-1.137	1.194
Rule of law	WGI	Extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.758	0.735	-0.434	2.013
Regulatory quality	WGI	Ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.905	0.609	-0.195	1.907
Voice and accountability	WGI	Extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Scores range between –2.5 and 2.5, with higher scores corresponding to better outcomes.	35	0.799	0.612	-0.860	1.609

Table 1.A.1 continued from previous page

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Public sector corruption index	V-Dem	Extent to which public sector employees engage in bribery and theft. The index ranges from 0 to 1 (highly corrupt).	35	0.215	0.222	0.001	0.764
Academic freedom index	V-Dem	Extent to which universities are autonomous from the state, and academics can research, teach, publish, express opinions, and participate in academic bodies without interference, censorship, or institutional restrictions. The index ranges from 0 to 1 (most free).	35	0.852	0.175	0.070	0.965
Deliberative democracy index	V-Dem	Extent to which universities are autonomous from the state, and academics can research, teach, publish, express opinions, and participate in academic bodies without interference, censorship, or institutional restrictions. The index ranges from 0 to 1 (most democratic).	35	0.620	0.208	0.097	0.876
Egalitarian democracy index	V-Dem	Extent of voting rights, the freedom and fairness of elections, freedoms of association and expression, and the protection of rights, access to power, and distribution of resources being equal. The index ranges from 0 to 1 (most democratic).	35	0.635	0.178	0.227	0.877
Liberal democracy index	V-Dem	Extent of voting rights, the freedom and fairness of elections, freedoms of association and expression, civil liberties, and executive constraints. The index ranges from 0 to 1 (most democratic).	35	0.646	0.198	0.119	0.882
Electoral democracy index	V-Dem	Extent to which political leaders are elected under comprehensive voting rights in free and fair elections, and freedoms of association and expression are guaranteed. The index ranges from 0 to 1 (most democratic).	35	0.744	0.169	0.288	0.915
Corruption	UNODC	Unlawful acts as defined in the United Nations Convention against corruption and other national and international legal instruments against corruption.	32	25.045	38.702	0.179	204.308
Other forms of corruption	UNODC	Other acts of corruption includes embezzlement, abuse of functions, trading in influence, illicit enrichment and all other acts of corruption not mentioned above.	28	23.811	39.765	0.000	200.468
Bribery	UNODC	Promising, offering, giving, soliciting, or accepting an undue advantage to or from a public official or a person who directs or works in a private sector entity, directly or indirectly, in order that the person act or refrain from acting in the exercise of his or her official duties.	31	5.075	11.389	0.099	63.330

Table 1.A.1 continued from previous page

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Fraud	UNODC	Obtaining money or other benefit, or evading a liability through deceit or dishonest conduct.	29	405.716	501.598	6.624	2479.049
EQI Corruption	EQI	Corruption pillar: citizens' perceptions and experiences of bribery, nepotism, or misuse of public office. Higher scores correspond to lower perceived or experienced corruption.	210	0.000	1.000	-2.177	2.010
EQI Corruption perceptions	EQI	Corruption sub-pillar: citizens' perceptions of corruption in key public services, such as police, education, and health care. Higher scores indicate lower perceived corruption in regional public institutions.	210	0.000	1.000	-2.655	2.177
EQI Corruption experiences	EQI	Corruption sub-pillar: citizens' perceptions of corruption in key public services, such as police, education, and health care. Higher scores indicate fewer personal experiences with corruption.	210	0.000	1.000	-3.295	1.449
EQI Quality	EQI	Quality pillar: how citizens perceive the quality of healthcare, education, and law enforcement. Higher scores reflect better perceived quality of public services.	210	0.000	1.000	-2.337	2.197
EQI Impartiality	EQI	Impartiality pillar: whether public services are perceived to be delivered without discrimination. Higher scores indicate more impartial service provision.	210	0.000	1.000	-2.223	2.410
EQI Index	EQI	Average of three EQI pillars.	210	0.000	1.000	-2.144	2.289
Control variables (WDI – yearly / Eurostat – yearly)							
Previous whistleblower law	Authors' compilation	Binary indicator for the presence of a horizontal (cross-sector) whistleblower protection law as of 2021, and time-varying dummy that switches to one in the year a horizontal whistleblower protection law was adopted (before 2021).	35	0.457	0.505	0	1
Total population (log)	WDI	Log of total population.	35	15.678	1.377	13.159	18.248
Rural population (%)	WDI	Rural population (% of total population).	35	29.364	13.882	1.883	57.002
Agriculture value added (%)	WDI	Agriculture, forestry, and fishing, value added (% of GDP).	35	3.550	3.472	0.191	18.358
GDP per capita (log)	WDI	Log of GDP per capita, PPP (constant 2021 international $).	35	10.630	0.498	9.660	11.835
Employment rate (%)	WDI	Employment to population ratio, 15+, total (% of total population, modeled ILO).	35	54.202	6.401	42.439	71.937

Table 1.A.1 continued from previous page

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Female employment rate (%)	WDI	Employment to population ratio, 15+, female (% of female population, modeled ILO).	35	48.206	8.258	27.934	70.243
Women in national parliaments (%)	WDI-IPU	Share of parliamentary seats held by women in the lower or single chamber of the national legislature, as reported by the Inter-Parliamentary Union.	35	30.574	9.857	13.100	47.000
Labor force with intermediate education (%)	WDI	Share of total working-age population with intermediate education.	32	63.256	5.060	56.724	75.188
Part-time female employment (%)	WDI	Part-time employment, female (% of female employment).	33	36.853	17.194	6.620	77.840
Participatory democracy index	WDI-V-Dem	Extent of voting rights, the freedom and fairness of elections, freedoms of association and expression, and citizens being able to engage in regional and local government, civil society organizations, and direct democracy. The index ranges from 0 to 1 (most democratic).	35	0.521	0.137	0.175	0.696
Total population (log)	Eurostat (regional)	Log of regional population.	210	14.166	0.913	10.313	16.702
Unemployment rate (%)	Eurostat (regional)	Unemployed regional population, total (% of total regional population).	206	6.542	4.180	1.600	21.400
Female unemployment rate (%)	Eurostat (regional)	Unemployed female regional population, female (% of female regional population).	194	7.421	5.329	1.500	24.100
Population w/ upper secondary education (%)	Eurostat (regional)	Regional population with upper secondary education, total (% of total regional population).	209	47.018	12.733	20.500	73.800
Female population w/ upper secondary education (%)	Eurostat (regional)	Female regional population with upper secondary education, female (% of female regional population).	209	44.287	11.696	19.000	69.900
GDP per capita (log)	Eurostat (regional)	Log of regional GDP per capita, PPP (2020).	205	10.254	0.377	9.116	11.366
Mechanisms and other mediators (Eurobarometer: corruption – 2013, 2017, 2019, 2022, 2023 / OECD – 2018, 2019–21 / Eurobarometer: institutional trust – 2018, 2020)							

Table 1.A.1 continued from previous page

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Knowledge of reporting channels	Eurobarometer	Share of respondents who said they know where to report corruption if they were to experience it.	27	0.426	0.093	0.270	0.650
Corruption reporting	Eurobarometer	Share of respondents who experienced or witnessed corruption and reported it.	27	0.185	0.120	0.030	0.440
Reporting barrier: unclear reporting channels	Eurobarometer	Share of respondents who selected "do not know where to report it" as one of the main reasons why people do not report corruption.	27	0.205	0.072	0.120	0.370
Reporting barrier: difficulty to prove	Eurobarometer	Share of respondents who selected "difficult to prove anything" as one of the main reasons why people do not report corruption.	27	0.447	0.096	0.250	0.610
Reporting barrier: lack of punishment	Eurobarometer	Share of respondents who selected "reporting it would be pointless because those responsible will not be punished" as one of the main reasons why people do not report corruption.	27	0.334	0.077	0.200	0.530
Reporting barrier: lack of protection	Eurobarometer	Share of respondents who selected "there is no protection for those who report corruption" as one of the main reasons why people do not report corruption.	27	0.297	0.075	0.150	0.490
Low tolerance to corruption	Eurobarometer	Share of respondents who consider corruption to be "unacceptable".	27	0.648	0.146	0.380	0.880
Presence of corruption	Eurobarometer	Share of respondents who believe corruption is "widespread" in their country.	27	0.726	0.216	0.220	0.970
Experience of corruption	Eurobarometer	Share of respondents who experienced or witnessed corruption.	27	0.071	0.034	0.020	0.150
Perceived effectiveness of government anti-corruption efforts	Eurobarometer	Share of respondents who agreed with the statement "government efforts to combat corruption are effective".	27	0.315	0.086	0.150	0.500
Perceived effectiveness of corruption prosecutions	Eurobarometer	Share of respondents who agreed with the statement "there are enough successful prosecutions in my country to deter people from corrupt practices".	27	0.339	0.095	0.170	0.580
Trust in the government	OECD, 2018	Share of respondents who reported having confidence in the national government.	26	0.424	0.153	0.220	0.760

Table 1.A.1 continued from previous page

Variable	Data source	Description	Obs.	Mean	Std. D.	Min	Max
Trust in the government	OECD, 2019-21 average	Share of respondents who reported having confidence in the national government.	26	0.444	0.159	0.214	0.760
Trust in the judiciary	Eurobarometer, 2018	Share of respondents who reported that they tend to trust the justice/legal system in their country.	27	0.496	0.196	0.190	0.870
Trust in the public administration	Eurobarometer, 2018	Share of respondents who reported that they tend to trust the public administration in their country.	27	0.510	0.176	0.180	0.770
Trust in the written press	Eurobarometer, 2018	Share of respondents who reported that they tend to trust the written press in their country.	32	0.482	0.136	0.250	0.760
My voice counts in national decision-making	Eurobarometer, 2019	Share of respondents who agreed with the statement "my voice counts in my country".	27	0.596	0.184	0.290	0.930
Political interest index	Eurobarometer, 2018	Share of respondents who showed strong political interest (Eurobarometer's calculation)	32	0.171	0.077	0.060	0.410
OECD member	Author's compilation	If member of OECD (as of 2021).	35	0.657	0.482	0.000	1.000
EU maturity	Author's compilation	N. of years of EU membership(as of 2021).	35	24.800	21.909	0.000	63.000
EU is going in the right direction	Eurobarometer, 2018	Share of respondents who feel that "things are going in the right direction in the EU".	32	0.369	0.143	0.180	0.840
EU is going in the right direction	Eurobarometer, 2020	Share of respondents who feel that "things are going in the right direction in the EU".	27	0.316	0.073	0.180	0.480
Trust in EU	Eurobarometer, 2018	Share of respondents who reported that they tend to trust the EU.	32	0.475	0.112	0.250	0.720
My voice counts in EU decision-making	Eurobarometer, 2019	Share of respondents who agreed with the statement "my voice counts in the EU".	27	0.488	0.157	0.230	0.770
European political matters are discussed with friends	Eurobarometer, 2018	Share of respondents who reported that they "frequently discuss European political matters with their friends".	32	0.137	0.059	0.040	0.240
Understanding of how the EU works	Eurobarometer, 2018	Share of respondents who agreed with the statement "I understand how the EU works".	32	0.626	0.094	0.440	0.820

Table 1.A.2: Balance of covariates (values for 2021)

Variable	Treatment group (mean)	Treatment group (obs.)	Control group EU members (mean)	Control group EU members (obs.)	P-value	Control group EU members & candidates (mean)	Control group EU members & candidates (obs.)	P-value
Country-level variables								
Previous whistleblower law	0.700	10	0.235	17	0.017**	0.360	25	0.072*
Total population	15.302	10	16.130	17	0.128	15.828	25	0.314
Rural population	25.530	10	26.324	17	0.883	30.897	25	0.309
Agriculture, forestry, and fishing, value added	2.047	10	2.185	17	0.783	4.151	25	0.106
GDP per capita	10.877	10	10.808	17	0.628	10.531	25	0.062*
Employment rate	56.484	10	54.210	17	0.270	53.290	25	0.186
Female employment rate	51.484	10	48.518	17	0.211	46.895	25	0.140
Share of women in parliament	30.420	10	30.718	17	0.943	30.636	25	0.954
Labor force with intermediate education	65.335	10	61.823	17	0.067*	62.311	22	0.119
Part time female employment	39.794	10	42.428	17	0.667	35.575	23	0.526
Participatory democracy index	0.602	10	0.562	17	0.258	0.489	25	0.027**
Regional-level variables								
Total population (log)	14.075	59	14.201	151	0.369	n/a	n/a	n/a
Unemployment rate (%)	6.481	58	6.566	148	0.896	n/a	n/a	n/a
Female unemployment rate (%)	6.693	56	7.717	138	0.226	n/a	n/a	n/a
Population w/ upper secondary ed. (%)	42.390	58	48.795	151	0.001	n/a	n/a	n/a
Female population w/ upper secondary ed. (%)	38.795	58	46.397	151	0.000	n/a	n/a	n/a
GDP per capita (log)	10.302	56	10.236	149	0.260	n/a	n/a	n/a

Notes: The table presents means, observations, and p-values from two-sample t-tests evaluating covariate balance between early adopters (treatment group) and two specifications of late adopters (control groups): (1) EU members only and (2) EU members and candidate countries. P-values are based on standard country-clustered errors. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.3: Robustness check: EQI indicators (placebo outcomes)

Variable	**EQI**			**EQI Quality**			**EQI Impartiality**		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Early transposition	0.024	0.022	-0.022	-0.038	-0.037	-0.068	0.128	0.055	0.021
	[0.084]	[0.092]	[0.115]	[0.115]	[0.118]	[0.119]	[0.086]	[0.090]	[0.114]
Post transposition of EU Directive	0.181**	0.127	0.125	0.101	0.178	0.047	0.047	-0.082	-0.062
	[0.078]	[0.099]	[0.191]	[0.109]	[0.115]	[0.236]	[0.162]	[0.196]	[0.205]
Observations	1050	933	571	1050	933	571	1050	933	571
R-squared	0.081	0.091	0.185	0.054	0.067	0.344	0.084	0.081	0.158
Control mean (2021)	9.52e-06	9.52e-06	9.52e-06	4.76e-06	4.76e-06	4.76e-06	1.43e-05	1.43e-05	1.43e-05
Number of regions	210	202	198	210	202	198	210	202	198
EU candidates included	NO	NO	NO	NO	NO	NO	NO	NO	NO
Country controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Regional controls	NO	YES	YES	NO	YES	YES	NO	YES	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	YES	NO	NO	YES
Clust. P-value	0.778	0.816	0.852	0.746	0.753	0.572	0.148	0.550	0.855
Boot. Std. Error	0.0840	0.0920	0.115	0.115	0.118	0.119	0.0859	0.0904	0.114
Boot. P-value	0.805	0.842	0.891	0.790	0.810	0.684	0.205	0.616	0.889

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on regional-level EQI dimensions of quality of public services such as education, healthcare, and police, and impartiality in their access. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) and (3) include time-varying regional controls too. Models (3) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.4: Robustness check: WGI indicators (placebo outcomes)

Variable	**Government effectiveness**			**Political stability**			**Rule of law**			**Regulatory quality**			**Voice & accountability**		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Early transposition	-0.022	-0.006	0.009	0.011	-0.020	-0.015	0.006	0.011	0.012	0.057	0.058	0.034	0.004	0.022	-0.010
	[0.068]	[0.064]	[0.031]	[0.055]	[0.061]	[0.037]	[0.056]	[0.052]	[0.020]	[0.066]	[0.064]	[0.031]	[0.029]	[0.034]	[0.014]
Post transposition of EU Directive	-0.419***	-0.255**		-0.284**	-0.333**	-0.127	-0.070	-0.031	0.025	-0.193	-0.063	0.023	0.079	0.115*	0.061**
	[0.117]	[0.116]		[0.129]	[0.128]	[0.081]	[0.092]	[0.086]	[0.024]	[0.115]	[0.092]	[0.044]	[0.047]	[0.065]	[0.029]
Observations	378	468	403	378	468	403	378	468	403	378	468	403	378	468	403
R-squared	0.332	0.264	0.596	0.365	0.252	0.537	0.404	0.388	0.671	0.151	0.203	0.470	0.398	0.486	0.774
Control mean (2021)	0.726	0.726	0.726	0.491	0.491	0.491	0.758	0.758	0.758	0.905	0.905	0.905	0.799	0.799	0.799
Number of countries	27	35	35	27	35	35	27	35	35	27	35	35	27	35	35
EU candidates included	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES
Country controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES
Clust. P-value	0.753	0.930	0.766	0.845	0.744	0.685	0.915	0.835	0.563	0.396	0.372	0.286	0.893	0.511	0.468
Boot. Std. Error	0.0681	0.0637	0.0305	0.0547	0.0611	0.0372	0.0559	0.0522	0.0204	0.0654	0.0637	0.0313	0.0293	0.0337	0.0138
Boot. P-value	0.773	0.933	0.776	0.851	0.753	0.711	0.919	0.844	0.570	0.453	0.425	0.313	0.902	0.556	0.472

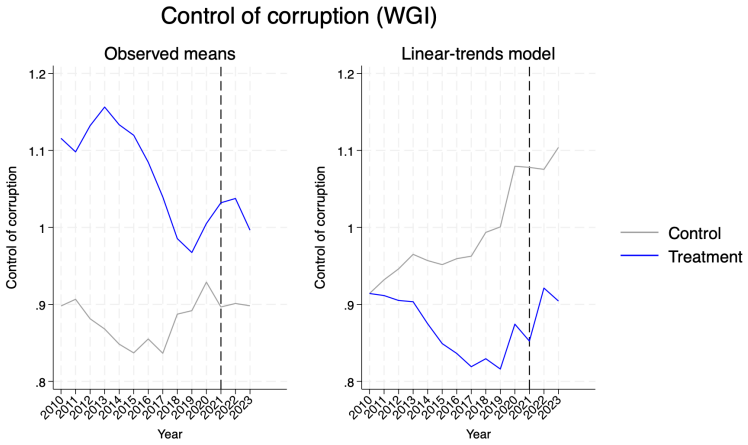
Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level governance measures, using the WGI dimensions of government effectiveness, political stability, rule of law, regulatory quality, and voice and accountability. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) and (3) include EU candidates too. Models (3) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.5: Robustness check: V-Dem democracy indicators (placebo outcomes)

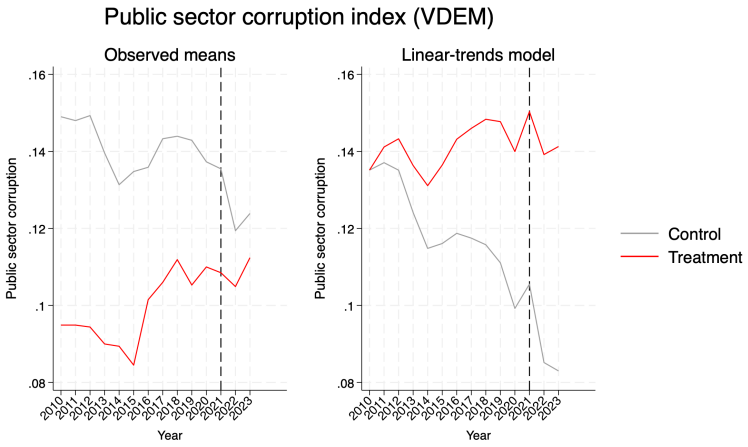
Variable	**Academic freedom**			**Deliberative democracy index**			**Egalitarian democracy index**			**Liberal democracy index**			**Electoral democracy index**		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Early transposition	0.014	0.012	-0.003	-0.006	-0.007	-0.008	-0.006	-0.005	-0.009	0.001	0.003	0.001	0.001	0.000	-0.001
	[0.014]	[0.016]	[0.010]	[0.013]	[0.012]	[0.011]	[0.007]	[0.007]	[0.005]	[0.008]	[0.007]	[0.005]	[0.006]	[0.006]	[0.004]
Post transposition of EU Directive	-0.035	0.009	-0.014	-0.045**	-0.036*	-0.028	-0.019	-0.018	-0.011	-0.023	-0.019	-0.008	-0.004	0.003	0.005
	[0.021]	[0.029]	[0.009]	[0.021]	[0.018]	[0.018]	[0.018]	[0.013]	[0.009]	[0.018]	[0.012]	[0.008]	[0.015]	[0.010]	[0.007]
Observations	378	468	403	378	468	403	378	468	403	378	468	403	378	468	403
R-squared	0.654	0.619	0.810	0.728	0.773	0.752	0.849	0.875	0.901	0.830	0.868	0.915	0.873	0.903	0.938
Control mean (2021)	0.852	0.852	0.852	0.620	0.620	0.620	0.635	0.635	0.635	0.646	0.646	0.646	0.744	0.744	0.744
Number of countries	27	35	35	27	35	35	27	35	35	27	35	35	27	35	35
EU candidates included	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES
Country controls	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Two lags of dependent variable included	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES	NO	NO	YES
Clust. P-value	0.344	0.452	0.763	0.646	0.533	0.466	0.456	0.456	0.127	0.877	0.709	0.834	0.846	0.948	0.873
Boot. Std. Error	0.0143	0.0156	0.00970	0.0134	0.0119	0.0108	0.00744	0.00697	0.00546	0.00775	0.00680	0.00542	0.00592	0.00586	0.00405
Boot. P-value	0.389	0.471	0.771	0.671	0.551	0.481	0.468	0.471	0.146	0.870	0.702	0.840	0.858	0.953	0.884

Note: This table reports DiD estimates of the transposition effect of the EU Whistleblower Protection Directive on country-level democracy measures, using the V-Dem dimensions of participatory, deliberative, egalitarian, liberal, and electoral democracy. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) and (3) include EU candidates too. Models (3) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Figure 1.A.2: Visual DiD diagnostics of pre-treatment trends in control of corruption (WGI) and public sector corruption index (V-Dem) when only EU members are included in the control group



(a) WGI
H0: Linear trends are parallel
F(1,26)=4.49
Prob > F = 0.0438



(b) V-Dem
H0: Linear trends are parallel
F(1,26)=4.06
Prob > F = 0.0544

Table 1.A.6: Heterogeneity analysis of reporting behaviour and corruption-related perceptions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
High knowledge of reporting channels (Eurobarometer 2019)				
Early transposition	-0.060	0.004	0.026	0.011
	[0.066]	[0.037]	[0.019]	[0.013]
Early transp. * High knowledge of reporting channels	0.034	-0.032	0.004	0.011
	[0.100]	[0.041]	[0.016]	[0.009]
Post transposition of EU Directive	0.176*	0.088	-0.052	-0.031*
	[0.103]	[0.065]	[0.033]	[0.018]
Observations	378	324	378	324
R-squared	0.363	0.711	0.397	0.628
Control mean (2019)	0.770	0.770	0.159	0.159
Coeff: T + T*Mediator = 0	-0.0263	-0.0275	0.0299	0.0215
P-value: T + T*Mediator = 0	0.761	0.325	0.0139**	0.023**
Unclear reporting channels highly perceived as a reporting barrier (Eurobarometer 2019)				
Early transposition	-0.099	-0.032	0.032**	0.021*
	[0.086]	[0.022]	[0.013]	[0.010]
Early transp. * Unclear reporting channels barrier	0.157	0.041	-0.010	-0.008
	[0.111]	[0.036]	[0.011]	[0.008]
Post transposition of EU Directive	0.187*	0.084	-0.052	-0.029
	[0.099]	[0.063]	[0.033]	[0.017]
Observations	378	324	378	324
R-squared	0.375	0.712	0.398	0.627
Control mean (2019)	0.500	0.500	0.192	0.192
Coeff: T + T*Mediator = 0	0.0582	0.0086	0.0227	0.0124
P-value: T + T*Mediator = 0	0.399	0.824	0.127	0.222
Difficulty to prove highly perceived as a reporting barrier (Eurobarometer 2019)				
Early transposition	-0.151	-0.005	0.031**	0.014
	[0.093]	[0.028]	[0.014]	[0.011]
Early transp. * Difficulty to prove barrier	0.221*	-0.022	-0.006	0.007
	[0.113]	[0.040]	[0.013]	[0.010]
Post transposition of EU Directive	0.148	0.087	-0.051	-0.030
	[0.110]	[0.068]	[0.033]	[0.018]
Observations	378	324	378	324
R-squared	0.384	0.711	0.397	0.627
Control mean (2019)	0.457	0.457	0.185	0.185
Coeff: T + T*Mediator = 0	0.0704	-0.0269	0.0259	0.0209
P-value: T + T*Mediator = 0	0.215	0.428	0.0715*	0.0714*
High perceived pervasiveness of corruption (Eurobarometer 2019)				
Early transposition	0.058	0.009	0.023	0.012
	[0.068]	[0.038]	[0.014]	[0.010]
Early transp. * High perceived corruption pervasiveness	-0.157	-0.041	0.010	0.008
	[0.111]	[0.036]	[0.011]	[0.008]
Post transposition of EU Directive	0.187*	0.084	-0.052	-0.029
	[0.099]	[0.063]	[0.033]	[0.017]
Observations	378	324	378	324
R-squared	0.375	0.712	0.398	0.627
Control mean (2019)	1.402	1.402	0.0858	0.0858
Coeff: T + T*Mediator = 0	-0.0986	-0.0320	0.0323	0.0209
P-value: T + T*Mediator = 0	0.261	0.162	0.0158**	0.0543*
High experience of corruption (Eurobarometer 2019)				
Early transposition	0.015	-0.005	0.023*	0.015
	[0.057]	[0.029]	[0.012]	[0.010]
Early transp. * High experience of corruption	-0.164	-0.034	0.019	0.008
	[0.141]	[0.034]	[0.014]	[0.010]
Post transposition of EU Directive	0.196*	0.086	-0.053	-0.030*
	[0.097]	[0.063]	[0.033]	[0.017]
Observations	378	324	378	324
R-squared	0.374	0.711	0.400	0.627
Control mean (2019)	1.312	1.312	0.0699	0.0699
Coeff: T + T*Mediator = 0	-0.149	-0.0395	0.0413	0.0230
P-value: T + T*Mediator = 0	0.254	0.154	0.0172**	0.0588*
Number of countries	27	27	27	27
EU candidates included	NO	NO	NO	NO
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline reporting behaviour and other corruption-related perceptions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.7: Heterogeneity analysis of readiness of and trust in national institutions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
Institutional preparedness (author-compiled measures)				
Early transposition	-0.11	-0.008	0.021	0.015**
	[0.131]	[0.033]	[0.015]	[0.007]
Early transposition * Institutional preparedness	0.139	0.003	0.008	0.001
	[0.137]	[0.036]	[0.015]	[0.008]
Post transposition of EU Directive	0.219**	0.116*	-0.102***	-0.040***
	[0.098]	[0.061]	[0.033]	[0.013]
Observations	468	403	468	403
R-squared	0.307	0.722	0.416	0.694
Control mean (2021)	0.817	0.817	0.192	0.192
Number of countries	35	35	35	35
Coeff: T + T*Mediator = 0	0.0291	-0.00525	0.0288	0.0159
P-value: T + T*Mediator = 0	0.531	0.822	0.0536*	0.126
High trust in the government (OECD 2018)				
Early transposition	0.061	-0.028	0.044**	0.026*
	[0.055]	[0.028]	[0.019]	[0.014]
Early transposition * High trust (2018)	-0.003	0.041	-0.026	-0.008
	[0.060]	[0.035]	[0.021]	[0.012]
Post transposition of EU Directive	0.272***	0.113*	-0.094**	-0.033*
	[0.094]	[0.063]	[0.044]	[0.019]
Observations	364	312	364	312
R-squared	0.442	0.680	0.428	0.654
Control mean (2018)	0.533	0.533	0.178	0.178
Number of countries	26	26	26	26
Coeff: T + T*Mediator = 0	0.058	0.0131	0.0182	0.0179
P-value: T + T*Mediator = 0	0.15	0.598	0.242	0.0255**
High trust in the government (OECD 2019–21 average)				
Early transposition	0.061	-0.028	0.044**	0.026*
	[0.055]	[0.028]	[0.019]	[0.014]
Early transposition * High trust (2019–21)	-0.003	0.041	-0.026	-0.008
	[0.060]	[0.035]	[0.021]	[0.012]
Post transposition of EU Directive	0.272***	0.113*	-0.094**	-0.033*
	[0.094]	[0.063]	[0.044]	[0.019]
Observations	364	312	364	312
R-squared	0.442	0.680	0.428	0.654
Control mean (2019–21 avg.)	0.464	0.464	0.187	0.187
Number of countries	26	26	26	26
Coeff: T + T*Mediator = 0	0.058	0.0131	0.0182	0.0179
P-value: T + T*Mediator = 0	0.15	0.598	0.242	0.0255**
EU candidates included	YES	YES	YES	YES
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline readiness of and trust in national institutions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.7 (continued): Heterogeneity analysis of readiness of and trust in national institutions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
Trust in the national justice system (Eurobarometer 2018)				
Early transposition	-0.095	-0.037*	0.027**	0.017
	[0.075]	[0.021]	[0.013]	[0.010]
Early transposition * High trust in judiciary	0.205*	0.075**	0.007	0.001
	[0.105]	[0.035]	[0.011]	[0.007]
Post transposition of EU Directive	0.195*	0.089	-0.051	-0.029*
	[0.098]	[0.062]	[0.032]	[0.017]
Observations	378	324	378	324
R-squared	0.380	0.713	0.397	0.626
Control mean (2018)	0.464	0.464	0.183	0.183
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	0.11	0.0382	0.0337	0.0185
P-value: T + T*Mediator = 0	0.126	0.327	0.0425**	0.0672*
My voice counts in my country (Eurobarometer 2019)				
Early transposition	-0.089	-0.038	0.026**	0.020*
	[0.094]	[0.024]	[0.011]	[0.010]
Early transp. * High perceived voice	0.104	0.043	0.005	-0.005
	[0.105]	[0.029]	[0.012]	[0.008]
Post transposition of EU Directive	0.193*	0.087	-0.051	-0.029*
	[0.100]	[0.063]	[0.032]	[0.017]
Observations	378	324	378	324
R-squared	0.368	0.712	0.397	0.627
Control mean (2019)	0.491	0.491	0.176	0.176
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	0.0147	0.00556	0.0311	0.015
P-value: T + T*Mediator = 0	0.812	0.856	0.0658*	0.163
Political interest (Eurobarometer 2018)				
Early transposition	0.025	-0.019	0.022	0.020*
	[0.055]	[0.024]	[0.014]	[0.010]
Early transp. * Strong political interest index	-0.1	0.019	0.008	-0.004
	[0.105]	[0.029]	[0.013]	[0.007]
Post transposition of EU Directive	0.208**	0.093	-0.103**	-0.034**
	[0.096]	[0.063]	[0.043]	[0.016]
Observations	436	375	436	375
R-squared	0.311	0.708	0.376	0.633
Control mean (2018)	0.666	0.666	0.174	0.174
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	-0.0745	0.000438	0.0295	0.0154
P-value: T + T*Mediator = 0	0.468	0.987	0.0441**	0.0414**
EU candidates included	YES	YES	YES	YES
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline readiness of and trust in national institutions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.8: Heterogeneity analysis of interest and trust in EU institutions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
OECD membership (author-compiled measures)				
Early transposition	-0.217**	-0.007	0.033**	0.014*
	[0.099]	[0.027]	[0.013]	[0.008]
Early transposition * OECD membership	0.286***	0.002	-0.01	0.002
	[0.104]	[0.029]	[0.016]	[0.007]
Post transposition of EU Directive	0.190*	0.115*	-0.103***	-0.040***
	[0.099]	[0.062]	[0.033]	[0.013]
Observations	468	403	468	403
R-squared	0.328	0.722	0.416	0.694
Control mean (2021)	-0.129	-0.129	0.402	0.402
Number of countries	35	35	35	35
Coeff: T + T*Mediator = 0	0.0691	-0.0057	0.0231	0.016
P-value: T + T*Mediator = 0	0.163	0.808	0.137	0.0872*
EU maturity (author-compiled measures)				
Early transposition	-0.04	0.028	0.051***	0.019
	[0.080]	[0.024]	[0.018]	[0.014]
Early transposition * High EU maturity	0.023	-0.039	-0.028	-0.004
	[0.083]	[0.035]	[0.017]	[0.011]
Post transposition of EU Directive	0.196*	0.118*	-0.101***	-0.040***
	[0.111]	[0.063]	[0.033]	[0.013]
Observations	468	403	468	403
R-squared	0.300	0.722	0.418	0.694
Control mean (2021)	-0.238	-0.238	0.474	0.474
Number of countries	35	35	35	35
Coeff: T + T*Mediator = 0	-0.0173	-0.0103	0.0232	0.0151
P-value: T + T*Mediator = 0	0.78	0.648	0.0822*	0.0729*
EU support (Eurobarometer 2018)				
Early transposition	-0.083	-0.023	0.017	0.009
	[0.100]	[0.035]	[0.013]	[0.007]
Early transposition * High EU support (2018)	0.115	0.022	0.015	0.016*
	[0.099]	[0.036]	[0.013]	[0.009]
Post transposition of EU Directive	0.219**	0.100	-0.099**	-0.031**
	[0.096]	[0.064]	[0.043]	[0.015]
Observations	436	375	436	375
R-squared	0.312	0.708	0.378	0.635
Control mean (2018)	1.188	1.188	0.125	0.125
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	0.0317	-0.000807	0.0316	0.025
P-value: T + T*Mediator = 0	0.548	0.971	0.042**	0.0235**
EU support (Eurobarometer 2020)				
Early transposition	-0.100	-0.027	0.017	0.008
	[0.101]	[0.036]	[0.012]	[0.008]
Early transposition * High EU support (2020)	0.110	0.019	0.021*	0.017*
	[0.098]	[0.037]	[0.011]	[0.009]
Post transposition of EU Directive	0.206*	0.087	-0.047	-0.025
	[0.102]	[0.066]	[0.033]	[0.017]
Observations	378	324	378	324
R-squared	0.368	0.711	0.401	0.630
Control mean (2020)	1.100	1.100	0.121	0.121
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	0.0106	-0.00812	0.038	0.0252
P-value: T + T*Mediator = 0	0.845	0.742	0.0143**	0.0473**
EU candidates included	YES	YES	YES	YES
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline interest and trust in EU institutions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 1.A.8. (continued): Heterogeneity analysis of interest and trust in EU institutions

Variable	Control of corruption (WGI)		Public sector corruption (V-Dem)	
	(1)	(2)	(1)	(2)
Trust in EU (Eurobarometer 2018)				
Early transposition	-0.101	-0.036	0.032*	0.015
	[0.100]	[0.025]	[0.016]	[0.010]
Early transp. * High trust in EU	0.143	0.043	-0.012	0.005
	[0.116]	[0.035]	[0.016]	[0.008]
Post transposition of EU Directive	0.206**	0.100	-0.102**	-0.034**
	[0.095]	[0.062]	[0.043]	[0.016]
Observations	436	375	436	375
R-squared	0.316	0.708	0.377	0.633
Control mean (2018)	0.390	0.390	0.234	0.234
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	0.0418	0.00772	0.0201	0.0199
P-value: T + T*Mediator = 0	0.496	0.79	0.142	0.0289**
My voice counts in EU (Eurobarometer 2019)				
Early transposition	-0.070	-0.051**	0.023*	0.018*
	[0.092]	[0.018]	[0.012]	[0.010]
Early transp. * High perceived voice in EU	0.067	0.075**	0.012	-0.001
	[0.108]	[0.029]	[0.013]	[0.008]
Post transposition of EU Directive	0.193*	0.097	-0.050	-0.029
	[0.100]	[0.064]	[0.032]	[0.017]
Observations	378	324	378	324
R-squared	0.365	0.714	0.398	0.626
Control mean (2019)	0.576	0.576	0.163	0.163
Number of countries	27	27	27	27
Coeff: T + T*Mediator = 0	-0.00286	0.0239	0.0349	0.0171
P-value: T + T*Mediator = 0	0.966	0.444	0.0353**	0.114
Interest in EU politics (Eurobarometer 2018)				
Early transposition	-0.012	-0.006	0.021	0.016
	[0.053]	[0.024]	[0.014]	[0.010]
Early transp. * High interest in EU politics	-0.014	-0.010	0.010	0.005
	[0.108]	[0.036]	[0.014]	[0.010]
Post transposition of EU Directive	0.200**	0.097	-0.103**	-0.036**
	[0.097]	[0.063]	[0.044]	[0.016]
Observations	436	375	436	375
R-squared	0.307	0.707	0.377	0.633
Control mean (2018)	0.699	0.699	0.184	0.184
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	-0.026	-0.0165	0.0308	0.0206
P-value: T + T*Mediator = 0	0.808	0.621	0.0402**	0.0285**
Understanding of how the EU works (Eurobarometer 2018)				
Early transposition	-0.052	-0.022	0.023	0.015
	[0.057]	[0.023]	[0.014]	[0.010]
Early transp. * High understanding of EU	0.068	0.023	0.004	0.006
	[0.086]	[0.030]	[0.016]	[0.009]
Post transposition of EU Directive	0.192*	0.095	-0.102**	-0.035**
	[0.103]	[0.063]	[0.044]	[0.016]
Observations	436	375	436	375
R-squared	0.309	0.708	0.376	0.633
Control mean (2018)	0.563	0.563	0.220	0.220
Number of countries	32	32	32	32
Coeff: T + T*Mediator = 0	0.0162	0.0015	0.0274	0.021
P-value: T + T*Mediator = 0	0.855	0.96	0.0807*	0.0214**
EU candidates included	YES	YES	YES	YES
Country controls	YES	YES	YES	YES
Two lags of dependent variable included	NO	YES	NO	YES

Note: The table reports heterogeneity tests examining whether the transposition effect of the EU Whistleblower Protection Directive on country-level perceptions of corruption (using the WGI and V-Dem indices) varies by baseline interest and trust in EU institutions. All models include country and year fixed effects and the full list of time-varying country controls. Models (2) additionally control for two lags of the dependent variable. Standard errors (in brackets) are clustered by country. P-values are reported from both cluster-robust and wild bootstrap tests with 20,000 replications. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Chapter 2

Combating corruption through (mis)perception: Integrity training and pluralistic ignorance in Ukraine

This paper examines the impact of formal integrity training on reducing corrupt behaviour among intermediaries in a culture with widespread rule-breaking. We collaborated with the USAID New Justice Program to offer an integrity training course to undergraduate law students in Ukraine. Using a randomized design, students were assigned to either receive the training or not. Afterwards, all students participated in a bribery game, where they acted as intermediaries in a potentially corrupt transaction. Some students also received information about their peers' high participation in the integrity training. Results show that the training alone did not significantly affect perceptions about corruption or corrupt behaviour. However, students exposed to the information treatment — regardless of whether they themselves received the training — mistakenly believed that corruption would be less common among trained peers and were thus significantly less likely to engage in it. These findings suggest a silver lining of pluralistic ignorance: perceiving that a training program shifts social norms, even when it does not, can still generate positive behavioural change

[0]This chapter is joint work with Oana Borcan, Stephanie Heger, and Amrish Patel.

2.1 Introduction

Dishonest, fraudulent, and corrupt practices impose heavy societal costs. Globally, corruption is estimated to cost more than $3.6 trillion annually — approximately 5% of global GDP (World Economic Forum, 2018). In recent global surveys, roughly 1 in 4 adults worldwide reported having paid a bribe to access public services (Transparency International, 2017). A significant portion of corruption is facilitated by intermediaries, with 75% of cases involving a "middle man" (OECD, 2020). This trend is alarming because intermediaries exacerbate corruption by reducing transaction costs (Bayar, 2005), lowering the risk of detection (Hasker and Okten, 2008), and diminishing the moral costs associated with bribery (Drugov et al., 2014). Moreover, detecting and punishing intermediaries is particularly difficult in countries where rule-breaking is deeply embedded in the culture.

To address this, many organizations advocate for formal integrity education programs, but little is known about their effectiveness (OECD, 2018). While early research on integrity programs focused on self-reported measures of moral reasoning and ethical judgment (Shaub, 1994; Desplaces et al., 2007), more recent research examines the effects on actual behaviour although with mixed findings. The impacts on behaviour in the lab are short-lived or are not reflected by real-world outcomes (Mayhew and Murphy, 2009; Banerjee and Mitra, 2018; Harris et al., 2022). Although integrity training may shape individual moral preferences or attitudes, depending on content and delivery, it is unclear whether this carries over to behaviour embedded in the social context, where beliefs and pressures from peers and authority may run counter to the desired actions. This raises the question of whether behaviour is shaped by integrity training interacted with social norms, which can play a key role in supporting corruption (Rothstein, 2000; Köbis et al., 2015; Abbink et al., 2018).

This paper addresses this question by examining the impact of formal integrity training and information about training prevalence on corrupt behaviour among intermediaries in Ukraine, a culture where rule-breaking is widespread — nearly 70% of Ukrainians report making informal payments to healthcare workers, education officials and law enforcement agents (Kiev International Institute of Sociology,

2015).[1] In collaboration with the USAID New Justice Program, we developed and implemented a program called "Training in Ethics, Integrity, and Anti-Corruption for Law Students" (referred to as "integrity training") for undergraduate law students in Ukraine in May 2021.[2] The program was based on the university modules on ethics, integrity, and anti-corruption developed by the United Nations Office on Drugs and Crime (UNODC), adapted by a team of national experts who delivered the training. To evaluate the program's effectiveness, we randomly assigned law students interested in the training into two groups: those who received an invitation to participate in the course and those who did not (our control group). Due to COVID-19, the course was delivered online.

Two months after the course, all participants were invited to participate in an online bribery game, where they played the role of intermediaries and decided whether to facilitate a corrupt transaction between a firm and a public official.[3] We then introduced a second intervention: an information treatment in which both trained and untrained students were randomly assigned either to receive information about their peers' high participation in the integrity training or to receive no additional information. Importantly, the information treatment did not provide explicit social norms about corruption but was intended to test whether participants' perceptions of the training prevalence affected their behaviour.

To structure our thinking on how integrity training and information impact players' decisions, we develop a dynamic game with incomplete information that models a corrupt transaction between a firm and a public official facilitated by an intermediary. Intermediaries decide whether to facilitate the transaction — by informing the firm about the size of the bribe demanded — and whether to report the true amount or inflate it to embezzle part of the payment. Their choices depend

[1]Ukraine ranks 104 out of 180 countries in the 2023 Transparency International Corruption Perception Index (CPI), due to endemic corruption (Denisova-Schmidt and Prytula, 2017).

[2]The United States Agency for International Development (USAID) New Justice Program was implemented in Ukraine from October 2016 to September 2021. The program was designed to support the Judiciary, the Government, the Parliament, the Bar, Law Schools, Civil Society, Media and Citizens to create the conditions for independent, accountable, transparent, and effective justice system that upholds the rule of law and to fight corruption. The program has been succeeded by the USAID Justice for All Activity in October 2021, which continues to actively support the Ukrainian Government and its non-governmental partners.

[3]Previous research has also studied the role of intermediaries on facilitating corrupt behavior (Abbink et al., 2002; Barr and Serra, 2010; Drugov et al., 2014). Unlike prior work where intermediaries were passive participants, we place their decisions at the center of the analysis.

on their moral costs and their beliefs about the average bribe demands of peers. The model predicts that higher moral costs, as well as beliefs that peers reduced bribery, lower both the likelihood of facilitating corruption and the size of the bribe demanded. Our interventions aimed to increase the moral costs of corruption and shift beliefs about prevailing norms.

Our experimental results suggest that the integrity training alone had little impact on reducing corrupt behaviour. We hypothesized that training would reduce corrupt behaviour by increasing awareness of the costs of corruption and by increasing the moral costs of engaging in corruption. To investigate these channels, we implemented a pre- and post- training survey which collected subjects' perceptions of corruption, and the moral foundations underlying their decisions and judgements (Haidt, 2007).[4] We find that training alone has no effect on their perceptions of corruption, attitudes about corruption, or the moral underpinnings of their decisions.

We also elicited subjects' beliefs about other subjects' behaviour. We find that subjects who received information about their peers' participation in the training *believed* that corruption would be less common and were thus also less likely to engage in corrupt behaviour to align with their misperceived social norm; that is, we document a form of pluralistic ignorance, a phenomenon by which misperceptions of the norm leads to behaviour consistent with the misperceived norm (Cialdini et al., 1990; Bicchieri, 2016; Tankard and Levy Paluck, 2016). Notably and similar to Harris et al. (2022), we find that this effect was driven by first-year law students (Junior students), who were less likely to facilitate corruption when informed about the prevalence of training (and also embezzled slightly less, conditional on facilitating). In our setting, the results suggest that younger students may be more susceptible to these (mis)perceived social norms. While training alone did not change the attitudes or moral considerations of the Junior students, information about the prevalence of such training effectively shifted perceptions, fostered more ethical norms, and reduced corrupt behaviour.

Our study contributes to three main strands of literature. First, we expand

[4]The Moral Foundations Theory measures how important various moral considerations are when people decide whether a behaviour is right or wrong. The five dimensions are: harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, purity/sanctity.

an emerging economics literature on the impacts of ethics-based interventions on behaviour. Very few studies have addressed such questions experimentally. Banerjee and Mitra (2018) find that an ethics module among Indian MBA students had a short-term effect on the likelihood of demanding a bribe, but no effect on bribe size. Mayhew and Murphy (2009)'s quasi-experimental study reveals no direct impact of training on students' behaviour. Harris et al. (2022) find a long-term positive effect of integrity training on the attitudes to corruption and individual behaviour of Ghanian traffic police officers, but the impact on field outcomes that are influenced by corruption (such as traffic violations sent to court) fades after 3 months. The effects were driven entirely by junior police officers, whose intrinsic motivation to serve were reactivated by the training (pointing to the role of moral costs). However, the disappearing effects on behaviour on the job raise questions regarding the limitations of integrity training when actors operate in a system of corrupt social norms. Our paper highlights that such programs may require several complementary approaches: (1) a training component to increase awareness of how corruption harms society, and (2) spreading information about the program to shift social norms surrounding corrupt practices.

Second, we contribute to a growing literature on social norms and corruption by highlighting the role of expectations about others' behaviour in shaping corrupt practices. Social norms reflect shared beliefs about what others do (descriptive norms) and what they approve of (injunctive norms), and these expectations can strongly influence behaviour. Individuals may engage in corruption not because they endorse it, but because they perceive it to be widespread or socially expected. Experimental and survey evidence shows that higher perceived prevalence of corruption is associated with higher perceived social acceptance of such behaviour and greater engagement in corrupt acts (Abbink et al., 2018; Köbis et al., 2015; Tankard and Levy Paluck, 2016; Köbis et al., 2022). Importantly, such expectations may be systematically misperceived. An emerging literature documents how pluralistic ignorance in a variety of contexts drives societies to undesirable outcomes such as support for racial segregation in the U.S. (O'Gorman, 1975), support among Saudi men for women working outside the home in Saudi Arabia (Bursztyn et al., 2020), the lack of action against climate change in the U.S. (Andre et al., 2024), and gender gaps in labour supply (Boneva et al., 2024). By contrast, we document a "positive"

side of pluralistic ignorance — students who learned about the high incidence of integrity training among their peers *overestimated* the effectiveness of the training and thus behaved more ethically to conform to the misperceived social norm.

Third, we contribute to a growing but still limited literature on the role of intermediaries in corruption. While anecdotal evidence suggests that middlemen are ubiquitous, economic research on their role remains scarce. Existing work shows that intermediaries can facilitate corruption by reducing uncertainty about whom to bribe and how much to pay, lowering detection risks, and diminishing the moral or psychological costs associated with corrupt exchanges (Bayar, 2005; Hasker and Okten, 2008; Drugov et al., 2014). Closest to our study, Drugov et al. (2014), experimentally show that the presence of intermediaries increases corruption by reducing the moral costs faced by both bribers and public officials. In their design, however, intermediaries play a passive role. In contrast, we study intermediaries as active decision-makers who can choose whether to facilitate corrupt transactions. We also provide, to our knowledge, the first test of an integrity intervention on incentivized corruption decisions among a natural population of future intermediaries—law students. By targeting intermediaries directly, our study highlights an important but understudied channel through which corruption can be sustained or curtailed.

The rest of the paper is organized as follows. Section 2.2 describes the game, the interventions and the experimental design. Section 2.3 describes the data. Section 2.4 outlines the main results and section 2.5 concludes.

2.2 Methods and procedures

In this section, we develop and characterize the equilibrium of a bribery game with an intermediary (Section 2.2.1). This game informs our experimental design in the field, which is meant to test how integrity training and information about the prevalence of integrity training among peers affect the behaviours of the intermediaries in the game (Section 2.2.2). Finally, based on the equilibrium predictions of the bribery game and the experimental design, we propose a set of hypotheses to be tested with the data (Section 2.2.3).

2.2.1 Bribery game with Intermediary

The set of players comprises an Intermediary (I), a Firm (F), a Public Official (PO) and Society (S). Their endowments are: Y_I, Y_F, Y_{PO} and Y_S, respectively.

The Firm is awarded a contract of value $v > 0$ with probability $p \in (0, 1)$. However, if the Firm pays the Public Official's Minimal Acceptable Bribe (MAB), then they are awarded the contract with certainty. The MAB is privately known by the Public Official and the Intermediary (we assume a common prior uniformly distributed over $[\underline{MAB}, \overline{MAB}] \in \mathbb{N}^+$). The Firm can only offer a bribe via the Intermediary.

In Stage 1, the Intermediary decides whether to facilitate a bribe between the Firm and the Public Official. If the Intermediary does facilitate, they can report the MAB truthfully ($b = \text{MAB}$) to the Firm or over-report ($b > \text{MAB}$), the latter case implying embezzlement. If the Intermediary does not facilitate, the Firm cannot pay a bribe (neither corruption nor embezzlement occur) and the game ends.

If the Intermediary plays "facilitate", the game moves to Stage 2 where the Firm decides whether to pay the bribe or not pay. If the Firm pays (i.e., corruption occurs), then the MAB is transferred to the Public Official, the Intermediary retains $b - \text{MAB}$ and Society suffers an externality cost of $E > 0$. If the Firm does not pay (i.e., corruption does not occur), they may still win the contract with probability p, and the Public Official, the Intermediary and Society receive their endowments.

In addition to the material incentives modeled above, we also capture two psychological incentives: players' moral and conformity incentives. The Firm and the Intermediary incur moral costs if they engage in corruption, reflecting feelings of guilt or awareness of acting immorally (Rose-Ackerman, 1978; Della Porta and Vannucci, 1999; Drugov et al., 2014). Moral costs thus capture the intrinsic disutility from engaging in corruption and are independent of others' behaviour. In our setting, the Firm incurs a cost if they pay a bribe, while the Intermediary incurs a cost if they inform the Firm and thus facilitate the corrupt transaction. Let m_I and m_F be each player's respective, privately known marginal moral cost (we assume uniform priors over their respective domains, $[0, \overline{m_I}]$ and $[0, \overline{m_F}]$). The Firm's moral cost is m_F multiplied by the damage to Society (E), while the Intermediary's moral cost is m_I multiplied by the sum of E and the amount embezzled ($b - \text{MAB}$).

The Intermediary is also assumed to have social conformity concerns $c \geq 0$, reflecting a preference for aligning their behaviour with that of others in their reference group (Akerlof, 1980; Bernheim, 1994; Cialdini and Goldstein, 2004; Bicchieri and Xiao, 2009; Tankard and Levy Paluck, 2016). Social conformity concerns thus capture the disutility from deviating from perceived expectations about others' behaviour, as individuals may experience discomfort or social pressure when their behaviour differs from what others are perceived to do or consider acceptable. In our setting, these concerns operate through the Intermediary's beliefs about others' behaviour, captured by the expected average bribe demanded by the reference group (i.e., $\tilde{b}$). If $c > 0$, the Intermediary suffers a social conformity cost that depends on how different their bribe ask b is from the average bribe they believe other Intermediaries ask for (i.e., $\tilde{b}$), which we model as a quadratic loss function (i.e., $(b - \tilde{b})^2$).

The game is depicted in Figure 2.1. As this is a dynamic game with incomplete information, we use Perfect Bayesian Equilibrium as a solution concept. Since our primary objective is to obtain comparative statics to motivate a set of hypotheses, we will focus only on interior solutions.[5]

Proposition 2.1. *There exists a Perfect Bayesian Equilibrium described by the following strategy profile (and associated beliefs).*

- *The Intermediary facilitates a bribe iff $m_I \leq m_i^*$, where m_i^* solves:*

$$m_i^* = \frac{\left[v(1-p) - b^*\right] / (E\overline{m_F})\,(b^* - MAB) - c(b^* - \tilde{b})^2 + c(\tilde{b}^2)}{(b^* - MAB + E)}$$

- *Conditional on facilitating a bribe, the Intermediary asks for the optimal bribe:*

$$b^*(m_I, MAB, \tilde{b}) = \frac{v(1-p) + MAB - m_I E\overline{m_F} + 2c\tilde{b}E\overline{m_F}}{2(1 + cE\overline{m_F})}$$

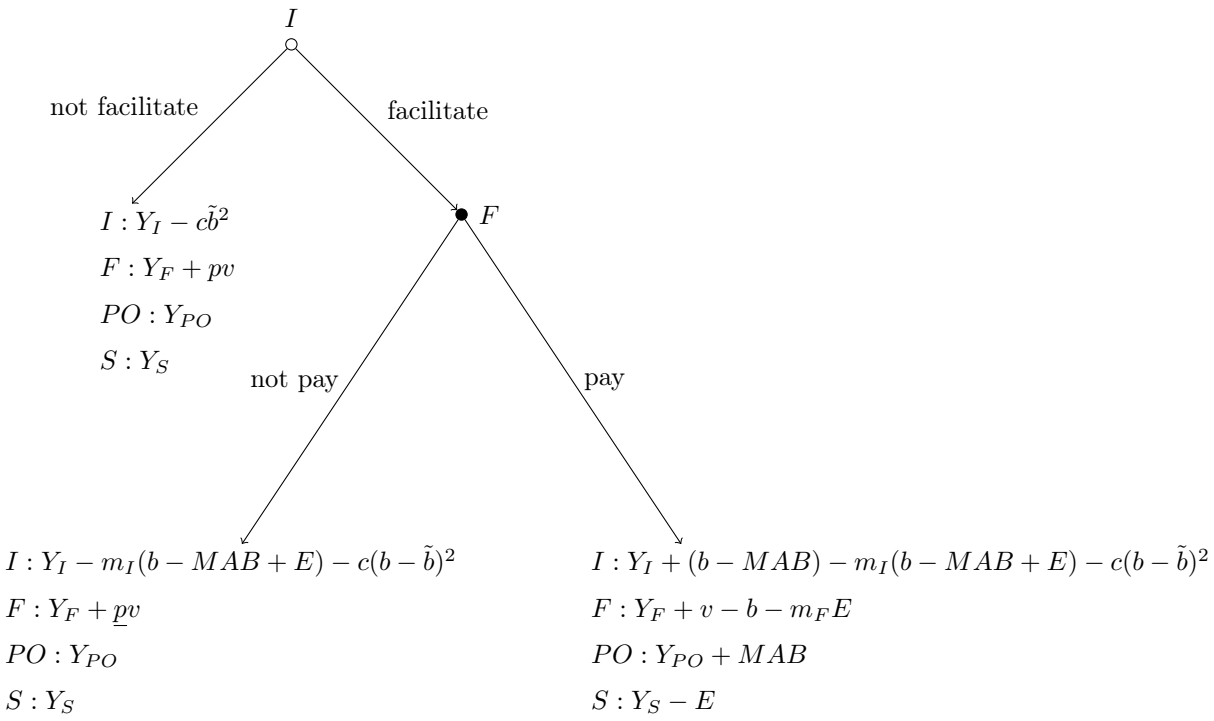
- *The Firm pays the bribe if:*

$$m_F \leq \frac{v(1-p) - b}{E}$$

[5] Our characterization does not solve for $\tilde{b}$, as the comparative statics motivate some of our experimental hypotheses.

and does not otherwise.

Figure 2.1: Bribery game tree



Proof. All proofs are found in Section 2.A in the Appendix. □

In summary, we have a closed form solution for the optimal bribe ask, b^*, which intuitively depends on exogenous variables. Its value is larger the larger are v, MAB and $\tilde{b}$, and the smaller are p and E. Moreover, the Intermediary decides to facilitate a bribe and ask for b^* only if its moral cost is sufficiently low ($m_I \leq m_i^*$), since only then does his expected monetary benefit outweigh his expected moral cost.[6]

[6] Analogously, the Firm pays the bribe only if its moral cost is sufficiently small ($m_F \leq \frac{v(1-p)-b}{E}$).

2.2.1.1 Comparative statics

Our research question asks whether integrity training and information about the prevalence of training among peers will cause Intermediaries to behave more ethically. In terms of our model, we can think of integrity training as increasing the Intermediary's marginal moral cost. This can be modelled as an increase in $\overline{m_I}$. Receiving information that a large share of peers were trained can be thought of as decreasing the Intermediary's belief of others' average bribe ask, $\tilde{b}$.

An increase in $\overline{m_I}$ and a decrease in $\tilde{b}$ have the following effects on equilibrium behaviour:

Corollary 2.1. *An increase in $\overline{m_I}$ reduces the expected bribe asked for by the Intermediary (conditional on informing).*

Corollary 2.2. *A decrease in $\tilde{b}$ reduces the expected bribe asked for by the Intermediary (conditional on informing).*

Corollary 2.3. *An increase in $\overline{m_I}$ reduces the probability that the Intermediary will inform.*

Corollary 2.4. *A decrease in $\tilde{b}$ reduces the probability that the Intermediary will inform.*

Corollary 2.5. *An increase in $\overline{m_I}$ reduces the ex-ante probability of the corrupt transaction taking place. The reduction is larger the larger is E and the smaller are p and v.*

Corollary 2.6. *A decrease in $\tilde{b}$ increases the ex-ante probability of the corrupt transaction taking place.*

2.2.2 Experimental design

Our field experiment has three main stages. First, the recruitment and baseline survey. Second, the training program. Third, the endline survey that included the experimental bribery game. We describe each stage below.

2.2.2.1 Baseline: Recruitment and survey

We collaborated with the USAID New Justice Program, who provided support with translation, advertising campaign and recruitment of participants for the study across 63 local universities. Interested students were invited to read the information received via email about the project and to consent to participate to take part in the integrity training and the corresponding study (Figure 2.B.1). We concluded the recruitment of law students in early May 2021, reaching 496 registrations. Of these, about half (242 randomly selected students) were offered the chance to participate in the integrity training and the remaining 254 students were assigned to the control group. All participants were informed of the opportunity to earn money (framed as gifts) for completing surveys.

Along with the consent form, we collected some basic characteristics of the participants: gender, age, university, year of study and main reason for participating (skills, curiosity for behavioural experiments, interest in gifts, or other reasons). Shortly after, we emailed the enrolled participants a more extensive baseline survey: Haidt's Moral Foundations, including questions pertaining to five considerations for moral judgments: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Purity/Sanctity, questions on perception and justifiability of corruption, and further personal characteristics, including income and career aspirations (Figure 2.B.2). The baseline survey was completed by 323 students, approximately equally split between the assigned treatment (49%) and control group (51%).[7] The survey was designed to generate baseline measures of students' moral values, as well as their perceptions of corruption in Ukraine.

Additionally, in early June — about two weeks after the integrity training — we sent all study participants a short quiz unrelated to the training as a way to reduce attrition. We received 214 complete responses.

2.2.2.2 The Integrity training

The "Training in Ethics, Integrity and Anti-Corruption for Law Students" program (henceforth, the integrity training) was co-designed in March-April 2021 with the

[7] Among the 242 law students randomly assigned to the treatment group, 158 (65%) completed the baseline survey. Of the 254 students assigned to the control group, 165 (65%) completed the survey.

support of the USAID New Justice Program and their higher education partners in Ukraine. The intervention was particularly relevant to the Ukrainian context, as the country ranked 122nd in 180 countries on Transparency International's Corruption Perception Index (CPI), due to widespread corruption across the public sector (Denisova-Schmidt and Prytula, 2017). A 2015 national survey showed reported bribes were commonplace in health care (70% of respondents), secondary schools (64%), police (51%) and higher education staff (49%) (Kiev International Institute of Sociology, 2015).

We based our integrity and anti-corruption training on a series of peer-reviewed University modules (14 modules on integrity and ethics and 13 modules on anti-corruption) developed in 2018 by the United Nations Office on Drugs and Crime (UNODC) in consultation with academic experts from over 30 countries.[8] These modules aim to enhance students' ethical awareness and understanding of corruption, equipping them with skills to behave with integrity. The modules cover both theoretical and practical perspectives, and use interactive teaching methods such as experiential learning and group-based work to engage students' critical thinking and problem solving. To the best of our knowledge, the effectiveness of such modules has not yet been experimentally evaluated.

To adapt the UNODC modules to the local context, we conducted two focus group discussions in March-April 2021 with representatives from 17 Ukrainian law schools, law students and practitioners. The discussions revealed specific challenges in combating corruption. First, Ukrainian law degrees offered ethics training of varying quality and over-focused on theory. Second, ethics modules were often led by faculty whose moral stance undercut their credibility. Third, absent role models and an abundance of stories of corrupt leaders' success led students to equate success with dishonesty and selfishness.

Our training integrated these insights, focusing on the UNODC themes like Challenges to Ethical Living, Ethics in Business and in Law, What is Corruption, Public and Private Sector Corruption. Over two weeks in May 2021, our team of

[8]These University modules have been developed as part of the UNODC Education for Justice (E4J) program, currently the GRACE initiative, following the Doha Declaration, which promotes education at all ages as a means to combat crime and corruption, and build a culture of lawfulness. They can be accessed at https://grace.unodc.org/grace/en/academia/module-series-on-integrity-and-ethics.html.

four Ukrainian instructors (experts in Law and Anti-corruption),[9] delivered six two-hour sessions via Zoom due to Covid-19 restrictions. These included two lectures focused on principles of ethics and anti-corruption, and four interactive workshops including group work, discussion boards (e.g., Padlet), case study presentations and Q&A sessions by guest legal experts with a national reputation as champions of integrity.[10] Their objective was to equip students with practical strategies to address everyday challenges to ethical behaviour and to demonstrate that career success is compatible with ethical behaviour. All teaching was conducted in Ukrainian.

While there was no formal assessment, students who attended at least 75% (four) of the sessions were awarded a participation certificate from the University of East Anglia and the USAID New Justice Program. Students were also informed about this prior to starting training, and that participation in the training was capped at 200 participants and had to be decided by a random draw.[11] Of the 242 randomly selected students who were invited to the training, 145 students attended at least one out of six sessions. In total, 110 students completed the minimum requirement for receiving participation certificates (i.e., four out of six sessions).

Post-training debriefing sessions with the instructors and student feedback revealed students were active (around 20% were consistently participating in conversations, with the rest participating in the chat or group discussions). The students particularly appreciated the presence of guest speakers, who illustrated that career success is compatible with moral integrity. Students anonymously rated the training 4.77 (out of 5) despite the intense pace and online format.

[9]The instructors were: Nataliya Gutorova (Professor in Criminal Law the Poltava Law Institute, Yaroslav Mudryi National Law University), Oleh Herasymchuk (Associate Professor of Law at the National University Ostroh Academy), Ms. Oleksandra Keudel (Post-Doctoral Fellow in Political Science at the Free University of Berlin) and Ms Iryna Shyba (Deputy head at EU Anti-Corruption Initiative and executive director of DEJURE Foundation).

[10]Invited guest speakers included Mr Roman Maselko (on behalf of Judge Sergiy Bodnarenko), Ms Antonina Prudko (Head of Secretariat at UNIC, Ukrainian Network of Integrity and Compliance), Mr Artem Krykun-Trush (Attorney specialized on white-collar crime, compliance and investigations practices at DLA Piper), Mr Oleg Klimov (President of the All-Ukrainian Pharmaceutical Chamber), Ms Viktoria Kozachenko (Head of the Integrity Office/National Agency on Corruption Prevention NAZK).

[11]For fairness considerations, interested non-participants were given access to the asynchronous online integrity training resources upon request, once the study was completed.

2.2.2.3 Endline: Bribery game and final survey

To minimize experimenter demand effects and capture medium-term effects of training on behaviour, the experiment was played in July, more than six weeks after the conclusion of the integrity training. We operationalized the game described in Section 2.2.1, which mimics the procurement process of public contracts. Like the training, the experiment was run online due to Covid-19 restrictions. All participants read the same instructions so that the set of actions for each player was common knowledge (see Figure 2.B.2 for the full instructions). Moreover, we chose to conduct a framed experiment to provide rich context and to make the decisions salient to both law students, who played the role of Intermediaries, and to non-law students, who played the role of Firms, Public Official and Member of Society (Charness et al., 2007). To mitigate experimenter demand effects, all choices were incentivized and made anonymously (Zizzo, 2010; De Quidt et al., 2018). Moreover, we generally avoided using morally loaded terms, opting for neutral wording such as "payment" instead of "bribe". One notable exception is the information treatment, which stated that the participant was "part of a group in which 75% of Intermediaries have just received an Integrity, Ethics and Anti-Corruption training". Although this information was factual and intended to shift participants' perceptions about the prevalence of training among peers, the use of terms such as "anti-corruption" may have not only affected beliefs about others' behaviour, but also indirectly conveyed cues about what behaviour is considered appropriate. As we did not directly elicit normative expectations in our design, we cannot empirically isolate this mechanism. Nevertheless, we expect that normative expectations — when activated — may reinforce empirical expectations and jointly shape behaviour, rather than operating fully independently.

In the experiment, there are four players — the Firm, an Intermediary, a Public Official and a Member of Society — who are randomly matched and partake in a one-shot interaction. Each of the four players has an initial endowment of 70 UAH (about 2.55 USD in July 2021). The Public Official and the Member of Society do not make any decisions in the game, but their final payoffs are impacted by the decisions of the Firm and the Intermediary. The Firm maximizes his payoff by winning a contract from the Public Official that is worth 50 UAH to the

Firm. Public Officials are willing to accept a bribe from the Firm. The minimum acceptable bribe for the Public Official is a randomly determined private value $\in \{5\text{UAH}, 10\text{UAH}, 15\text{UAH}, 20\text{UAH}, 25\text{UAH}, 30\text{UAH}, 35\text{UAH}, 40\text{UAH}\}$.[12] The experiment unfolds in four stages.

Stage 1 Intermediaries are randomly assigned to one of two information conditions: truthfully informed that they are part of a group in which 75% of Intermediaries have received an Integrity, Ethics and Anti-Corruption training or no additional information. This novel approach of manipulating descriptive social norms in the lab — already proposed by Krupka and Weber (2009); Bicchieri and Xiao (2009); Abbink et al. (2018) — allowed us to control what subjects believed about the pervasiveness of integrity training among peers — and, in turn, their expectations about others' typical behaviour (Abbink et al., 2018). Thus, we have four treatment conditions: (1) Control, (2) Training Only, (3) Information Only, and (4) Training + Information.

Stage 2 The Intermediary learns the Public Official's minimum acceptable bribe that the Firm can pay to ensure they win the contract. Without the bribe, the Firm will win the contract with a 5% chance.

Stage 3 The Intermediary decides whether to facilitate a bribe by informing the Firm of this confidential information. Importantly, the Intermediary can inform the Firm of a number equal to or higher than the true minimum acceptable bribe. For example, if the Intermediary learns that the Public Official requires a bribe of 5 and the Intermediary chooses to facilitate a bribe, the Intermediary can inform the Firm of an amount equal to or greater than 5 (but not smaller) by choosing from the following set: 5 UAH, 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH, 40 UAH. If the Intermediary chooses to inform the Firm of an amount greater than the minimum acceptable bribe and the Firm chooses to pay, then

[12]The minimum acceptable bribe is exogenously and randomly determined to introduce variation in the cost of corruption across interactions while maintaining experimental control. This randomization also implements the common prior assumed in the model, ensuring that minimum bribe levels are uniformly distributed over $[\underline{MAB}, \overline{MAB}] \in \mathbb{N}^+$. Allowing the Public Official to choose the bribe would have resulted in a single common threshold for all Intermediaries in our design, as only one Public Official was present in the game, and would have introduced additional strategic considerations beyond the focus of our study.

the Intermediary passes the minimum acceptable bribe to the Public Official and embezzles the difference.

Stage 4 If the Intermediary "Facilitates" a bribe, then the Firm can act on that information and "Pay" the Public Official or "Not Pay".[13] If the Firm "Pays", the Firm wins the contract for sure and the Member of Society incurs a cost of 35 UAH from their initial endowment to mimic the negative externality associated with corruption. There are no partial bribes. If the Firm decides to Pay the Public Official, the Firm must transfer the full amount reported by the Intermediary. If the Firm does "Not Pay", the Firm will win the contract with a 5% chance. However, if the Intermediary chooses "Not to Facilitate" a bribe, then the Firm has a 5% chance of winning the contract.

At the conclusion of the experiment, we randomly matched Firms and Intermediaries to calculate the earnings from the activity for all four roles. Given our experimental design, we did not need as many Firms, Public Officials and Members of Society as Intermediaries. In terms of the actual participation in the experiment, we had 252 Intermediaries play. In addition to this pool of participants, for the experiment we independently enrolled a further 82 students from other degrees in Ukraine to play the role of Firms in the game, of which 40 actually played the game and were paid based on the outcome of one randomly chosen Intermediary match. Finally, 1 Public Official and 1 Member of Society were also paid based on a random pairing with an Intermediary.

In September 2021, we sent an invitation to a final (endline) survey which concluded in October 2021, with 186 responses in total, equally split between the treatment (48%) and control group (52%). The endline questionnaire was identical to the baseline survey, covering moral foundations, corruption perceptions and justifiability (Figure 2.B.2). Overall, 158 students completed both the baseline and endline surveys, and participated in the experiment.

[13]Due to Covid-19 and the asynchronous online format, we elicit the Firm's choices via the strategy method, i.e., they are asked to make a choice between Pay and Not Pay for each potential Informed Payment Amount.

2.2.3 Hypotheses

Our model posits two channels through which the behaviour of the intermediaries may be affected: the moral costs of engaging in corrupt behaviour and the beliefs about other intermediaries' behaviour. Our experiment is designed to exogenously vary both. Moral costs are shifted through the random assignment of the integrity training course. Beliefs are shifted through the random assignment of information about the prevalence of training among peers, which affects participants' expectations about others' behaviour. Combined, our model and experimental design generate two sets of hypotheses.

Hypothesis 2.1. *Integrity training will increase the moral cost of facilitating a bribe and thus will*

1. *reduce the likelihood that an intermediary facilitates a bribe;*
2. *and conditional on facilitating a bribe, will reduce the amount the intermediary embezzles.*

Hypothesis 2.2. *Information about the high prevalence of integrity training will reduce the intermediary's beliefs about the likelihood of other intermediaries facilitating a bribe and thus will*

1. *reduce the likelihood that an intermediary facilitates a bribe;*
2. *and conditional on facilitating a bribe, will reduce the amount the intermediary embezzles.*

To test these hypotheses, we will focus on estimating Intention-To-Treat effects using the following specification

$$\begin{aligned} Y_i = & \\ & \beta_0 \quad +\beta_1 \mathbf{1}[\text{Training Only Treatment}] \\ & \quad\;\; +\beta_2 \mathbf{1}[\text{Information Only Treatment}] \\ & \quad\;\; +\beta_3 \mathbf{1}[\text{Training + Information Treatment}] \\ & \quad\;\; +\beta_4 \Psi_i + \varepsilon_i \end{aligned} \tag{2.1}$$

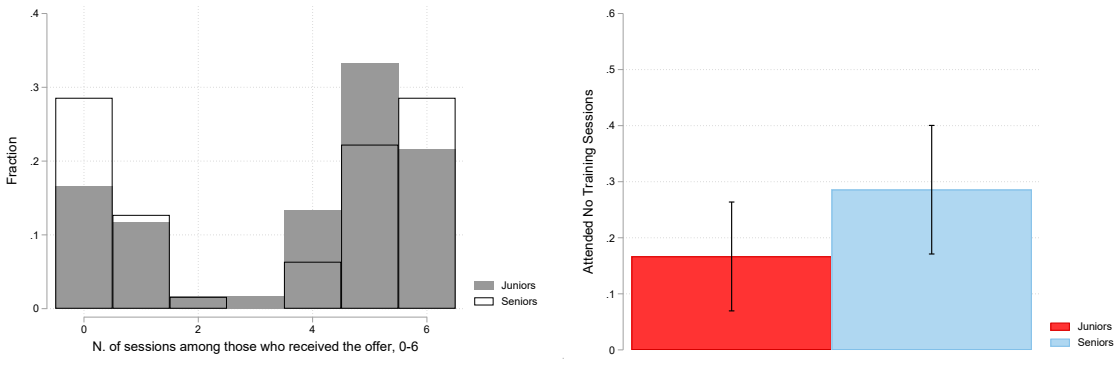
where Y_i is the outcome variable of subject i (e.g., Y_i = Prob[Facilitate a bribe], Y_i = Belief about other intermediaries' behaviour) which we regress on treatment

dummies ($\beta_1, \beta_2, \beta_3$) and a vector of individual characteristics (β_4). The omitted group is the Control treatment unless otherwise specified.

2.3 Data

Our experiment employs a between-subject 2×2 design, in which we randomly vary exposure to the integrity training and the information treatment. As our objective is to estimate Intention-to-Treat (ITT) effects, we first focus on assignment to the training offer, regardless of actual participation. Of the 496 law students who initially agreed to participate in the study, 242 were randomly assigned to receive an invitation to the training, while 254 were not. Among those invited, 145 students (60%) attended at least one of the six sessions, and 110 attended at least four sessions, thereby meeting the minimum requirement for receiving a participation certificate. Figure 2.2a shows the distribution of training sessions attended among students assigned to the training by cohort (Juniors = first-year students; Seniors = students in later years), while Figure 2.2b compares non-compliance rates across cohorts. In both figures, Senior students appear more likely to attend zero training sessions, although this difference is not statistically significant.

Figure 2.2: Compliance with training treatment



(a) Training sessions attended

(b) No sessions

Note: Figure 2.2a shows the distribution of training sessions attended for subjects assigned to receive training. Figure 2.2b shows rate of non-compliance for Junior and Senior students. Senior students were 12 percentage points more likely to attend zero training sessions (two-sided Mann-Whitney test: p-value=0.12).

The distinction between Junior and Senior students, which we revisit throughout the analysis, is motivated by prior evidence suggesting that individuals at earlier stages of their professional development may respond differently to ethics-based interventions (Harris et al., 2022). In our context, this distinction reflects differences in stages of socialization (Weidman, 1989). Junior students, as first-year entrants, are in a transition phase between prior and new social environments. While they may still be influenced by norms and expectations formed before university, they are newly exposed to the university setting, where they encounter new information through both formal instruction and observation of peers' behaviour. This stage is typically characterised by ongoing identity formation and the updating of beliefs about appropriate conduct as students adjust to new reference groups. In contrast, Senior students have had longer exposure to the same institutional and peer environment and are therefore more likely to have internalized its prevailing norms and developed more stable beliefs. As a result, they may be less responsive to new information or interventions that seek to shift moral preferences, beliefs and norms. At the same time, differences between these groups may also reflect selection, as students who progress further in their studies may differ systematically in their attitudes or motivations. While our design does not allow us to disentangle these mechanisms, this distinction provides a useful lens to explore heterogeneous treatment effects.

Turning to participation in the bribery game, 252 Intermediaries took part. Of these, 123 were offered the training, of whom 95 (77%) attended at least one session. Within the trained group, 61 also received the information treatment, while 62 did not. Among the 254 students who were not offered the training, 129 participated in the game; of these, 64 received the information treatment and 65 did not (see Table 2.1). In addition, we independently recruited a further 82 students from other degree programs in Ukraine to play the role of Firms, of whom 40 ultimately participated in the game.

2.4 Main results

In this section, we report results from the framed field experiment. We are interested in two behavioural measures — (1) whether the intermediary was willing to

Table 2.1: Summary statistics by treatment assignment

	All	Control	Training Only	Information	Training + Information
Prior ethics training	0.37	0.36	0.32	0.42	0.44
	(0.48)	(0.48)	(0.47)	(0.50)	(0.50)
Junior	0.45	0.42	0.44	0.39	0.54
	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)
Female	0.71	0.74	0.61	0.75	0.75
	(0.45)	(0.44)	(0.49)	(0.44)	(0.43)
Enroll for Skills	0.85	0.88	0.79	0.91	0.87
	(0.35)	(0.33)	(0.41)	(0.29)	(0.34)
Enroll for Prize	0.00	0.00	0.00	0.00	0.02
	(0.06)	(0.00)	(0.00)	(0.00)	(0.13)
Observations	252	65	62	64	61

Note: This table displays means and standard deviations of the main control variables used in the analysis. We collected additional data that we use in Section 2.4.4 and report summary statistics for those additional variables in 2.B.1.

facilitate a bribe by reporting the public official's minimum acceptable bribe, and (2) conditional on facilitating a bribe, how much the intermediary embezzles for himself — and beliefs about how the other intermediaries in the experiment will behave.

2.4.1 Behaviour

In this section, we test Hypotheses 2.1 and 2.2.

In Figure 2.3, we display the average willingness to facilitate a bribe across treatments for our entire sample as well as separately for Juniors and Seniors with 95% confidence intervals. Figure 2.3 shows no significant effects of the Training or Information treatments on average willingness to facilitate a bribe. This null result is replicated in columns (1) - (4) of Table 2.2 in regression analysis with and without additional demographic controls.

Given our null Intention-to-Treat effect of the training treatment, we conduct an ex-post power analysis as conducted in Spantig (2021). The study closest to ours is Banerjee and Mitra (2018) who find that ethics training increases the probability of behaving ethically by 97% (20% of participants in the control and 39.4% of participants in the training treatment). Setting $\alpha = .05$ and $1 - \beta = .80$ and given our sample sizes (123 in the Control and 129 subjects in the Training treatment), we would be able to detect an increase in ethical behaviour of 17.48 percentage

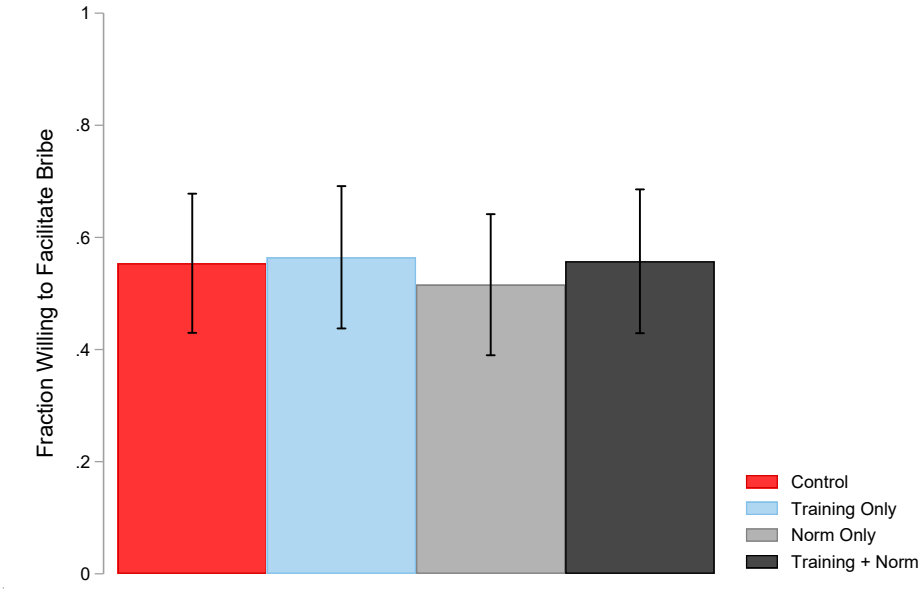
points, which corresponds to a 37.6% increase in ethical behaviour. Given that our minimal detectable effect size of 37.6% is smaller than the effect in the previous study (97%), we are powered enough to detect the increase found in the previous literature.

However, as discussed above, differences in stages of socialisation may lead Junior and Senior students to respond differently to the interventions. Consistent with this, the null result in the pooled sample masks substantial heterogeneity across cohorts, which can be clearly seen in Figure 2.3b and is replicated in columns (4) - (8) in Table 2.2. We find that Juniors in the Information Only and the Training + Information treatments are 20 to 25 percentage points (approximately 33 percent) less likely to facilitate a bribe than Juniors in the Control treatment. For these students, the information treatment also slightly reduced embezzlement, conditional on facilitating corruption. However, there is no significant effect of treatments on the behaviour of the Senior students.

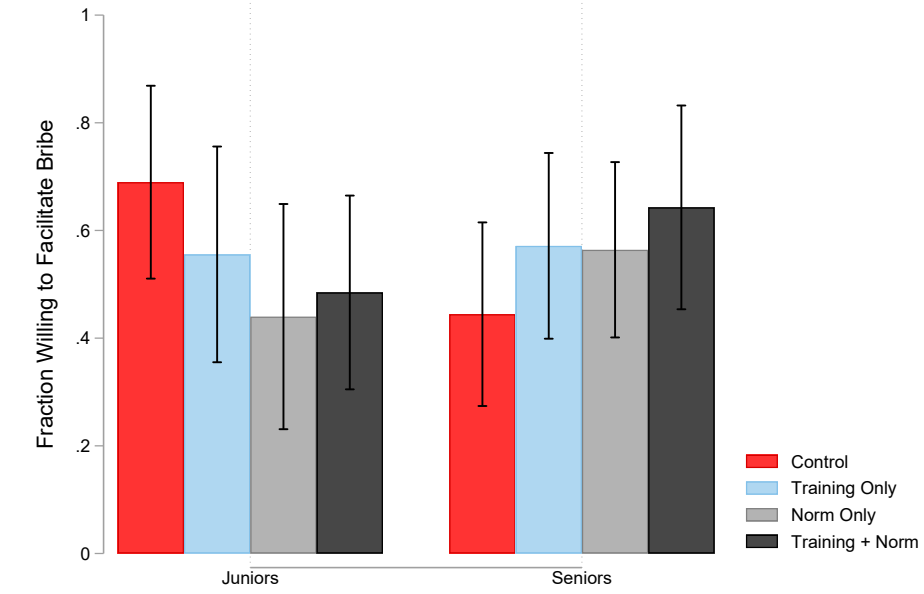
It is possible that the difference between Juniors and Seniors in the effects of the Training + Information treatment is driven by differences in Training attendance (see Figure 2.2). To investigate this, we estimate the effect of the treatments on the treated using a Local Average Treatment Effect (Imbens and Angrist, 1994), in which we instrument for the number of training sessions attended using the treatment assignment. We present these results in Table 2.B.2 and obtain coefficient estimates that are qualitatively equivalent to those in Table 2.2.

However, closer inspection of Figure 2.3 and Tables 2.2 and 2.B.2 reveal an important difference between Juniors and Seniors: Juniors in the Control treatment are 25 percentage points (p-value= 0.049) more likely to facilitate a bribe than Seniors assigned to the Control treatment. The treatments appear to make the Junior students behave more like Senior students.

Figure 2.3: Intention-to-Treat effects



(a) Pooled sample



(b) By cohort

Note: Figure 2.3 shows Intention-to-Treat effects with 95% confidence bars and Figure 2.3b shows Intention-to-Treat effects for Junior and Seniors separately.

Table 2.2: Intention-to-Treat effects on behaviour

	Probability of facilitating a bribe						Amount embezzled	
	All students				Jrs Only	Srs Only	Jrs Only	Srs Only
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Train Only Treatment	0.01	0.01	0.13	0.12	-0.15	0.12	1.18	-1.16
	(0.09)	(0.09)	(0.12)	(0.12)	(0.13)	(0.12)	(0.89)	(0.83)
Information Only Treatment	-0.04	-0.04	0.12	0.1	-0.25*	0.1	-2.73**	-0.1
	(0.09)	(0.09)	(0.12)	(0.12)	(0.14)	(0.12)	(1.14)	(0.79)
Train + Information Treatment	0.004	0.002	0.2	0.17	-0.21*	0.17	1.58*	0.81
	(0.09)	(0.09)	(0.13)	(0.13)	(0.12)	(0.13)	(0.84)	(0.81)
Prior ethics training	.	0.01	-0.006	0.12	-0.18*	0.12	2.54***	-1.17**
		(0.07)	(0.07)	(0.09)	(0.1)	(0.09)	(0.71)	(0.6)
Junior	.	-0.01	0.25**	0.35***	.	.	.	.
		(0.06)	(0.12)	(0.13)				
Jr X Training Only Treatment	.	.	-0.27	-0.27	.	.	.	.
			(0.18)	(0.18)				
Jr X Information Only Treatment	.	.	-0.38**	-0.35**	.	.	.	.
			(0.18)	(0.18)				
Jr X Train + Norm Treatment	.	.	-0.41**	-0.38**	.	.	.	.
			(0.18)	(0.18)				
Jr X with prior ethics training	.	.	.	-0.28**	.	.	.	.
				(0.13)				
Female	.	-0.01	0.002	0.02	0.07	-0.03	4.94***	-0.07
		(0.07)	(0.07)	(0.07)	(0.1)	(0.1)	(0.91)	(0.63)
Enroll for Skills	.	0.05	0.06	0.05	0.009	0.09	0.26	-3.28***
		(0.09)	(0.09)	(0.09)	(0.12)	(0.13)	(0.91)	(0.75)
Constant	0.55***	0.53***	0.39***	0.36***	0.7***	0.37**	.	.
	(0.06)	(0.12)	(0.13)	(0.13)	(0.15)	(0.16)		
Observations	252	252	252	252	114	138	62	76
R^2	0.001	0.003	0.03	0.05	0.06	0.04	.	.

Note: OLS regression coefficients. In columns [3], [6] & [9] we collapse all three treatments into one variable. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

2.4.1.1 Malleability of attitudes and morals

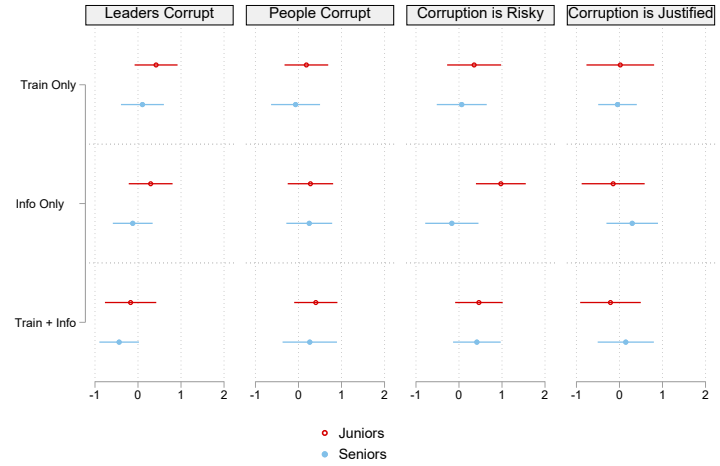
We hypothesized that training would change behaviour by increasing the awareness of and changing attitudes towards corruption as well as increase the moral costs of engaging in corrupt behaviour. In both the baseline and endline surveys, we collected data about the subjects' attitudes about corruption and we conducted the 20-item Moral Foundations questionnaire to assess moral foundations at the beginning of the study and at the conclusion.[14] Ex ante we did not know through which of the attitudes or moral foundations the training treatment would operate. Figure 2.4 shows coefficient plots of the treatment effects on attitudes and morals. In these regressions, we include the corresponding baseline measure of attitudes or morals as well as our set of demographic controls. Figure 2.4a and Figure 2.4b show that the Training Only treatment had no significant effect on either Junior or Senior subjects' attitudes towards corruption or their moral foundations of decision-making, respectively. This provides further evidence consistent with the null finding on the effect of training alone.

Notably, it is not just that attitudes and morals are not malleable. By contrast to the effects of training, subjects who received Information (either alone or in conjunction with Training) display changes in attitudes and morals. For example, Juniors who received Information have an *increased* perception that corruption is risky and, to a lesser extent, an *increased* perception that people are corrupt. These Juniors also *decrease* the emphasis they put on "Harm/Care" and "Ingroup Loyalty" concerns when making decisions or moral judgements about behaviour.[15] Thus, attitudes and moral foundations show some degree of malleability within our study but they are *not* significantly affected by the integrity training program.

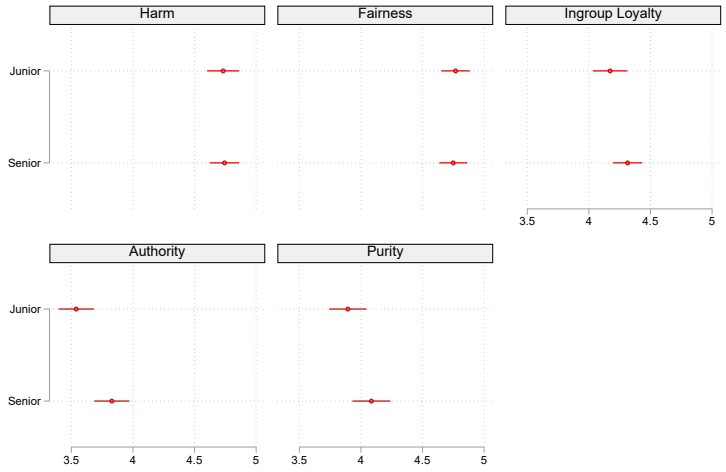
[14] In total, we have 13 variables about attitudes towards corruption. To more concisely analyze the data and to incorporate as much of the collected data as possible, we conducted a principal component analysis to summarize the data in 4 variables.

[15] Consistent with the treatments having no significant effect on Senior students, we find only one marginally significant result across 9 regressions: Information causes a marginally significant increase in the emphasis Senior students put on the "Harm/Care" foundation, but has no other effect on the other 8 outcomes.

Figure 2.4: Treatment effects on attitudes and morals



(a) Attitudes



(b) Morals

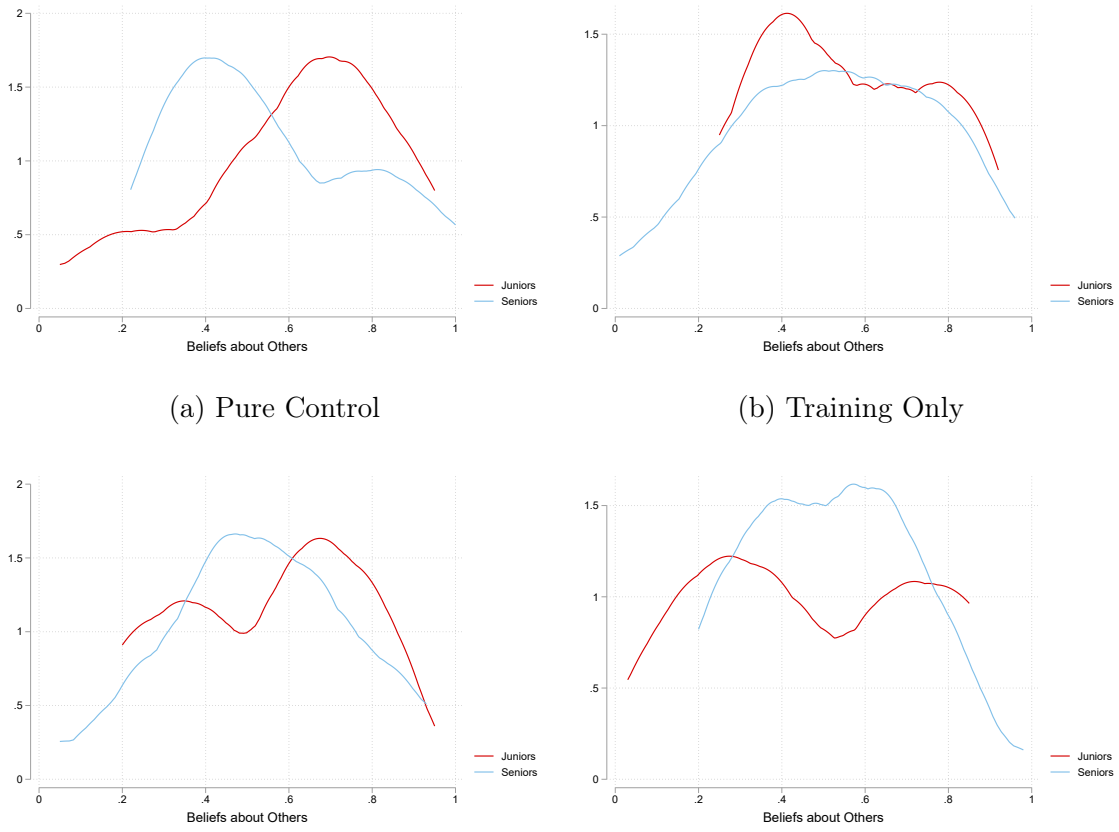
Note: The OLS coefficients show treatment effects relative to the control group. All regressions control for baseline attitudes/moral preferences and the set of demographic characteristics.

2.4.2 Beliefs

In addition to behavioural measures, we also collected intermediaries' beliefs about the behaviour of other intermediaries in the experiment. We plot the distributions of beliefs by treatment in Figure 2.5. The most striking feature of the data is the difference between the beliefs of Junior and Senior students in the Control treatment and the attenuation of these differences between Junior and Senior students in

the Training and Information treatments. This finding mirrors what we find for behaviour in Section 2.4.1.

Figure 2.5: Beliefs about the probability of intermediaries facilitating a bribe



(a) Pure Control

(b) Training Only

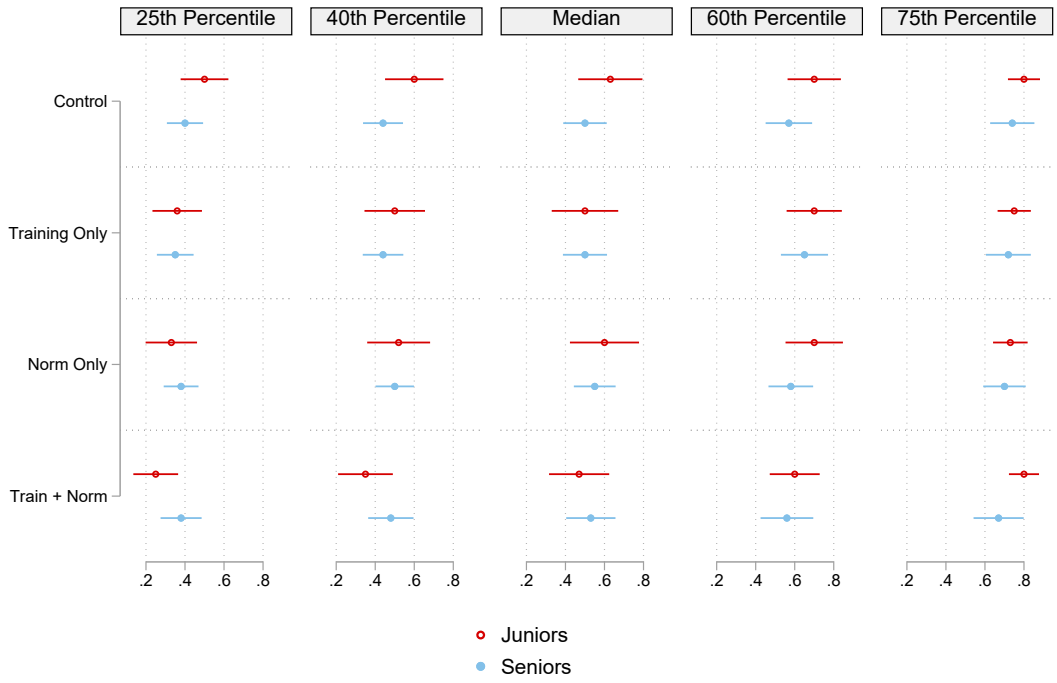
(c) Info Only

(d) Train + Info

Note: The figures show the distribution of beliefs about the probability that Intermediaries will facilitate a bribe, by treatment and by Seniority. The median belief of Juniors and Seniors differs only in the Pure Control ($\chi^2 = 3.45$, p-value= 0.063). The median beliefs of Juniors and Seniors do not significantly differ across the three other treatments.

To better understand the effect of treatments on beliefs for Juniors and Seniors we conducted quantile regression analysis to investigate how the treatments affected different parts of the distribution. We report the coefficient plot from those regressions in Figure 2.6.

Figure 2.6: Effect of treatments on distribution of beliefs



Note: The figure shows treatment effects on beliefs at different points in the belief distribution.

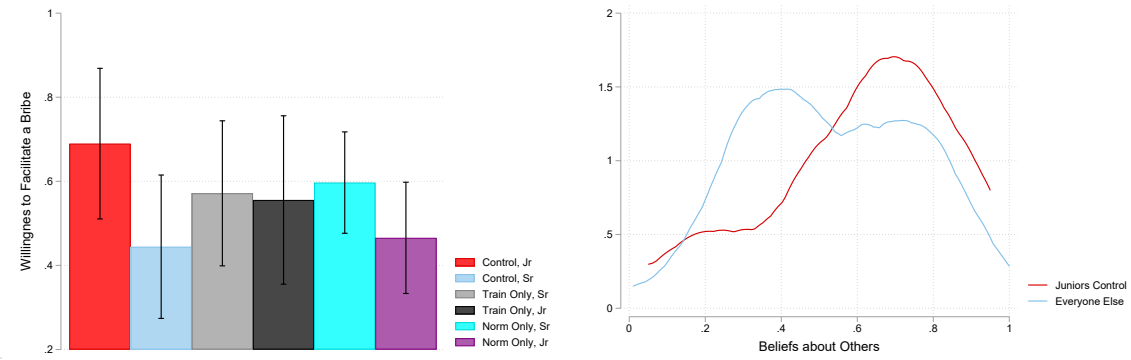
There are three important results to highlight here. First, there is a significant difference in beliefs between Juniors and Seniors in the control treatment. In terms of median beliefs, Figure 2.5a shows that the median beliefs of Juniors in the control is significantly larger than the median beliefs of Seniors in the pure control ($\chi^2 = 3.45$, p-value$= 0.063$). In fact, we find that this pattern holds at various percentiles across the lower half of the distribution (see the Control panel of Figure 2.6). Second, among Junior students, the Information Only treatment and the Training + Information treatment significantly reduce the 25th and 40th percentile beliefs relative to the Control, but no treatments have any effect at the higher moments of the distribution. Thus, for Juniors the treatments effectively reduced beliefs about the probability for those students who already had relatively low beliefs. Third, there is no effect of treatments on beliefs among Senior students.

In sum, the analysis on beliefs reveals the same pattern as the analysis on behaviour: the Information Only and Training + Information treatments serve to bring Junior students more in-line with Senior students.

2.4.3 The importance of social conformity

We find that integrity training alone has no average effect on law students' behaviour in Ukraine. However, we find that our Information treatment does affect beliefs and ethical behaviour, but only among Junior students. In fact, treated Junior students behave and hold beliefs that are indistinguishable from Senior students, while Junior students in the Control treatment are significantly more likely to engage in corrupt behaviour (p-value=0.08) and *believe* others are doing the same ($\chi^2 = 6.59$, p-value= 0.01) relative to everyone else (i.e., all Seniors and treated Juniors) (Figure 2.7).

Figure 2.7: Convergence of beliefs and behaviour



(a) Behaviour (b) Beliefs

Note: The figures compare the beliefs and behaviour of Juniors in the Control with everyone else.

The fact that the training program did not have an impact on moral preferences (or at least not one captured by corruption perceptions and considerations underlying moral decisions), whilst information about a majority of peers being trained led junior students to revise down their expectations about the incidence of corruption and adjust their behaviours accordingly, indicates the most important mechanism at play is social conformity. While the information treatment was purely factual and did not spell out a descriptive or injunctive social norm, it likely led students to update their expectations about the social norm and activated their desire to

conform. This mechanism is similar to that in Abbink et al. (2018), who show in a collusive bribery game that firms offered bribes twice as often to officials known to be from a predominantly corrupt group compared to those from a mostly honest group.

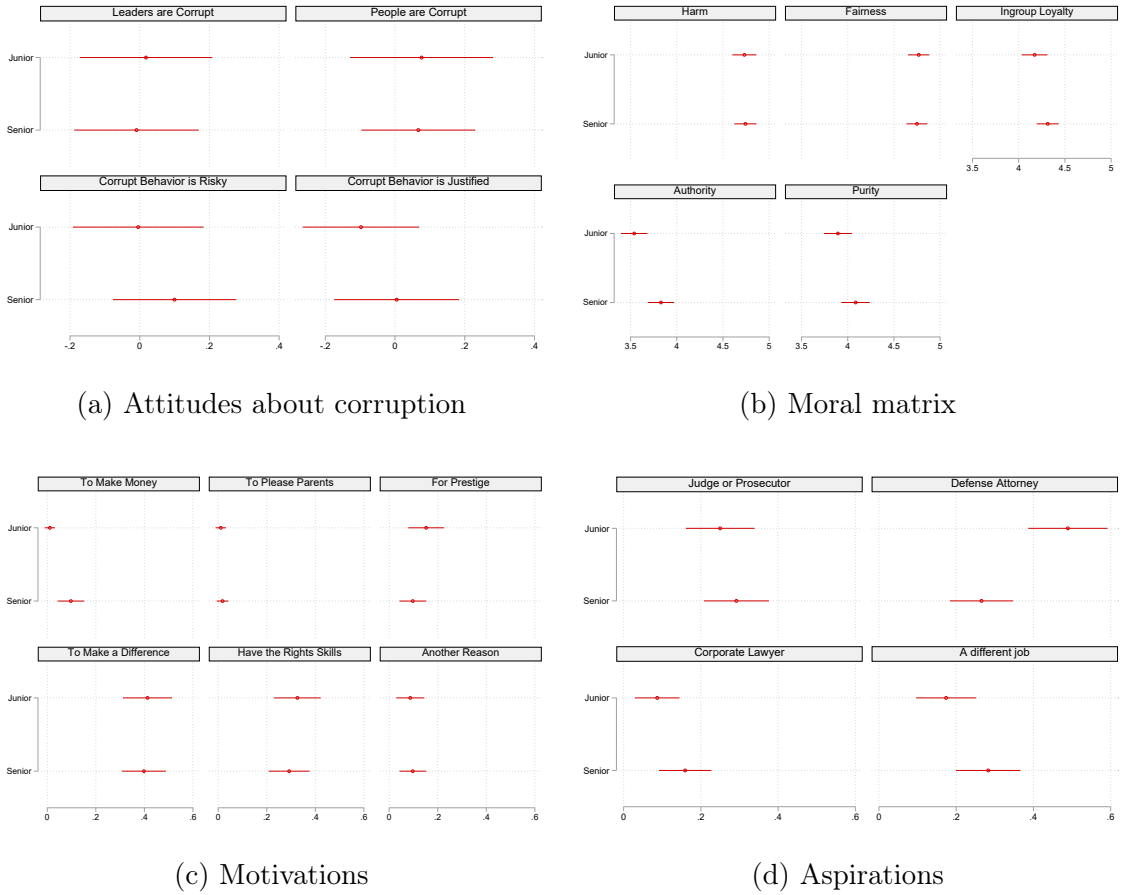
These findings underscore that integrity is not shaped in isolation, but in relation to what individuals perceive others around them to be doing. This suggests that shifting collective expectations may sometimes be more effective than direct moral instruction.

2.4.4 What drives the difference between Juniors and Seniors?

The main driving force behind our Intention-To-Treat results for behaviour and beliefs is that Juniors in the Control treatment think and behave significantly differently from Seniors in the Control treatment, and that our treatments — particularly the Information treatment — bring the Juniors' beliefs and behaviour in line with those of Seniors. Thus, what drives this baseline difference between Juniors and Seniors that the treatments appear to mitigate?

At baseline, before the treatments were administered, we collected several measures, including students' moral foundations, their views of corruption and their career motivations and aspirations. Figure 2.8 shows the average responses for Juniors and Seniors across several baseline measures and we find minimal differences. Of the 19 variables constructed and tested, we find differences between Juniors and Seniors across 4 variables — (1) Juniors put significantly less emphasis on the importance of maintaining or respecting authority when judging the morality of a situation; (2) Juniors are less likely to report "to make money" as their motivation for law school and (3) more likely to report "to make a difference" than Seniors; (4) Juniors are also more likely to report that they aspire to work as a defense attorney after graduation than Seniors.

Figure 2.8: Juniors versus Seniors at baseline



(a) Attitudes about corruption

(b) Moral matrix

(c) Motivations

(d) Aspirations

Note: Coefficients from OLS regression estimates with robust standard errors.

These stated motivations and aspirations make Junior students appear more idealistic and driven to make a change, possibly reflecting a more pessimistic assessment of the prevalence of corruption than that held by Senior students. Those with greater exposure to their institutional environment may have had time to naturally correct these misperceptions and adopt a more pragmatic outlook. Our treatment likely prompted Junior students to revise their perceptions and align their behavior with their updated beliefs about the prevailing social norms.

2.5 Conclusion

We develop a dynamic game with incomplete information to model a corrupt transaction between a firm and a public official facilitated by an intermediary. Intermediaries decide whether to facilitate the transaction — by informing the firm about the size of the bribe demanded — and whether to report the true amount or inflate it to embezzle part of the payment. Their choices depend on their moral costs and their beliefs about the average bribe demands of peers. The model predicts that higher moral costs, as well as beliefs that peers demand smaller bribes, reduce both the likelihood of facilitating corruption and the size of the bribe demanded. Consequently, interventions that increase the moral costs of corruption or shift beliefs about prevailing norms are expected to lower the incidence of corrupt transactions.

We test these predictions in an online experiment with law students in Ukraine. Participants were randomly assigned to receive an interactive integrity and ethics training over six sessions across two weeks. A month later, they played the role of intermediaries in a simulated bribery game. To examine the effect of perceived norms, we further randomized half of both trained and untrained participants to receive information that 75% of their peers had completed the integrity training. Following prior studies, we analyze treatment effects separately for Junior (first-year) and Senior students, who may differ in their perceptions of corruption, prevailing social norms, and motivations for ethical behaviour.

The results reveal several surprising insights. First, the integrity training alone did not affect behaviour, moral attitudes, or beliefs — neither overall nor within student subgroups. Second, the information treatment also had no aggregate effect, but it significantly reduced Juniors' willingness to facilitate bribes. This reduction operated through revised beliefs about the prevalence of corruption and the activation of social conformity. Junior students who initially overestimated the incidence of corruption adjusted their expectations downward after learning that a majority of peers had undergone integrity training, and in turn behaved more ethically. For these students, the information treatment also slightly reduced embezzlement conditional on facilitating corruption.

These findings highlight the role of pluralistic ignorance in sustaining corrupt practices: individuals overestimate the extent of corruption among peers and con-

form to these misperceived norms. While integrity training alone did not alter moral preferences in the short run, information about the prevalence of such training effectively corrected misperceptions, fostered more ethical norms, and reduced corrupt behaviour.

Importantly, the informational intervention works only because integrity training exists: without real integrity initiatives to reference, information about their prevalence would lack credibility. Both interventions are therefore essential but act through different channels: training builds the ethical foundation, while information shifts social perceptions and activates conformity with those norms. Together, they form a complementary strategy to reduce corruption.

References

Abbink, K., Freidin, E., Gangadharan, L. and Moro, R.: 2018, The effect of social norms on bribe offers, *The Journal of Law, Economics, and Organization* **34**(3), 457–474.

Abbink, K., Irlenbusch, B. and Renner, E.: 2002, An experimental bribery game, *Journal of Law, Economics, and Organization* **18**(2), 428–454.

Akerlof, G. A.: 1980, A theory of social custom, of which unemployment may be one consequence, *The Quarterly Journal of Economics* **94**(4), 749–775.

Andre, P., Boneva, T., Chopra, F. and Falk, A.: 2024, Misperceived social norms and willingness to act against climate change, *Review of Economics and Statistics* pp. 1–46.

Banerjee, R. and Mitra, A.: 2018, On monetary and non-monetary interventions to combat corruption, *Journal of Economic Behavior & Organization* **149**, 332–355.

Barr, A. and Serra, D.: 2010, Corruption and culture: An experimental analysis, *Journal of Public Economics* **94**(11–12), 862–869.

Bayar, G.: 2005, The role of intermediaries in corruption, *Public Choice* **122**, 277–298.

Bernheim, B. D.: 1994, A theory of conformity, *Journal of Political Economy* **102**(5), 841–877.

Bicchieri, C.: 2016, *Norms in the wild: how to diagnose, measure, and change social norms*, Oxford University Press.

Bicchieri, C. and Xiao, E.: 2009, Do the right thing: but only if others do so, *Journal of Behavioral Decision Making* **22**(2), 191–208.

Boneva, T., Brás-Monteiro, A., Golin, M. and Rauh, C.: 2024, Are men's preferences for couple equity misperceived? evidence from six countries, *Technical report*, CESifo Working Paper Series No. 11536.

Bursztyn, L., González, A. L. and Yanagizawa-Drott, D.: 2020, Misperceived social norms: Women working outside the home in saudi arabia, *American Economic Review* **110**(10), 2997–3029.

Charness, G., Haruvy, E. and Sonsino, D.: 2007, Social distance and reciprocity: an internet experiment, *Journal of Economic Behavior & Organization* **63**(1), 88–103.

Cialdini, R. B. and Goldstein, N. J.: 2004, Social influence: Compliance and conformity, *Annual Review of Psychology* **55**(1), 591–621.

Cialdini, R. B., Reno, R. R. and Kallgren, C. A.: 1990, A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places, *Journal of Personality and Social Psychology* **58**(6), 1015–1026.

De Quidt, J., Haushofer, J. and Roth, C.: 2018, Measuring and bounding experimenter demand, *American Economic Review* **108**(11), 3266–3302.

Della Porta, D. and Vannucci, A.: 1999, *Corrupt exchanges: actors, resources, and mechanisms of political corruption*, Routledge.

Denisova-Schmidt, E. and Prytula, Y.: 2017, Ukraine: endemic higher education corruption, *International Higher Education* (90), 16–18.

Desplaces, D. E., Melchar, D. E., Beauvais, L. L. and Bosco, S. M.: 2007, The impact of business education on moral judgment competence: an empirical study, *Journal of Business Ethics* **74**, 73–87.

Drugov, M., Hamman, J. and Serra, D.: 2014, Intermediaries in corruption: an experiment, *Experimental Economics* **17**(1), 78–99.

Haidt, J.: 2007, The new synthesis in moral psychology, *Science* **316**(5827), 998–1002.

Harris, D., Borcan, O., Serra, D., Telli, H., Schettini, B. and Dercon, S.: 2022, Proud to belong: the impact of ethics training on police officers in ghana, *University of Oxford Working Paper* .

Hasker, K. and Okten, C.: 2008, Intermediaries and corruption, *Journal of Economic Behavior & Organization* **67**(1), 103–115.

Imbens, G. W. and Angrist, J. D.: 1994, Identification and estimation of local average treatment effects, *Econometrica* **62**(2), 467–475.

Kiev International Institute of Sociology: 2015, Corruption in ukraine: comparative analysis of national surveys 2007, 2009, 2011, 2015, *Technical report*, Kiev International Institute of Sociology.

Krupka, E. and Weber, R. A.: 2009, The focusing and informational effects of norms on pro-social behavior, *Journal of Economic Psychology* **30**(3), 307–320.

Köbis, N. C., Troost, M., Brandt, C. O. and Soraperra, I.: 2022, Social norms of corruption in the field: social nudges on posters can help to reduce bribery, *Behavioural Public Policy* pp. 507–624.

Köbis, N. C., Van Prooijen, J.-W., Righetti, F. and Van Lange, P. A. M.: 2015, "who doesn't?"—the impact of descriptive norms on corruption, *PLOS ONE* **10**(6), e0131830.

Mayhew, B. W. and Murphy, P. R.: 2009, The impact of ethics education on reporting behavior, *Journal of Business Ethics* **86**, 397–416.

OECD: 2018, Education for integrity: teaching on anti-corruption, values and the rule of law, *Technical report*, OECD.

OECD: 2020, Foreign bribery and the role of intermediaries, managers and gender, *Technical report*, OECD.

O'Gorman, H. J.: 1975, Pluralistic ignorance and white estimates of white support for racial segregation, *Public Opinion Quarterly* **39**(3), 313–330.

Rose-Ackerman, S.: 1978, *Corruption: a study in political economy*, Academic Press.

Rothstein, B.: 2000, Trust, social dilemmas and collective memories, *Journal of Theoretical Politics* **12**(4), 477–501.

Shaub, M. K.: 1994, An analysis of the association of traditional demographic variables with the moral reasoning of auditing students and auditors, *Journal of Accounting Education* **12**(1), 1–26.

Spantig, L.: 2021, Cash in hand and savings decisions, *Journal of Economic Behavior & Organization* **188**, 1206–1220.

Tankard, M. E. and Levy Paluck, E.: 2016, Norm perception as a vehicle for social change, *Social Issues and Policy Review* **10**(1), 181–211.

Transparency International: 2017, Global corruption barometer: citizens' voices from around the world.

Weidman, J.: 1989, *Undergraduate socialization: A conceptual approach*, Higher education: Handbook of theory and research.

World Economic Forum: 2018, Corruption is costing the global economy $3.6 trillion dollars every year. Written by Stephen Johnson, originally published by Big Think.

Zizzo, D. J.: 2010, Experimenter demand effects in economic experiments, *Experimental Economics* **13**(1), 75–98.

Appendix

2.A Proofs

Proposition 2.2. *There exists a Perfect Bayesian Equilibrium described by the following strategy profile (and associated beliefs):*

- Conditional on informing, the Intermediary asks for the optimal bribe of:

$$b^*(m_I, MAB) = \frac{v(1-p) + MAB - m_I E\overline{m_F} + 2c\tilde{b}E\overline{m_F}}{2(1 + cE\overline{m_F})}$$

- The Intermediary informs iff $m_I \leq m_i^*$, where m_i^* solves:

$$m_i^* = \frac{\left[v(1-p) - b^*\right]/(E\overline{m_F})(b^* - MAB) - c(b^* - \tilde{b})^2 + c\tilde{b}^2}{(b^* - MAB + E)}$$

- The Firm pays the bribe if:

$$m_F \leq \frac{v(1-p) - b}{E}$$

  and does not otherwise.

Proof. We use backward induction to characterise the equilibrium. First consider the Firm's optimal strategy.

Following a history of b, the Firm will pay if and only if the expected utility from paying is greater than that from not:

$$Y_F + v - b - m_F E \geq Y_F + pv,$$

$$m_F \leq (v(1-p) - b)/E.$$

Therefore, if m_F takes values between $[0, (v(1-p) - b)/E]$ the Firm will pay the bribe; if m_F takes values between $[(v(1-p) - b)/E, \overline{m_F}]$ it will not.

Given the Firm's strategy above, we now turn to the Intermediary's optimal strategy. Given the common uniform prior over m_F, the Intermediary believes that

if he asks for a bribe b the probability that the Firm will pay is $(v(1-p)-b)/(E\overline{m_F})$.

For a given b, the Intermediary will inform if and only if the expected utility of informing is greater than that of not informing:

$$\begin{aligned}&\left(\frac{v(1-p)-b}{E\overline{m_F}}\right)\left[Y_I+(b-MAB)-m_I(b-MAB+E)-c(b-\tilde{b})^2\right]\\&\quad+\left(1-\frac{v(1-p)-b}{E\overline{m_F}}\right)\left[Y_I-m_I(b-MAB+E)-c(b-\tilde{b})^2\right]\\&\geq Y_I-c\tilde{b}^2.\end{aligned}\tag{2.2}$$

$$\frac{v(1-p)-b}{E\overline{m_F}}(b-MAB)-m_I(b-MAB+E)-c(b-\tilde{b})^2+c\tilde{b}^2\geq 0.\tag{2.3}$$

$$\frac{\dfrac{v(1-p)-b}{E\overline{m_F}}(b-MAB)-c(b-\tilde{b})^2+c\tilde{b}^2}{b-MAB+E}\;\geq\;m_I.\tag{2.4}$$

To find the Intermediary's optimal b, we identify the value of b that maximises the Intermediary's expected utility (conditional on informing):

$$\begin{aligned}0=\frac{\partial}{\partial b}\Bigg[&\left(\frac{v(1-p)-b}{E\overline{m_F}}\right)\left(Y_I+(b-MAB)-m_I(b-MAB+E)-c(b-\tilde{b})^2\right)\\&+\left(1-\frac{v(1-p)-b}{E\overline{m_F}}\right)\left(Y_I-m_I(b-MAB+E)-c(b-\tilde{b})^2\right)\Bigg].\end{aligned}\tag{2.5}$$

$$\begin{aligned}0=&-\frac{1}{E\overline{m_F}}\left[Y_I+(b-MAB)-m_I(b-MAB+E)-c(b-\tilde{b})^2\right]\\&+\left(\frac{v(1-p)-b}{E\overline{m_F}}\right)\left[1-m_I-2c(b-\tilde{b})\right]\\&+\frac{1}{E\overline{m_F}}\left[Y_I-m_I(b-MAB+E)-c(b-\tilde{b})^2\right]\\&+\left(1-\frac{v(1-p)-b}{E\overline{m_F}}\right)\left[-m_I-2c(b-\tilde{b})\right].\end{aligned}\tag{2.6}$$

$$-\frac{1}{E\overline{m_F}}(b-MAB)+\frac{v(1-p)-b}{E\overline{m_F}}-m_I-2c(b-\tilde{b})=0,\tag{2.7}$$

$$\frac{2b}{E\overline{m_F}} + 2bc = \frac{MAB}{E\overline{m_F}} + 2c\tilde{b} + \frac{v(1-p)}{E\overline{m_F}} - m_I, \tag{2.8}$$

$$b^* = \frac{MAB + 2E\overline{m_F}c\tilde{b} + v(1-p) - m_I E\overline{m_F}}{2\big(1 + E\overline{m_F}c\big)}. \tag{2.9}$$

Returning to the inequality determining whether the Intermediary informs, we can then state that he will ask for bribe $b^*(m_I, MAB)$ if the following inequality holds:

$$m_I \leq \left(\frac{[v(1-p) - b^*]/(E\overline{m_F})(b^* - MAB) - c(b^* - \tilde{b})^2 + c\tilde{b}^2}{(b^* - MAB + E)} \right)$$

□

Corollary 2.7. *An increase in $\overline{m_I}$ reduces the expected bribe asked for by the Intermediary (conditional on informing).*

Proof. An increase in $\overline{m_I}$ implies an increase in the m_I of a randomly drawn Intermediary. Note that

$$\frac{\partial b^*}{\partial m_I} = -\frac{E\overline{m_F}}{2(1 + cE\overline{m_F})} < 0.$$

Thus, the bribe asked for by a randomly drawn Intermediary will be strictly lower.

□

Corollary 2.8. *An increase in $\tilde{b}$ increases the expected bribe asked for by the Intermediary (conditional on informing).*

Proof. An increase in $\tilde{b}$ implies an increase in a randomly drawn Intermediary's belief about the average bribe asked for by other Intermediaries in the population of interest. Note that

$$\frac{\partial b^*}{\partial \tilde{b}} = \frac{2cE\overline{m_F}}{2(1 + cE\overline{m_F})} > 0.$$

Thus, the bribe asked for by a randomly drawn Intermediary will be strictly greater. This is in line with our assumption that the greater the difference between the Intermediary's actual bribe ask b and the belief about the average bribe asked for by other Intermediaries $\tilde{b}$, the larger the social conformity cost they will incur. If $\tilde{b}$ increases, b will increase accordingly. □

Corollary 2.9. *An increase in $\overline{m_I}$ reduces the probability that the Intermediary will inform.*

Proof. Based on our previous results, the Intermediary informs iff $m_I \leq m_I^*$. An increase in $\overline{m_I}$ does not affect the incentives of any given m_I, thus m_I^* is independent of $\overline{m_I}$. An increase in $\overline{m_I}$ reduces the probability mass of the Intermediary that satisfies the condition $m_I \leq m_I^*$, thus the probability of the Intermediary informing is lower. □

Corollary 2.10. *An increase in $\tilde{b}$ increases the probability that the Intermediary will inform.*

Proof. Based on our previous results, the Intermediary informs iff $m_I \leq m_I^*$, where

$$m_I^* = \frac{[v(1-p) - b^*]/(E\overline{m_F})(b^* - MAB) - c(b^* - \tilde{b})^2 + c\tilde{b}^2}{(b^* - MAB + E)}.$$

If m_I takes values between $[0,\, m_I^*]$ the Intermediary will inform; if m_I takes values between $(m_I^*,\, \overline{m_I}]$ it will not. Therefore, given the common uniform prior over m_I, the Intermediary's probability of informing is $\frac{m_I^*}{\overline{m_I}}$. Note that

$$\frac{\partial\left(\frac{m_I^*}{\overline{m_I}}\right)}{\partial\tilde{b}} = \frac{2cb^*}{\overline{m_I}(b^* - MAB + E)} > 0.$$

Thus, the probability that the Intermediary will inform will be strictly greater. □

2.B Additional tables and figures

Figure 2.B.1: Participant invitation flyer

Figure 2.B.2: Questionnaires and game instructions

Baseline and endline questionnaire

Survey information

Thank you for participating. Your participation is voluntary and you may withdraw at anytime. The survey has three sections: 1) your views on different social issues; 2) your views on corruption in the Ukrainian society; 3) some questions about yourself. The survey will take approximately 20 minutes to complete. If you complete the questionnaire, you will receive a gift worth 100 UAH as a shopping voucher.

Data collected will not be shared with anyone. To protect your identity, you have been assigned a numerical identifier and your answers to this survey will be linked to your number and not to your name. Data will be stored using the assigned number and will thus be de-identified. The **de-identified data** will be analysed by the researchers and will be stored afterwards in the School of Economics at University of East Anglia (UEA), United Kingdom or in a journal repository upon publication, where it may be downloaded by other researchers for reanalysis.

You have signed a general consent form to participate in this research. You will shortly be asked to sign a specific consent form for this survey. Your participation will be recorded in the School of Economics, UEA.

You will receive the gift within three weeks of completing the survey. Furthermore, if you complete this questionnaire and ALL surveys and quizzes during this study, you will be entered in a draw for a chance to win one of three prizes worth 1400 UAH.

If you have any concerns about this survey, please contact Dr Oana Borcan at o.borcan@uea.ac.uk, who will endeavour to respond within 10 working days. If you are still unhappy and wish to make a formal complaint, please email the chair of the Ethics Committee at the School of Economics, UEA, Dr *David Hugh-Jones (*D.Hugh-Jones@uea.ac.uk*)*.

Consent Form

I confirm that I have read and understood the "Survey Information" above. I understand that the data generated during the experiment will be entirely anonymous, so that my name will not be linked to the data that is generated. The anonymous data will be analysed by the researchers and subsequently stored at the University of East Anglia, where it may be downloaded by other researchers. I understand that my participation is voluntary and that I am free to withdraw at any time but I will receive a gift worth 100 UAH only if I complete the survey in full.

☐ YES, I agree to take part in this survey.
☐ NO, I do not agree to take part in this survey.

Participant numeric identifier: …

Date: …

General views on social issues	
Question	**Coding Category**
When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement below:	
Q1. Whether or not someone suffered emotionally	1 Not at all relevant (This consideration has nothing to do with my judgments of right and wrong) 2 Not very relevant

	3 Slightly relevant 4 Somewhat relevant 5 Very relevant 6 Extremely relevant (This is one of the most important factors when I judge right and wrong)
Q2. Whether or not some people were treated differently than others	
Q3. Whether or not someone's action showed love for his or her country	
Q4. Whether or not someone showed a lack of respect for authority	
Q5. Whether or not someone violated standards of purity and decency	
Q6. Whether or not someone was good at math	
Q7. Whether or not someone cared for someone weak or vulnerable	
Q8. Whether or not someone acted unfairly	
Q9. Whether or not someone did something to betray his or her group	
Q10. Whether or not someone conformed to the traditions of society	
Q11. Whether or not someone did something disgusting	
Q12. Whether or not someone was cruel	
Q13. Whether or not someone was denied his or her rights	
Q14. Whether or not someone showed a lack of loyalty	
Q15. Whether or not an action caused chaos or disorder	
Q16. Whether or not someone acted in a way that God would approve of	
Please read the following sentences and indicate your agreement or disagreement:	
Q17. Compassion for those who are suffering is the most crucial virtue.	1 Strongly disagree 2 Moderately disagree 3 Slightly disagree 4 Slightly agree 5 Moderately agree 6 Strongly agree
Q18. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.	
Q19. I am proud of my country's history.	
Q20. Respect for authority is something all children need to learn.	
Q21. People should not do things that are disgusting, even if no one is harmed.	
Q22. It is better to do good than to do bad.	
Q23. One of the worst things a person could do is hurt a defenseless animal.	

Q24. Justice is the most important requirement for a society.	
Q25. People should be loyal to their family members, even when they have done something wrong.	
Q26. Men and women each have different roles to play in society.	
Q27. I would call some acts wrong on the grounds that they are unnatural.	
Q28. It can never be right to kill a human being.	
Q29. I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.	
Q30. It is more important to be a team player than to express oneself.	
Corruption in Ukraine	
Q31. Now I would like you to tell us your views on corruption – when people pay a bribe, give a gift or do a favor to other people in order to get the things they need done or the services they need. How would you place your views on corruption in your country on a 10-point scale where "1" means "there is no corruption in my country" and "10" means there is abundant corruption in my country". If you think there is an intermediate level of corruption, choose the appropriate number in between.	1 There is no corruption in my country 2 3 4 5 6 7 8 9 10 There is abundant corruption in my country
Among the following groups of people, how many people do you believe are involved in corruption? Tell me for each group if you believe it is none of them, few of them, or all of them?	
Q32. State authorities	
Q33. Business executives	
Q34. Local authorities	
Q35. The judiciary	
Q36. Civil service providers (police, civil servants, doctors, teachers)	
Q37. Journalists and media	
Q38. We want to know about your experience with local officials and service providers, like police officers, lawyers, doctors, teachers and civil servants in your community. How often do you think ordinary people like yourself or people from your neighbourhood have to pay a bribe, give a gift, or do a favor to these people in order to get the service you need? Does it happen never, rarely, frequently or always?	1 Never 2 Rarely 3 Frequently 4 Always
Q39. Can you tell me how strongly you agree or disagree with the following statement: "on the whole, women are less corrupt than men"?	1 Strongly agree 2 Agree 3 Disagree 4 Strongly disagree 0 Hard to say

Q40. How high is the risk in this country to be held accountable for giving or receiving a bribe, gift, favor in return for public service? To indicate your opinion, use a 10-point scale, where "1" means "no risk at all" and "10" means "very high risk".	1 No risk at all 2 3 4 5 6 7 8 9 10 Very high risk
Q41. When a foreign firm pays a bribe, offers a gift, or does a favor to local officials in order to get some business or service they need, who do you think should be held accountable?	1 The foreign firm 2 The local officials 3 The local middlemen who facilitate the payment, gift, or favor. 4 All of the above 5 None of the above
In your opinion, can the situations below be justified?	
Q42. Is it justifiable that student cheats on their exam because most of their colleagues are cheating?	1 Never 2 Rarely 3 Frequently 4 Always
Q43. Is it justifiable that a judge who is about to retire accepts a bribe from a defendant?	1 Never 2 Rarely 3 Frequently 4 Always
Q44. Is it justifiable that a public hospital manager purchases the products of a pharmaceutical company without a bid for tender, after attending a conference sponsored by the pharmaceutical company?	1 Never 2 Rarely 3 Frequently 4 Always
Respondent profile	
Q45. Household income? Please tell us about the approximate household income of your primary family? Select one of the categories below (Put into bins that are reasonable for Ukraine).	1 We need to save money for food 2 We have enough money for food, but we need to save or borrow money for buying clothes and shoes 3 We have enough money for food and necessary clothing and shoes, but we need to save or borrow money for other purchases like a good suit, a mobile phone, or a vacuum cleaner 4 We have enough money for food, clothing, shoes, and other purchases, but we need to save or borrow money for purchasing more expensive things (e.g., appliances) 5 We have enough money for food, clothes, shoes, and expensive purchases, but we need to save or borrow money for purchases like a car or an apartment 6 I We can buy anything at any time 7 Difficult to answer
Q46. Do you currently have any employment?	0 No 1 Yes
Q46a. If your answer to Q46 was "Yes", what is your current income? Select one of the categories below.	Up to 3000,0 3000,1–8000,0 8000,1–12000,0 12000,1-17000 17000,1-20000 20000,1 and more Prefer not to say
Q47. Are you paying tuition fees for your university studies?	0 No 1 Yes

Q48. Are you receiving a scholarship for these studies?	0 No 1 Yes
Q49. What job do you aspire to have?	1 Judge or prosecutor 2 Defense attorney 3 Corporate lawyer 4 Other
Q49a. If answered "Other" in Q49, can you tell us your choice?	
Q50. What is the main reason you chose a degree in Law?	1 I want to make money 2 It was my parents' aspiration for me 3 It is a career that has prestige and power in society 4 I believe I can make a difference in this justice system 5 I have the talents and skills to pursue this career. 6 Other
Q50a. If answered "Other" in Q50, can you tell us your choice?	[text]
Q51. In the past year, how often have you cheated (or plagiarized) in university exams?	1 Never 2 A few times 3 Many times 4 Always
Q52. In the past year, how often have you made payments, given gifts or favours to obtain to improve your school or university marks?	
Q53. In the past 4 weeks, have you had to make a payment, give a gift or a favour to any public servant (for example teachers, doctors or the police) to obtain a service that should be free.	1 No 2 Yes, a few times 3 Yes, many times
Q54. In the past year, how often have you witnessed a colleague or a university teacher behave unethically?	1 No 2 Yes, a few times 3 Yes, many times
Q55. In the past year, have you reported unethical behaviour of a colleague or a university teacher?	1 Never 2 A few times 3 Many times

Experimental instructions and belief elicitation

General information

Thank you for participating. Your participation is voluntary, and you may withdraw at any time without penalty. The activity will take approximately 25 minutes to complete. If you decide to participate in this activity, you will receive a gift of 50 UAH as mobile top-up. You will also have the opportunity to receive an additional gift depending on your choices and chance, to compensate you for your time and effort. You will receive detailed instructions that will explain your choices and how those choices will affect your final earnings.

Data collected will not be shared with anyone. No participant will know the identity of other participants. To protect participants, you have been assigned a numerical identifier and your decisions in this experiment will be linked to your number and not to your name. Data will be stored using the assigned number and will thus be de-identified. The de-identified data will be analysed by the researchers and will be stored afterwards in the School of Economics at University of East Anglia (UEA), United Kingdom or in a journal repository upon publication, where it may be downloaded by other researchers for reanalysis.

You have signed a general consent form to participate in this research. You will shortly be asked to sign a consent form for this activity. You will receive the gifts in the form of a mobile top-up within 3 weeks of completing the activity. Your participation will be recorded in the School of Economics, UEA.

We have a strict "no deception policy" and thus everything in these instructions is true and accurate. If you have any questions about this or wish to know more about the research behind this activity, please contact Dr Oana Borcan at o.borcan@uea.ac.uk.

If you have any complaints or concerns about this research, please email the chair of the Ethics Committee at the School of Economics, UEA, Dr David Hugh-Jones (D.Hugh-Jones@uea.ac.uk)

Consent form

I confirm that I have read and understood the "General information" sheet dated … which I may keep for my records. I agree to take part in this experiment by undertaking a number of computer-based tasks. I understand that the data generated during the experiment will be entirely anonymous, so that my name will not be linked to the data that is generated. The anonymous data will be analysed by the researchers and subsequently stored at the University of East Anglia, where it may be downloaded by other researchers. I understand that my participation is voluntary and that I am free to withdraw at any time.

Participant numeric identifier: …

Date: …

General instructions

You will be playing a role in a simple activity with other participants.

This activity involves four players, each playing a different role. Your role will be either a Firm, an Intermediary, a Public Official or a Member of Society. You will be matched with three other participants, and roles will be anonymously and randomly assigned. At no point will you learn the identity of the other members from your group.

The activity is meant to mimic the procurement process of public contracts. Each of the four players in your game will have a budget of 70 UAH to start. The Public Official and the Member of Society do not make any decisions in the game, but their final payoffs are impacted by the decisions of the Firm and the Intermediary.

The Firm wants to win a contract from the Public Official that is worth 50 UAH to the Firm. The Intermediary will choose whether to help the Firm win the contract by informing the Firm of a piece of confidential information. If the Firm receives the confidential information from the Intermediary, the Firm decides whether or not to act upon this information.

The game proceeds as follows:

First, the Intermediary learns a piece of confidential information from the Public Official. This piece of confidential information is the amount the Public Official must receive as a payment to award the public contract to the Firm with a 100% chance. Without this payment, the Firm will win the contract with a 5% chance. The amount the Public Official must receive as a payment is determined randomly. A random number generator chooses one of the following amounts randomly and each amount has an equal chance of being chosen: 5 UAH, 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH, 40 UAH.

Second, the Intermediary decides whether or not to inform the Firm of this confidential information. The choice here is to Inform or Not Inform. Important Note: The Intermediary can inform the Firm of a number equal to or higher than the amount required by the Public Official. For example, if the Intermediary learns that the Public Official requires a payment of 5 and the Intermediary chooses to Inform the Firm, the Intermediary can choose 5, or a number greater than 5 (but not smaller) from the following set: 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH, 40 UAH. We call the "5" the True Payment Amount and for example the "10" the Informed Payment Amount. If the Intermediary chooses to inform the Firm of 10 and the Firm chooses to Pay, then the Intermediary passes the 5 to the Public Official and can keep the remaining difference 10-5.

Third, the Firm's action depends on the Intermediary's action. If the Intermediary informs the Firm about the confidential information, then the Firm can act on that information and Pay the Public Official or Not Pay. If the Firm Pays, the Firm wins the contract for sure. If the Firm does Not Pay, the Firm will win the contract with a 5% chance. Moreover, the decision to Pay has a negative effect on the Member of Society's payoff: if the Firm decides to Pay, then Society loses 35 UAH (half) of their endowment. Important Note: There are no partial payments. If the Firm decides to Pay the Public Official, the Firm must transfer the full Informed Payment Amount reported by the Intermediary. If the Intermediary does Not Inform the Firm, then the Firm can only choose to Not Pay and there is a 5% chance of winning the contract. The final amount of the gift you receive will be rounded to the next integer up.

Specific instructions

Intermediary

Your role

You have been assigned to the role of Intermediary in today's activity.

To begin with, you have 70 UAH as initial budget, like all the other players participating in the activity.

You will shortly receive a piece of confidential information, that is the True Payment Amount which the Public Official must receive to award the public contract worth 50 UAH to the Firm.

The experiment begins with your decision as Intermediary: you have to decide whether you want to inform the Firm or not. If you want to inform, you will have to choose what amount you would like to report.

Possible scenarios

If you choose not to inform the Firm, there will not be a deal between the two of you and the Firm will win the public contract with a 5% chance. If that is the case, earnings will be as follows:

- You: 70 UAH
- The Firm: 70 UAH plus a 5% chance of an additional 50 UAH
- The Public Official: 70 UAH
- The Member of Society: 70 UAH

If you choose to inform the Firm, earnings will depend both on the Firm's and your decisions. If the Firm decides to pay the Informed Payment Amount, the deal is done: the Firm has to pay the full amount and will win the public contract with a 100% chance. The Public Official will receive their True Payment Amount and you will keep the difference between the Informed Payment Amount and the True Payment Amount. The Member of Society will lose half of their initial endowment, i.e. 35 UAH. In other words, if you decide to inform the Firm, and they accept to pay the Informed Payment Amount, earnings will be as follows:

- You: 70 UAH + Informed Payment Amount – True Payment Amount
- The Firm: 70 UAH – Informed Payment Amount + 50 UAH
- The Public Official: 70 UAH + True Payment Amount
- The Member of Society: 70 UAH – 35 UAH

If the Firm refuses to pay the Informed Payment Amount, there will not be a deal and the Firm will win the public contract with a 5% chance. If that is the case, earnings will be as follows:

- You: 70UAH
- The Firm: 70UAH plus a 5% chance of an additional 50UAH
- The Public Official: 70UAH
- The Member of Society: 70UAH

Example

Please select a value of your choice for both the True and the Informed Payment Amount to see how your final earnings will change accordingly.

Remember that you can inform the Firm of a number equal to or higher than the True Payment Amount.

- True Payment Amount (dropdown menu with all values)
- Informed Payment Amount (dropdown menu with all values)

Your final earning are: (formula that calculates final payoff)

Questions to check your comprehension

The following questions will help you understand the consequences associated with the different options you choose in this experiment.

1) What are your monetary earnings if you decide not to inform the Firm? (70)
2) The Intermediary can inform the Firm of a number equal to or higher than the True Payment Amount. (TRUE)
3) Let's assume that the Public Official's True Payment Amount is 15 UAH, that you choose to report to the Firm 30 UAH as Informed Payment Amount, and that the Firm accepts to pay it. What are the resulting monetary earnings for the Member of Society? (35)

Now make your decision…

We will now reveal the piece of confidential information. We will then match your decision to those of the Firm you have been paired with to calculate the earnings you are entitled to. Don't forget that your decisions here are binding and you cannot change your mind once you have made your choice.

(Descriptive social norm variation)

Just to let you know, you are part of a group in which 75% of Intermediaries have just received an Integrity, Ethics and Anti-Corruption training / No information

The True Payment Amount which the Public Official must receive to serve the Firm is (random number).

Would you like to inform the Firm about this confidential information? Please select one of the two options below:

- Yes
- No

(If Yes)

Please select the Informed Payment Amount that you want to report to the Firm:
5UAH, 10UAH, 15UAH, 20UAH, 25UAH, 30UAH, 35UAH, 40UAH.

We want to hear from you

In this section, we will ask you some questions about your opinion of what the other Intermediaries in the game have decided to do. You have an opportunity to earn additional bonuses if you answer the following 5 questions. The closer your guesses are to the actual decisions of the other players, the higher your bonus will be. Note that only one question will be picked at random to calculate your additional bonus.

1. We have asked all the Intermediaries to decide whether they want to inform the Firm. Guess what share of all Intermediaries who took part in this game decided to "Inform"? Remember, the closer your guess is to the true percentage, the more bonus you earn.

   Slider with 0-100% in increments of 10%

   Your bonus will be calculated as follows:

If the difference between your answer and the true share of Intermediaries who "Informed" is:	Larger than 70%	Between 51 % and 70%	Between 30% and 50%	Between 10 and 30%	Between 1% and 10%	Exactly 0%
You earn 40 UAH with probability:	10%	30%	50%	70%	90%	100%

2. We have asked all the Intermediaries to guess what share of Intermediaries who took part in this game they think they decided to "Inform". Guess what the average Intermediary answered to this question. Remember, the closer your guess is to the true percentage, the more bonus you earn.

   Slider with 0-100% in increments of 10%

   Your bonus will be calculated as follows:

If the difference between your answer and the true share of Intermediaries who "Informed" is:	Larger than 70%	Between 51 % and 70%	Between 30% and 50%	Between 10 and 30%	Between 1% and 10%	Exactly 0%
You earn 40 UAH with probability:	10%	30%	50%	70%	90%	100%

3. All Intermediaries who chose to inform were asked how much they wanted to inform. Guess how large was the amount overreported on average? (Example: if the True Payment Amount is 20 UAH and the Intermediary informed 30 UAH, the amount overreported is 30-20 = 10 UAH). Remember, the closer your guess is to the true percentage, the more bonus you earn.
   0 UAH, 5 UAH, 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH.

   Your bonus will be calculated as follows:

If the difference between your answer and the true average amount overreported is:	Larger than 30 UAH	Between 26 and 30 UAH	Between 16 and 25 UAH	Between 6 and 15 UAH	Between 1 and 5 UAH	Exactly 0
You earn 40 UAH with probability	10%	30%	50%	70%	90%	100%

4. We asked all the Intermediaries to guess how much other Intermediaries overreported on average. Guess what the average Intermediary answered to this question? Remember, the closer your guess is to the true percentage, the more bonus you earn.

   0 UAH, 5 UAH, 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH.

   Your bonus will be calculated as follows:

If the difference between your answer and the true average amount overreported is:	Larger than 30 UAH	Between 26 and 30 UAH	Between 16 and 25 UAH	Between 6 and 15 UAH	Between 1 and 5 UAH	Exactly 0
You earn 40 UAH with probability	10%	30%	50%	70%	90%	100%

5. We have asked all the Firms to decide whether they want to pay the amount informed by the Intermediary. Guess how much the average Firm was willing to pay at most? Remember, the closer your guess is to the true percentage, the more bonus you earn.

   0 UAH, 5 UAH, 10 UAH, 15 UAH, 20 UAH, 25 UAH, 30 UAH, 35 UAH, 40 UAH.

   Your bonus will be calculated as follows:

If the difference between your answer and the true average amount overreported is:	Larger than 30 UAH	Between 26 and 30 UAH	Between 16 and 25 UAH	Between 6 and 15 UAH	Between 1 and 5 UAH	Exactly 0
You earn 40 UAH with probability:	10%	30%	50%	70%	90%	100%

Thank you for your participation!

At the end of the game, you will be randomly matched with a Firm, a Public Officer and a Member of Society and your respective choices will be enacted. We will let you know the gift value you have earned, and you will receive your gift by email in the next 14 days.

Firm

Your role

You have been assigned to the role of Firm in today's activity.

To begin with, you have 70 UAH as initial budget, like all the other players participating in the activity.

The decision situation begins with the Intermediary's choice of whether to Inform or Not Inform you about the amount that the Public Official needs to receive in order to award you a contract worth 50 UAH.

Possible scenarios…

On one hand, if the Intermediary chooses not to inform, there is no possible deal between you and the Intermediary, and you will win the public contract with a 5% chance. If that is the case, earnings will be as follows:

- You: 70 UAH plus a 5% chance of an additional 50 UAH
- The Intermediary: 70 UAH
- The Public Official: 70 UAH
- The Member of Society: 70 UAH

On the other hand, if the Intermediary chooses to inform, you have then two options: 1) you can accept to pay the Intermediary's Informed Payment Amount; or 2) you can refuse to make the payment.

If you choose to refuse to pay the Informed Payment Amount, there will not be a deal between you and the Intermediary and you will win the public contract with a 5% chance. If that is the case, earnings will be as follows:

- You: 70 UAH plus a 5% chance of an additional 50 UAH
- The Intermediary: 70 UAH
- The Public Official: 70 UAH
- The Member of Society: 70 UAH

If you decide to accept to pay the Informed Payment Amount, it means that the deal is done: you have to pay the full amount informed by the Intermediary and will win the public contract with a 100% chance. The Public Official will receive his True Payment Amount and the Intermediary will keep the difference between the Informed Payment Amount and the True Payment Amount. The Member of Society will lose half of their initial endowment, that is 35 UAH. In other words, if you decide to accept to pay the Intermediary, earnings will be as follows:

- You: 70 UAH – Informed Payment Amount + 50 UAH
- The Intermediary: 70 UAH + Informed Payment Amount – True Payment Amount
- The Public Official: 70 UAH + True Payment Amount
- The Member of Society: 70 UAH – 35 UAH

Example

Please select a value of your choice for both the True and the Informed Payment Amount to see how your final earnings will change accordingly.

Remember that the Intermediary can inform you about a number equal to or higher than the True Payment Amount.

- True Payment Amount (dropdown menu with all values)
- Informed Payment Amount (dropdown menu with all values)

Your final earning are (formula that calculates final payoff)

Questions to check your comprehension

The following questions will help you understand the consequences associated with the different options you choose in this experiment.

1. What are your monetary earnings if the Intermediary decides not to inform you? (70 UAH plus a 5% chance of an additional 50 UAH)
2. The Intermediary can inform the Firm of a number equal to or higher than the True Payment Amount. (TRUE)

3. Let's assume that the Public Official's True Payment Amount is 15 UAH, that the Intermediary chooses to inform you that the Informed Payment Amount is 30 UAH, and that you accept to pay it. What are the resulting monetary earnings for the Member of Society? (35)

Now make your decision…

We would now like to ask you whether you would accept or refuse to pay any possible amount that the Intermediary might report to you. We will then match your decisions to those of the Intermediary you have been paired with to calculate the gift you are entitled to.

Your decisions here are binding and you cannot change your mind once you have made your choice. We will not allow inconsistencies in the decisions, i.e. cannot reject 10 UAH but accept 15 UAH.

Please tick the appropriate buttons below to let us know whether you would pay or not the Informed Payment Amount, for all possible amounts:

Informed Payment Amount	Your decision	
5 UAH	Pay	Not Pay
10 UAH	Pay	Not Pay
15 UAH	Pay	Not Pay
20 UAH	Pay	Not Pay
25 UAH	Pay	Not Pay
30 UAH	Pay	Not Pay
35 UAH	Pay	Not Pay
40 UAH	Pay	Not Pay

Thank you for your participation!

At the end of the game, you will be randomly matched with an Intermediary, a Public Officer and a Member of Society and your respective choices will be enacted. We will let you know the gift value you have earned, and you will receive your gift by email in the next 14 days.

Public Officer

Your role

You have been assigned to the role of Public Officer in today's activity.

To begin with, you have 70 UAH as initial budget, like all the other players participating in the activity.

The True Payment Amount that you must receive to award the public contract to the Firm with a 100% chance is (random number).

Thank you for your participation!

At the end of the game, you will be randomly matched with a Firm, an Intermediary and a Member of Society and your respective choices will be enacted. We will let you know the gift value you have earned and you will receive your gift by email in the next 14 days.

Member of Society

Your role

You have been assigned to the role of Member of Society in today's activity.

To begin with, you have 70 UAH as initial budget, like all the other players participating in the activity.

Thank you for your participation!

At the end of the game, you will be randomly matched with a Firm, an Intermediary and a Public Official and your respective choices will be enacted. We will let you know the gift value you have earned, and you will receive your gift by email in the next 14 days.

Table 2.B.1: Summary statistics by treatment assignment

	All	Control	Training Only	Information	Training + Information
Prior ethics training	0.37	0.36	0.32	0.42	0.44
	(0.48)	(0.48)	(0.47)	(0.50)	(0.50)
Junior	0.45	0.42	0.44	0.39	0.54
	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)
Female	0.71	0.74	0.61	0.75	0.75
	(0.45)	(0.44)	(0.49)	(0.44)	(0.43)
Enroll for Skills	0.85	0.88	0.79	0.91	0.87
	(0.35)	(0.33)	(0.41)	(0.29)	(0.34)
Enroll for Prize	0.00	0.00	0.00	0.00	0.02
	(0.06)	(0.00)	(0.00)	(0.00)	(0.13)
Corrupt Leaders	0.00	-0.08	0.10	-0.14	0.09
	(0.95)	(1.00)	(0.88)	(1.03)	(0.90)
Corrupt People	0.07	0.01	0.15	-0.14	0.12
	(0.94)	(0.97)	(0.82)	(0.88)	(0.98)
Corruption is Risky	0.05	0.21	-0.15	0.27	-0.09
	(0.94)	(0.89)	(1.01)	(0.87)	(0.94)
Corruption is Justified	-0.04	-0.03	-0.00	-0.00	-0.11
	(0.91)	(0.96)	(0.84)	(0.88)	(0.86)
Make Money	0.06	0.05	0.12	0.06	0.02
	(0.24)	(0.21)	(0.33)	(0.24)	(0.14)
For Parents	0.01	0.02	0.02	0.04	0.00
	(0.12)	(0.14)	(0.14)	(0.20)	(0.00)
For Prestige	0.12	0.11	0.16	0.10	0.10
	(0.33)	(0.32)	(0.37)	(0.30)	(0.31)
Make a Difference	0.40	0.39	0.37	0.37	0.48
	(0.49)	(0.49)	(0.49)	(0.49)	(0.50)
Have Right Skills	0.31	0.32	0.24	0.31	0.33
	(0.46)	(0.47)	(0.43)	(0.47)	(0.48)
Other Reason	0.09	0.11	0.08	0.12	0.06
	(0.29)	(0.32)	(0.28)	(0.33)	(0.24)
Judge or Prosecutor	0.27	0.26	0.27	0.25	0.31
	(0.45)	(0.44)	(0.45)	(0.44)	(0.47)
Defense Attorney	0.37	0.34	0.43	0.37	0.35
	(0.48)	(0.48)	(0.50)	(0.49)	(0.48)

Continued on next page

Table 2.B.1 (continued)

	All	Control	Training Only	Information	Training + Information
Corporate Attorney	0.13	0.17	0.08	0.14	0.08
	(0.33)	(0.37)	(0.28)	(0.35)	(0.28)
Care	4.74	4.82	4.60	4.80	4.70
	(0.64)	(0.63)	(0.67)	(0.68)	(0.63)
Fairness	4.76	4.76	4.77	4.75	4.75
	(0.59)	(0.63)	(0.47)	(0.64)	(0.64)
Ingroup Loyalty	4.25	4.24	4.23	4.22	4.29
	(0.66)	(0.69)	(0.60)	(0.69)	(0.68)
Authority	3.70	3.74	3.64	3.81	3.67
	(0.75)	(0.81)	(0.65)	(0.78)	(0.72)
Purity	4.00	4.10	3.87	4.14	3.90
	(0.80)	(0.83)	(0.76)	(0.85)	(0.76)
Observations	252	129	62	64	61

Note: This table displays means and standard deviations of the main control variables used in the analysis.

Table 2.B.2: Local Average Treatment Effects

	All students			Juniors Only			Seniors Only		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Attend Training	0.01	0.01	0.02	-0.16	-0.17	-0.18*	0.17	0.15	0.14
	(0.11)	(0.11)	(0.07)	(0.15)	(0.15)	(0.1)	(0.16)	(0.16)	(0.09)
Norm Only Treatment	-0.04	-0.04	.	-0.25*	-0.26*	.	0.12	0.1	.
	(0.09)	(0.09)		(0.13)	(0.13)		(0.11)	(0.12)	
Attend Training + Norm Treatment	0.005	0.003	.	-0.25*	-0.26*	.	0.29	0.25	.
	(0.12)	(0.12)		(0.15)	(0.15)		(0.18)	(0.18)	
Treated	.	.	-0.09	.	.	-0.3	.	.	0.01
			(0.15)			(0.27)			(0.17)
Junior	.	-0.01	-0.003	.	.	.	.	.	.
		(0.06)	(0.06)						
Prior ethics training	.	0.01	0.02	.	-0.17*	-0.18*	.	0.12	0.14
		(0.07)	(0.07)		(0.1)	(0.1)		(0.09)	(0.09)
Enroll for Skills	.	0.05	0.05	.	0.03	0.03	.	0.07	0.07
		(0.09)	(0.09)		(0.12)	(0.12)		(0.13)	(0.13)
Female	.	-0.01	-0.02	.	0.07	0.07	.	-0.03	-0.04
		(0.07)	(0.07)		(0.1)	(0.1)		(0.1)	(0.1)
Constant	0.55***	0.53***	0.57***	0.69***	0.68***	0.73***	0.44***	0.38**	0.45**
	(0.06)	(0.11)	(0.13)	(0.09)	(0.14)	(0.18)	(0.08)	(0.16)	(0.18)
Observations	252	252	252	114	114	114	138	138	138
R^2	0.001	0.003	.	0.05	0.07	0.05	0.003	0.02	0.03

Note: Two stage least squares regression coefficients. Columns [1] & [2] report LATE coefficients pooled across all students and in columns [3] & [4] and [5] & [6] we report effects for Juniors only and Seniors only, respectively. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 2.B.3: Intention to Treat: Quantity Embezzled

	Conditional on facilitating a bribe			Two-step selection model		
	All students	Jr Only	Sr Only	All students	Jr Only	Sr Only
	[1]	[2]	[3]	[4]	[5]	[6]
Train Only Treatment	-0.01	0.18	-0.21	-0.03	0.18	-0.27
	(0.1)	(0.13)	(0.15)	(0.3)	(0.39)	(0.47)
Norm Only Treatment	-0.07	-0.38**	-0.02	-12.09	1.25	-1.74
	(0.1)	(0.17)	(0.14)	(22.13)	(5.26)	(1.09)
Train + Norm Treatment	0.2**	0.32**	0.14	0.19	0.24	0.21
	(0.09)	(0.13)	(0.14)	(0.31)	(0.39)	(0.47)
Prior ethics training	0.08	0.36***	-0.21**	3.02	2.70	-25.25
	(0.07)	(0.1)	(0.11)	(5.48)	(7.14)	(15.45)
Female	0.31***	0.76***	-0.01	-3.15	-0.2	6.63
	(0.08)	(0.13)	(0.11)	(6.45)	(2.88)	(4.10)
Enroll for Skills	-0.36***	-0.09	-0.59***	11.49	0.21	-12.95*
	(0.09)	(0.14)	(0.13)	(21.92)	(0.65)	(7.56)
Inverse mills ratio	.	.	.	-156.71	8.07	101.85
				(289.16)	(24.94)	(62.91)
Constant	1.72***	1.22***	2.33***	133.42	-6.69	-72.36
	(0.13)	(0.17)	(0.16)	(242.82)	(24.12)	(46.21)
Observations	138	62	76	138	62	76

Note: Poisson regression coefficients. Cols [1] - [3] regress the quantity embezzled conditional on the subject choosing to facilitate a bribe. Cols [4] - [6] present results of a two-step model that first estimates the probability of facilitating a bribe (see Table 2.2) and accounts for the selection into bribery when estimating the amount embezzled. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Chapter 3

Mind how you frame your compliance demands: The limits of collaboration and incentives in multi-tier supply chains

Severe human rights violations in global supply chains often occur upstream with limited oversight. Buyers rely on direct suppliers to monitor sub-suppliers, but largely uncompensated top-down compliance demands may be perceived as unfair, weakening engagement. This paper tests whether framing and incentivization of compliance requests shape first-tier supplier behaviour through fairness perceptions. In a pre-registered online experiment simulating a multi-tier supply chain, we compare a baseline deterrence approach with two treatments: deterrence plus collaborative framing, and deterrence plus collaborative framing with a conditional financial incentive. We measure effort allocation across monitoring sub-suppliers, own compliance, and production. Deterrence sustains high monitoring, while collaborative framing reduces strict monitoring, and incentives further erode fairness perceptions and compliance. Effects vary across participants: private-sector participants shift from strict to moderate monitoring, whereas public-sector and highly prosocial participants reduce compliance. Findings highlight risks of hybrid approaches combining deterrence with collaborative framing and financial incentives.

[0]This chapter is joint work with Oana Borcan, Amrish Patel, and Theodore Turocy.

3.1 Introduction

Beyond canonical forms of abuse of entrusted power, such as corruption and fraud, other forms of rule-breaking in principal–agent settings can generate substantial organizational and societal costs. Multinational enterprises (MNEs) rely on globally dispersed suppliers and sub-suppliers to produce goods while requiring social compliance with international labour standards, typically enforced through deterrence-based mechanisms such as codes of conduct and third-party audits.[1] However, monitoring remains concentrated on first-tier suppliers, even though violations are often most severe among sub-suppliers, as illustrated by the 2013 Rana Plaza collapse (Aßländer et al., 2016; Wilhelm et al., 2016; Fontana and Egels-Zandén, 2019).[2] MNEs have therefore increasingly delegated compliance responsibilities to first-tier suppliers, expecting them to both comply and monitor sub-suppliers. Yet this model faces key limitations: while production is rewarded, compliance and monitoring are not, and requirements are imposed top-down, leading suppliers to perceive an unfair distribution of costs and responsibilities and weakening incentives to engage beyond minimum requirements (Locke et al., 2009; Aßländer et al., 2016; Soundararajan and Brammer, 2018; Amengual et al., 2019; Kuruvilla et al., 2020). With the adoption of the EU's 2024 Corporate Sustainability Due Diligence Directive (European Union, 2024), understanding how to motivate compliance beyond first-tier suppliers has become increasingly urgent.

In this paper, we examine whether modifying how MNEs' social compliance requests are communicated and incentivized — by making them more collaborative or by introducing conditional rewards — affects first-tier suppliers' perceptions of fairness and, in turn, their monitoring of sub-suppliers and voluntary compliance. To do

[1] A supplier code of conduct is a legal document drafted by MNEs that includes a list of social and environmental standards by which suppliers need to abide and explains the penalty for non-compliance. To ensure these standards are implemented rather than symbolic, MNEs typically enforce them through audit-based monitoring strategy, creating top-down compliance pressures throughout their supply chains (Goerzen and Van Assche, 2023).

[2] On April 24, 2013, the Rana Plaza building in Dhaka, Bangladesh, collapsed, killing 1,134 people and injuring over 2,500, making it the deadliest disaster in the history of the garment industry. The building housed several garment factories producing for major European and North American brands but was structurally unsound and not designed to support heavy industrial machinery. Despite visible cracks and evacuation warnings issued the day before, workers were instructed to return to work, and the building collapsed the following morning (Koenig and Poncet, 2022).

so, we study hybrid governance approaches that combine credible enforcement with collaborative framing and incentives. Specifically, we test whether a buyer's ("Alpha") compliance request that complements deterrence with a collaborative framing increases first-tier suppliers' ("Beta") perceptions of fairness and, through this channel, encourages them to increase monitoring of sub-suppliers ("Gamma") and improve their own compliance, relative to a purely deterrence-based approach. We further examine whether introducing a conditional financial incentive for detecting and reporting sub-supplier non-compliance amplifies these effects.

We study these mechanisms in a pre-registered online experiment with 427 participants from South Africa. Participants acted as first-tier suppliers, deciding how to allocate their time and effort across production, their own compliance, and monitoring sub-suppliers under one of three conditions: (i) a baseline deterrence approach (Control), (ii) deterrence combined with collaborative framing emphasizing shared responsibility and "leading by example" (Treatment 1), and (iii) deterrence combined with collaborative framing and a financial incentive conditional on detecting and reporting sub-supplier non-compliance (Treatment 2).

Our findings show that, contrary to our initial assumptions, a pure deterrence-based approach (C) sustains relatively high levels of monitoring, even in the presence of only a modest probability of detection and sanction. In contrast, the addition of collaborative framing (T1) does not improve average fairness perceptions or compliance relative to the control condition, but reduces the likelihood of selecting the strictest inspection policy, suggesting a weakening of monitoring intensity. This pattern is consistent with the interpretation that collaborative framing may alter the perceived appropriateness of strict monitoring. The introduction of conditional financial incentives (T2) yields more adverse effects. While average monitoring effort remains unchanged, participants are less likely to perceive Alpha's requests as fair or effective, and compliance outcomes decline, consistent with a crowding-out of intrinsic or relational motivations.

Heterogeneity analysis reveals distinct motivational patterns across groups. Private sector participants tend to substitute strict monitoring with more moderate inspection strategies while maintaining relatively stable compliance levels, suggesting greater responsiveness to the relational framing of the task. In contrast, public-sector and highly prosocial participants maintain stable monitoring but reduce

compliance under both treatments, consistent with the observed decline in fairness perceptions. Among highly prosocial individuals in particular, both collaborative framing and conditional incentives reduce the likelihood of selecting strict monitoring and weaken compliance outcomes, in line with a crowding-out of intrinsic and relational motivations. At the supply chain level, these dynamics are reinforced by robust evidence that T2 reduces both fairness perceptions and the likelihood of joint compliance. Taken together, the results suggest that hybrid governance approaches combining deterrence with collaborative signals and financial elements may generate conflicting cues that undermine perceptions of fairness and, in turn, weaken monitoring and compliance in different ways across groups.

Our study makes several contributions to the existing literature. First, it makes a methodological contribution by bridging insights from regulatory theory, supply chain management, and behavioural and experimental economics. Prior regulatory research distinguishes between deterrence-based approaches, which rely on monitoring and sanctions, and voluntary compliance frameworks grounded in trust, legitimacy, and fairness (Feldman, 2025). Evidence shows that while deterrence can reduce non-compliance, it may also undermine intrinsic motivation when perceived as unfair or distrustful, making people less likely to go beyond the minimum required (Fehr et al., 1997; Fehr and Schmidt, 2004; Schildberg-Hörisch and Strassmair, 2012; Kessler and Leider, 2016; Feldman, 2025). By contrast, trust-based approaches can enhance voluntary compliance, but primarily among individuals already inclined to comply. Hybrid approaches that combine enforcement with trust-building have been shown to improve compliance in domains such as tax enforcement, yet their effectiveness in social compliance, particularly in multi-tier global supply chains, remains largely unexplored in experimental settings (Feldman, 2025). Existing research in this area is predominantly qualitative or correlational, limiting causal inference (Locke et al., 2009; Aßländer et al., 2016; Wilhelm et al., 2016; Soundararajan and Brammer, 2018; Amengual et al., 2019; Kuruvilla et al., 2020). This study addresses these gaps by using a controlled experiment to causally identify how different combinations of enforcement, trust-building, and incentives shape fairness perceptions and, in turn, influence both compliance and delegated monitoring across supply-chain tiers – mechanisms that are difficult to observe in real-world settings.

Second, we extend traditional deterrence and inspection models to a multi-task principal-agent setting. Unlike standard inspection games where the principal and agent interact strategically over a single monitored task (Rauhut, 2009; Nosenzo et al., 2010; Fandel and Trockel, 2013; Rauhut, 2015), our design includes a principal (Alpha) who performs a non-strategic compliance activity and an agent (Beta) responsible for multiple tasks — production, self-compliance, and monitoring of sub-suppliers (Gamma) — with only production being financially rewarded. By incorporating a modest probability of detection for both Beta and Gamma, combined with credible commercial sanctions on Beta, we provide new experimental evidence that credible deterrence can stimulate effort in otherwise unrewarded tasks and reinforce vertical enforcement mechanisms. These findings complement recent quasi-experimental evidence from Amengual and Distelhorst (2025), who show that credible penalty threats improved compliance among direct suppliers narrowly failing audits in Gap Inc.'s supply chain. Our experiment extends this evidence by demonstrating that deterrence can not only sustain compliance among tier-1 suppliers but also incentivize them to monitor sub-suppliers — an enforcement structure largely absent in current practice but potentially more relevant under emerging due-diligence regimes.

Third, we contribute to research on leadership, group identity, tone at the top, and fairness perceptions in shaping cooperative behaviour and compliance. Evidence from social psychology and experimental economics indicates that followers' behaviour is strongly influenced by leaders' actions and signals (Potters et al., 2007; Figuières et al., 2012; Gächter et al., 2012), and that leadership effectiveness is strengthened by perceived social similarity between leaders and followers (Tajfel and Turner, 1986; Burger et al., 2004; Cialdini and Goldstein, 2004; Drouvelis and Nosenzo, 2013). Complementing this, the "tone at the top" literature in corruption and organizational ethics highlights how leaders shape ethical conduct also through communication that signals trust, fairness, and shared norms (Graf Lambsdorff, 2015). A positive tone expressed through example, recognition, or inclusive communication can foster intrinsic motivation and reciprocity, whereas signals of distrust may undermine them. Informed by this evidence, our first treatment tests whether fairness perceptions can be enhanced through a relational, trust-building, and "leading by example" framing, in which the principal (Alpha) signals its own

compliance efforts and uses inclusive, group-identity language. Unlike most leader-follower experiments where leaders make strategic decisions, our design isolates the effect of communication by manipulating only the framing of the leader's message. This allows us to provide new evidence on the limits of fairness-based persuasion through "leading by example" messages in hierarchical, deterrence-oriented contexts.

Finally, we extend this analysis by examining how the combination of deterrence, trust-building framing, and conditional financial incentives shapes fairness perceptions and compliance. In our setting, introducing a conditional financial incentive for reporting sub-supplier non-compliance amplifies negative effects, reducing perceived fairness and compliance. This finding aligns with behavioural evidence on motivational crowding out, which shows that external incentives — particularly when perceived as controlling — can undermine intrinsic and relational motivations (Gneezy and Rustichini, 2000; Bowles and Polania-Reyes, 2012; Fiorin, 2023). Importantly, these negative effects are not uniform across participants. The decline in compliance is primarily driven by public-sector and more prosocial individuals – groups typically more responsive to intrinsic and relational motivations – suggesting that financial incentives may crowd out precisely the motivations that sustain voluntary cooperation. Our results also connect to the growing literature on whistleblower rewards, which examines how financial incentives shape individuals' willingness to report wrongdoing (Nyreröd and Spagnolo, 2021). Experimental studies show that while rewards can increase reporting (Schmolke and Utikal, 2018; Butler et al., 2019), they may also backfire when reporting entails moral costs — as in Fiorin (2023)'s field experiment in Afghanistan, where monetary rewards for peer reporting among public-school employees reduced whistleblowing when reports could lead to punishment of colleagues. Our study differs in focus: we did not aim to measure willingness to report per se, but to examine how conditional incentives affect compliance and monitoring efforts. We show that in a deterrence-oriented setting, combining deterrence mechanisms with "leading by example" messages and conditional rewards backfires, reducing Beta's perceived fairness and compliance efforts. This likely reflects the perception of Betas as "Alpha's police" which undermined trust and relational reciprocity within the supply chain (Frenkel and Scott, 2002; Dyer and Chu, 2003).

The remainder of the paper is structured as follows. Section 3.2 presents the

conceptual framework, followed by the experimental design in section 3.3. Section 3.4 describes outcome measures and hypotheses. Section 3.5 presents the data and methodology. Section 3.6 describes the results, and section 3.7 concludes.

3.2 Conceptual framework

Supplier codes of conduct and third-party audits are private governance mechanisms commonly used by multinational enterprises (MNEs) to induce social compliance with labour standards in global supply chains. These tools operate under conditions of limited observability, asymmetric information, and misaligned incentives, and are designed to deter violations by increasing the likelihood that misconduct is detected and penalized.

However, their effectiveness is limited. Monitoring remains largely concentrated on first-tier suppliers, while labour violations are often most severe further down the supply chain (Fontana and Egels-Zandén, 2019). In response, MNEs have increasingly delegated social compliance responsibilities to first-tier suppliers. These suppliers are expected to take on a triple agency role: meeting production targets, ensuring compliance in their own operations, and monitoring sub-suppliers. Yet, while production is directly rewarded, compliance and monitoring are typically not. This structural imbalance creates persistent tension between economic and social objectives and contributes to perceptions that buyers' compliance demands are unrealistic or unfair (Locke et al., 2009; Aßländer et al., 2016; Amengual et al., 2019; Kuruvilla et al., 2020).

We conceptualize this governance structure as giving rise to two sources of perceived unfairness. First, compliance requirements are imposed in a top-down manner. Second, compliance and monitoring tasks impose costs that are not directly compensated. Together, these features can lead suppliers to perceive compliance demands as involving an unfair distribution of both responsibilities and costs of social compliance, weakening incentives to engage beyond minimum requirements. This raises the central research question of this paper: can modifying how compliance requests are communicated and incentivized — by making them more collaborative or by introducing conditional rewards — improve suppliers' perceptions of fairness and, in turn, increase monitoring of sub-suppliers and voluntary compliance?

Building on insights from regulatory theory, supply chain management, and behavioural economics, we conceptualize a buyer's (Alpha's) compliance request as triggering a chain of perceptual and behavioural responses. First, the way compliance expectations are designed and communicated shapes how they are perceived. Second, first-tier suppliers (Beta) form views about the fairness of these expectations. Third, these perceptions influence behaviour, including how effort is allocated between monitoring sub-suppliers and ensuring own compliance.

A key assumption underlying this framework is that compliance and monitoring are perceived as legitimate contributions to a shared objective — namely, improving working conditions along the supply chain. If this assumption holds, fairness-enhancing signals can reinforce monitoring and voluntary compliance beyond minimum levels. If it does not, similar signals may instead alter how tasks are interpreted without increasing perceived fairness.

To examine these mechanisms in a controlled but realistic setting, we design a contextualized online experiment in which participants assigned to the role of Beta make real-effort decisions under three conditions that vary how Alpha's compliance expectations are communicated and enforced:

1. **Deterrence-based approach (Control).** This condition is designed to mimic a stylized real-world setting where social auditing among first-tier and second-tier suppliers is possible but limited. Compliance is enforced through a credible but relatively low probability of detection and sanction. Moreover, the top-down nature of the social compliance request and the lack of compensation for monitoring and compliance tasks may render it unfair, thereby limiting incentives to engage beyond minimum monitoring and compliance.

2. **Hybrid approach combining deterrence with collaborative framing (Treatment 1).** This condition seeks to mitigate the perception of top-down imposition by introducing a relational framing of compliance. The buyer emphasizes shared responsibility through inclusive "we"-language and signals its own compliance efforts ("leading by example" and "tone at the top"). Such framing may enhance perceived fairness by reducing the sense of unilateral imposition.

3. **Hybrid approach combining deterrence with collaborative framing**

and a conditional financial incentive (Treatment 2). This condition builds on Treatment 1 by introducing a conditional financial incentive for detecting and reporting sub-supplier non-compliance. This aims to mitigate also the second source of unfairness – namely, the absence of compensation for monitoring and compliance – by linking these activities to a monetary reward.

Based on this framework, we derive the following testable hypotheses.

Behavioural outcomes (monitoring and compliance):

- **H1:** Collaborative framing increases suppliers' monitoring effort relative to a traditional deterrence-based approach by improving perceptions of fairness and legitimacy of the buyer's compliance expectations;
- **H2:** Adding conditional financial incentives increases monitoring effort beyond the effect of trust-building framing alone by strengthening the expected returns to monitoring;
- **H3:** Collaborative framing may affect suppliers' own compliance effort, but the direction of this effect is theoretically ambiguous. While fairness and "leading by example" may promote reciprocity and increase effort, increased monitoring demands may crowd out the time and attention allocated to own compliance .

Mechanism (fairness perceptions):

- **H4:** Collaborative framing increases suppliers' perceptions of fairness of the buyer's compliance expectations.

Finally, we explore whether responses differ across participant types. In particular, private- and public-sector participants, as well as individuals with different levels of prosocial orientation, may perceive compliance requests differently and respond to relational and incentive-based signals in distinct ways.

3.3 Experimental design

3.3.1 Decision environment

We study these mechanisms through a contextualized online experiment simulating a stylized three-tier global supply chain that produces a fictitious good, XYZ.[3] The lead firm, Alpha (a European multinational), procures XYZ from Beta, a medium-sized South African first-tier supplier. To meet Alpha's production order, Beta outsources part of its output to smaller local manufacturers. Gamma represents one of these subcontractors, acting as a second-tier supplier in the chain. Alpha's objectives are to achieve the production target and ensure compliance with labour standards across the entire supply chain.

This Global North-South structure reflects the organization of many global supply chains, where MNEs outsource labour-intensive production to suppliers in emerging economies. To capture this context, participants playing the role of Beta and Gamma were recruited from South Africa via the online platform Prolific. At the time of data collection, South Africa was one of the few emerging economies on the platform with a sufficiently large and active participant pool. This approach enhances contextual realism while allowing for a sufficiently large and heterogeneous sample within a single emerging-country context, while maintaining a controlled and anonymous decision environment.

The game unfolds as follows. Alpha issues an urgent production order to Beta, requesting the delivery of up to 100 units of good XYZ and requiring that both Beta and its subcontractors comply with at least 5 out of 15 labour standards. Alpha also informs Beta that it conducts compliance inspections in 1 out of every 10 orders, auditing both direct suppliers and subcontractors.[4] If an inspection reveals

[3]The pre-analysis plan for this experiment was registered on the AEA RCT Registry (RCT ID AEARCTR-0015674) on 30 March 2025. Unless otherwise specified, our analysis follows the approach set out in the pre-analysis plan. Ethical approval for this study was granted by the School of Economics Research Ethics Subcommittee (ETH2425-0106) at the University of East Anglia.

[4]In real-world settings, buyers (Alpha) conduct frequent audits of their tier-1 suppliers (Beta). Auditing at the tier-2 level (Gamma) is becoming more common but remains far less systematic, while inspections rarely extend beyond that tier (Bloemer and Minner, 2025). To keep the experimental design manageable while maintaining credibility in detecting violations at both levels, we assume that Alpha can also audit Gamma. However, rather than assigning different detection probabilities to each tier, we apply a single moderate probability (10%) across the chain to cap-

that either Beta or any of its subcontractors fail to meet the compliance threshold, Alpha terminates the business relationship with Beta, and no payment is made for any of the units delivered.

Because the order is urgent, Beta must outsource half of the production (50 units) to a subcontractor, Gamma. Beta is therefore responsible for three things:

1. **Inspection**, by committing to inspect a random share of subcontractors (10%, 30%, or 50%), where higher inspection rates entail greater time costs.

2. **Compliance**, by using the remaining time to meet Alpha's minimum requirement of complying with at least 5 out of 15 labour standards. This task has no direct financial reward but is necessary to maintain Alpha's business relationship.

3. **Production**, by using the remaining time to produce up to 50 units of XYZ. Each correctly produced unit yields a monetary reward, with total earnings depending on the combined production of Beta and Gamma.

Beta's challenge is to balance these three competing activities within a fixed time budget of 180 seconds, knowing that failure to comply — either by Beta or by Gamma — can result in termination of the contract.

Gamma, in turn, receives a subcontracting order from Beta for 50 units of XYZ and faces the same compliance requirement as Beta. Gamma must decide how to divide their available time between production and compliance, knowing the probability of being inspected by Beta. If Gamma is inspected and found non-compliant, Beta terminates the subcontract and Gamma receives no payment. Detailed game instructions can be found in Figure 3.A.2.

3.3.2 Tasks

While Beta is the primary decision-maker of interest, the inclusion of active players Alpha and Gamma was essential to maintain realism, ensure consistency in instructions and incentives, and avoid deception.

ture the lower overall likelihood of detection in real supply networks while preserving simplicity for participants.

Participants assigned to the role of Alpha were given 60 seconds to perform a compliance activity. The task was operationalized as a transcription exercise: participants had to correctly transcribe at least 5 out of 15 images, each displaying a sequence of five letters, representing the 15 labour standards. Alphas were not required to make any strategic decisions. However, their performance mattered in Treatments 1 and 2. In these treatments, the best Alpha score (12 correctly transcribed images in our case) was disclosed to Betas and Gammas as a signal of Alpha's own commitment to compliance with labour standards.[5]

Participants assigned to the role of Beta were allocated 180 seconds, which they had to distribute across three competing activities:

1. **Inspection**, which determines Beta's ability to monitor how many labour standards Gamma complies with. Before allocating time to compliance and production, Beta was required to select one of three inspection policies to determine the probability of inspecting Gamma (Figure 3.1). Each policy incurred a time cost, reducing time available for production and compliance:

   - **10% inspection policy**: Inspects 1 of 10 subcontractors (costing 55 seconds)
   - **30% inspection policy**: Inspects 3 of 10 subcontractors (costing 75 seconds)
   - **50% inspection policy**: Inspects 5 of 10 subcontractors (costing 95 seconds)

   After choosing the inspection policy, Beta was prompted to divide the remaining time between compliance and production (Figure 3.2).

2. **Compliance**, which simulates Beta's efforts to comply with the labour standards. The activity was implemented as a transcription task, similar to the one performed by Alpha (Figure 3.3).[6] Correctly transcribing at least 5 out of

[5]For Alpha only, the compliance task was linked to a donation to a charity fighting child labour, with the size of the donation proportional to compliance performance. To simplify the instructions, the compliance task for Beta and Gamma did not involve any donation.

[6]Pilot data showed low performance on the compliance task across roles. We therefore simplified the task for Betas and Gammas, reducing the number of letters in each word from five to four and making them not case-sensitive.

15 images satisfied Alpha's minimum compliance requirements. This task offered no direct financial reward but was essential for maintaining the business relationship with Alpha (upon inspection) and securing the extra earnings from the production activity. If an inspection by Alpha revealed that either Beta or Gamma failed to meet the compliance threshold, Alpha terminated the business relationship with Beta, and no payment was made for any units delivered. The only exception applied in T2: if Beta inspected Gamma, detected non-compliance, and reported it to Alpha, Beta was spared this sanction.

3. **Production**, which simulates Beta's efforts to produce units of good XYZ. The activity was operationalized as a slider task, in which participants were required to position 50 sliders to a target value between 0-100 (63 in our case), representing the 50 units of good XYZ to be produced by Beta (Figure 3.4). Each correctly positioned slider earned 1 South African Rand (R).[7] Beta's total production earnings depended on the combined production performance of both Beta and Gamma.

Participants assigned to the role of Gamma were allocated 120 seconds, which they had to distribute across two competing activities:

1. **Compliance**, which simulates Gamma's efforts to comply with the labour standards (operationalized as above).

2. **Production**, which simulates Gamma's efforts to produce units of good XYZ (operationalized as above). Gamma's earnings depended solely on their own production and compliance performance. If Gamma was inspected by Beta and found non-compliant, the subcontract was terminated and no payment was made. Alpha's inspections did not directly affect Gamma's payoff.

Before the game, participants completed a 20-second practice round of their activities and a short comprehension quiz on the rules. These steps familiarized participants with the tasks and ensured understanding of the setup. The number of quiz attempts was recorded as a measure of attentiveness. At the end of the game, participants completed a short survey eliciting demographic information (gender,

[7] At the time of data collection (March-May 2025), R1 = £0.04 approximately.

Figure 3.1: Inspection task

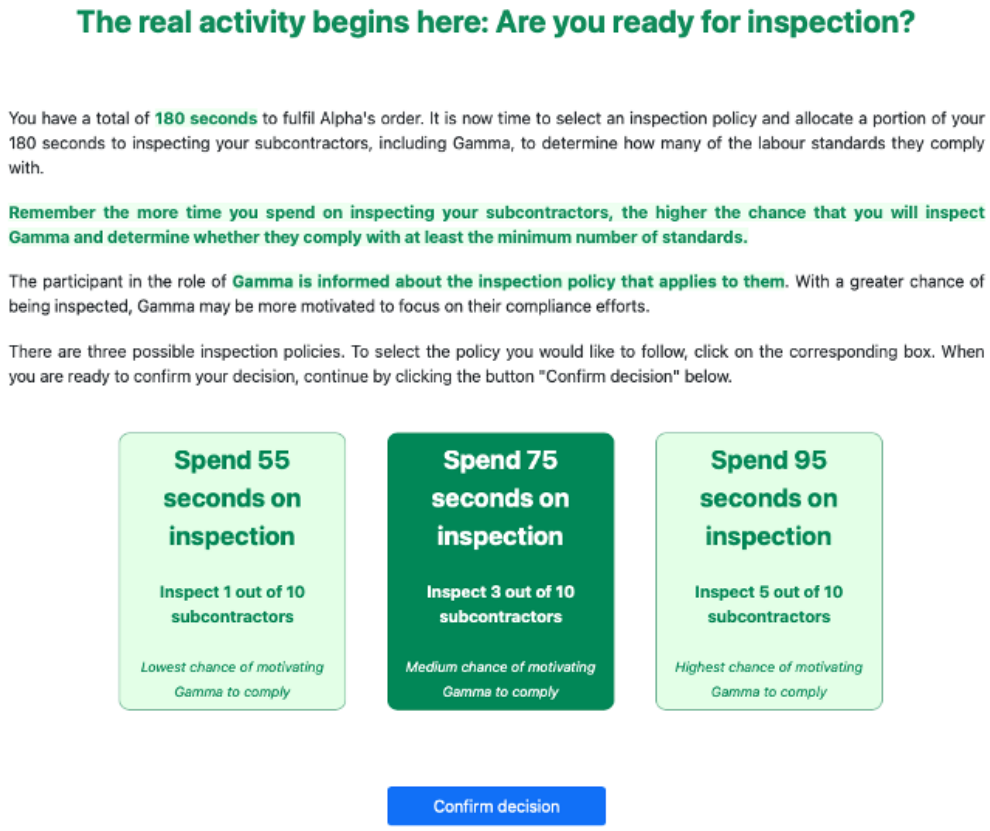
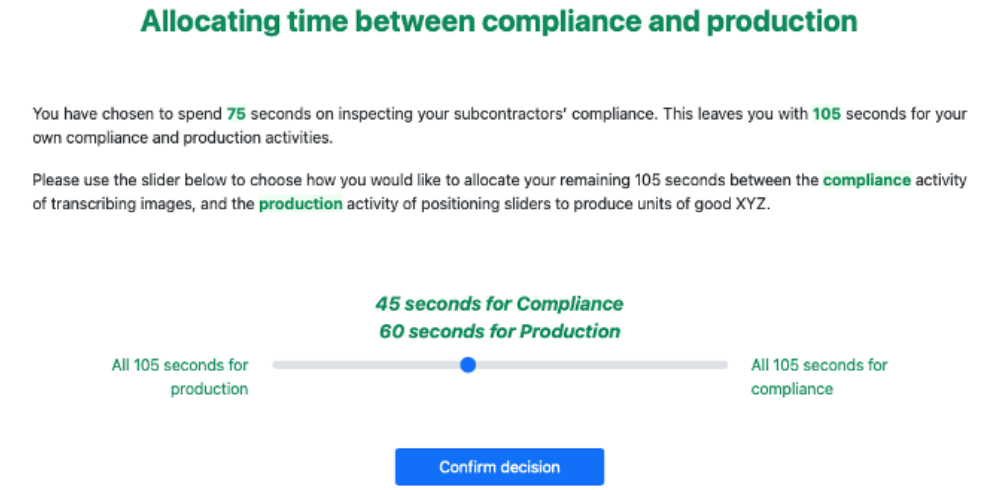


Figure 3.2: Time allocation between compliance and production



age, residence, employment) and non-incentivized preference items (risk, time, and other social preferences) following Falk et al. (2022). Figure 3.5 provides an overview of the experimental design for each player.

Figure 3.3: Compliance task

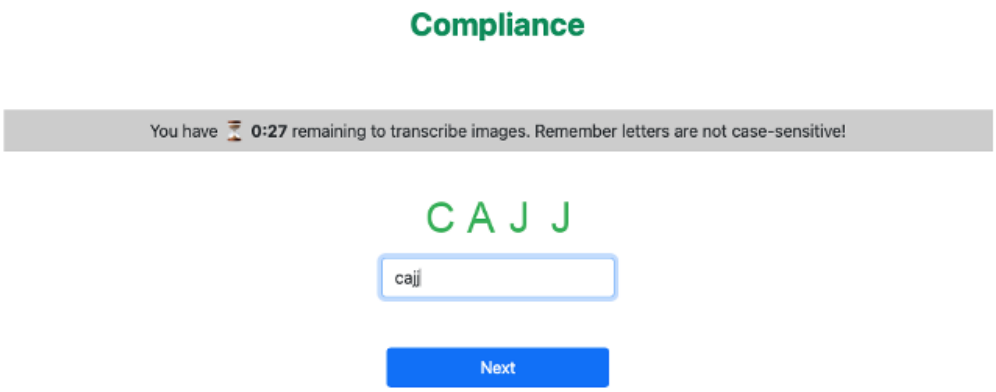
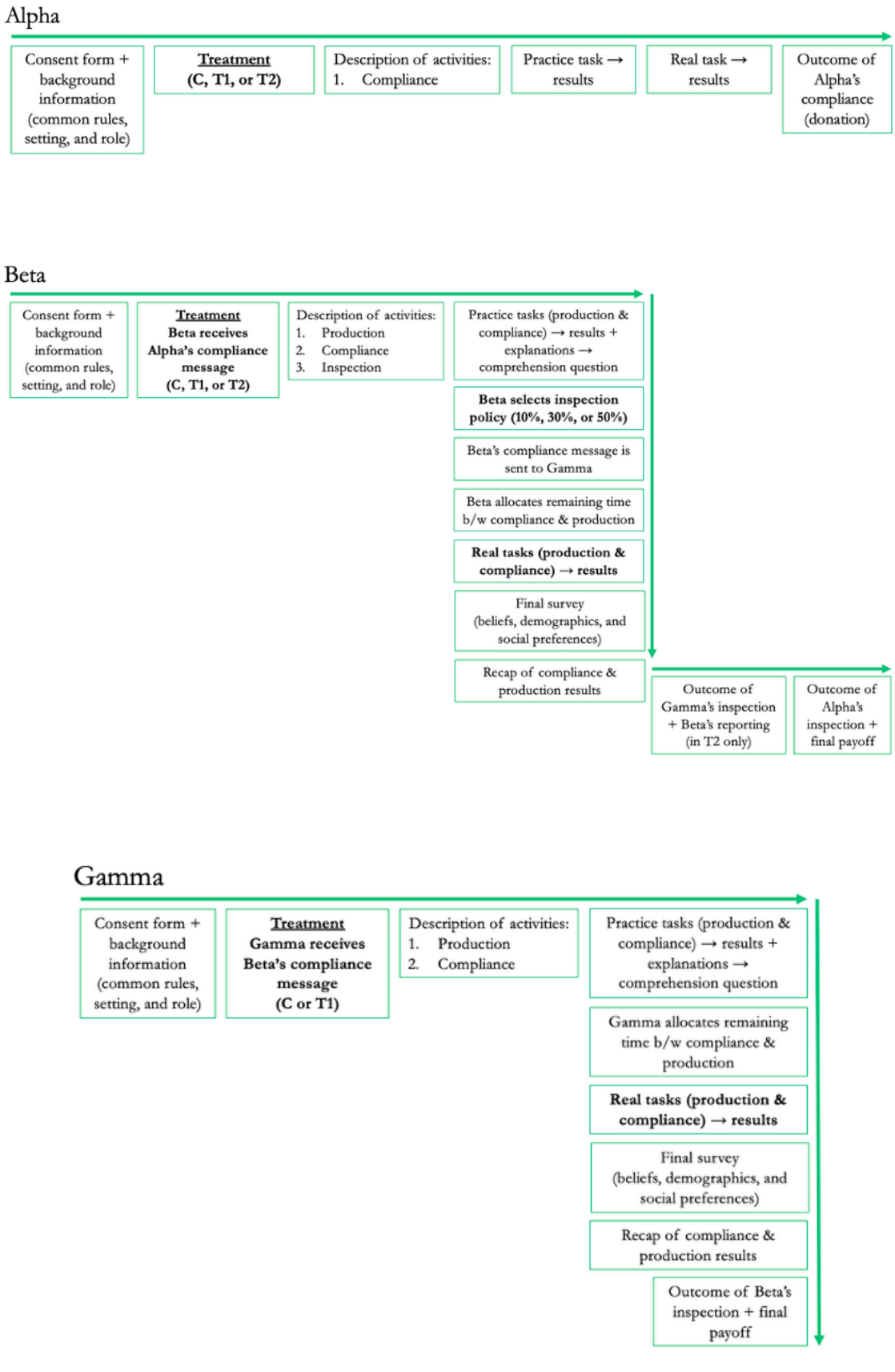


Figure 3.4: Production task

Figure 3.5: Experimental design flow



3.3.3 Treatments

In global supply chains, buyers typically communicate compliance requirements to their direct suppliers through codes of conduct or other formal communication. Inspired by a real-world example of such governance tools (Figure 3.A.1), we designed three treatment conditions to examine how Alpha's incentive structure and communication strategy affect Beta's allocation of effort between monitoring, compliance, and production activities. In our experiment, participants assigned to the role of Beta were randomly allocated to one of three treatment conditions, which differed in how Alpha communicated and enforced its compliance expectations (Figure 3.6):

1. **Deterrence-based approach (Control, C)**: A traditional top-down message in which Alpha emphasizes its compliance expectations and warns that the business relationship with Beta is terminated if non-compliance is detected — either at Beta's or Gamma's level — with a 10% probability of detection.

2. **Hybrid approach which combines deterrence with collaborative framing (Treatment 1, T1)**: Retains the sanction threat in the Control condition but adds a more collaborative, trust-building and reciprocity-based framing that aims to reduce the perception of unilateral, top-down imposition. This is achieved by emphasizing shared responsibility through inclusive "we"-language and by signaling that the lead firm itself engages in compliance efforts ("leading by example" and "tone at the top"), thereby reinforcing fairness and legitimacy.

3. **Hybrid approach which combines deterrence with collaborative framing and conditional financial incentives (Treatment 2, T2)**: Builds on T1 by adding a conditional financial incentive. If Beta detects and reports Gamma's non-compliance before Alpha's own inspection, Alpha's business relationship with Beta is maintained. This aims to reduce the material burden associated with monitoring and compliance.

Figure 3.6: Alpha's compliance messages in each treatment condition

Control

Dear Beta,

Our company Alpha would like to place an order for **up to 100 units of good XYZ** to be produced as soon as possible.

You will be paid **R1.0 per unit** produced.

Please note that we are **committed** to producing good XYZ in compliance with 15 labour standards.

We **expect** that you carry out your production activities in compliance with **at least 5 of those standards**, and that you ensure that your subcontractors do the same.

We at Alpha have a policy of conducting **thorough inspections** of our suppliers' compliance, covering both the direct suppliers and any subcontractors, **in one out of every 10 orders**. If we inspect your activities and find out that you or your subcontractors do not comply with at least 5 of the standards, we will terminate our business relationship with Beta, and **your company will not be paid** for any units of good XYZ you deliver.

Regards,
Alpha

Treatment 1

Dear Beta,

Our company Alpha would like to place an order for **up to 100 units of good XYZ** to be produced as soon as possible.

You will be paid **R1.0 per unit** produced.

Please note that we are **committed to working together with our entire supply chain** to produce good XYZ in compliance with 15 labour standards.

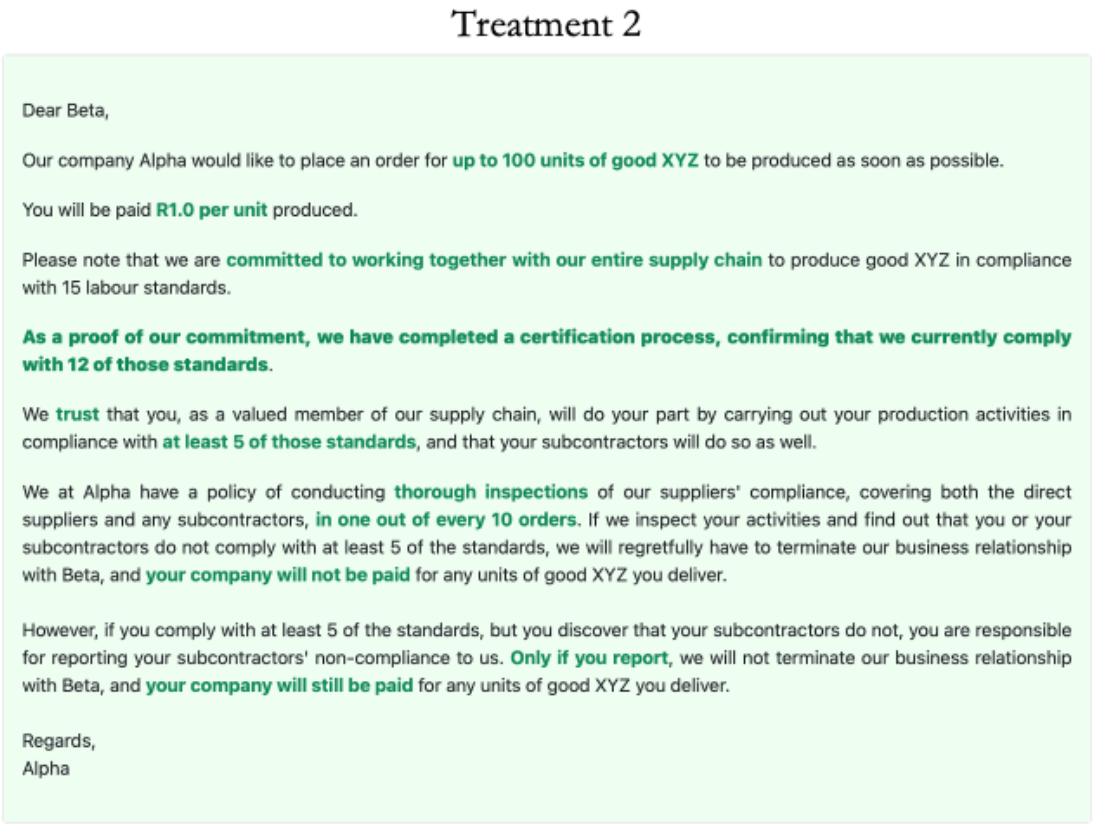
As a proof of our commitment, we have completed a certification process, confirming that we currently comply with 12 of those standards.

We **trust** that you, as a valued member of our supply chain, will do your part by carrying out your production activities in compliance with **at least 5 of those standards**, and that your subcontractors will do so as well.

We at Alpha have a policy of conducting **thorough inspections** of our suppliers' compliance, covering both the direct suppliers and any subcontractors, **in one out of every 10 orders**. If we inspect your activities and find out that you or your subcontractors do not comply with at least 5 of the standards, we will regretfully have to terminate our business relationship with Beta, and **your company will not be paid** for any units of good XYZ you deliver.

Regards,
Alpha

Figure 3.6 (continued): Alpha's compliance messages in each treatment condition



3.3.4 Implementation and matching procedures

The experiment was implemented in oTree and deployed via Prolific. Eligibility criteria included age (≥ 18), residence (South Africa for Betas and Gammas, and European countries for Alphas), English proficiency, and non-student status.

Data collection ran from 26 March to 28 May 2025, yielding 561 participants: 8 Alphas, 126 Gammas, and 427 Betas.[8] As we had no precedent to inform a minimum detectable effect size due to the novelty of the design, Beta's final sample size was maximized against the budget and checked against preliminary power calculations.[9] On average, Alphas completed the study in 5-10 minutes, Betas in 20-25

[8]Before registering the pre-analysis plan and beginning final data collection, we conducted three pilot sessions (on 2 September 2024 for Alpha, 31 January 2025 for Gamma, and 5 March 2025 for Beta) to test whether the instructions were clear and the tasks functioned as intended. Pilot results indicated that participants had difficulty understanding the instructions, as shown by low quiz accuracy and limited engagement with the production and compliance tasks. Accordingly, the compliance task was simplified (five letters reduced to four), and the instructions were clarified and segmented into more manageable components.

[9]Before concluding data collection, we ran preliminary power calculations (not pre-specified in the pre-analysis plan) using baseline means from a sample of 289 Betas (roughly evenly split

minutes, and Gammas in 15-20 minutes. Participants received a participation fee (EUR1.75=£1.5 for Alphas, R60=£2.5 for Betas, and R47=£2 for Gammas), with Betas and Gammas earning additional payoffs based on production performance (up to R100=£4 and R50=£2 respectively).[10]

To facilitate implementation, the experiment was conducted asynchronously. Data from Alphas and Gammas were collected before the Beta sessions, and their recorded outcomes were matched ex-post with Betas' decisions. This approach ensured realistic interactions and payoffs while avoiding deception. Treatment assignment was not handled by Prolific directly but through oTree configuration files prepared by the research team. These files pre-specified treatment conditions, inspection policy (for Gamma sessions only), and sequence order, ensuring balance across conditions. Because participant entry into Prolific sessions is effectively random, the procedure approximated random assignment. Data collection proceeded in three stages:

- **Alpha session (September 2024):** We conducted one Prolific session and collected data from 9 European participants, evenly distributed across the three treatment arms. They were asked to perform only the compliance task.[11] In T1 and T2, participants were also informed that their performance could influence supplier compliance. The compliance score of the top-performing Alpha (12/15) was later disclosed to Betas and Gammas in those treatments.

- **Gamma sessions (March-May 2025):** We ran three Prolific sessions and collected data from 126 South African participants. Gammas were assigned to one of two framing treatments (C or T1) and to one of three inspection conditions (10%, 30%, or 50%). T2 was excluded at the Gamma level since the

across treatment conditions) to assess whether additional recruitment was necessary and to verify that randomization produced balanced covariates across groups. We examined three primary outcomes: (i) selection of inspection policy 5, (ii) number of words correctly transcribed, and (iii) compliance with at least minimum standards (≥ 5 words correctly transcribed). For the first outcome (mean = 0.55) and third outcome (mean = 0.58), the calculations indicated that detecting a 10 percentage-point effect would require between 780 and 1,000 total participants, which was not feasible due to budget constraints. For the second outcome (mean = 5), the calculations suggested that detecting a one-unit increase (from 5 to 6 words) would require between 400 and 550 participants, approximately in line with our final sample size.

[10] At the time of data collection, the minimum hourly payment allowed on Prolific was £6.

[11] As Alphas were not the main actors in our design, and the pilot data showed no issues with instructions or task implementation, we retained these observations as the final Alpha data rather than running additional sessions.

messages for Gamma were identical in T1 and T2. To reflect Alpha's policy of inspecting 1 out of every 10 production orders, Gammas were organised into blocks of 6 (2 treatment arms x 3 inspection policy conditions), which were repeated 10 times to yield the complete set of 60 configurations. Dropouts were replaced in subsequent sessions until the full set was obtained.[12] The dataset of Gammas, including compliance and production outcomes, served as the input for matching Beta's decisions in the game and to generate Beta's payoffs.

- **Beta sessions (April-May 2025):** We conducted five Prolific sessions and collected data from 427 South African participants. Betas were assigned to one of the three framing treatments (C, T1, or T2). They were then matched ex post on treatment and inspection policy with Gamma data according to the pre-specified configuration files including pre-recorded data on Gamma's treatment, assigned inspection policy, and Alpha's inspection outcome for payoff calculation. Dropouts were systematically replaced by re-running the corresponding configurations until the target sample was reached. Each Gamma was assigned to two or three Betas.

3.4 Measurement and hypotheses

3.4.1 Outcome measures and controls

Building on the previous section, the analysis focuses on two primary outcomes (POs) related to Beta's monitoring and compliance efforts, and two secondary outcomes (SOs) capturing underlying mechanisms.

1. **PO_1: Monitoring effort.** Measured as the number of seconds Beta allocates to the inspection task, determined by their selected inspection policy (10% policy = 55 seconds, 30% policy = 75 seconds, or 50% policy = 95 seconds). For the main specification, we construct a binary indicator equal to one if Beta selects more than 55 seconds (i.e., 30% or 50% inspection policy), and

[12] Across both Gamma and Beta sessions, the average response rate was approximately 70%.

zero otherwise. We also examine the probability of selecting each inspection policy separately.

2. **PO_2: Compliance effort.** Measured as Beta's performance in the compliance task, defined by the number of correctly transcribed words. We also construct a binary indicator equal to one if Beta correctly transcribes at least 5 words (i.e., complies with at least minimum labour standards), and zero otherwise.

To explore mechanisms, we elicited the following subjective perceptions after task completion:

1. **SO_1: Perceived fairness.** Beta's agreement with the statement "I found Alpha's compliance requests fair", measured on a Likert scale from 0 (completely disagree) to 10 (completely agree).

2. **SO_2: Perceived impact of monitoring.** Beta's belief about whether their chosen inspection policy influenced Gamma's compliance ("In your opinion, how does your chosen inspection policy influence the likelihood that Gamma meets at least five labour standards?"), measured on a Likert scale from 0 (no influence at all) to 10 (complete influence).

To account for individual heterogeneity in preferences and attentiveness, we included the following controls: participants' demographics (gender, age, residence, and employment in private vs public sector), preferences (risk, time, and social preferences from Falk et al. (2022)) and the number of attempts required to correctly answer the pre-game comprehension quiz.

3.4.2 Hypotheses

Our main hypotheses are summarised as follows.

In terms of monitoring effort (PO_1):

- **Hypothesis 1:** Receiving a social compliance demand framed in a more collaborative manner increases the time Beta allocates to the inspection task compared to traditional top-down framing (Hp 1a: T1 – C $>$ 0; Hp 1b: T2 – C $>$ 0);

- **Hypothesis 2:** Adding a financial incentive to report Gamma's non-compliance (T2) increases Beta's monitoring time beyond the effect of collaborative framing alone (Hp 2: T2 – T1 > 0).

In terms of compliance effort (PO_2):

- **Hypothesis 3:** Collaborative framing by Alpha — including disclosure of Alpha's own compliance performance — may influence Beta's own compliance performance, but the direction of this effect is ambiguous. On one hand, it may promote reciprocity and effort. On the other hand, increased monitoring effort may crowd out the time and attention available for Beta's compliance task (Hp 3a: $T1 - C \neq 0$; Hp 3b: $T2 - C \neq 0$; Hp 3c: $T2 - T1 \neq 0$).

In terms of mechanisms (SO_1):

- **Hypothesis 4:** Receiving a social compliance demand framed in a more collaborative manner increases Beta's perception of the fairness of the buyer's compliance request. (Hp 4a: T1 – C > 0; Hp 4b: T2 – C > 0).

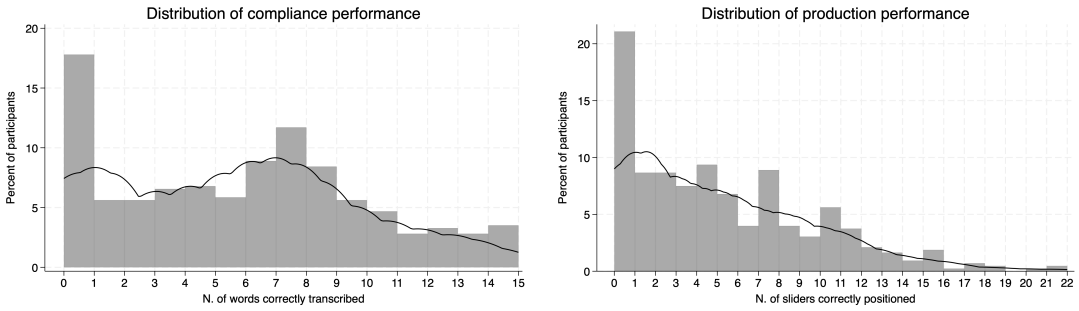
3.5 Data and methodology

3.5.1 Data description

Summary statistics for outcome variables and controls are reported in Table 3.A.1. Among Beta participants, 86% selected either inspection policy 3 or 5, with 51% choosing the highest policy. On average, Betas allocated their remaining time almost evenly between compliance and production, devoting about 55% to compliance. They correctly transcribed an average of 5.5 words, with 58% meeting at least the minimum standard (≥ 5 words). As shown in Figure 3.7a, the distribution of compliance performance is bimodal, with one peak at very low performance (0 words correctly transcribed) and a second, smaller peak just above the minimum compliance threshold (around 6-8 correctly transcribed words). In the production task, Betas correctly positioned an average of 4.9 sliders. The production distribution is right-skewed, with a concentration of participants at very low production levels and progressively fewer reaching higher output (3.7b). Belief measures were generally

Figure 3.7: Distribution of compliance and production performance

(a) (b)



high, with average scores in the upper range of the 10-point scale. The perceived fairness of Alpha's compliance requests averaged 7. The perceived effectiveness of the chosen inspection policy in influencing Gamma's compliance behaviour averaged 7.8.

Turning to demographics, among the 427 participants in the Beta role, about 60% were female, 63% were aged 18-34, and 32% lived in an urban area. Only 26 participants (6%) reported being unemployed, while most were employed in the private sector (58%). Social preference measures were also generally high: willingness to take risks (7.8), willingness to delay gratification (7.2), optimism about others' intentions (6.5), willingness to share without expecting return (8.4), and willingness to punish unfair behaviour (5.8). Generosity, measured on a 6-point scale, also scored relatively high (4.2). On average, participants required 1.5 attempts for the production quiz, 1.2 for the compliance quiz, and 1.4 for the inspection quiz.

Balance tests across treatment arms are reported in Table 3.A.2 in the Appendix. Only a few variables differed significantly between groups, namely private-sector employment and the number of attempts on the production and inspection quizzes, indicating that randomisation was largely successful.

3.5.2 Empirical methodology

To test our hypotheses, we estimate the following equation using OLS regression with robust standard errors:

$$Y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \gamma X_i + \delta_s + \varepsilon_i \tag{3.1}$$

where Y_i is one of the primary or secondary outcome measures for individual i, representing Beta's decisions and effort. $T1_i$ and $T2_i$ represent indicators for assignment to Treatment 1 or 2, respectively. $\mathbf{X_i}$ is a vector of individual covariates including demographics (dummies for female, early-career, urban residence, unemployed, and private-sector employment), social preferences (risk tolerance, patience, trust, prosociality, generosity, inequity aversion), as well as the number of attempts required to pass the comprehension quiz, which captures participants' focus and understanding of the rules. δ_s represents session fixed effects to account for unobserved heterogeneity across data collection sessions, and ε_i is the error term.

To explore heterogeneity, we augment equation (3.1) by interacting the treatment indicators with key demographic and preference variables. We are particularly interested in differences between participants from the private and public sectors, as well as differences linked to their level of prosociality. These subgroup analyses, as well as the analysis of secondary outcomes, are treated as exploratory. Results are therefore interpreted with caution, given concerns about limited statistical power and the increased risk of false discoveries.

3.6 Main results

3.6.1 Primary outcome: Monitoring efforts (PO_1)

As shown in Figure 3.8, the share of participants selecting either inspection policy 3 or 5 was consistently high (above 80 percent), with negligible variation across treatments.

Figure 3.9 provides a more detailed breakdown of policy choices by treatment. Around half of the participants in each group selected the highest inspection policy (policy 5), with the largest share in C (57 percent) and the smallest in T1 (45 percent). Choices for policy 3 were relatively stable across treatments, ranging from 32 to 39 percent, while policy 1 was chosen by about 10 to 16 percent of the sample.

Figure 3.8: Probability of selecting policy 3 or 5 by treatment

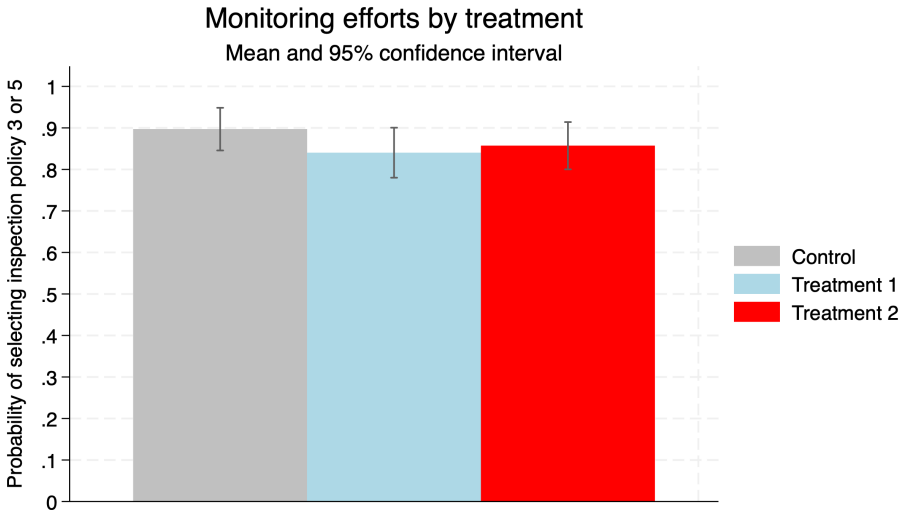
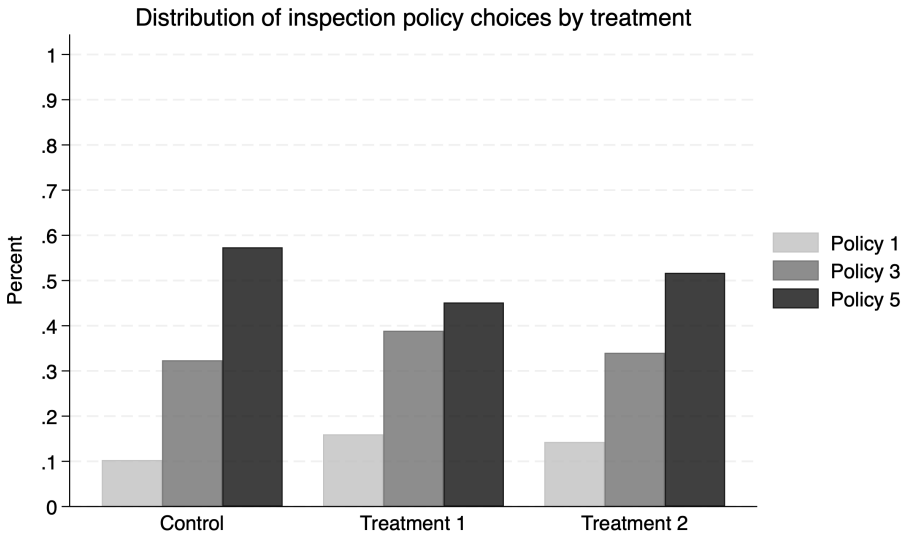


Figure 3.9: Distribution of inspection choices by treatment



Regression estimates reported in Table 3.1 show that neither T1 nor T2 had a significant effect on the likelihood of selecting inspection policies 3 or 5, nor on the probability of choosing policy 3. However, compared to the control group, participants in T1 were significantly less likely to select policy 5, with a marginal effect of -12.3 percentage points (approximately a 20% reduction compared to the control share; $p < 0.05$). These results are consistent with ordered probit estimates, which further indicate that T1 shifts participants away from the highest inspection policy toward lower and intermediate levels. Detailed results are reported in Table 3.A.3.

Taken together, the results reveal two key patterns. First, the data do not support our prior assumption that a pure deterrence-based approach imposing compliance requirements in a top-down manner would induce only limited monitoring. Instead, the control condition — combining a 10 percent probability of detection with the threat of financial loss — sustains relatively high levels of monitoring. Second, contrary to Hypotheses 1 and 2, neither collaborative framing (T1) nor its combination with conditional financial incentives (T2) increases monitoring effort. If anything, T1 reduces the likelihood of selecting the strictest inspection policy. shifting choices toward lower and intermediate levels.

Table 3.1: Treatment effects on Beta's monitoring, compliance, and production efforts

Variable	**Inspection policy 3 or 5**	**Inspection policy 3**	**Inspection policy 5**	**Compl. seconds (share)**	**Compl. perform.**	**If complied with min. standard**	**Prod. perform.**
Treatment 1	-0.059	0.064	-0.123**	-0.010	-0.003	-0.086	0.549
	[0.041]	[0.058]	[0.060]	[0.028]	[0.441]	[0.055]	[0.361]
Treatment 2	-0.042	0.014	-0.056	-0.034	-0.531	-0.145**	0.000
	[0.040]	[0.057]	[0.060]	[0.027]	[0.463]	[0.059]	[0.356]
Compliance seconds				0.067***	0.006***		
				[0.010]	[0.001]		
Gender	-0.011	-0.034	0.023	0.032	-0.657*	-0.007	-0.975***
	[0.035]	[0.049]	[0.050]	[0.023]	[0.383]	[0.047]	[0.322]
Private sector	0.053	0.047	0.006	-0.063**	0.925**	0.098*	0.917***
	[0.037]	[0.051]	[0.053]	[0.026]	[0.417]	[0.052]	[0.306]
Observations	427	427	427	427	427	427	427
R-squared	0.036	0.047	0.078	0.146	0.202	0.158	0.553
Control mean	0.897	0.324	0.574	0.568	5.669	0.647	4.272
Hp 1a (p-value): T1 – C = 0	0.150	0.269	0.0397	0.722	0.994	0.121	0.129
Hp 1b (p-value): T2 – C = 0	0.296	0.805	0.353	0.213	0.252	0.0137	1
Hp 2 (p-value): T2 – T1 = 0	0.680	0.384	0.254	0.387	0.252	0.284	0.143

Note: This table reports OLS estimates of the treatment effects on Beta's monitoring, compliance, and production effort. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

3.6.2 Primary outcome: Compliance efforts (PO_2)

Hypothesis 3 anticipated ambiguous effects on compliance efforts. On one hand, collaborative framing, including disclosure of Alpha's own compliance performance ("leading by example"), could have promoted reciprocity and higher compliance effort. On the other hand, additional monitoring efforts risked crowding out Betas' time and attention for compliance. Since the results above show that Betas did not increase monitoring under T1 or T2, we would expect to see higher compliance efforts.

Figure 3.10 shows that participants allocated their post-inspection time almost evenly between compliance and production, with no meaningful differences across treatments (C = 57%, T1 = 56%, T2 = 52%). Regression estimates in Table 3.1 confirm that neither collaborative framing alone nor the addition of financial incentives significantly affected the share of time devoted to compliance. A visual breakdown by treatment and inspection policy choice suggests a consistent pattern: participants selecting the most intensive policy (policy 5) devoted slightly more of their residual time to compliance than those choosing lighter inspection policies, although this effect appears weaker under T2 (Figure 3.11). This suggests that monitoring and compliance are complements, rather than substitutes.

Figure 3.10: Average compliance seconds by treatment

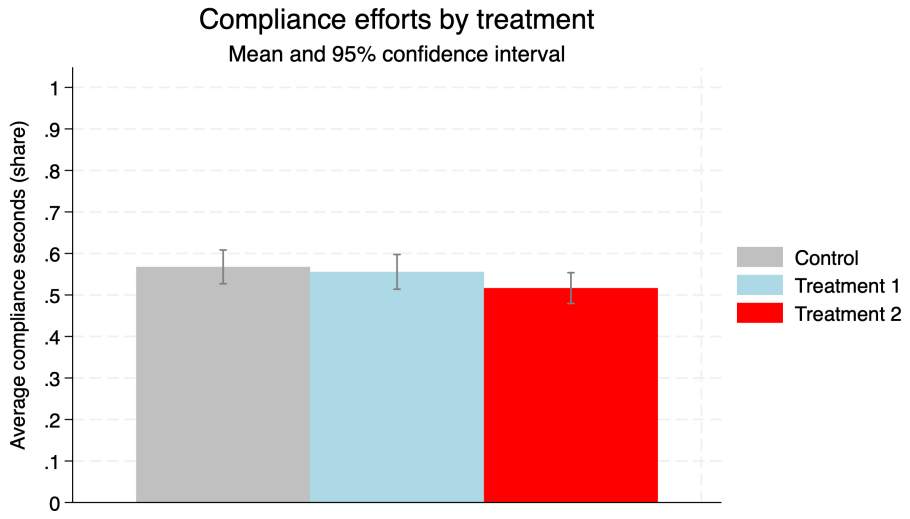
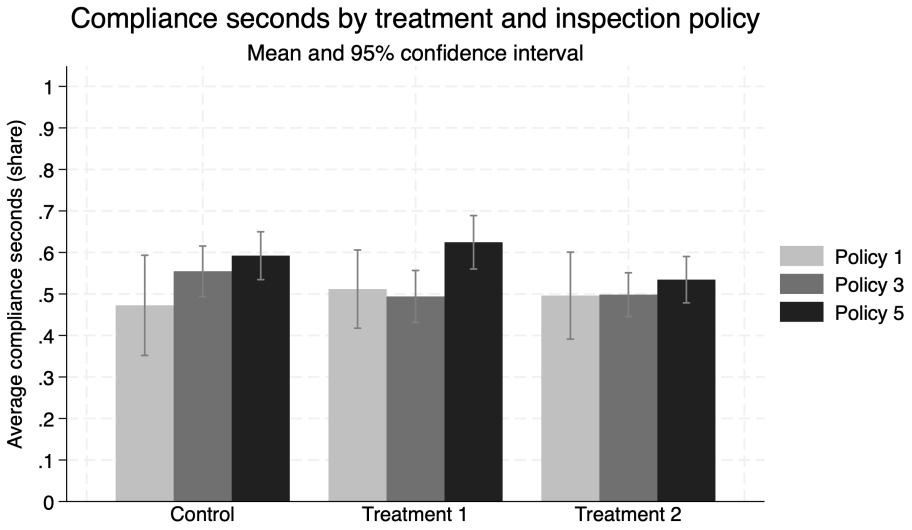


Figure 3.11: Average compliance seconds by treatment and inspection policy



Turning to compliance performance measured by the number of correctly transcribed words, we again find little evidence of treatment effects: participants averaged 5.7 words in C, 5.8 in T1, and 5.1 in T2, with OLS estimates showing no statistically significant differences (Table 3.1).

When compliance is instead measured as the probability of meeting at least minimum standards (i.e., correctly transcribing at least five words), we find a negative effect of T2: Betas in this group were around 14.5 percentage points (or almost 25%) less likely than those in the control to meet minimum standards ($p < 0.05$), even after controlling for available compliance time. We also observe systematic differences across participant types. Holding other variables constant, private-sector players were more compliant and productive overall, transcribing one additional word ($p < 0.05$) and positioning one extra slider on average ($p < 0.01$). Finally, neither T1 nor T2 significantly influenced production performance.

Taken together, these results indicate that collaborative framing, with or without financial incentives, did not strengthen compliance efforts. If anything, the addition of conditional financial incentives in T2 appears to weaken compliance, as reflected in the lower likelihood of meeting minimum standards.

3.6.3 Secondary outcome (mechanisms): perceived fairness (SO_1) and effectiveness of monitoring (SO_2)

We now examine how treatments shape Beta's beliefs about (i) the fairness of Alpha's compliance requests and (ii) the effectiveness of their own inspection policy in influencing Gamma's compliance behaviour.

Recall that Treatment 1 ("collaborative framing") introduced a relational, trust-building and reciprocity-based framing emphasising shared responsibility through inclusive "we"-language and signaling that the lead firm itself engaged in compliance efforts ("leading by example"). We hypothesized that such framing would reinforce perceived fairness and legitimacy relative to the control condition, thereby increasing monitoring and, potentially, compliance. However, as shown in the previous section, these behavioural effects did not materialize. Monitoring did not increase and, if anything, declined in T1, while compliance remained overall unchanged, suggesting that the fairness mechanism we aimed to activate may not have operated as intended in this setting.

Regression estimates in Table 3.2 are consistent with this pattern. On average, T1 did not affect perceived fairness and, if anything, weakly reduced perceived effectiveness. One possible explanation for the lack of effects on perceived fairness and the reduction in monitoring is that collaborative framing may have redefined the social meaning of monitoring: by emphasising shared responsibility and group identity, it might have made the inspection of sub-suppliers appear less appropriate or less legitimate, thereby reducing willingness to engage in strict enforcement. In this sense, "we-thinking" may have weakened monitoring by increasing its perceived relational cost. Alternative explanations, such as heterogeneous responses across participants, cannot be ruled out and are explored in the next section.

Treatment 2 ("collaborative framing with conditional financial incentives") built on T1's relational framing by introducing a conditional financial incentive for detecting and reporting sub-supplier non-compliance. We hypothesized that this would reinforce perceived fairness and legitimacy relative to the control condition by reducing the material burden associated with monitoring and compliance, thereby increasing monitoring and, potentially, compliance. Again, as shown in the previous section, these behavioural effects did not materialize. Monitoring remained

overall unchanged, while compliance declined in T2, suggesting that the fairness mechanism not only failed to operate but may have worked in the opposite direction.

Regression estimates in Table 3.2 show that Betas in T2 were 11.2 percentage points less likely than those in the control to rate Alpha's requests as highly fair (above the sample median) ($p < 0.05$). Similarly, T2 participants were 11.7 percentage points less likely to believe that their inspection policy strongly influenced Gamma's compliance (above the sample median) ($p < 0.05$). One possible explanation is that by tying rewards to the detection and reporting of others' misconduct, the incentive shifted the interpretation of the task toward a more instrumental or enforcement-oriented role. In particular, asking Betas to report sub-supplier non-compliance may have positioned them as "policing" others, potentially crowding out intrinsic or relational motivations. While most participants complied with this requirement when non-compliance was detected (42 out of 46 cases), the arrangement may nonetheless have weakened perceptions of fairness and cooperation. Although the data do not allow us to distinguish between these mechanisms, the results point to tensions in how incentive structures interact with fairness perceptions.

Table 3.2: Treatment effects on Beta's secondary outcomes (mechanisms)

Variable	**Fairness**	**Effectiveness**
Treatment 1	-0.078	-0.112*
	[0.055]	[0.058]
Treatment 2	-0.112**	-0.117**
	[0.054]	[0.059]
Gender	-0.019	-0.062
	[0.045]	[0.050]
Private sector	0.070	0.083
	[0.047]	[0.052]
Observations	427	427
R-squared	0.138	0.131
Control mean	0.368	0.544
Hp 1a (p-value): T1 – C = 0	0.161	0.0548
Hp 1b (p-value): T2 – C = 0	0.0408	0.0473
Hp 2 (p-value): T2 – T1 = 0	0.515	0.922

Note: This table reports OLS estimates of the treatment effects on Beta's fairness and effectiveness beliefs. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

3.6.4 Heterogeneity analysis

We also conducted exploratory heterogeneity analysis to assess whether treatment effects varied with demographic characteristics or social preferences (Table 3.3). Two dimensions were particularly salient in our supply chain setting: participants' occupational background and their prosocial orientation.

We find clear heterogeneity by occupation, both in inspection choices and compliance. Among public-sector participants, neither T1 nor T2 significantly affected inspection choices, but both treatments led to substantial reductions in compliance with minimum standards (-25.6 pp, $p < 0.01$; -20.8 pp, $p < 0.05$). By contrast, private-sector participants adjusted their inspection behaviour across treatments. Relative to the control condition — where they were more likely than public-sector participants to select the strictest inspection policy — they shifted toward less stringent monitoring under both T1 and T2. In T1, they were more likely to choose the intermediate policy (+18.1 pp, $p < 0.05$) and less likely to select the strictest option (-20.7 pp, $p < 0.05$); under T2, they were again less likely to select the strictest policy (-19.6 pp, $p < 0.05$). In terms of compliance, private-sector participants in T1 showed a positive and significant response that offset the negative effect observed for the public sector, suggesting that the "leading by example" treatment may resonate more strongly with this group. However, the overall effect of T1 on their compliance was small and not statistically significant. Under T2, their compliance did not differ significantly from the control.

Taken together, these results suggest distinct behavioural responses across groups. Public-sector participants maintained stable monitoring but reduced compliance under both treatments, consistent with the decline in fairness perceptions documented above. In contrast, private-sector participants maintained stable compliance but reduced monitoring, possibly suggesting greater responsiveness to the relational framing of the task. While these interpretations remain tentative, they point to potentially different mechanisms across participant types.

Table 3.3: Beta's heterogeneity analysis

Variable	Inspection policy 3 or 5	Inspection policy 3	Inspection policy 5	Compl. seconds (share)	If complied with min. standard	Prod. perform.
Private sector						
T1	-0.110	-0.064	-0.046	-0.043	-0.256***	0.277
	[0.068]	[0.084]	[0.087]	[0.044]	[0.083]	[0.504]
T2	0.036	-0.093	0.129	-0.049	-0.208**	-0.023
	[0.059]	[0.087]	[0.090]	[0.043]	[0.090]	[0.528]
Private sector	0.064	-0.103	0.167*	-0.095**	-0.055	0.718
	[0.053]	[0.082]	[0.085]	[0.041]	[0.086]	[0.507]
T1 × Private sector	0.083	0.245**	-0.161	0.061	0.315***	0.496
	[0.084]	[0.114]	[0.119]	[0.057]	[0.111]	[0.737]
T2 × Private sector	-0.122	0.203*	-0.325***	0.032	0.134	0.084
	[0.078]	[0.115]	[0.119]	[0.055]	[0.118]	[0.727]
Observations	427	427	427	427	427	427
R-squared	0.050	0.058	0.094	0.149	0.174	0.553
Control mean	0.873	0.366	0.507	0.616	0.690	3.423
Coeff: T1 + T1×Priv = 0	-0.0262	0.181	-0.207	0.0183	0.0596	0.773
P-value: T1 + T1×Priv = 0	0.596	0.0216	0.0116	0.616	0.420	0.142
Coeff: T2 + T2×Priv = 0	-0.0863	0.109	-0.196	-0.0174	-0.0739	0.0608
P-value: T2 + T2×Priv = 0	0.0996	0.149	0.0138	0.614	0.335	0.901
Prosociality						
T1	-0.055	0.025	-0.080	0.023	-0.088	1.009**
	[0.052]	[0.074]	[0.075]	[0.032]	[0.069]	[0.505]
T2	-0.023	-0.050	0.027	0.009	-0.121	0.413
	[0.051]	[0.075]	[0.077]	[0.030]	[0.075]	[0.464]
Prosociality	-0.007	-0.170*	0.163	0.105**	-0.067	0.568
	[0.068]	[0.097]	[0.102]	[0.051]	[0.104]	[0.586]
T1 × Prosociality	-0.012	0.106	-0.117	-0.090	0.004	-1.300*
	[0.085]	[0.118]	[0.124]	[0.063]	[0.117]	[0.718]
T2 × Prosociality	-0.051	0.168	-0.220*	-0.113*	-0.067	-1.137
	[0.082]	[0.116]	[0.120]	[0.059]	[0.117]	[0.722]
Observations	427	427	427	427	427	427
R-squared	0.038	0.054	0.086	0.157	0.162	0.557
Control mean	0.885	0.391	0.494	0.517	0.644	4.839
Coeff: T1 + T1×Pro = 0	-0.0668	0.131	-0.197	-0.0675	-0.0844	-0.291
P-value: T1 + T1×Pro = 0	0.318	0.157	0.0452	0.213	0.370	0.549
Coeff: T2 + T2×Pro = 0	-0.0741	0.118	-0.192	-0.104	-0.187	-0.724
P-value: T2 + T2×Pro = 0	0.256	0.181	0.0412	0.0415	0.0416	0.191

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Beta's monitoring, compliance, and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Prosocial orientation, measured through willingness to share with others, also shaped treatment responses. For participants with lower prosociality (i.e., below the median), neither T1 nor T2 had significant effects on inspection choices or compliance outcomes, though T1 did increase production performance. Among participants with higher prosociality preferences, clear patterns emerged. They were significantly less likely to select the strictest inspection policy under both T1 and T2 (-19.7pp, $p < 0.05$; -19.2pp, $p < 0.05$). Given that these individuals are more likely to place weight on intrinsic motivations and relational considerations, they may be more responsive to the relational framing of the task, whereby monitoring sub-suppliers is perceived as less compatible with a cooperative, "we"-oriented interaction. Under T2, they also devoted less time to the compliance task and were less likely to meet minimum standards (both $p < 0.05$), consistent with the decline in fairness perceptions documented above. This pattern is in line with a crowding-out mechanism, whereby conditional incentives weaken intrinsic and relational motivations. Finally, while T1 raised production performance among participants with lower prosociality preferences, this gain was offset by a significant reduction among high-prosociality participants.

As shown in Table 3.A.4, heterogeneity patterns across other demographic and social preference dimensions were mixed. No clear effects emerged for urban participants. Under T1, female participants were significantly less likely to adopt the strictest inspection policy (-18.1pp, $p < 0.05$), and early-career participants were more likely to choose the intermediate option (+16.0pp, $p < 0.05$) and less likely to select the strictest one (–20.4pp, $p < 0.01$). Under T2, early-career participants were also less likely to comply with minimum standards (–17.2pp, $p < 0.05$). No systematic effects were found for participants reporting high risk tolerance, patience, or trust, but participants with stronger inequity aversion complied less under T2 (-18.2pp, $p < 0.05$). Perceptions also mattered: those who saw Alpha's request as highly fair were less likely to select policy 5 under T1 (-19.9pp, $p < 0.10$), and the positive effect of T1 on production performance among those rating the request as less fair was offset by a negative interaction among those perceiving it as fair. Finally, participants who considered Beta's inspection policy highly effective showed significantly lower compliance under both T1 and T2.

3.6.5 Exploratory analysis of Gamma's outcomes

While Beta is the primary decision-maker of interest, we also conducted an exploratory analysis of Gamma's compliance and production outcomes.

Summary statistics for Gamma's outcome variables and controls are reported in Table 3.A.5. Gamma participants were evenly allocated across inspection policy 1 (36%), 3 (31%), and 5 (33%). On average, Gammas devoted about 58% of their time to compliance, almost 3 percentage points more than Betas. They correctly transcribed an average of 6 words, with 56% meeting at least the minimum standard ($\geq$ 5 words). In the production task, Gammas correctly positioned an average of 4.8 sliders. In terms of beliefs, the perceived fairness of Alpha's compliance requests averaged 7.

Turning to control variables, among the 126 Gamma participants, 76% were female, 64% were aged 18-34, and 32% lived in an urban area. About 10% reported being unemployed, while over half (54%) were employed in the private sector. As with Betas, social preference measures were on the higher side of the scale, with average scores in the upper range of the 10-point scale: risk tolerance (7.8), patienece (7.1), trust (6.4), prosociality (8.6), and inequity aversion (6.1). Generosity, measured on a 6-point scale, also scored relatively high (4.3). On average, participants required 1.2 attempts at the comprehension questions for the production quiz and 1.3 for the compliance quiz.

Balance tests across treatment arms are reported in Table 3.A.6. Only a few variables differed significantly between groups, namely gender and inequity aversion, indicating that randomization was largely successful. As shown in Tables 3.A.7 and 3.A.8, we found no significant effects on either our primary outcomes (compliance and production performance) or our secondary outcome (perceived fairness of Alpha's compliance requests). This is unsurprising given the relatively small sample size. Similarly, no treatment effects were detected in our heterogeneity analysis (3.A.9), with one exception: respondents working in the private sector. Participants from public-oriented sectors performed significantly worse in the production task under T1, but this was offset by the significantly higher performance of private-sector participants, displaying an overall effect of 18 percentage points ($p < 0.10$) on production performance under T1.

3.6.6 Exploratory analysis of the supply chain's outcomes

We further conducted an exploratory analysis of treatment effects on compliance and production efforts aggregated to the supply chain level.[13] Specifically, we examined four main outcomes: (i) average compliance time, (ii) average compliance performance, (iii) joint minimum compliance (whether both Beta and Gamma met at least the minimum standard), and (iv) overall production performance. In addition, we constructed a joint beliefs measure, capturing whether both Beta and Gamma rated Alpha's compliance requests above the median fairness score.

Table 3.A.10 reports treatment effects on supply chain-level primary outcomes. Columns (1) show results without Gamma weights, while Columns (2) apply probability weights to account for multiple Beta-Gamma matches. All specifications control for Beta's and Gamma's demographics and social preferences, with standard errors clustered at the Gamma level. For compliance time, Beta-Gamma pairs under T2 devote less time to compliance relative to their counterparts under C ($p < 0.10$) and T1 ($p < 0.05$), though these differences lose significance once weights are applied. For minimum compliance, T2 pairs perform significantly worse than C pairs under the weighted specification ($p < 0.05$). Moreover, across all specifications of compliance performance, whether measured by the number of correctly transcribed words or the probability of meeting the minimum standard, T2 pairs consistently perform worse than T1, with consistently statistically significant differences. Finally, for production performance, both T1 and T2 pairs outperform C pairs under the unweighted specification, but these effects vanish once weights are applied. Nonetheless, significant differences remain between T1 and T2, with T2 generally underperforming.

Taken together, these findings suggest that, once Gamma weights are applied, collaborative framing with financial incentives reduced Beta-Gamma pairs' likelihood of meeting at least minimum compliance standards by 9.9 percentage points,

[13] To construct supply chain-level outcomes, we matched each Beta with Gammas who were assigned to the same treatment and inspection policy. This procedure generated a many-to-many structure in which each Gamma was paired with multiple Betas, resulting in 8,532 Beta-Gamma observations. As Gammas are observed repeatedly, we treated them as clusters in the analysis and estimated two specifications: one unweighted, and one using probability weights equal to the inverse of the number of Beta matches per Gamma. This ensures that each Gamma carries equal weight in the estimation, irrespective of the number of pairings.

a result significant at conventional levels ($p < 0.05$). Overall, T2 pairs perform significantly worse than T1 pairs across most outcomes. For the joint beliefs measure, whether both Beta and Gamma rated Alpha's compliance requests above the median fairness score, T2 pairs are also less likely to report high fairness ($p < 0.10$) (Table 3.A.11).

3.7 Conclusion

Hybrid compliance approaches that combine enforcement with trust-building have been shown to improve compliance in domains such as tax enforcement, yet their effectiveness in social compliance, particularly in multi-tier global supply chains, remains largely unexplored in experimental settings. This study addresses this gap by using a pre-registered, contextualized online experiment that simulates a multi-tier supply chain to causally identify how different combinations of enforcement, trust-building, and incentives shape fairness perceptions and, in turn, influence both compliance and delegated monitoring across supply-chain tiers – mechanisms that are difficult to observe in real-world settings.

The results provide four main insights. First, contrary to our initial assumptions, a pure deterrence-based approach sustains relatively high levels of monitoring, even in the presence of only a modest probability of detection and sanction. Second, the addition of collaborative framing does not improve average fairness perceptions or compliance relative to the control condition, but reduces the likelihood of selecting the strictest inspection policy, suggesting a weakening of monitoring intensity. This pattern is consistent with the interpretation that relational framing may alter the perceived appropriateness of strict monitoring within a supply chain setting. Third, the introduction of conditional financial incentives yields more adverse effects. While average monitoring effort remains unchanged, participants are less likely to perceive compliance requests as fair or effective, and compliance outcomes decline, consistent with a crowding-out of intrinsic or relational motivations. This is consistent with the literature on motivational crowding out (Gneezy and Rustichini, 2000; Bowles and Polania-Reyes, 2012) and the moral costs of incentivized reporting (Fiorin, 2023). Finally, heterogeneity analysis reveals that responses are shaped by motivational orientation: private-sector participants tend to substitute strict

monitoring with more moderate inspection strategies while maintaining relatively stable compliance levels, suggesting greater responsiveness to the relational framing of the task. In contrast, public-sector and highly prosocial participants maintain stable monitoring but reduce compliance under both treatments, consistent with the observed decline in fairness perceptions. Among highly prosocial individuals in particular, both collaborative framing and conditional incentives reduce the likelihood of selecting strict monitoring and weaken compliance outcomes, in line with a crowding-out of intrinsic and relational motivations. Taken together, the results suggest that hybrid governance approaches combining deterrence with collaborative signals and financial elements may generate conflicting cues that undermine perceptions of fairness and, in turn, weaken monitoring and compliance in different ways across groups. This echoes prior work suggesting that multinational enterprises often combine weak carrots and small sticks (LeBaron and Lister, 2015; Goerzen and Van Assche, 2023), and that misaligned enforcement mechanisms can erode trust in buyer-supplier relations (Frenkel and Scott, 2002; Dyer and Chu, 2003). Effective private governance may thus require greater coherence between tone, incentives, and enforcement mechanisms — possibly relying more on deterrence tools in arm's-length relationships and on reciprocity-based strategies where stronger relational and trust-based ties exist (Amengual and Distelhorst, 2025).

Several limitations should be acknowledged. First, to simplify instructions and ensure participants clearly understood the task, we made several simplifying assumptions that may have reduced realism and inadvertently calibrated the control condition to feel stronger than intended. We assumed that Alpha could inspect both tier-1 and tier-2 suppliers with a single detection probability across tiers (10%), even though in reality audit coverage and effectiveness decline sharply upstream. Buyers frequently audit tier-1 suppliers — often with near-certainty in countries with weak public governance — but tier-2 and other sub-suppliers are monitored far less systematically. We also assumed perfect inspection mechanisms. Social audits are known to suffer from reliability issues, including incomplete reporting and "false compliance" practices including training workers on what to say, terminating troublesome workers, subcontracting to uncertified units, and bribing auditors (Goerzen and Van Assche, 2023). Moreover, despite the relatively modest probability of detection in the experiment, the threat of terminating business with Beta may have

felt disproportionately strong, as buyers rarely impose such direct consequences for sub-supplier violations and typically keep sanction language vague. These design choices likely elevated perceived enforcement strength in the control condition. Future research could address these limitations by testing alternative deterrence calibrations — such as tier-specific detection rates or alternative penalty structures — to disentangle which elements most strongly drive sustained levels of monitoring and compliance behaviour.

Second, while our collaborative treatment was designed to enhance fairness through inclusive "we" language and "leading-by-example" framing, it still conveyed a top-down request, potentially violating principles of procedural fairness (Leventhal, 1980). Future research could examine whether greater participatory voice — such as jointly setting compliance targets — strengthens perceived fairness and cooperative responses.

Third, the design of the financial incentive treatment presents an important limitation. Because the bonus was conditional on reporting sub-supplier non-compliance, we cannot disentangle whether the backfiring effect was driven by the financial incentive, the reporting mechanism, or their interaction. Moreover, while such reporting-based incentives rarely exist in real-world buyer-supplier relations — where buyers typically expect first-tier suppliers to ensure compliance rather than to report violations — our aim was to explore, in a counterfactual manner, whether explicit and contractible incentives could enhance delegated monitoring. Finally, although this study focused primarily on fairness perceptions as the mediating channel, future research could collect complementary measures such as moral costs or specific dimensions of fairness (procedural, distributive, or interactional) to better understand the mechanisms underlying these behavioural responses.

Beyond these contributions, the study demonstrates the value of experimental methods for testing policy-relevant interventions in supply chains before large-scale implementation. While experiments cannot substitute for field studies, they provide an important and cost-effective starting point for investigating (and anticipating) behavioural responses under emerging due diligence regimes such as the EU's 2024 Corporate Sustainability Due Diligence Directive or similar governance policies. This highlights the potential of experimental approaches to enrich debates on how best to balance deterrence, collaboration, and incentives in the pursuit of more

sustainable and equitable supply chains.

References

Amengual, M. and Distelhorst, G.: 2025, Cooperation and punishment in managing social performance: labor standards in the Gap Inc. supply chain, *Strategic Management Journal* **46**(11), 2663–2689.

Amengual, M., Distelhorst, G. and Tobin, D.: 2019, Global purchasing as labor regulation: the missing middle, *ILR Review* **73**(4), 817–840.

Aßländer, M. S., Roloff, J. and Zamantili Nayır, D.: 2016, Suppliers as stewards? managing social standards in first- and second-tier suppliers, *Journal of Business Ethics* **139**, 661–683.

Bloemer, A. and Minner, S.: 2025, Auditing and training to incentivize sustainability in multi-tier supply chains: substitutes or complements?, *European Journal of Operational Research* .

Bowles, S. and Polania-Reyes, S.: 2012, Economic incentives and social preferences: substitutes or complements?, *Journal of Economic Literature* **50**(2), 368–425.

Burger, J. M., Messian, N., Anderson, C., Del Prado, A. and Patel, S.: 2004, What a coincidence! the effects of incidental similarity on compliance, *Personality and Social Psychology Bulletin* **30**(1), 35–43.

Butler, J., Serra, D. and Spagnolo, G.: 2019, Motivating whistleblowers, *Management Science* **66**.

Cialdini, R. B. and Goldstein, N. J.: 2004, Social influence: Compliance and conformity, *Annual Review of Psychology* **55**(1), 591–621.

Drouvelis, M. and Nosenzo, D.: 2013, Group identity and leading-by-example, *Journal of Economic Psychology* **39**, 414–425.

Dyer, J. H. and Chu, W.: 2003, The role of trustworthiness in reducing transaction costs and improving performance: empirical evidence from the united states, japan, and korea, *Organization Science* **14**(1), 57–68.

European Union: 2024, Directive (eu) 2024/1760 on corporate sustainability due diligence (csddd). Official Journal of the European Union.

Falk, A., Becker, A., Dohmen, T., Huffman, D. and Sunde, U.: 2022, The preference survey module: A validated instrument for measuring risk, time, and social preferences, *Management Science* **69**(4), 935–1950.

Fandel, G. and Trockel, J.: 2013, Applying a one-shot and infinite repeated inspection game to materials management, *Central European Journal of Operations Research* **21**, 495–506.

Fehr, E., Gächter, S. and Kirchsteiger, G.: 1997, Reciprocity as a contract enforcement device: experimental evidence, *Econometrica* **65**(4), 833–860.

Fehr, E. and Schmidt, K. M.: 2004, Fairness and incentives in a multi-task principal–agent model, *Journal of Economics* **106**(3), 453–474.

Feldman, Y.: 2025, The state's relationship with its citizens: From coercion to compliance and back, *Bar Ilan University Faculty of Law Research Paper* **No. 5967915**.

Figuières, C., Masclet, D. and Willinger, M.: 2012, Vanishing leadership and declining reciprocity in a sequential contribution experiment, *Economic Inquiry* **50**(3), 567–584.

Fiorin, S.: 2023, Reporting peers' wrongdoing: evidence on the effect of incentives on morally controversial behavior, *Journal of the European Economic Association* **21**(3), 1033–1071.

Fontana, E. and Egels-Zandén, N.: 2019, Non sibi, sed omnibus: influence of supplier collective behaviour on corporate social responsibility in the bangladeshi apparel supply chain, *Journal of Business Ethics* **159**, 1047–1064.

Frenkel, S. J. and Scott, D.: 2002, Compliance, collaboration, and codes of labor practice: the ADIDAS connection, *California Management Review* **45**(1), 29–49.

Gächter, S., Nosenzo, D., Renner, E. and Sefton, M.: 2012, Who makes a good leader? cooperativeness, optimism and leading-by-example, *Economic Inquiry* **50**(4), 853–967.

Gneezy, U. and Rustichini, A.: 2000, Pay enough or don't pay at all, *The Quarterly Journal of Economics* **115**(3), 791–810.

Goerzen, A. and Van Assche, A.: 2023, Cascading compliance to achieve improved GVC sustainability: what is it and why does it fail?, *Research handbook on international corporate social responsibility*, pp. 127–137.

Graf Lambsdorff, J.: 2015, Preventing corruption by promoting trust: Insights from behavioral science, *Passauer Diskussionspapiere - Volkswirtschaftliche Reihe, Universität Passau* **No. V-69-15**.

Kessler, J. B. and Leider, S.: 2016, Procedural fairness and the cost of control, *The Journal of Law, Economics, and Organization* **32**(4), 685–718.

Koenig, P. and Poncet, S.: 2022, The effects of the rana plaza collapse on the sourcing choices of french importers, *Journal of International Economics* **137**.

Kuruvilla, S., Liu, M., Chen, W. and Li, C.: 2020, Field opacity and practice-outcome decoupling: private regulation of labor standards in global supply chains, *ILR Review* **73**(4), 841–872.

LeBaron, G. and Lister, J.: 2015, Benchmarking global supply chains: the power of the 'ethical audit' regime, *Review of International Studies* **41**(5), 905–924. Special Issue: The Politics of Numbers.

Leventhal, G. S.: 1980, What should be done with equity theory?, *in* K. J. Gergen, M. S. Greenberg and R. H. Willis (eds), *Social exchange*, pp. 27–55.

Locke, R., Amengual, M. and Mangla, A.: 2009, Virtue out of necessity? compliance, commitment, and the improvement of labor conditions in global supply chains, *Politics & Society* **37**(3), 319–351.

Nosenzo, D., Offerman, T., Sefton, M. and Van der Veen, A.: 2010, Inducing good behavior: bonuses versus fines in inspection games, *Discussion Paper 2010-21*, Centre for Decision Research and Experimental Economics, University of Nottingham.

Nyreröd, T. and Spagnolo, G.: 2021, A fresh look at whistleblower rewards, *SSRN Electronic Journal* .

Potters, J., Sefton, M. and Vesterlund, L.: 2007, Leading-by-example and signaling in voluntary contribution games: an experimental study, *Economic Theory* **33**(1), 169–182.

Rauhut, H.: 2009, Higher punishment, less control?, *Rationality and Society* **21**(3), 359–392.

Rauhut, H.: 2015, Stronger inspection incentives, less crime? further experimental evidence on inspection games, *Rationality and Society* **27**(4), 414–454.

Schildberg-Hörisch, H. and Strassmair, C.: 2012, An experimental test of the deterrence hypothesis, *The Journal of Law, Economics, and Organization* **28**(3), 447–459.

Schmolke, K. U. and Utikal, V.: 2018, Whistleblowing: incentives and situational determinants, *SSRN Electronic Journal* .

Soundararajan, V. and Brammer, S.: 2018, Developing country sub-supplier responses to social sustainability requirements of intermediaries: exploring the influence of framing on fairness perceptions and reciprocity, *Journal of Operations Management* **58/59**, 42–58.

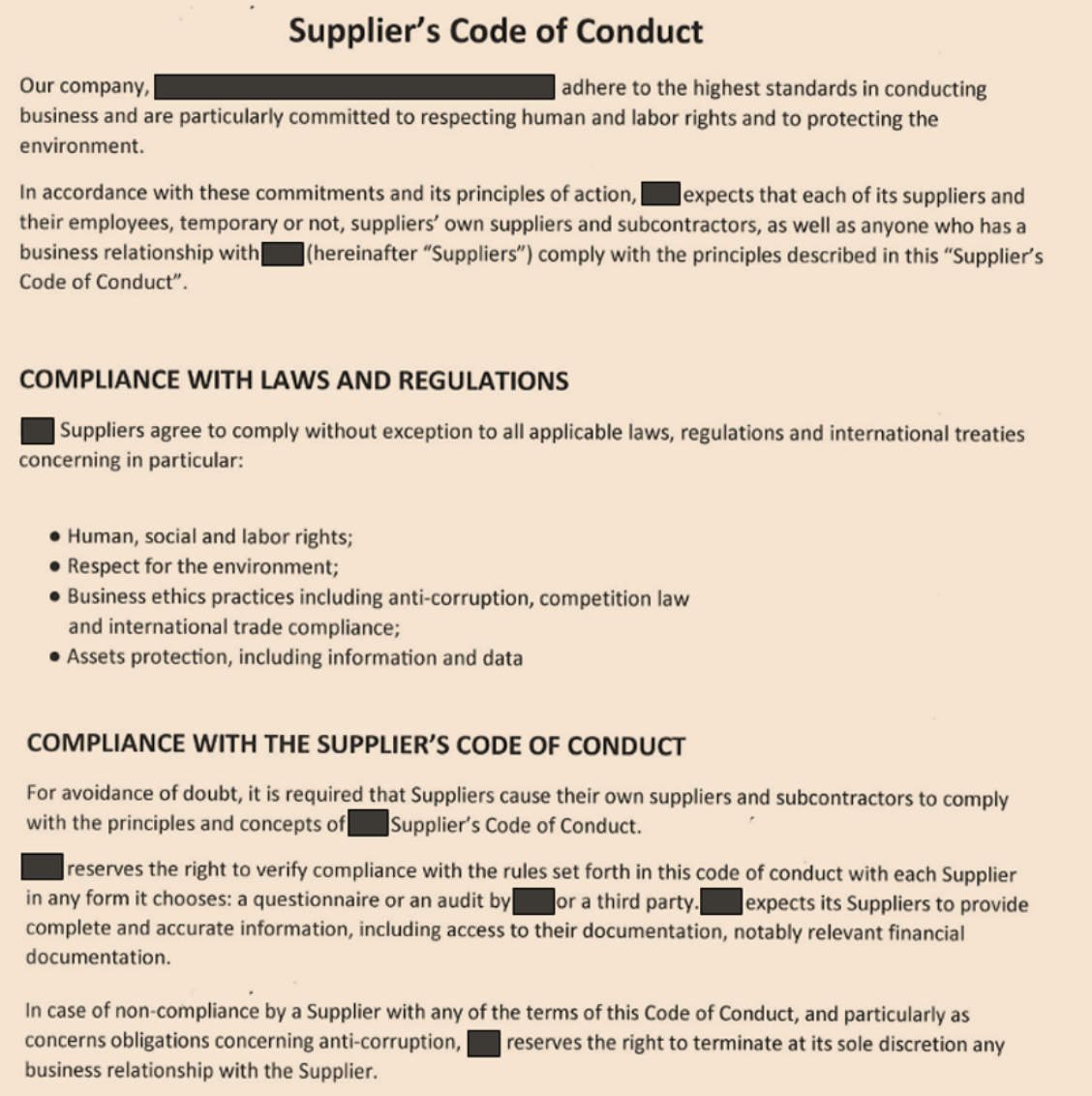
Tajfel, H. and Turner, J. C.: 1986, The social identity theory of intergroup behavior, *in* S. Worchel and W. G. Austin (eds), *Psychology of intergroup relations*, pp. 7–24.

Wilhelm, M. M., Blome, C., Bhakoo, V. and Paulraj, A.: 2016, Sustainability in multi-tier supply chains: understanding the double agency role of the first-tier supplier, *Journal of Operations Management* **41**, 42–60.

Appendix

3.A Additional tables and figures

Figure 3.A.1: Example of a supplier's code of conduct



Supplier's Code of Conduct

Our company, [redacted] adhere to the highest standards in conducting business and are particularly committed to respecting human and labor rights and to protecting the environment.

In accordance with these commitments and its principles of action, [redacted] expects that each of its suppliers and their employees, temporary or not, suppliers' own suppliers and subcontractors, as well as anyone who has a business relationship with [redacted] (hereinafter "Suppliers") comply with the principles described in this "Supplier's Code of Conduct".

COMPLIANCE WITH LAWS AND REGULATIONS

[redacted] Suppliers agree to comply without exception to all applicable laws, regulations and international treaties concerning in particular:

- Human, social and labor rights;
- Respect for the environment;
- Business ethics practices including anti-corruption, competition law and international trade compliance;
- Assets protection, including information and data

COMPLIANCE WITH THE SUPPLIER'S CODE OF CONDUCT

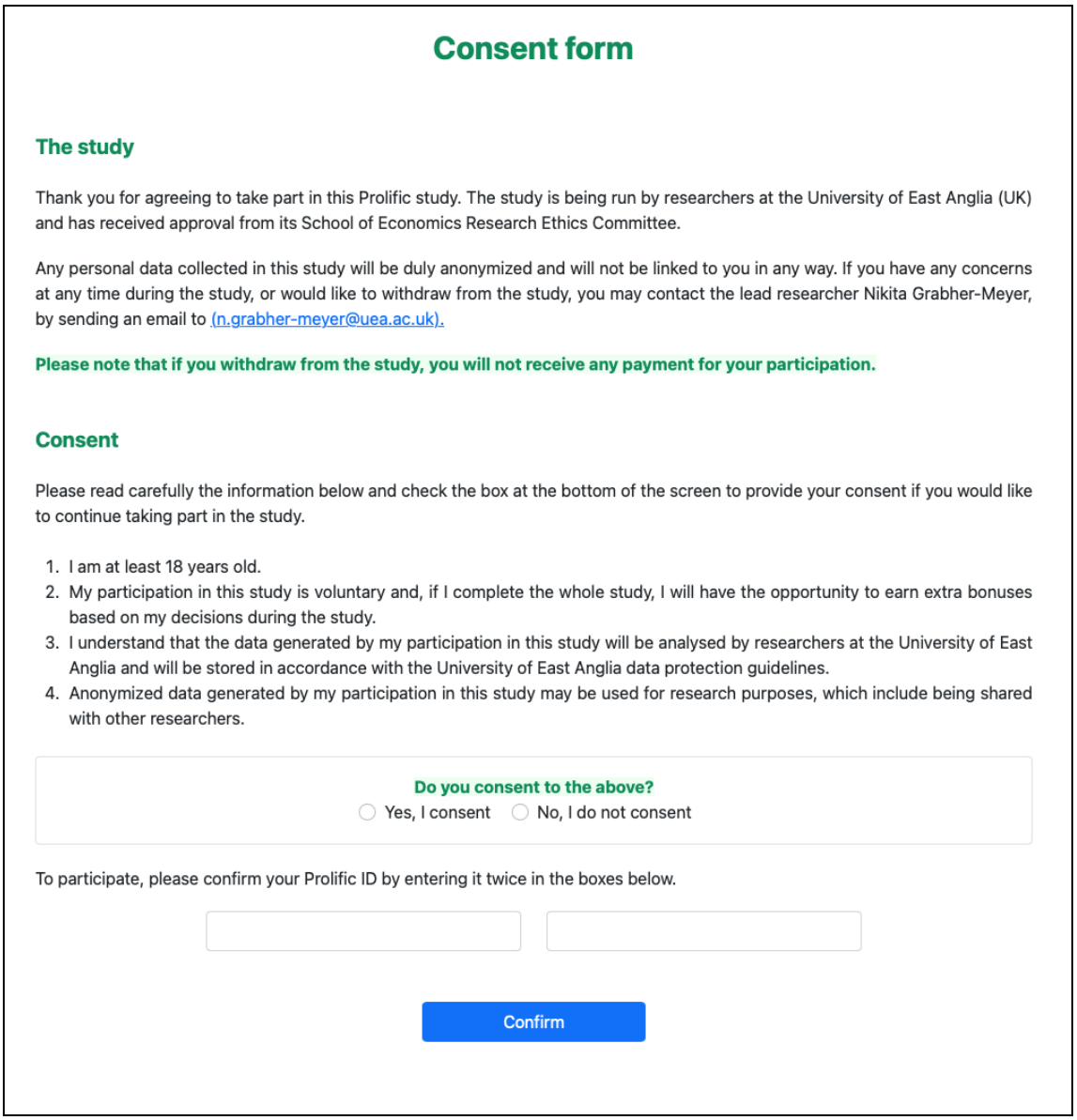
For avoidance of doubt, it is required that Suppliers cause their own suppliers and subcontractors to comply with the principles and concepts of [redacted] Supplier's Code of Conduct.

[redacted] reserves the right to verify compliance with the rules set forth in this code of conduct with each Supplier in any form it chooses: a questionnaire or an audit by [redacted] or a third party. [redacted] expects its Suppliers to provide complete and accurate information, including access to their documentation, notably relevant financial documentation.

In case of non-compliance by a Supplier with any of the terms of this Code of Conduct, and particularly as concerns obligations concerning anti-corruption, [redacted] reserves the right to terminate at its sole discretion any business relationship with the Supplier.

Source: This figure reproduces an anonymized excerpt from an actual supplier code of conduct that one of the researchers had access to through prior business experience.

Figure 3.A.2: Beta's game instructions and survey



Consent form

The study

Thank you for agreeing to take part in this Prolific study. The study is being run by researchers at the University of East Anglia (UK) and has received approval from its School of Economics Research Ethics Committee.

Any personal data collected in this study will be duly anonymized and will not be linked to you in any way. If you have any concerns at any time during the study, or would like to withdraw from the study, you may contact the lead researcher Nikita Grabher-Meyer, by sending an email to (n.grabher-meyer@uea.ac.uk).

Please note that if you withdraw from the study, you will not receive any payment for your participation.

Consent

Please read carefully the information below and check the box at the bottom of the screen to provide your consent if you would like to continue taking part in the study.

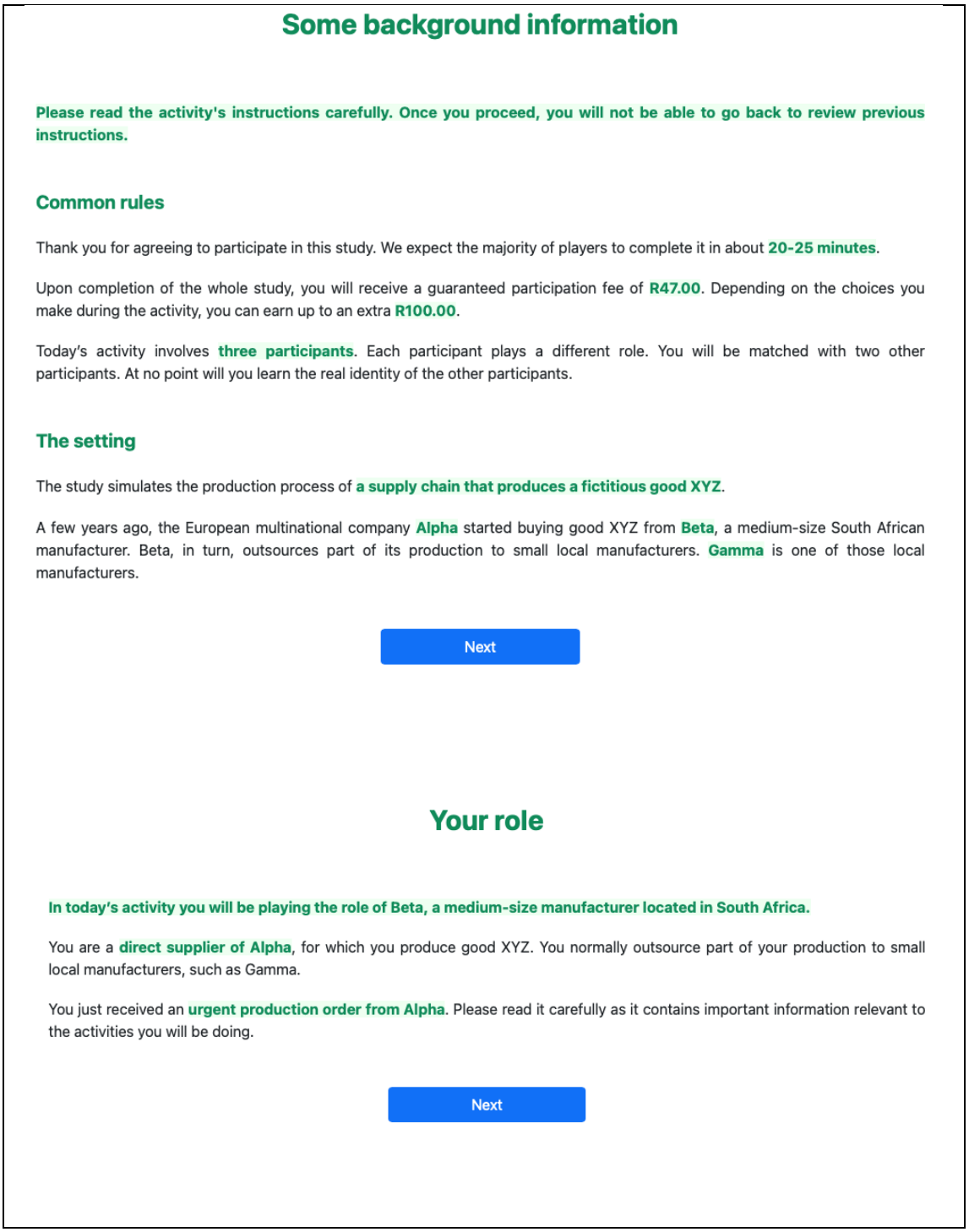
1. I am at least 18 years old.
2. My participation in this study is voluntary and, if I complete the whole study, I will have the opportunity to earn extra bonuses based on my decisions during the study.
3. I understand that the data generated by my participation in this study will be analysed by researchers at the University of East Anglia and will be stored in accordance with the University of East Anglia data protection guidelines.
4. Anonymized data generated by my participation in this study may be used for research purposes, which include being shared with other researchers.

Do you consent to the above?

○ Yes, I consent ○ No, I do not consent

To participate, please confirm your Prolific ID by entering it twice in the boxes below.

Confirm


Some background information
Please read the activity's instructions carefully. Once you proceed, you will not be able to go back to review previous instructions.
Common rules
Thank you for agreeing to participate in this study. We expect the majority of players to complete it in about 20-25 minutes.
Upon completion of the whole study, you will receive a guaranteed participation fee of R47.00. Depending on the choices you make during the activity, you can earn up to an extra R100.00.
Today's activity involves three participants. Each participant plays a different role. You will be matched with two other participants. At no point will you learn the real identity of the other participants.
The setting
The study simulates the production process of a supply chain that produces a fictitious good XYZ.
A few years ago, the European multinational company Alpha started buying good XYZ from Beta, a medium-size South African manufacturer. Beta, in turn, outsources part of its production to small local manufacturers. Gamma is one of those local manufacturers.
Next
Your role
In today's activity you will be playing the role of Beta, a medium-size manufacturer located in South Africa.
You are a direct supplier of Alpha, for which you produce good XYZ. You normally outsource part of your production to small local manufacturers, such as Gamma.
You just received an urgent production order from Alpha. Please read it carefully as it contains important information relevant to the activities you will be doing.
Next

Urgent production order

(C)

Dear Beta,

Our company Alpha would like to place an order for **up to 100 units of good XYZ** to be produced as soon as possible.

You will be paid **R1.0 per unit** produced.

Please note that we are **committed** to producing good XYZ in compliance with 15 labour standards.

We **expect** that you carry out your production activities in compliance with **at least 5 of those standards**, and that you ensure that your subcontractors do the same.

We at Alpha have a policy of conducting **thorough inspections** of our suppliers' compliance, covering both the direct suppliers and any subcontractors, **in one out of every 10 orders**. If we inspect your activities and find out that you or your subcontractors do not comply with at least 5 of the standards, we will terminate our business relationship with Beta, and **your company will not be paid** for any units of good XYZ you deliver.

Regards,
Alpha

Next

(T1)

Dear Beta,

Our company Alpha would like to place an order for **up to 100 units of good XYZ** to be produced as soon as possible.

You will be paid **R1.0 per unit** produced.

Please note that we are **committed to working together with our entire supply chain** to produce good XYZ in compliance with 15 labour standards.

As a proof of our commitment, we have completed a certification process, confirming that we currently comply with 12 of those standards.

We **trust** that you, as a valued member of our supply chain, will do your part by carrying out your production activities in compliance with **at least 5 of those standards**, and that your subcontractors will do so as well.

We at Alpha have a policy of conducting **thorough inspections** of our suppliers' compliance, covering both the direct suppliers and any subcontractors, **in one out of every 10 orders**. If we inspect your activities and find out that you or your subcontractors do not comply with at least 5 of the standards, we will regretfully have to terminate our business relationship with Beta, and **your company will not be paid** for any units of good XYZ you deliver.

Regards,
Alpha

(T2)

Dear Beta,

Our company Alpha would like to place an order for **up to 100 units of good XYZ** to be produced as soon as possible.

You will be paid **R1.0 per unit** produced.

Please note that we are **committed to working together with our entire supply chain** to produce good XYZ in compliance with 15 labour standards.

As a proof of our commitment, we have completed a certification process, confirming that we currently comply with 12 of those standards.

We **trust** that you, as a valued member of our supply chain, will do your part by carrying out your production activities in compliance with **at least 5 of those standards**, and that your subcontractors will do so as well.

We at Alpha have a policy of conducting **thorough inspections** of our suppliers' compliance, covering both the direct suppliers and any subcontractors, **in one out of every 10 orders**. If we inspect your activities and find out that you or your subcontractors do not comply with at least 5 of the standards, we will regretfully have to terminate our business relationship with Beta, and **your company will not be paid** for any units of good XYZ you deliver.

However, if you comply with at least 5 of the standards, but you discover that your subcontractors do not, you are responsible for reporting your subcontractors' non-compliance to us. **Only if you report**, we will not terminate our business relationship with Beta, and **your company will still be paid** for any units of good XYZ you deliver.

Regards,
Alpha

Your activities

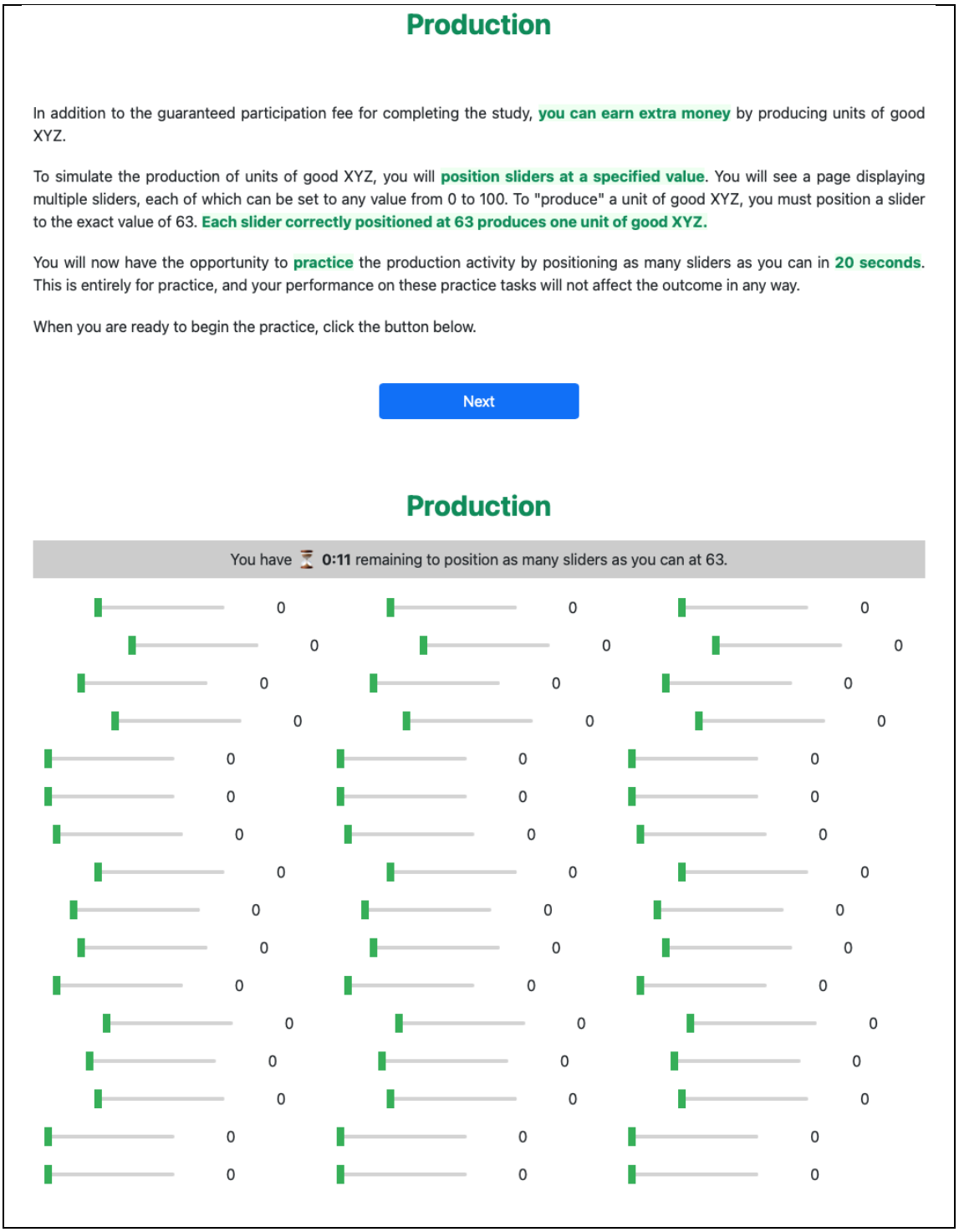
Because of the urgency of Alpha's order, you have decided to **outsource** some of the production of good XYZ to your subcontractor **Gamma**.

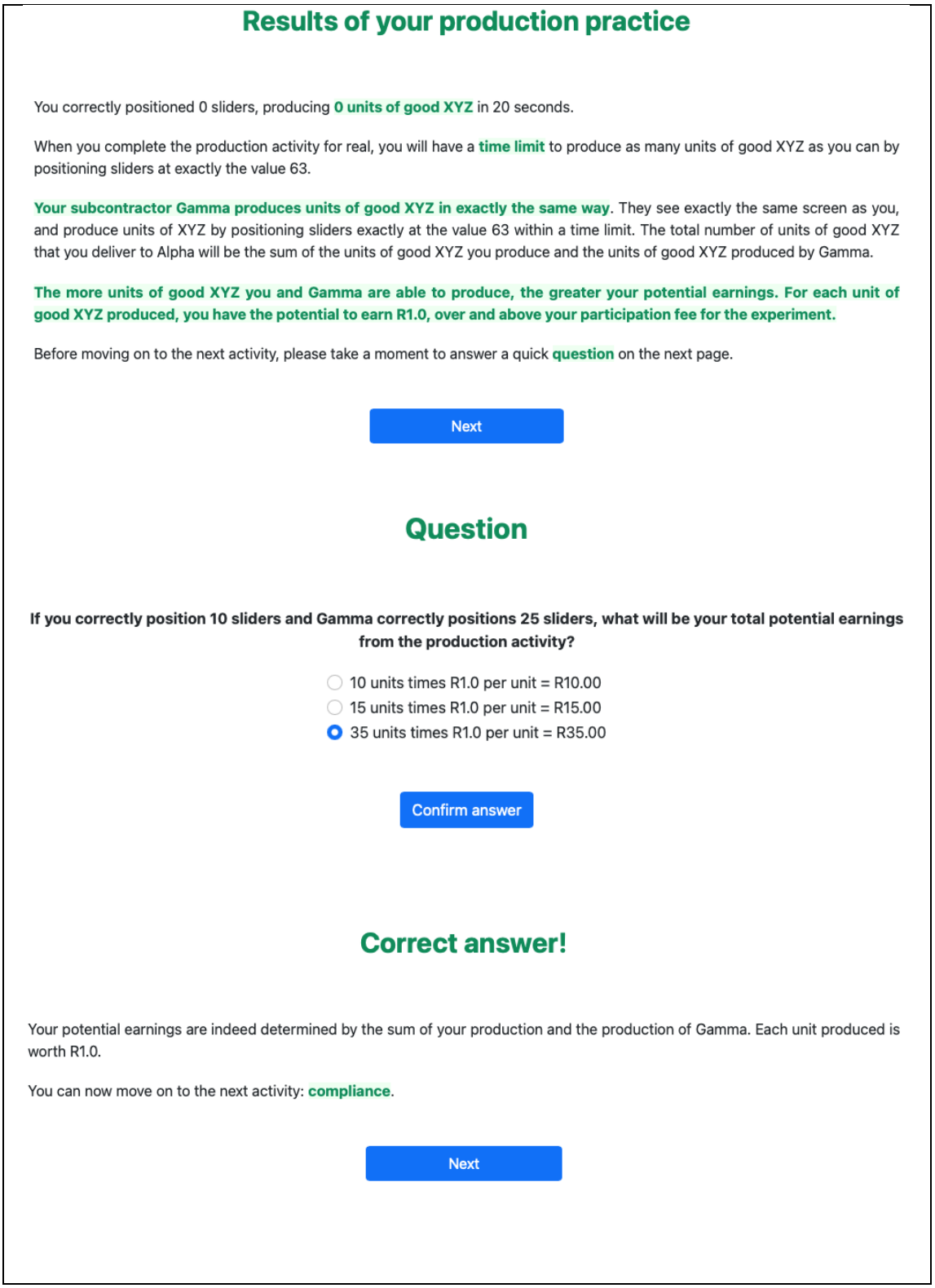
In order to fulfil Alpha's order, you will need to divide your time among **three activities**:

1. **Production**, which simulates your firm's efforts to produce units of good XYZ.
2. **Compliance**, which simulates your firm's efforts to comply with the labour standards.
3. **Inspection**, which determines your firm's ability to monitor how many labour standards Gamma complies with.

We will describe each of these activities in more detail.

Next


Production
In addition to the guaranteed participation fee for completing the study, you can earn extra money by producing units of good XYZ.
To simulate the production of units of good XYZ, you will position sliders at a specified value. You will see a page displaying multiple sliders, each of which can be set to any value from 0 to 100. To "produce" a unit of good XYZ, you must position a slider to the exact value of 63. Each slider correctly positioned at 63 produces one unit of good XYZ.
You will now have the opportunity to practice the production activity by positioning as many sliders as you can in 20 seconds. This is entirely for practice, and your performance on these practice tasks will not affect the outcome in any way.
When you are ready to begin the practice, click the button below.
Next
Production
You have 0:11 remaining to position as many sliders as you can at 63.


Results of your production practice
You correctly positioned 0 sliders, producing 0 units of good XYZ in 20 seconds.
When you complete the production activity for real, you will have a time limit to produce as many units of good XYZ as you can by positioning sliders at exactly the value 63.
Your subcontractor Gamma produces units of good XYZ in exactly the same way. They see exactly the same screen as you, and produce units of XYZ by positioning sliders exactly at the value 63 within a time limit. The total number of units of good XYZ that you deliver to Alpha will be the sum of the units of good XYZ you produce and the units of good XYZ produced by Gamma.
The more units of good XYZ you and Gamma are able to produce, the greater your potential earnings. For each unit of good XYZ produced, you have the potential to earn R1.0, over and above your participation fee for the experiment.
Before moving on to the next activity, please take a moment to answer a quick question on the next page.
Next
Question
If you correctly position 10 sliders and Gamma correctly positions 25 sliders, what will be your total potential earnings from the production activity?
10 units times R1.0 per unit = R10.00
15 units times R1.0 per unit = R15.00
35 units times R1.0 per unit = R35.00
Confirm answer
Correct answer!
Your potential earnings are indeed determined by the sum of your production and the production of Gamma. Each unit produced is worth R1.0.
You can now move on to the next activity: compliance.
Next

Compliance

Alpha is committed to ensuring that the production of good XYZ adheres to 15 labour standards.

Alpha **expects** that you carry out your production activities in compliance with **at least 5 of those standards**.

To simulate your compliance with a labour standard, you will **transcribe letters**. You will see a series of images, each showing four uppercase letters. Each image represents one standard. To "comply" with a standard, you must correctly transcribe the letters exactly as they appear in that image. Please note that the letters are not case-sensitive.

You will now have the opportunity to **practice** the compliance activity by transcribing as many sequences of letters as you can in **20 seconds**. This is entirely for practice, and your performance on these practice tasks will not affect the outcome in any way.

When you are ready to begin the practice, click the button below.

Next

Compliance

You have ⏳ **0:11** remaining to transcribe images. Remember letters are not case-sensitive!

Z B S V

Type the text

Next

Results of your compliance practice

You correctly transcribed 0 images, complying with **0 labour standards** in 20 seconds.

When you complete the compliance activity for real, you will have a **time limit** to comply with as many labour standards as you can by transcribing up to 15 images.

Your subcontractor Gamma complies with labour standards in exactly the same way. They see exactly the same screens as you, and comply with labour standards by correctly transcribing up to 15 images within a time limit.

Your performance in this activity will help you **maintain the business relationship with Alpha and secure your extra earnings from production**. Alpha has a policy of conducting thorough inspections of their suppliers' compliance in one out of every 10 orders. If Alpha inspects your firm and finds out that you do not meet at least 5 of the standards, Alpha will terminate the business relationship with your firm Beta, and you will receive no payment for any units of good XYZ that you and Gamma produced.

Before moving on to the next activity, please take a moment to answer a quick **question** on the next page.

Next

Question

Suppose you correctly transcribe 13 images. Which of the following is true?

- ○ This will result in no benefit to you
- ○ You may be found not to meet Alpha's expectations if you are inspected
- ◉ You will be found to meet Alpha's expectations if you are inspected

Confirm answer

Correct answer!

If you are inspected, you will be found to meet Alpha's expectations if, and only if, you have transcribed at least 5 images correctly.

You can now move on to the next activity: **inspection**.

Next

Inspection

Your subcontractor Gamma has 120 seconds in which to produce units of good XYZ for you and comply with labour standards. Like you, the participant playing the role of Gamma has the opportunity to earn extra money by producing units of good XYZ. They decide how to allocate their 120 seconds between completing the production and compliance activities.

Alpha expects that you carry out your production activities in compliance with at least 5 of the labour standards, and that you will ensure that Gamma does the same.

Remember that Alpha has committed to inspecting one out of every 10 orders to ensure that production is being carried out in compliance with the required number of standards. If Alpha chooses your order for inspection, if either you or Gamma do not meet the minimum number of standards, Alpha will terminate the business relationship with your firm Beta, and you will receive no payment for any units of good XYZ you deliver.

In other words, your company Beta is responsible not only for your own compliance, but for Gamma's compliance as well.

Your firm Beta monitors its subcontractors by setting its own **inspection policy**. The inspection policy commits to inspecting a fraction of its subcontractors at random. If Gamma is chosen to be inspected, you will find out the exact number of labour standards that Gamma has complied with.

Inspection comes at a **cost**: if you choose to inspect more subcontractors, you will spend more of your time inspecting, and this takes away from the time you have to complete your own production and compliance activities.

Next

Inspection

Your choice of your inspection policy will have **three important effects**:

1. **Influencing how Gamma divides their time between compliance and production.** Before Gamma decides how to divide their time between the compliance and production activities, they are informed of the chance of being inspected. With a greater chance of being inspected, Gamma may be more motivated to focus on ensuring they comply with at least the minimum number of standards.
2. **Increasing your chance of observing Gamma's compliance.** The more time you allocate to inspecting, the greater the chance that you will inspect Gamma and determine if they do or do not comply with the minimum number of standards. If you find out that Gamma does not comply with the minimum number of standards, Gamma will receive no payment for their production activities.
3. **Decreasing the time you have for completing your other activities.** The more time you allocate to inspecting Gamma, the less time you will have to complete your own compliance and production activities.

Before moving on to the real activity, please take a moment to answer a quick **question** on the next page.

Next

Question

Once you have decided how much time to spend on inspection, how does this decision affect the amount of time you have for your own compliance and production?

- (●) It reduces the total amount of time you have available, but you can still choose how much to allocate to your compliance and your production
- ( ) Time spent on inspection reduces the time you have for your own compliance only
- ( ) Time spent on inspection reduces the time you have for your own production only

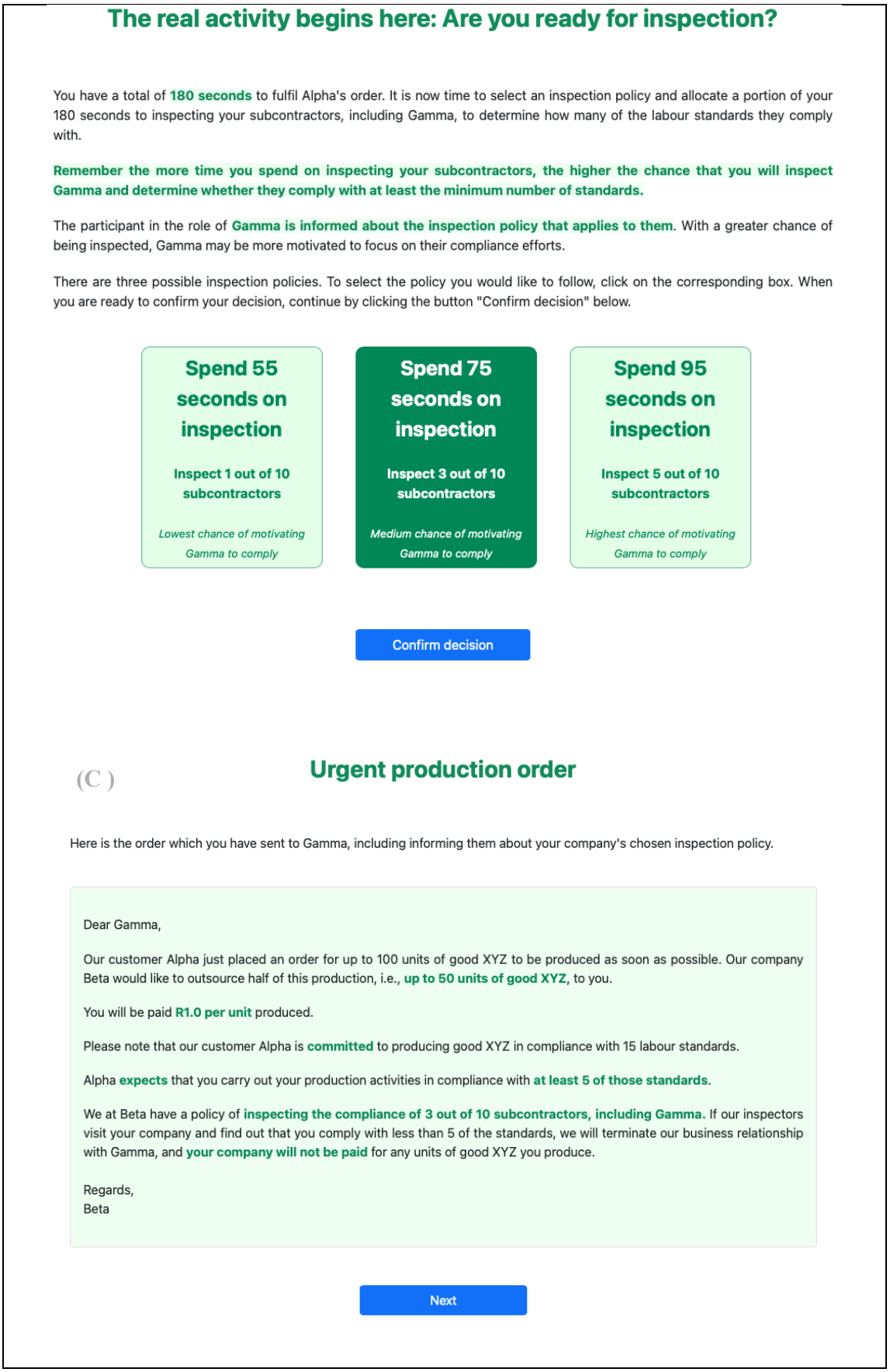
Confirm answer

Correct answer!

The total seconds available for your own compliance and production are indeed determined by subtracting the time spent inspecting subcontractors from the total time allocated to fulfil Alpha's order.

You can now move on to the **real activity**.

Next


The real activity begins here: Are you ready for inspection?
You have a total of 180 seconds to fulfil Alpha's order. It is now time to select an inspection policy and allocate a portion of your 180 seconds to inspecting your subcontractors, including Gamma, to determine how many of the labour standards they comply with.
Remember the more time you spend on inspecting your subcontractors, the higher the chance that you will inspect Gamma and determine whether they comply with at least the minimum number of standards.
The participant in the role of Gamma is informed about the inspection policy that applies to them. With a greater chance of being inspected, Gamma may be more motivated to focus on their compliance efforts.
There are three possible inspection policies. To select the policy you would like to follow, click on the corresponding box. When you are ready to confirm your decision, continue by clicking the button "Confirm decision" below.
Spend 55 seconds on inspection
Inspect 1 out of 10 subcontractors
Lowest chance of motivating Gamma to comply
Spend 75 seconds on inspection
Inspect 3 out of 10 subcontractors
Medium chance of motivating Gamma to comply
Spend 95 seconds on inspection
Inspect 5 out of 10 subcontractors
Highest chance of motivating Gamma to comply
Confirm decision
(C)
Urgent production order
Here is the order which you have sent to Gamma, including informing them about your company's chosen inspection policy.
Dear Gamma,
Our customer Alpha just placed an order for up to 100 units of good XYZ to be produced as soon as possible. Our company Beta would like to outsource half of this production, i.e., up to 50 units of good XYZ, to you.
You will be paid R1.0 per unit produced.
Please note that our customer Alpha is committed to producing good XYZ in compliance with 15 labour standards.
Alpha expects that you carry out your production activities in compliance with at least 5 of those standards.
We at Beta have a policy of inspecting the compliance of 3 out of 10 subcontractors, including Gamma. If our inspectors visit your company and find out that you comply with less than 5 of the standards, we will terminate our business relationship with Gamma, and your company will not be paid for any units of good XYZ you produce.
Regards,
Beta
Next

(T1)

Dear Gamma,

Our customer Alpha just placed an order for up to 100 units of good XYZ to be produced as soon as possible. Our company Beta would like to outsource half of this production, i.e., **up to 50 units of good XYZ**, to you.

You will be paid **R1.0 per unit** produced.

Please note that our customer Alpha is **committed** to working together with its entire supply chain to produce good XYZ in compliance with 15 labour standards.

As a proof of its commitment, Alpha has completed a certification process, confirming that they currently comply with 12 of those standards.

We are fully committed to following Alpha's example, and we **trust** that you, as a valued member of our supply chain, will do your part by carrying out your production activities in compliance with at least 5 of those standards.

We at Beta have a policy of **inspecting the compliance of 1 out of 10 subcontractors, including Gamma**. If our inspectors visit your company and find out that you do not comply with at least 5 of the standards, we will regretfully have to terminate our business relationship with Gamma, and **your company will not be paid** for any units of good XYZ you produce.

Regards,
Beta

Allocating time between compliance and production

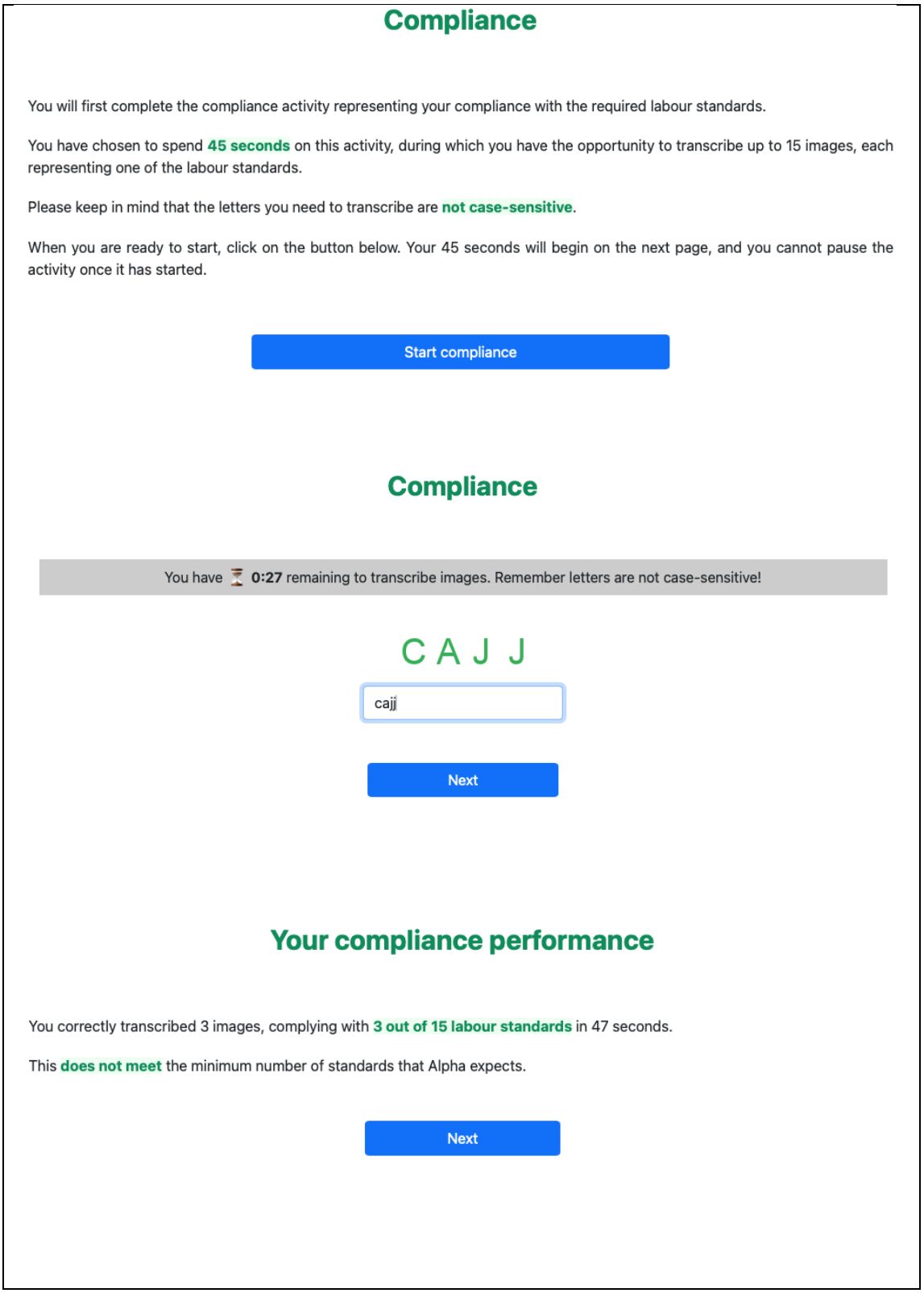
You have chosen to spend **75** seconds on inspecting your subcontractors' compliance. This leaves you with **105** seconds for your own compliance and production activities.

Please use the slider below to choose how you would like to allocate your remaining 105 seconds between the **compliance** activity of transcribing images, and the **production** activity of positioning sliders to produce units of good XYZ.

45 seconds for Compliance
60 seconds for Production

All 105 seconds for production — All 105 seconds for compliance

Confirm decision


Compliance
You will first complete the compliance activity representing your compliance with the required labour standards.
You have chosen to spend 45 seconds on this activity, during which you have the opportunity to transcribe up to 15 images, each representing one of the labour standards.
Please keep in mind that the letters you need to transcribe are not case-sensitive.
When you are ready to start, click on the button below. Your 45 seconds will begin on the next page, and you cannot pause the activity once it has started.
Start compliance
Compliance
You have 0:27 remaining to transcribe images. Remember letters are not case-sensitive!
C A J J
cajj
Next
Your compliance performance
You correctly transcribed 3 images, complying with 3 out of 15 labour standards in 47 seconds.
This does not meet the minimum number of standards that Alpha expects.
Next

Production

You will now complete the production activity representing the production of units of good XYZ.

You have chosen to spend **58 seconds** on this activity, during which you have the opportunity to position up to 50 sliders, each representing one unit of good XYZ.

Please keep in mind that:

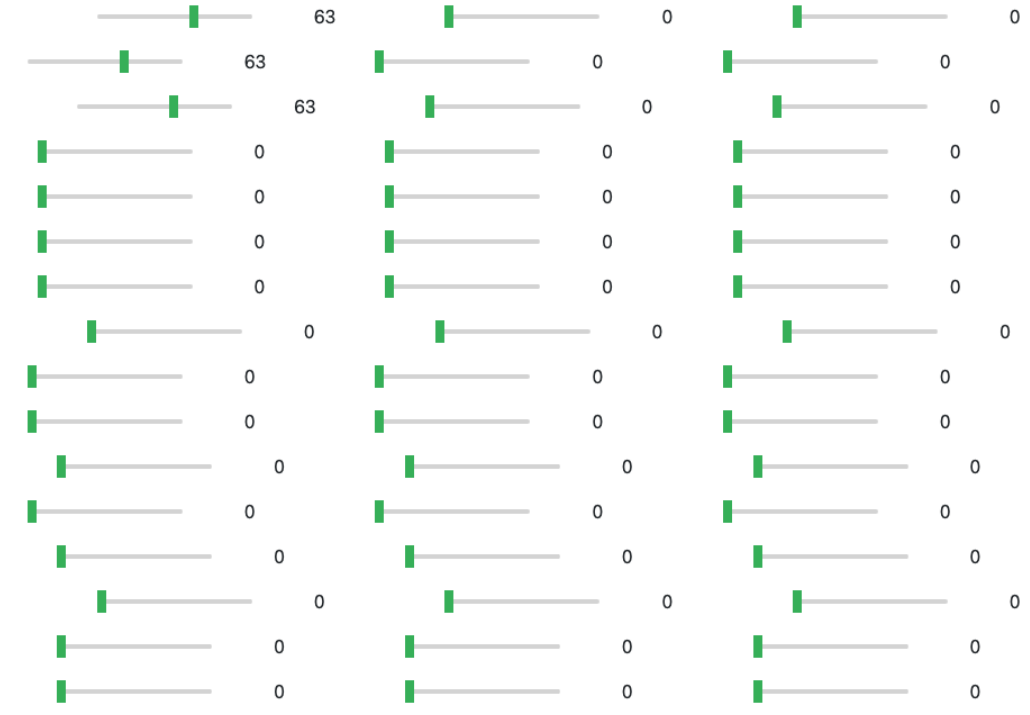
- To produce a unit of good XYZ, you must position a slider to the exact value of **63**. Each slider correctly positioned at 63 produces one unit of good XYZ.
- The production order from Alpha sets a payment of **R1.0** for each unit of good XYZ produced.

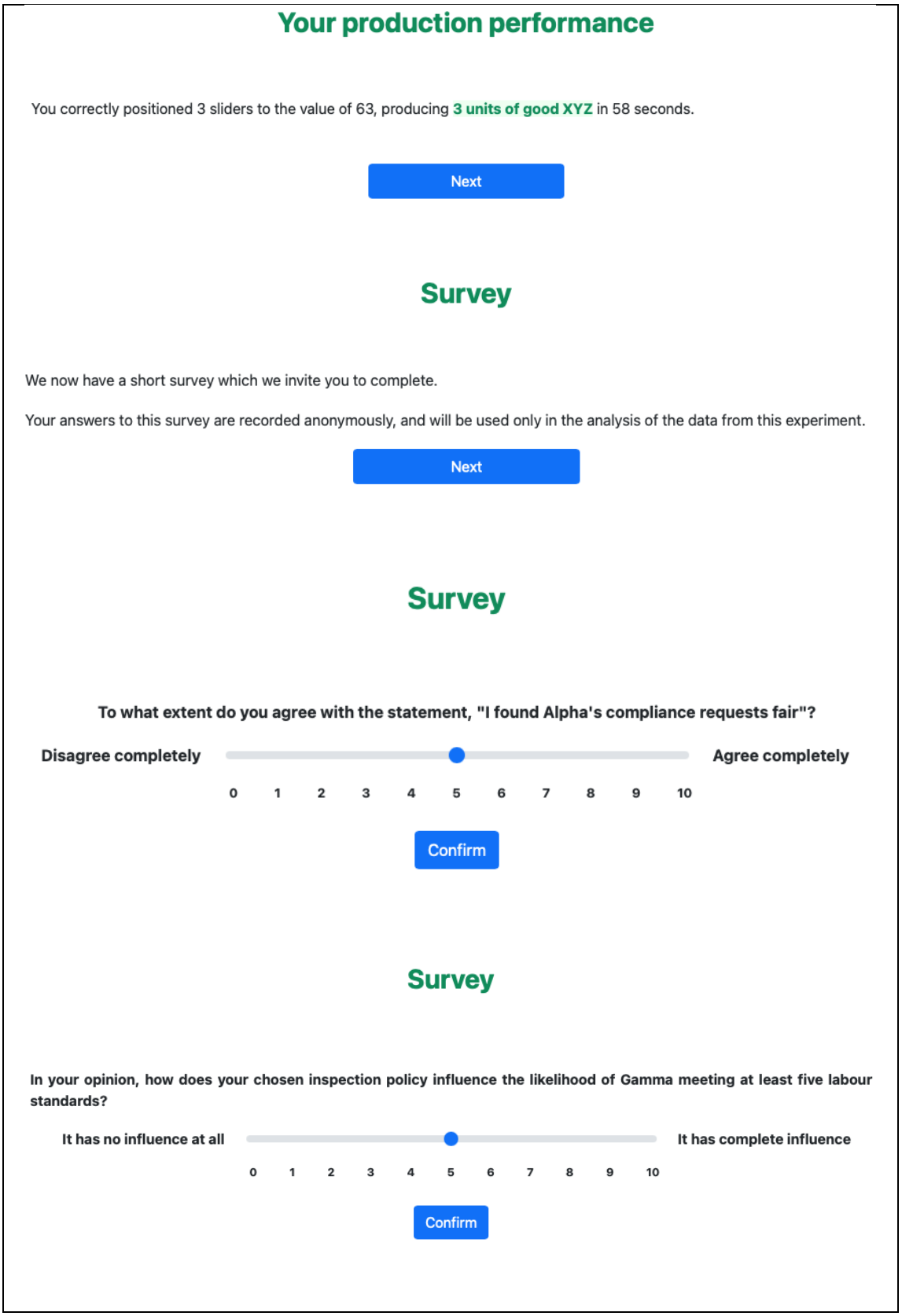
When you are ready, click on the button below. Your 58 seconds will begin on the next page, and you cannot pause the activity once it has started.

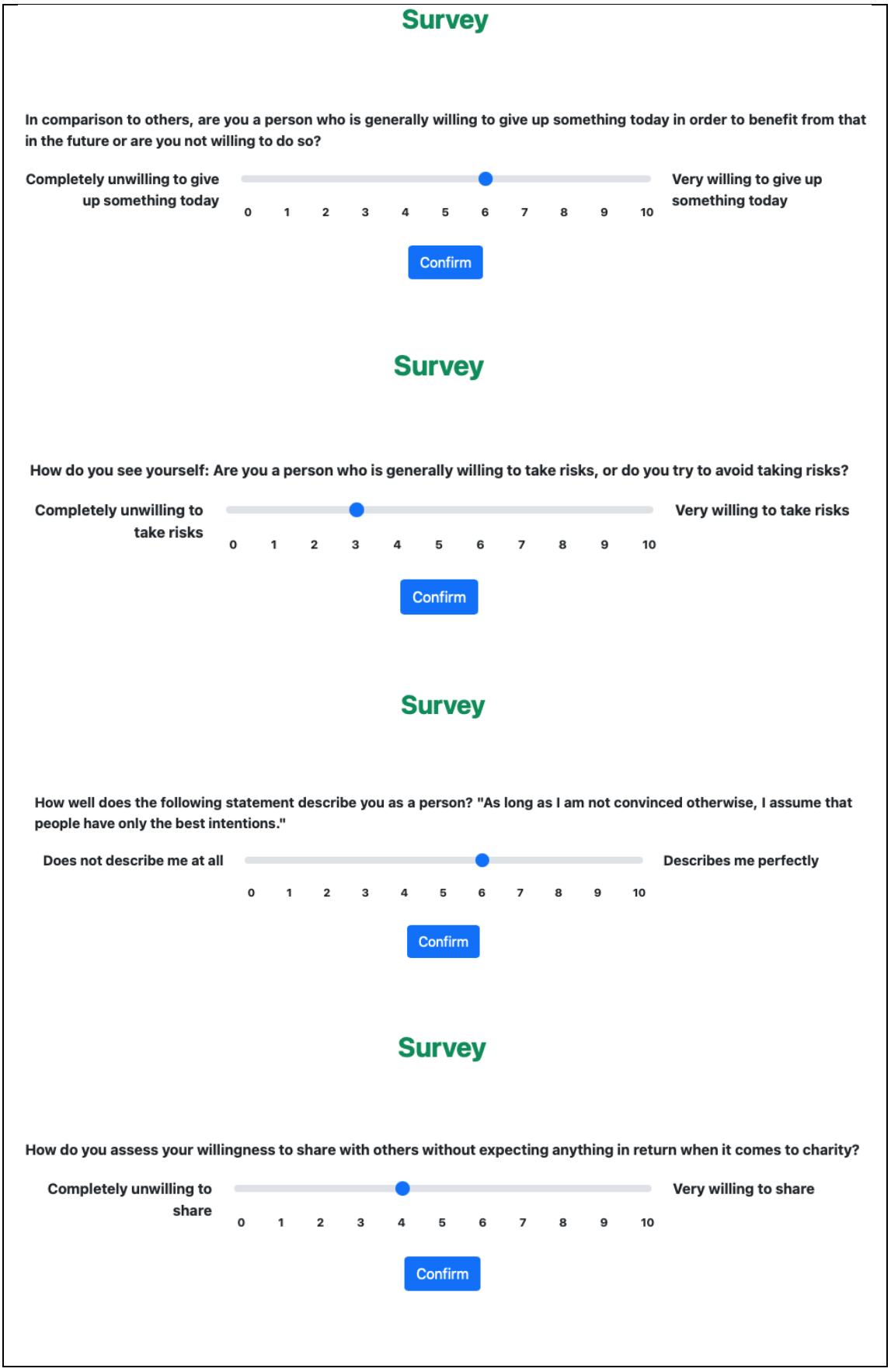
Start production

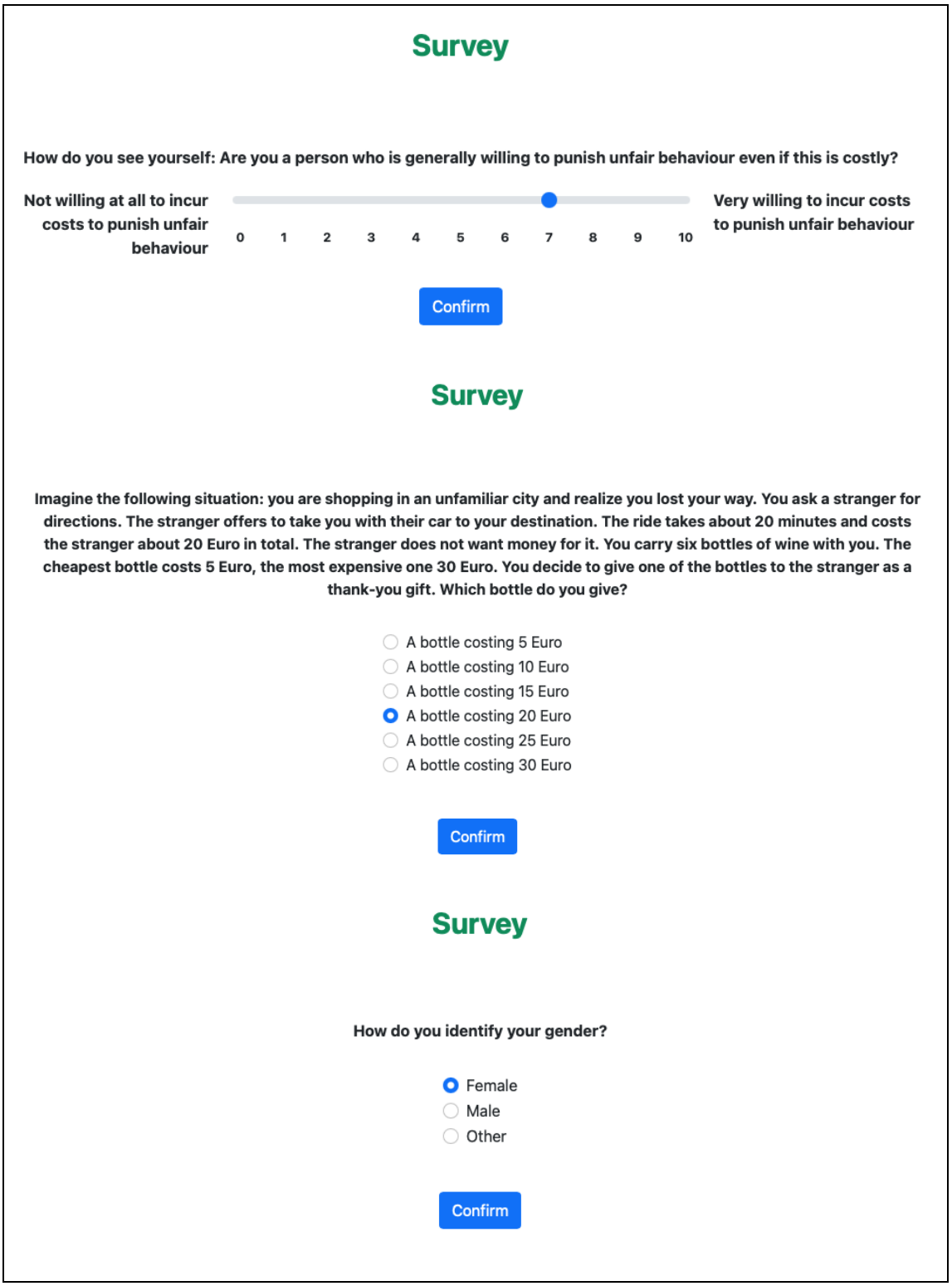
Production

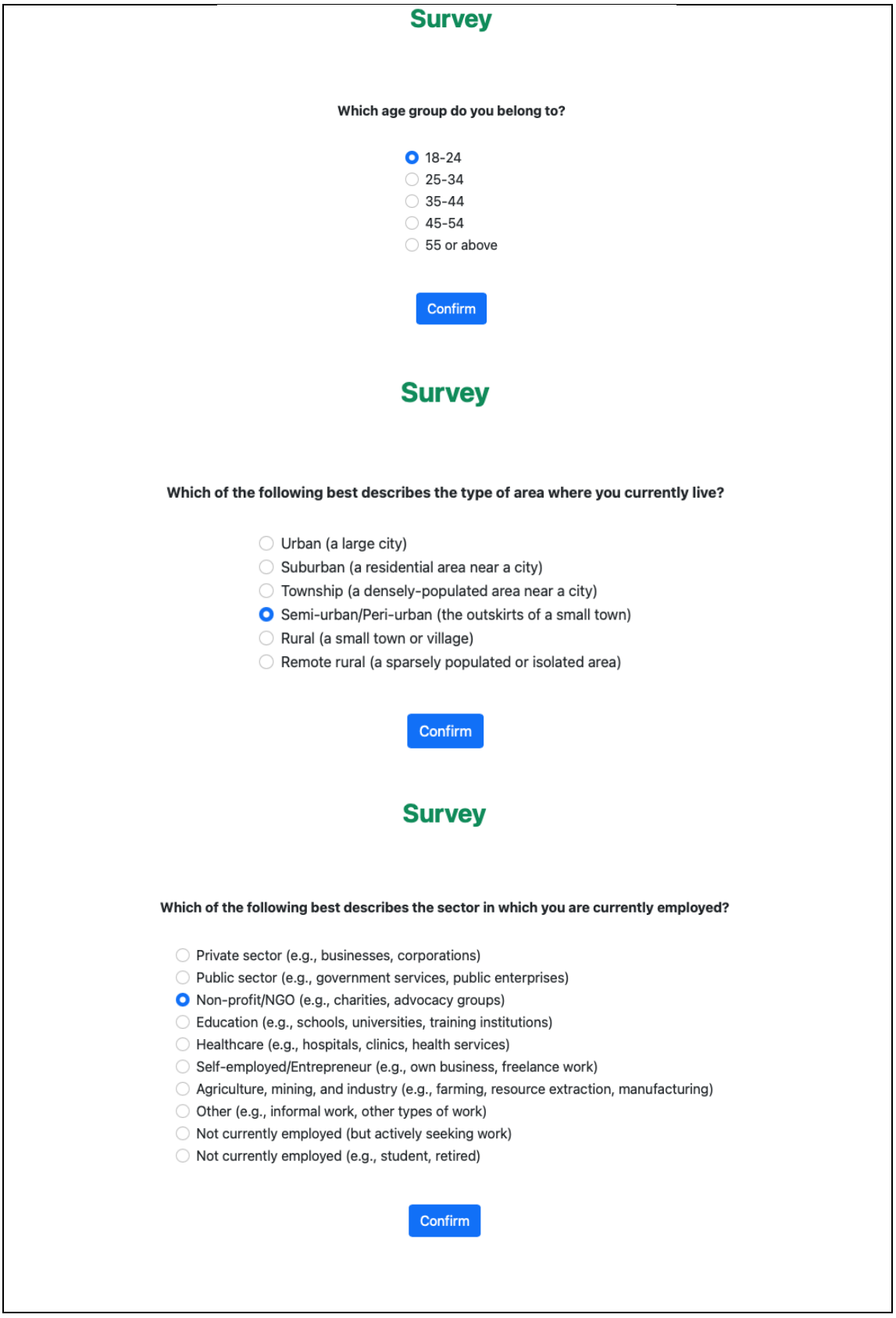
You have ⏳ **0:34** remaining to position as many sliders as you can at 63.




Your production performance
You correctly positioned 3 sliders to the value of 63, producing 3 units of good XYZ in 58 seconds.
Next
Survey
We now have a short survey which we invite you to complete.
Your answers to this survey are recorded anonymously, and will be used only in the analysis of the data from this experiment.
Next
Survey
To what extent do you agree with the statement, "I found Alpha's compliance requests fair"?
Disagree completely
Agree completely
0 1 2 3 4 5 6 7 8 9 10
Confirm
Survey
In your opinion, how does your chosen inspection policy influence the likelihood of Gamma meeting at least five labour standards?
It has no influence at all
It has complete influence
0 1 2 3 4 5 6 7 8 9 10
Confirm


Survey
In comparison to others, are you a person who is generally willing to give up something today in order to benefit from that in the future or are you not willing to do so?
Completely unwilling to give up something today
0 1 2 3 4 5 6 7 8 9 10
Very willing to give up something today
Confirm
Survey
How do you see yourself: Are you a person who is generally willing to take risks, or do you try to avoid taking risks?
Completely unwilling to take risks
0 1 2 3 4 5 6 7 8 9 10
Very willing to take risks
Confirm
Survey
How well does the following statement describe you as a person? "As long as I am not convinced otherwise, I assume that people have only the best intentions."
Does not describe me at all
0 1 2 3 4 5 6 7 8 9 10
Describes me perfectly
Confirm
Survey
How do you assess your willingness to share with others without expecting anything in return when it comes to charity?
Completely unwilling to share
0 1 2 3 4 5 6 7 8 9 10
Very willing to share
Confirm


Survey
How do you see yourself: Are you a person who is generally willing to punish unfair behaviour even if this is costly?
Not willing at all to incur costs to punish unfair behaviour
0 1 2 3 4 5 6 7 8 9 10
Very willing to incur costs to punish unfair behaviour
Confirm
Survey
Imagine the following situation: you are shopping in an unfamiliar city and realize you lost your way. You ask a stranger for directions. The stranger offers to take you with their car to your destination. The ride takes about 20 minutes and costs the stranger about 20 Euro in total. The stranger does not want money for it. You carry six bottles of wine with you. The cheapest bottle costs 5 Euro, the most expensive one 30 Euro. You decide to give one of the bottles to the stranger as a thank-you gift. Which bottle do you give?
A bottle costing 5 Euro
A bottle costing 10 Euro
A bottle costing 15 Euro
A bottle costing 20 Euro
A bottle costing 25 Euro
A bottle costing 30 Euro
Confirm
Survey
How do you identify your gender?
Female
Male
Other
Confirm


Survey
Which age group do you belong to?
18-24
25-34
35-44
45-54
55 or above
Confirm
Survey
Which of the following best describes the type of area where you currently live?
Urban (a large city)
Suburban (a residential area near a city)
Township (a densely-populated area near a city)
Semi-urban/Peri-urban (the outskirts of a small town)
Rural (a small town or village)
Remote rural (a sparsely populated or isolated area)
Confirm
Survey
Which of the following best describes the sector in which you are currently employed?
Private sector (e.g., businesses, corporations)
Public sector (e.g., government services, public enterprises)
Non-profit/NGO (e.g., charities, advocacy groups)
Education (e.g., schools, universities, training institutions)
Healthcare (e.g., hospitals, clinics, health services)
Self-employed/Entrepreneur (e.g., own business, freelance work)
Agriculture, mining, and industry (e.g., farming, resource extraction, manufacturing)
Other (e.g., informal work, other types of work)
Not currently employed (but actively seeking work)
Not currently employed (e.g., student, retired)
Confirm

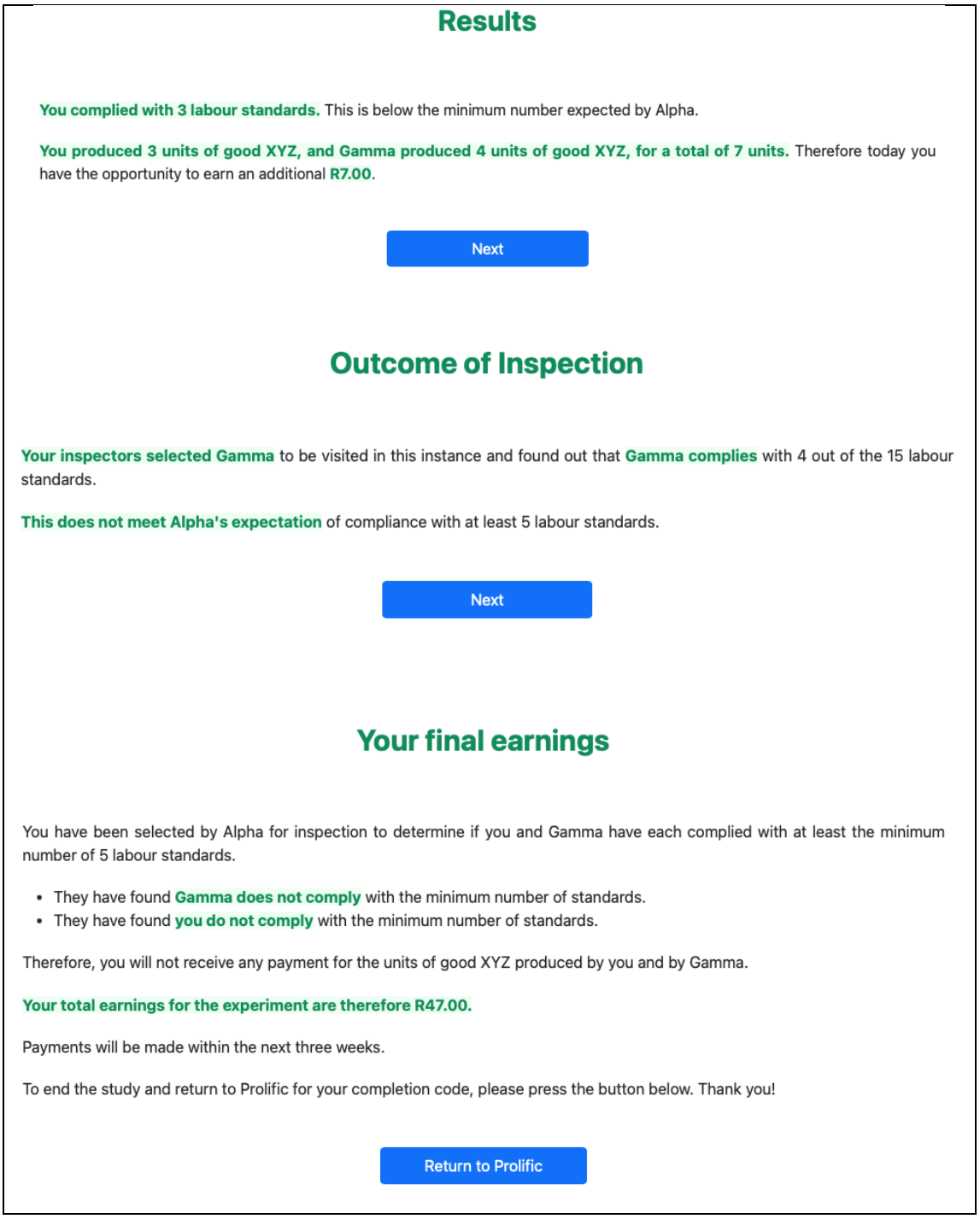

Results
You complied with 3 labour standards. This is below the minimum number expected by Alpha.
You produced 3 units of good XYZ, and Gamma produced 4 units of good XYZ, for a total of 7 units. Therefore today you have the opportunity to earn an additional R7.00.
Next
Outcome of Inspection
Your inspectors selected Gamma to be visited in this instance and found out that Gamma complies with 4 out of the 15 labour standards.
This does not meet Alpha's expectation of compliance with at least 5 labour standards.
Next
Your final earnings
You have been selected by Alpha for inspection to determine if you and Gamma have each complied with at least the minimum number of 5 labour standards.
They have found Gamma does not comply with the minimum number of standards.
They have found you do not comply with the minimum number of standards.
Therefore, you will not receive any payment for the units of good XYZ produced by you and by Gamma.
Your total earnings for the experiment are therefore R47.00.
Payments will be made within the next three weeks.
To end the study and return to Prolific for your completion code, please press the button below. Thank you!
Return to Prolific

Table 3.A.1: Beta's summary statistics

Variable	Question / Coding	Obs.	Mean	Std. Dev.	Min	Max
Outcome variables						
Inspection policy 3 or 5	Dummy: 1 if respondent selected either inspection policy 3 or 5	427	0.864	0.343	0	1
Inspection policy 1	Dummy: 1 if respondent selected inspection policy 1	427	0.136	0.343	0	1
Inspection policy 3	Dummy: 1 if respondent selected inspection policy 3	427	0.351	0.478	0	1
Inspection policy 5	Dummy: 1 if respondent selected inspection policy 5	427	0.513	0.500	0	1
Compliance seconds	Number of seconds devoted to compliance (after inspection)	427	52.717	23.510	0	125
Share of seconds devoted to compliance	Share of seconds devoted to compliance (after inspection)	427	0.546	0.242	0	1
Compliance performance	Number of words correctly transcribed	427	5.506	4.109	0	15
If complied with at least minimum standards (>=5)	Dummy: 1 if respondent correctly transcribed at least 5 words	427	0.576	0.495	0	1
If complied with minimum standards (=5)	Dummy: 1 if respondent correctly transcribed 5 words	427	0.059	0.235	0	1
Production seconds	Number of seconds devoted to production (after inspection)	427	44.742	25.250	0	125
Share of seconds devoted to production	Share of seconds devoted to production (after inspection)	427	0.454	0.242	0	1
To what extent Alpha's compliance requests are fair	Agreement with: "I found Alpha's compliance requests fair" (0–10)	427	7.002	2.607	0	10
To what extent Beta's inspection policy influences Gamma's compliance	"How much does your chosen inspection policy influence Gamma meeting minimum standards?" (0–10)	427	7.108	2.199	0	10
Control variables						
Gender	Dummy: 1 if respondent is female	427	0.597	0.491	0	1
Age	Age group (1: 18–24; 2: 25–34; 3: 35–44; 4: 45–54; 5: 55+)	427	2.499	1.110	1	5
Early-career	Dummy: 1 if younger than 35 (group 1 or 2)	427	0.630	0.483	0	1
Residence	Type of area lived in (1: Urban; 2: Suburban; 3: Township; 4: Rural; 5: Semi-urban; 6: Remote rural)	427	2.000	0.947	1	6
Urban area	Dummy: 1 if respondent lives in urban area (category 1)	427	0.321	0.467	0	1
Occupation	Sector of employment (1: Student/retired; ...; 10: Informal/other)	427	4.246	2.039	1	10
Unemployed	Dummy: 1 if unemployed (categories 1 or 2)	427	0.061	0.239	0	1
Private-oriented sector	Dummy: 1 if employed in private sector (3, 7, or 9)	427	0.576	0.495	0	1
If willing to take risks	Self-reported willingness to take risks (0–10)	427	7.761	2.294	0	10
If willing to give up something today for future benefit	Willingness to delay gratification (0–10)	427	7.197	2.686	0	10

Table 3.A.1 continued from previous page

Variable	Question / Coding	Obs.	Mean	Std. D.	Min	Max
If thinks people have best intentions	Agreement with "People have the best intentions" (0–10)	427	6.386	2.739	0	10
If willing to share with others without expecting anything	Willingness to share (0–10)	427	8.365	1.758	0	10
Wine scenario generosity	Value of wine bottle given as gift (1–6; 5–30€)	427	4.204	1.525	1	6
If willing to punish unfair behaviour even if costly	Willingness to punish unfair behaviour (0–10)	427	5.794	2.976	0	10
Compliance practice	Correct words transcribed in practice activity	427	1.445	1.172	0	5
Production practice	Correct sliders positioned in practice activity	427	1.136	1.232	0	5
Compliance quiz attempts	Attempts on compliance comprehension question	427	1.234	0.579	1	7
Production quiz attempts	Attempts on production comprehension question	427	1.522	0.823	1	8
Inspection quiz attempts	Attempts on inspection comprehension question	427	1.391	0.649	1	3

Table 3.A.2: Beta's balance test across control group (C) and treatment groups (T1, T2)

	Means			Differences		
Variable	**C**	**T1**	**T2**	**T1–C**	**T2–C**	**T1–T2**
Gender (0–1)	0.618	0.583	0.592	-0.034	-0.026	-0.009
	[0.488]	[0.495]	[0.493]	[0.059]	[0.058]	[0.058]
Age (1–5)	2.463	2.562	2.469	0.099	0.006	0.093
	[1.067]	[1.151]	[1.112]	[0.133]	[0.130]	[0.133]
Early-career (0–1)	0.654	0.583	0.653	-0.071	-0.001	-0.070
	[0.477]	[0.495]	[0.478]	[0.058]	[0.057]	[0.057]
Residence (1–6)	2.059	1.958	1.986	-0.100	-0.072	-0.028
	[0.933]	[0.876]	[1.027]	[0.108]	[0.117]	[0.112]
Urban area (0–1)	0.294	0.319	0.347	0.025	0.053	-0.027
	[0.457]	[0.468]	[0.478]	[0.055]	[0.056]	[0.055]
Occupation (1–10)	4.118	4.444	4.170	0.327	0.052	0.274
	[1.921]	[2.111]	[2.072]	[0.242]	[0.238]	[0.245]
Unemployed (0–1)	0.081	0.049	0.054	-0.032	-0.026	-0.006
	[0.274]	[0.216]	[0.228]	[0.029]	[0.030]	[0.026]
Private-oriented sector (0–1)	0.478	0.590	0.653	0.112*	0.175***	-0.063
	[0.501]	[0.493]	[0.478]	[0.059]	[0.058]	[0.057]
Willing to take risks (0–10)	7.875	7.597	7.816	-0.278	-0.059	-0.219
	[2.202]	[2.387]	[2.291]	[0.275]	[0.268]	[0.274]
Very willing to take risks (0–1)	0.471	0.444	0.463	-0.026	-0.008	-0.018
	[0.501]	[0.499]	[0.500]	[0.060]	[0.060]	[0.059]
Willing to delay gratification (0–10)	6.853	7.306	7.408	0.453	0.555*	-0.103
	[2.832]	[2.666]	[2.550]	[0.329]	[0.320]	[0.306]
Very willing to delay gratification (0–1)	0.353	0.389	0.381	0.036	0.028	0.008
	[0.480]	[0.489]	[0.487]	[0.058]	[0.058]	[0.057]
Optimistic about others' intentions (0–10)	6.412	6.340	6.408	-0.071	-0.004	-0.068
	[2.763]	[2.705]	[2.767]	[0.327]	[0.329]	[0.321]
Very optimistic (0–1)	0.368	0.396	0.442	0.028	0.075	-0.046
	[0.484]	[0.491]	[0.498]	[0.058]	[0.058]	[0.058]
Willing to share (0–10)	8.272	8.319	8.497	0.047	0.225	-0.177
	[1.868]	[1.823]	[1.585]	[0.221]	[0.205]	[0.200]
Very willing to share (0–1)	0.360	0.354	0.381	-0.006	0.021	-0.027
	[0.482]	[0.480]	[0.487]	[0.057]	[0.058]	[0.057]
Wine scenario generosity (1–5)	4.103	4.181	4.320	0.078	0.217	-0.139
	[1.625]	[1.508]	[1.448]	[0.187]	[0.183]	[0.173]
Very generous (0–1)	0.412	0.417	0.469	0.005	0.058	-0.053
	[0.494]	[0.495]	[0.501]	[0.059]	[0.059]	[0.058]
Willing to punish unfair behaviour (0–10)	5.824	5.785	5.776	-0.039	-0.048	0.009
	[3.008]	[3.048]	[2.895]	[0.362]	[0.351]	[0.348]
Very willing to punish unfair behaviour (0–1)	0.485	0.444	0.429	-0.041	-0.057	0.016
	[0.502]	[0.499]	[0.497]	[0.060]	[0.059]	[0.058]
Compliance quiz attempts (1–3)	1.272	1.250	1.184	-0.022	-0.088	0.066
	[0.537]	[0.548]	[0.641]	[0.065]	[0.071]	[0.070]
Production quiz attempts (1–3)	1.522	1.632	1.415	0.110	-0.107	0.217**
	[0.750]	[0.817]	[0.882]	[0.094]	[0.098]	[0.100]
Inspection quiz attempts (1–3)	1.412	1.465	1.299	0.054	-0.112	0.166**
	[0.683]	[0.678]	[0.578]	[0.081]	[0.075]	[0.074]
Observations	136	144	147	280	283	291

Note: The table reports means and pairwise differences in baseline covariates across C, T1, and T2. Standard deviations are reported under group means and standard errors are reported under differences. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 3.A.3: Treatment effects on Beta's monitoring efforts

Variable	**Pr(Low policy)**	**Pr(Medium policy)**	**Pr(High policy)**
Treatment 1	0.066**	0.055**	-0.121**
	[0.029]	[0.025]	[0.053]
Treatment 2	0.032	0.032	-0.064
	[0.028]	[0.028]	[0.055]
Gender	-0.002	-0.002	0.005
	[0.025]	[0.020]	[0.045]
Private-oriented sector	-0.019	-0.015	0.034
	[0.027]	[0.022]	[0.048]
Observations	427	427	427

Note: This table reports Ordered Probit estimates of the treatment effects on Beta's monitoring, compliance, and production effort. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.4: Beta's heterogeneity analysis

Variable	Inspection policy 3 or 5	Inspection policy 3	Inspection policy 5	Compliance seconds (share)	Min. compliance	Prod. perform.
Gender (female)						
T1	0.024	0.065	-0.041	-0.015	-0.162**	0.799
	[0.063]	[0.092]	[0.096]	[0.046]	[0.082]	[0.676]
T2	-0.043	0.065	-0.108	-0.027	-0.220**	-0.001
	[0.070]	[0.093]	[0.095]	[0.045]	[0.086]	[0.656]
Female	0.033	-0.004	0.037	0.033	-0.091	-0.840
	[0.058]	[0.085]	[0.089]	[0.042]	[0.080]	[0.520]
T1 × Female	-0.139*	0.001	-0.140	0.008	0.125	-0.420
	[0.082]	[0.118]	[0.123]	[0.059]	[0.112]	[0.804]
T2 × Female	0.007	-0.085	0.091	-0.011	0.122	0.013
	[0.085]	[0.119]	[0.122]	[0.058]	[0.116]	[0.794]
Observations	427	427	427	427	427	427
R-squared	0.045	0.048	0.087	0.147	0.161	0.553
Control mean	0.885	0.346	0.538	0.541	0.712	5.346
Coeff: T1 + T1×Female = 0	-0.115	0.0658	-0.181	-0.00661	-0.037	0.379
P-value: T1 + T1×Female = 0	0.0314	0.380	0.0174	0.855	0.623	0.359
Coeff: T2 + T2×Female = 0	-0.0366	-0.0201	-0.0165	-0.0382	-0.0984	0.0129
P-value: T2 + T2×Female = 0	0.443	0.783	0.831	0.272	0.209	0.976
Age (early-career)						
T1	-0.079	-0.088	0.010	-0.008	-0.036	0.745
	[0.075]	[0.092]	[0.096]	[0.045]	[0.092]	[0.527]
T2	0.006	0.004	0.002	-0.044	-0.094	-0.283
	[0.075]	[0.100]	[0.102]	[0.049]	[0.100]	[0.618]
Early-career	0.058	-0.060	0.117	0.028	0.143	0.610
	[0.060]	[0.087]	[0.090]	[0.045]	[0.089]	[0.508]
T1 × Early-career	0.035	0.249**	-0.214*	-0.005	-0.078	-0.338
	[0.087]	[0.118]	[0.122]	[0.059]	[0.116]	[0.736]
T2 × Early-career	-0.074	0.014	-0.088	0.015	-0.078	0.437
	[0.089]	[0.122]	[0.126]	[0.061]	[0.124]	[0.762]
Observations	427	427	427	427	427	427
R-squared	0.040	0.060	0.085	0.147	0.159	0.554
Control mean	0.872	0.362	0.511	0.549	0.574	4.191
Coeff: T1 + T1×Early-career = 0	-0.0438	0.160	-0.204	-0.0122	-0.114	0.407
P-value: T1 + T1×Early-career = 0	0.350	0.032	0.00714	0.742	0.102	0.413
Coeff: T2 + T2×Early-career = 0	-0.0678	0.0177	-0.0855	-0.0282	-0.172	0.154
P-value: T2 + T2×Early-career = 0	0.155	0.800	0.247	0.403	0.0189	0.726
Residence (urban area)						
T1	-0.042	0.109	-0.151**	-0.015	-0.122*	0.581
	[0.048]	[0.072]	[0.072]	[0.033]	[0.065]	[0.465]
T2	-0.056	-0.016	-0.040	-0.022	-0.195***	0.041
	[0.048]	[0.069]	[0.072]	[0.033]	[0.068]	[0.437]
Urban area	-0.033	-0.051	0.018	0.018	-0.093	-0.344
	[0.064]	[0.089]	[0.093]	[0.044]	[0.100]	[0.556]
T1 × Urban area	-0.054	-0.138	0.084	0.015	0.120	-0.105
	[0.095]	[0.122]	[0.129]	[0.064]	[0.127]	[0.773]
T2 × Urban area	0.045	0.095	-0.050	-0.035	0.158	-0.130
	[0.087]	[0.123]	[0.128]	[0.057]	[0.131]	[0.781]
Observations	427	427	427	427	427	427
R-squared	0.039	0.055	0.081	0.148	0.161	0.553
Control mean	0.906	0.344	0.562	0.549	0.656	4.615
Coeff: T1 + T1×Urban = 0	-0.0953	-0.0291	-0.0662	-0.000196	-0.00213	0.476
P-value: T1 + T1×Urban = 0	0.236	0.766	0.538	0.997	0.984	0.425
Coeff: T2 + T2×Urban = 0	-0.0115	0.0784	-0.0899	-0.0577	-0.0367	-0.0892
P-value: T2 + T2×Urban = 0	0.875	0.443	0.401	0.227	0.744	0.889
Occupation (private sector)						
T1	-0.110	-0.064	-0.046	-0.043	-0.256***	0.277
	[0.068]	[0.084]	[0.087]	[0.044]	[0.083]	[0.504]
T2	0.036	-0.093	0.129	-0.049	-0.208**	-0.023
	[0.059]	[0.087]	[0.090]	[0.043]	[0.090]	[0.528]
Private sector	0.064	-0.103	0.167*	-0.095**	-0.055	0.718
	[0.053]	[0.082]	[0.085]	[0.041]	[0.086]	[0.507]
T1 × Private sector	0.083	0.245**	-0.161	0.061	0.315***	0.496
	[0.084]	[0.114]	[0.119]	[0.057]	[0.111]	[0.737]
T2 × Private sector	-0.122	0.203*	-0.325***	0.032	0.134	0.084
	[0.078]	[0.115]	[0.119]	[0.055]	[0.118]	[0.727]
Observations	427	427	427	427	427	427
R-squared	0.050	0.058	0.094	0.149	0.174	0.553
Control mean	0.873	0.366	0.507	0.616	0.690	3.423
Coeff: T1 + T1×Private = 0	-0.0262	0.181	-0.207	0.0183	0.0596	0.773
P-value: T1 + T1×Private = 0	0.596	0.0216	0.0116	0.616	0.420	0.142
Coeff: T2 + T2×Private = 0	-0.0863	0.109	-0.196	-0.0174	-0.0739	0.0608
P-value: T2 + T2×Private = 0	0.0996	0.149	0.0138	0.614	0.335	0.901

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Beta's monitoring, compliance, and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance: *** p<0.01, ** p<0.05, * p<0.10.

Table 3.A.4 (continued): Beta's heterogeneity analysis

Variable	Inspection policy 3 or 5	Inspection policy 3	Inspection policy 5	Compliance seconds (share)	Min. compliance	Prod. perform.
Risk tolerance (very willing to take risks)						
T1	-0.077	0.021	-0.098	-0.058*	-0.106	0.922*
	[0.055]	[0.082]	[0.082]	[0.034]	[0.072]	[0.524]
T2	-0.052	-0.026	-0.025	-0.035	-0.160**	0.331
	[0.051]	[0.083]	[0.083]	[0.031]	[0.077]	[0.474]
Risk tolerance	-0.142*	-0.268***	0.126	-0.004	-0.180*	0.880
	[0.076]	[0.099]	[0.102]	[0.047]	[0.100]	[0.637]
T1 × Risk tolerance	0.046	0.105	-0.059	0.103*	0.050	-0.825
	[0.087]	[0.113]	[0.120]	[0.060]	[0.113]	[0.724]
T2 × Risk tolerance	0.025	0.093	-0.069	0.002	0.037	-0.725
	[0.082]	[0.112]	[0.118]	[0.056]	[0.115]	[0.727]
Observations	427	427	427	427	427	427
R-squared	0.048	0.066	0.082	0.157	0.167	0.555
Control mean	0.917	0.431	0.486	0.545	0.708	4.653
Coeff: T1 + T1×Risk = 0	-0.0316	0.125	-0.157	0.0453	-0.0557	0.0972
P-value: T1 + T1×Risk = 0	0.623	0.113	0.0753	0.343	0.518	0.845
Coeff: T2 + T2×Risk = 0	-0.0271	0.067	-0.0941	-0.0327	-0.124	-0.394
P-value: T2 + T2×Risk = 0	0.667	0.375	0.274	0.482	0.156	0.472
Patience (very willing to delay gratification)						
T1	-0.143***	0.050	-0.193***	-0.019	-0.098	1.292***
	[0.050]	[0.073]	[0.072]	[0.035]	[0.070]	[0.475]
T2	-0.074*	0.027	-0.101	-0.036	-0.170**	0.298
	[0.044]	[0.074]	[0.075]	[0.033]	[0.073]	[0.427]
Patience	-0.177**	-0.084	-0.093	-0.039	-0.040	0.990
	[0.074]	[0.102]	[0.105]	[0.047]	[0.103]	[0.612]
T1 × Patience	0.221**	0.031	0.189	0.022	0.031	-1.986***
	[0.086]	[0.120]	[0.125]	[0.059]	[0.119]	[0.740]
T2 × Patience	0.079	-0.045	0.124	0.002	0.064	-0.803
	[0.089]	[0.119]	[0.126]	[0.056]	[0.122]	[0.794]
Observations	427	427	427	427	427	427
R-squared	0.056	0.052	0.084	0.149	0.158	0.560
Control mean	0.932	0.352	0.580	0.565	0.625	3.852
Coeff: T1 + T1×Patience = 0	0.0776	0.0814	-0.0038	0.00304	-0.0662	-0.694
P-value: T1 + T1×Patience = 0	0.270	0.393	0.971	0.949	0.485	0.209
Coeff: T2 + T2×Patience = 0	0.0045	-0.0181	0.0226	-0.0339	-0.106	-0.505
P-value: T2 + T2×Patience = 0	0.953	0.845	0.823	0.454	0.283	0.441
Trust (very optimistic about others' intentions)						
T1	-0.069	0.090	-0.158**	-0.030	-0.117*	0.256
	[0.054]	[0.075]	[0.075]	[0.033]	[0.067]	[0.474]
T2	-0.052	-0.019	-0.033	-0.017	-0.180**	-0.175
	[0.052]	[0.074]	[0.077]	[0.033]	[0.074]	[0.436]
Trust	-0.032	0.069	-0.101	0.029	-0.099	-0.406
	[0.074]	[0.110]	[0.113]	[0.052]	[0.112]	[0.719]
T1 × Trust	0.028	-0.071	0.099	0.047	0.089	0.784
	[0.084]	[0.118]	[0.123]	[0.062]	[0.119]	[0.783]
T2 × Trust	0.028	0.066	-0.038	-0.041	0.096	0.474
	[0.084]	[0.118]	[0.123]	[0.057]	[0.120]	[0.759]
Observations	427	427	427	427	427	427
R-squared	0.036	0.052	0.083	0.153	0.160	0.554
Control mean	0.895	0.349	0.547	0.528	0.663	5.012
Coeff: T1 + T1×Trust = 0	-0.0409	0.0184	-0.0593	0.0172	-0.028	1.040
P-value: T1 + T1×Trust = 0	0.527	0.841	0.548	0.743	0.774	0.0839
Coeff: T2 + T2×Trust = 0	-0.0238	0.0465	-0.0703	-0.0585	-0.0838	0.300
P-value: T2 + T2×Trust = 0	0.716	0.611	0.468	0.213	0.386	0.632

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Beta's monitoring, compliance, and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance: *** p<0.01, ** p<0.05, * p<0.10.

Table 3.A.4 (continued): Beta's heterogeneity analysis

Variable	Inspection policy 3 or 5	Inspection policy 3	Inspection policy 5	Compliance seconds (share)	Min. compliance	Prod. perform.
Prosociality (very willing to share with others)						
T1	-0.055	0.025	-0.080	0.023	-0.088	1.009**
	[0.052]	[0.074]	[0.075]	[0.032]	[0.069]	[0.505]
T2	-0.023	-0.050	0.027	0.009	-0.121	0.413
	[0.051]	[0.075]	[0.077]	[0.030]	[0.075]	[0.464]
Prosociality	-0.007	-0.170*	0.163	0.105**	-0.067	0.568
	[0.068]	[0.097]	[0.102]	[0.051]	[0.104]	[0.586]
T1 × Prosociality	-0.012	0.106	-0.117	-0.090	0.004	-1.300*
	[0.085]	[0.118]	[0.124]	[0.063]	[0.117]	[0.718]
T2 × Prosociality	-0.051	0.168	-0.220*	-0.113*	-0.067	-1.137
	[0.082]	[0.116]	[0.120]	[0.059]	[0.117]	[0.722]
Observations	427	427	427	427	427	427
R-squared	0.038	0.054	0.086	0.157	0.162	0.557
Control mean	0.885	0.391	0.494	0.517	0.644	4.839
Coeff: T1 + T1×Prosociality = 0	-0.0668	0.131	-0.197	-0.0675	-0.0844	-0.291
P-value: T1 + T1×Prosociality = 0	0.318	0.157	0.0452	0.213	0.370	0.549
Coeff: T2 + T2×Prosociality = 0	-0.0741	0.118	-0.192	-0.104	-0.187	-0.724
P-value: T2 + T2×Prosociality = 0	0.256	0.181	0.0412	0.0415	0.0416	0.191
Inequity aversion (very willing to punish unfair behaviour)						
T1	-0.056	0.034	-0.090	-0.028	-0.055	1.253**
	[0.054]	[0.079]	[0.082]	[0.038]	[0.077]	[0.508]
T2	-0.100*	-0.059	-0.040	-0.046	-0.113	0.485
	[0.056]	[0.077]	[0.081]	[0.033]	[0.080]	[0.489]
Inequity aversion	-0.078	-0.229**	0.152	0.036	0.053	0.684
	[0.073]	[0.109]	[0.112]	[0.050]	[0.108]	[0.701]
T1 × Inequity aversion	-0.013	0.049	-0.062	0.044	-0.065	-1.499**
	[0.084]	[0.114]	[0.118]	[0.057]	[0.113]	[0.754]
T2 × Inequity aversion	0.127	0.144	-0.017	0.034	-0.069	-1.035
	[0.080]	[0.114]	[0.118]	[0.054]	[0.117]	[0.738]
Observations	427	427	427	427	427	427
R-squared	0.045	0.059	0.084	0.153	0.159	0.557
Control mean	0.914	0.357	0.557	0.554	0.586	3.957
Coeff: T1 + T1×Inequity = 0	-0.0687	0.0828	-0.152	0.0152	-0.120	-0.246
P-value: T1 + T1×Inequity = 0	0.284	0.322	0.0796	0.720	0.144	0.644
Coeff: T2 + T2×Inequity = 0	0.0274	0.0846	-0.0572	-0.0122	-0.182	-0.550
P-value: T2 + T2×Inequity = 0	0.628	0.320	0.519	0.779	0.0358	0.310
Beliefs − Fairness						
T1	-0.071	0.019	-0.090	-0.035	-0.048	1.192***
	[0.051]	[0.071]	[0.073]	[0.034]	[0.069]	[0.438]
T2	-0.067	-0.039	-0.029	-0.034	-0.176**	0.571
	[0.052]	[0.070]	[0.074]	[0.031]	[0.073]	[0.403]
Perceived fairness	-0.014	-0.033	0.020	0.023	0.175**	1.879***
	[0.059]	[0.085]	[0.089]	[0.044]	[0.083]	[0.538]
T1 × Fairness	0.038	0.148	-0.109	0.089	-0.076	-1.705**
	[0.086]	[0.122]	[0.127]	[0.062]	[0.117]	[0.775]
T2 × Fairness	0.088	0.184	-0.096	0.014	0.176	-1.390
	[0.086]	[0.123]	[0.128]	[0.060]	[0.110]	[0.846]
Observations	427	427	427	427	427	427
R-squared	0.039	0.057	0.082	0.162	0.200	0.565
Control mean	0.895	0.349	0.547	0.549	0.581	3.907
Coeff: T1 + T1×Fairness = 0	-0.0326	0.167	-0.199	0.0542	-0.123	-0.514
P-value: T1 + T1×Fairness = 0	0.640	0.0982	0.0572	0.288	0.191	0.412
Coeff: T2 + T2×Fairness = 0	0.0208	0.145	-0.124	-0.0197	-0.536	-0.819
P-value: T2 + T2×Fairness = 0	0.754	0.153	0.237	0.708	1.000	0.260
Beliefs − Effectiveness						
T1	-0.060	0.107	-0.166*	-0.046	0.019	1.342**
	[0.060]	[0.085]	[0.086]	[0.040]	[0.082]	[0.575]
T2	-0.034	-0.018	-0.016	-0.041	-0.049	0.847*
	[0.057]	[0.083]	[0.086]	[0.037]	[0.085]	[0.503]
Perceived effectiveness	0.005	-0.058	0.063	0.021	0.279***	1.779***
	[0.055]	[0.085]	[0.088]	[0.038]	[0.083]	[0.509]
T1 × Effectiveness	0.002	-0.109	0.111	0.085	-0.165	-1.334*
	[0.085]	[0.118]	[0.122]	[0.055]	[0.112]	[0.762]
T2 × Effectiveness	-0.016	0.055	-0.071	0.023	-0.144	-1.442**
	[0.079]	[0.116]	[0.120]	[0.053]	[0.114]	[0.720]
Observations	427	427	427	427	427	427
R-squared	0.036	0.057	0.089	0.164	0.189	0.565
Control mean	0.887	0.355	0.532	0.556	0.500	3.516
Coeff: T1 + T1×Effectiveness = 0	-0.0573	-0.00243	-0.0549	0.0391	-0.146	0.00773
P-value: T1 + T1×Effectiveness = 0	0.332	0.976	0.521	0.308	0.0547	0.987
Coeff: T2 + T2×Effectiveness = 0	-0.0498	0.0366	-0.0864	-0.0181	-0.193	-0.595
P-value: T2 + T2×Effectiveness = 0	0.376	0.654	0.306	0.644	0.0124	0.251

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Beta's monitoring, compliance, and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance: *** p<0.01, ** p<0.05, * p<0.10.

Table 3.A.5: Gamma's summary statistics

Variable	Question / Coding	Obs.	Mean	Std. Dev.	Min	Max
Outcome variables						
Inspection policy 3 or 5	Dummy: 1 if respondent was assigned to inspection policy 3 or 5	126	0.643	0.481	0	1
Inspection policy 1	Dummy: 1 if respondent was assigned to inspection policy 1	126	0.357	0.481	0	1
Inspection policy 3	Dummy: 1 if respondent was assigned to inspection policy 3	126	0.310	0.464	0	1
Inspection policy 5	Dummy: 1 if respondent was assigned to inspection policy 5	126	0.333	0.473	0	1
Compliance seconds	Number of seconds devoted to compliance (after inspection)	126	69.333	32.148	0	120
Share of seconds devoted to compliance	Share of seconds devoted to compliance (after inspection)	126	0.578	0.268	0	1
Compliance performance	Number of words correctly transcribed	126	6.048	5.146	0	15
If complied with at least minimum standards (5)	Dummy: 1 if correctly transcribed at least 5 words	126	0.556	0.499	0	1
If complied with minimum standards (=5)	Dummy: 1 if correctly transcribed exactly 5 words	126	0.048	0.214	0	1
Production seconds	Number of seconds devoted to production (after inspection)	126	50.667	32.148	0	120
Share of seconds devoted to production	Share of seconds devoted to production (after inspection)	126	0.422	0.268	0	1
Production performance	Number of sliders correctly positioned	126	4.849	4.731	0	23
To what extent Alpha's compliance requests are fair	Agreement with: "I found Alpha's compliance requests fair" (0–10)	126	6.960	2.872	0	10
Control variables						
Gender	Dummy: 1 if respondent is female	126	0.762	0.428	0	1
Age	Age group (1: 18–24; 2: 25–34; 3: 35–44; 4: 45–54; 5: 55+)	126	2.508	1.282	1	5
Early-career	Dummy: 1 if respondent is younger than 35 (categories 1,2)	126	0.635	0.483	0	1
Residence	Type of area lived in (1: Urban; 2: Suburban; 3: Township; 4: Rural; 5: Semi-urban; 6: Remote rural)	126	2.000	1.004	1	5
Urban area	Dummy: 1 if respondent lives in urban area (category 1)	126	0.317	0.467	0	1
Occupation	Sector of employment (1: Student/retired; ...; 10: Informal/other)	126	4.103	2.093	1	10
Unemployed	Dummy: 1 if respondent is unemployed (categories 1,2)	126	0.095	0.295	0	1
Private-oriented sector	Dummy: 1 if respondent works in private sector (categories 3,7,9)	126	0.540	0.500	0	1
If willing to take risks	Self-reported willingness to take risks (0–10)	126	7.802	2.234	0	10
If willing to give up something today for future benefit	Willingness to delay gratification (0–10)	126	7.143	2.785	0	10
If thinks people have always the best intentions	Agreement with: "People have the best intentions" (0–10)	126	6.421	2.303	1	10

Table 3.A.5 continued from previous page

Variable	Question / Coding	Obs.	Mean	Std. Dev.	Min	Max
If willing to share with others without expecting anything in return	Willingness to share (0–10)	126	8.619	1.893	0	10
Wine scenario generosity	Value of wine bottle selected as gift (1–6; 5–30€)	126	4.286	1.649	1	6
If willing to punish unfair behaviour even if costly	Willingness to punish unfair behaviour (0–10)	126	6.079	2.697	0	10
Compliance practice	Correct words transcribed during practice activity	126	1.262	1.147	0	5
Production practice	Correct sliders positioned during practice activity	126	0.960	1.148	0	5
Compliance quiz attempts	Attempts on compliance comprehension question	126	1.254	0.455	1	3
Production quiz attempts	Attempts on production comprehension question	126	1.238	0.625	1	6

Table 3.A.6: Gamma's balance test across control group (C) and treatment group (T1)

	Means		Difference
Variable	**C**	**T1**	**T1–C**
Gender (0–1)	0.824	0.690	-0.134*
	[0.384]	[0.467]	[0.076]
Age (1–5)	2.529	2.483	-0.047
	[1.240]	[1.341]	[0.230]
Early-career (0–1)	0.618	0.655	0.038
	[0.490]	[0.479]	[0.087]
Residence (1–6)	1.926	2.086	0.160
	[0.919]	[1.097]	[0.180]
Urban area (0–1)	0.338	0.293	-0.045
	[0.477]	[0.459]	[0.084]
Occupation (1–10)	4.132	4.069	-0.063
	[2.065]	[2.143]	[0.376]
Unemployed (0–1)	0.059	0.138	0.079
	[0.237]	[0.348]	[0.052]
Private-oriented sector (0–1)	0.603	0.466	-0.137
	[0.493]	[0.503]	[0.089]
Willing to take risks (0–10)	7.853	7.741	-0.112
	[2.371]	[2.082]	[0.401]
Very willing to take risks (0–1)	0.515	0.414	-0.101
	[0.503]	[0.497]	[0.089]
Willing to give up something today for future benefit (0–10)	7.221	7.052	-0.169
	[2.911]	[2.652]	[0.500]
Very willing to delay gratification (0–1)	0.426	0.362	-0.064
	[0.498]	[0.485]	[0.088]
Thinks people have best intentions (0–10)	6.544	6.276	-0.268
	[2.256]	[2.368]	[0.413]
Very optimistic about others' intentions (0–1)	0.485	0.500	0.015
	[0.503]	[0.504]	[0.090]
Willing to share without expecting return (0–10)	8.603	8.638	0.035
	[1.994]	[1.784]	[0.340]
Very willing to share with others (0–1)	0.559	0.414	-0.145
	[0.500]	[0.497]	[0.089]
Wine scenario generosity (1–5)	4.309	4.259	-0.050
	[1.499]	[1.822]	[0.296]
Very generous (0–1)	0.397	0.483	0.086
	[0.493]	[0.504]	[0.089]
Willing to punish unfair behaviour (0–10)	6.324	5.793	-0.530
	[3.073]	[2.166]	[0.482]
Very willing to punish unfair behaviour (0–1)	0.588	0.379	-0.209**
	[0.496]	[0.489]	[0.088]
Compliance quiz attempts (1–3)	1.294	1.207	-0.087
	[0.490]	[0.409]	[0.081]
Production quiz attempts (1–3)	1.250	1.224	-0.026
	[0.720]	[0.497]	[0.112]
Observations	68	58	126

Note: The table reports means and pairwise differences in baseline covariates across C, and T1. Standard deviations are reported under group means and standard errors are reported under differences. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 3.A.7: Treatment effects on Gamma's compliance and production efforts

Variable	**Compl. seconds (share)**	**Compl. perform.**	**If complied with at least min. stand.**	**Prod. perform.**
Treatment 1	-0.028	-0.192	-0.003	-0.163
	[0.045]	[0.852]	[0.090]	[0.722]
Compliance seconds		0.049***	0.003**	
		[0.014]	[0.001]	
Gender	0.022	-0.155	-0.024	-1.453
	[0.054]	[1.049]	[0.104]	[1.036]
Private-oriented sector	-0.168***	-0.129	0.028	1.853**
	[0.056]	[0.977]	[0.100]	[0.767]
Observations	126	126	126	126
R-squared	0.171	0.341	0.284	0.522
Control mean	0.583	6.412	0.588	4.882
Clust. P-value	0.544	0.822	0.976	0.821
Boot. Std. Error	0.0454	0.852	0.0897	0.722
Boot. P-value	0.538	0.823	0.976	0.825

Note: This table reports OLS estimates of the treatment effects on Gamma's compliance and production effort. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.8: Treatment effects on Gamma's secondary outcomes (mechanisms)

Variable	**Fairness**
Treatment 1	-0.041
	[0.082]
Gender	-0.057
	[0.089]
Private-oriented sector	0.038
	[0.086]
Observations	126
R-squared	0.268
Control mean	0.397
Clust. P-value	0.622
Boot. Std. Error	0.0823
Boot. P-value	0.618

Note: This table reports OLS estimates of the treatment effects on Gamma's fairness beliefs. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.9: Gamma's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	Min. compliance	Production performance
Gender (female)				
T1	-0.096	0.176	-0.056	-2.212
	[0.091]	[2.005]	[0.186]	[2.056]
Gender	-0.026	0.104	-0.062	-2.893
	[0.074]	[1.721]	[0.157]	[1.914]
T1 × Gender	0.088	-0.473	0.068	2.633
	[0.114]	[2.197]	[0.207]	[2.132]
Observations	126	126	126	126
R-squared	0.175	0.342	0.285	0.534
Control mean	0.562	5.583	0.583	7.833
Coeff: T1 + T1×Gender = 0	-0.00799	-0.297	0.0125	0.421
P-value (clust.)	0.888	0.750	0.901	0.537
P-value (boot.)	0.560	0.948	0.948	0.474
Age (early-career)				
T1	0.009	-0.027	0.070	-0.892
	[0.092]	[1.647]	[0.164]	[1.393]
Early-career	0.066	0.571	0.156	-0.077
	[0.074]	[1.330]	[0.126]	[1.273]
T1 × Early-career	-0.056	-0.254	-0.112	1.123
	[0.108]	[1.868]	[0.187]	[1.564]
Observations	126	126	126	126
R-squared	0.174	0.341	0.287	0.525
Control mean	0.555	6.538	0.538	5.462
Coeff: T1 + T1×Age = 0	-0.0472	-0.281	-0.042	0.231
P-value (clust.)	0.368	0.766	0.675	0.770
P-value (boot.)	0.666	0.957	0.828	0.784
Residence (urban area)				
T1	-0.009	-0.514	0.004	0.340
	[0.056]	[1.093]	[0.110]	[0.919]
Urban area	-0.013	0.053	0.072	-0.569
	[0.077]	[1.238]	[0.128]	[0.928]
T1 × Urban area	-0.058	1.002	-0.020	-1.565
	[0.108]	[1.673]	[0.183]	[1.325]
Observations	126	126	126	126
R-squared	0.174	0.343	0.285	0.527
Control mean	0.580	6.489	0.556	5
Coeff: T1 + T1×Residence = 0	-0.067	0.488	-0.0161	-1.225
P-value (clust.)	0.453	0.705	0.915	0.232
P-value (boot.)	0.740	0.829	0.995	0.451
Occupation (private-oriented sector)				
T1	-0.022	-1.225	-0.030	-1.970*
	[0.075]	[1.457]	[0.146]	[1.012]
Private sector	-0.163**	-1.039	0.004	0.261
	[0.078]	[1.423]	[0.135]	[1.021]
T1 × Private sector	-0.010	1.834	0.048	3.207**
	[0.104]	[1.769]	[0.176]	[1.297]
Observations	126	126	126	126
R-squared	0.171	0.348	0.285	0.546
Control mean	0.685	7.407	0.630	3.259
Coeff: T1 + T1×Occupation = 0	-0.0321	0.609	0.0183	1.237
P-value (clust.)	0.614	0.544	0.864	0.181
P-value (boot.)	0.828	0.593	0.964	0.0568

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Gamma's compliance and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.9 (continued): Gamma's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	Min. compliance	Production performance
Risk preferences				
T1	-0.025	0.070	-0.036	-0.748
	[0.062]	[1.178]	[0.121]	[0.979]
Very willing to take risks	0.070	0.768	0.056	-1.040
	[0.079]	[1.618]	[0.174]	[1.356]
T1 × Risk	0.004	-0.509	0.088	1.232
	[0.103]	[1.651]	[0.170]	[1.213]
Observations	126	126	126	126
R-squared	0.178	0.343	0.290	0.526
Control mean	0.542	5.879	0.576	5.636
Coeff: T1 + T1×Risk = 0	-0.0207	-0.439	0.0513	0.484
P-value (clust.)	0.788	0.715	0.689	0.589
P-value (boot.)	0.878	0.932	0.874	0.593
Future orientation				
T1	0.010	0.355	-0.009	-0.452
	[0.057]	[0.946]	[0.097]	[0.944]
Very willing to delay gratification	0.094	-0.078	-0.090	-0.631
	[0.074]	[1.722]	[0.165]	[1.217]
T1 × Future	-0.090	-1.580	0.002	0.721
	[0.098]	[1.736]	[0.184]	[1.277]
Observations	126	126	126	126
R-squared	0.180	0.348	0.287	0.523
Control mean	0.539	5.872	0.590	5.487
Coeff: T1 + T1×Future = 0	-0.0805	-1.225	-0.00692	0.269
P-value (clust.)	0.300	0.423	0.966	0.794
P-value (boot.)	0.568	0.658	0.996	0.841
Intentions / beliefs about others				
T1	-0.034	0.087	0.035	-0.050
	[0.066]	[1.167]	[0.126]	[0.993]
Very optimistic about others' intentions	0.063	-0.673	0.113	-2.100
	[0.118]	[2.023]	[0.189]	[1.277]
T1 × Intentions	0.001	-0.401	-0.087	0.137
	[0.100]	[1.708]	[0.174]	[1.199]
Observations	126	126	126	126
R-squared	0.175	0.344	0.288	0.534
Control mean	0.581	6.171	0.543	6.057
Coeff: T1 + T1×Intentions = 0	-0.0329	-0.313	-0.0525	0.0868
P-value (clust.)	0.638	0.804	0.679	0.920
P-value (boot.)	0.770	0.968	0.876	0.993

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Gamma's compliance and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.9 (continued): Gamma's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	Min. compliance	Production performance
Charity / altruism				
T1	0.017	-0.489	-0.078	-0.456
	[0.068]	[1.146]	[0.113]	[1.028]
Very willing to share with others	0.077	-0.621	-0.064	-0.578
	[0.101]	[1.667]	[0.151]	[1.104]
T1 × Charity	-0.082	0.488	0.169	0.498
	[0.102]	[1.706]	[0.175]	[1.220]
Observations	126	126	126	126
R-squared	0.177	0.342	0.291	0.523
Control mean	0.527	5.433	0.533	5.900
Coeff: T1 + T1×Charity = 0	-0.0644	-0.00157	0.0913	0.0415
P-value (clust.)	0.366	0.999	0.518	0.964
P-value (boot.)	0.653	0.910	0.624	0.895
Willingness to punish unfair behaviour				
T1	-0.006	-0.694	-0.135	0.562
	[0.064]	[1.149]	[0.118]	[1.089]
Very willing to punish unfair behaviour	0.123	-1.511	-0.165	2.502
	[0.116]	[1.896]	[0.195]	[1.669]
T1 × Punishment	-0.028	0.872	0.265	-1.214
	[0.108]	[1.644]	[0.169]	[1.502]
Observations	126	126	126	126
R-squared	0.183	0.345	0.298	0.534
Control mean	0.517	5.821	0.536	5.393
Coeff: T1 + T1×Punishment = 0	-0.0343	0.178	0.129	-0.652
P-value (clust.)	0.653	0.880	0.311	0.512
P-value (boot.)	0.890	0.824	0.303	0.725
Beliefs: Alpha's request perceived as fair				
T1	0.032	-0.772	-0.112	-0.496
	[0.055]	[1.031]	[0.108]	[0.917]
Alpha's request perceived as fair	0.140**	1.271	0.041	0.276
	[0.068]	[1.378]	[0.127]	[0.891]
T1 × Fairness	-0.158	1.822	0.322*	0.997
	[0.097]	[1.601]	[0.165]	[1.291]
Observations	126	126	126	126
R-squared	0.200	0.376	0.330	0.528
Control mean	0.524	5.390	0.537	5.488
Coeff: T1 + T1×Fairness = 0	-0.125	1.049	0.210	0.502
P-value (clust.)	0.113	0.407	0.109	0.623
P-value (boot.)	0.246	0.529	0.157	0.741

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on Gamma's compliance and production effort vary with participants' demographic characteristics and social preferences. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.10: Supply chain's main results: treatment effects on compliance and production efforts

	Compliance seconds (share)		Compliance performance		Compliance: at least minimum		Production performance	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Treatment 1	-0.032	-0.022	0.165	-0.112	-0.039	-0.064	1.998**	1.249
	[0.024]	[0.022]	[0.456]	[0.408]	[0.056]	[0.053]	[0.884]	[0.908]
Treatment 2	-0.044*	-0.024	-0.232	-0.544	-0.075	-0.099**	1.608*	0.534
	[0.024]	[0.021]	[0.457]	[0.411]	[0.053]	[0.049]	[0.863]	[0.889]
	[0.020]	[0.018]	[0.258]	[0.296]	[0.034]	[0.041]	[0.454]	[0.453]
Observations	8,532	8,532	8,532	8,532	8,532	8,532	8,532	8,532
R-squared	0.182	0.174	0.246	0.199	0.163	0.140	0.268	0.219
Control mean	0.586	0.586	6.071	6.071	0.390	0.390	8.919	8.919
Beta's & Gamma's controls	YES	YES	YES	YES	YES	YES	YES	YES
Gamma's weights	NO	YES	NO	YES	NO	YES	NO	YES
Hp 1a: T1 – C = 0	0.178	0.309	0.717	0.784	0.490	0.224	0.0255	0.172
Hp 1b: T2 – C = 0	0.0636	0.261	0.612	0.188	0.161	0.0467	0.0648	0.549
Hp 2: T2 – T1 = 0	0.0152	0.610	0.000	0.000	0.009	0.004	0.001	0.000

Note: This table reports OLS estimates of the treatment effects on the supply chain's compliance and production effort. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Models (1) show results without Gamma weights, while Models (2) apply probability weights to account for multiple Beta-Gamma matches. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.11: Supply chain's additional results: treatment effects on secondary outcomes

	Fairness	
	(1)	(2)
Treatment 1	-0.034	-0.027
	[0.029]	[0.024]
Treatment 2	-0.045	-0.045*
	[0.029]	[0.024]
Observations	8,532	8,532
R-squared	0.107	0.113
Control mean	0.150	0.150
Beta's & Gamma's controls	YES	YES
Gamma's weights	NO	YES
Hp 1a: T1 - C = 0	0.236	0.261
Hp 1b: T2 - C = 0	0.123	0.0675
Hp 2: T2 - T1 = 0	0.000	0.000

Note: This table reports OLS estimates of the treatment effects on the supply chain's fairness beliefs. All models include robust standard errors, individual covariates (demographics, social preferences, and comprehension-quiz attempts), and session fixed effects. Model (1) shows results without Gamma weights, while Model (2) apply probability weights to account for multiple Beta-Gamma matches. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.12: Supply chain's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	At least min. compl.	Production performance
Gender (female)				
T1	-0.008	-0.154	-0.133**	0.334
	[0.022]	[0.421]	[0.057]	[0.887]
T2	-0.013	-0.962**	-0.140**	-0.290
	[0.022]	[0.416]	[0.056]	[0.921]
Female	0.028***	-0.624***	-0.087***	-2.275***
	[0.004]	[0.075]	[0.018]	[0.177]
T1 × Female	-0.023**	0.078	0.113***	1.509***
	[0.009]	[0.150]	[0.026]	[0.387]
T2 × Female	-0.019***	0.727***	0.070***	1.414***
	[0.006]	[0.097]	[0.025]	[0.229]
Observations	8532	8532	8532	8532
R-squared	0.175	0.201	0.142	0.222
Control mean	0.572	6.517	0.429	9.996
Coeff: T1 + T1×Female = 0	-0.0312	-0.0763	-0.0203	1.844
P-value	0.167	0.853	0.699	0.0538
Coeff: T2 + T2×Female = 0	-0.0322	-0.234	-0.0703	1.125
P-value	0.133	0.571	0.132	0.201
Age (early-career)				
T1	-0.030	-0.764*	-0.074	1.220
	[0.022]	[0.409]	[0.058]	[0.931]
T2	-0.035	-1.046**	-0.158***	0.002
	[0.022]	[0.442]	[0.054]	[0.883]
Early-career	0.005	-0.202	0.025	-0.252
	[0.006]	[0.148]	[0.029]	[0.201]
T1 × Early-career	0.012	1.104***	0.013	0.023
	[0.008]	[0.176]	[0.034]	[0.256]
T2 × Early-career	0.018**	0.809***	0.092**	0.829***
	[0.007]	[0.209]	[0.043]	[0.208]
Observations	8532	8532	8532	8532
R-squared	0.174	0.203	0.141	0.219
Control mean	0.575	5.872	0.344	8.869
Coeff: T1 + T1×Early-career = 0	-0.0173	0.340	-0.0608	1.243
P-value	0.442	0.422	0.258	0.172
Coeff: T2 + T2×Early-career = 0	-0.0176	-0.237	-0.0661	0.831
P-value	0.420	0.565	0.215	0.358
Residence (urban area)				
T1	-0.022	-0.534	-0.103*	1.270
	[0.023]	[0.417]	[0.054]	[0.927]
T2	-0.016	-0.766*	-0.137***	0.333
	[0.022]	[0.409]	[0.052]	[0.907]
Urban area	0.031***	-0.785***	-0.096***	-1.347***
	[0.006]	[0.106]	[0.019]	[0.208]
T1 × Urban area	-0.001	1.293***	0.120***	-0.054
	[0.011]	[0.184]	[0.028]	[0.347]
T2 × Urban area	-0.026***	0.664***	0.113***	0.611**
	[0.008]	[0.110]	[0.027]	[0.247]
Observations	8532	8532	8532	8532
R-squared	0.175	0.204	0.143	0.219
Control mean	0.577	6.127	0.396	9.255
Coeff: T1 + T1×Urban = 0	-0.0232	0.758	0.0164	1.215
P-value	0.295	0.0722	0.761	0.186
Coeff: T2 + T2×Urban = 0	-0.0416	-0.102	-0.0236	0.944
P-value	0.0535	0.810	0.629	0.284

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on supply chain compliance and production effort vary with participants' demographic characteristics and social preferences. All models include individual covariates (demographics, social preferences, and comprehension-quiz attempts), session fixed effects, and sampling weights. Standard errors (in brackets) are clustered at the Gamma group level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.12 (continued): Supply chain's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	At least min. compl.	Production performance
Occupation (private sector)				
T1	-0.044*	-0.825*	-0.136**	1.690*
	[0.023]	[0.430]	[0.056]	[0.940]
T2	-0.036	-1.038**	-0.170***	0.463
	[0.023]	[0.439]	[0.053]	[0.876]
Private sector	-0.034***	-0.611***	-0.070***	1.473***
	[0.011]	[0.178]	[0.023]	[0.548]
T1 × Private sector	0.049***	1.600***	0.160***	-0.994
	[0.014]	[0.217]	[0.035]	[0.621]
T2 × Private sector	0.024**	0.969***	0.136***	0.088
	[0.011]	[0.205]	[0.029]	[0.582]
Observations	8532	8532	8532	8532
R-squared	0.177	0.208	0.146	0.220
Control mean	0.609	6.162	0.407	7.935
Coeff: T1 + T1×Private = 0	0.00497	0.775	0.0239	0.696
P-value	0.824	0.0618	0.671	0.478
Coeff: T2 + T2×Private = 0	-0.012	-0.0694	-0.0342	0.550
P-value	0.579	0.865	0.499	0.576
Risk tolerance (very willing to take risks)				
T1	-0.048**	-0.285	-0.094	2.166**
	[0.022]	[0.403]	[0.057]	[0.912]
T2	-0.030	-0.662	-0.134**	0.963
	[0.022]	[0.405]	[0.052]	[0.888]
Risk tolerance	-0.006	-0.857***	-0.114***	0.719***
	[0.007]	[0.121]	[0.014]	[0.273]
T1 × Risk tolerance	0.056***	0.435***	0.072***	-2.021***
	[0.009]	[0.148]	[0.020]	[0.331]
T2 × Risk tolerance	0.011	0.272*	0.076***	-0.904***
	[0.009]	[0.154]	[0.021]	[0.288]
Observations	8532	8532	8532	8532
R-squared	0.178	0.204	0.145	0.222
Control mean	0.574	6.185	0.433	9.257
Coeff: T1 + T1×Risk tolerance = 0	0.00808	0.151	-0.0225	0.145
P-value	0.717	0.725	0.644	0.877
Coeff: T2 + T2×Risk tolerance = 0	-0.0189	-0.390	-0.0574	0.0589
P-value	0.394	0.367	0.232	0.949
Patience (very willing to delay gratification)				
T1	-0.019	0.074	-0.061	2.221**
	[0.023]	[0.404]	[0.052]	[0.927]
T2	-0.018	-0.548	-0.120**	0.940
	[0.022]	[0.413]	[0.047]	[0.875]
Patience	-0.020*	0.287	0.003	2.662***
	[0.010]	[0.211]	[0.027]	[0.354]
T1 × Patience	-0.011	-0.493**	-0.012	-2.393***
	[0.012]	[0.214]	[0.031]	[0.431]
T2 × Patience	-0.019	0.053	0.052	-0.581
	[0.012]	[0.244]	[0.040]	[0.428]
Observations	8532	8532	8532	8532
R-squared	0.177	0.200	0.141	0.232
Control mean	0.586	5.969	0.381	8.480
Coeff: T1 + T1×Patience = 0	-0.0299	-0.419	-0.0728	-0.172
P-value	0.179	0.350	0.223	0.851
Coeff: T2 + T2×Patience = 0	-0.0369	-0.495	-0.0675	0.358
P-value	0.0938	0.268	0.270	0.707

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on supply chain compliance and production effort vary with participants' demographic characteristics and social preferences. All models include individual covariates (demographics, social preferences, and comprehension-quiz attempts), session fixed effects, and sampling weights. Standard errors (in brackets) are clustered at the Gamma group level. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 3.A.12 (continued): Supply chain's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	At least min. compl.	Production performance
Trust (very optimistic about others' intentions)				
T1	-0.033	-0.581	-0.119**	1.216
	[0.022]	[0.411]	[0.054]	[0.927]
T2	-0.019	-1.316***	-0.178***	-0.054
	[0.021]	[0.427]	[0.053]	[0.885]
Trust	0.017***	-0.290**	-0.045	-1.045***
	[0.005]	[0.124]	[0.029]	[0.200]
T1 × Trust	0.026***	1.266***	0.149***	0.196
	[0.006]	[0.168]	[0.028]	[0.251]
T2 × Trust	-0.014***	1.833***	0.190***	1.563***
	[0.005]	[0.243]	[0.040]	[0.256]
Observations	8532	8532	8532	8532
R-squared	0.176	0.213	0.148	0.222
Control mean	0.565	6.078	0.399	9.650
Coeff: T1 + T1×Trust = 0	-0.00685	0.685	0.030	1.412
P-value	0.758	0.108	0.592	0.114
Coeff: T2 + T2×Trust = 0	-0.0337	0.518	0.0121	1.509
P-value	0.128	0.227	0.820	0.0985
Prosociality (very willing to share with others)				
T1	-0.013	-0.093	-0.073	1.392
	[0.022]	[0.408]	[0.055]	[0.911]
T2	-0.008	-0.588	-0.091*	0.266
	[0.021]	[0.416]	[0.052]	[0.884]
Prosociality	0.039***	-0.421***	-0.032	-0.846***
	[0.006]	[0.124]	[0.027]	[0.208]
T1 × Prosociality	-0.028***	-0.029	0.027	-0.346
	[0.007]	[0.153]	[0.033]	[0.260]
T2 × Prosociality	-0.044***	0.127	-0.019	0.715*
	[0.007]	[0.165]	[0.028]	[0.381]
Observations	8532	8532	8532	8532
R-squared	0.178	0.200	0.141	0.221
Control mean	0.558	6.046	0.391	9.496
Coeff: T1 + T1×Prosociality = 0	-0.0417	-0.122	-0.0466	1.045
P-value	0.0645	0.776	0.393	0.262
Coeff: T2 + T2×Prosociality = 0	-0.0519	-0.461	-0.110	0.981
P-value	0.0221	0.277	0.0263	0.301
Inequity aversion (very willing to punish unfair behaviour)				
T1	-0.020	-0.019	-0.063	1.989**
	[0.022]	[0.410]	[0.050]	[0.924]
T2	-0.027	-0.622	-0.098**	1.346
	[0.022]	[0.404]	[0.049]	[0.933]
Inequity aversion	0.022***	-0.219*	-0.016	-0.664***
	[0.006]	[0.125]	[0.015]	[0.217]
T1 × Inequity aversion	-0.005	-0.205**	-0.003	-1.561***
	[0.007]	[0.099]	[0.019]	[0.261]
T2 × Inequity aversion	0.010	0.173	-0.006	-2.048***
	[0.008]	[0.141]	[0.026]	[0.322]
Observations	8532	8532	8532	8532
R-squared	0.175	0.199	0.140	0.227
Control mean	0.580	5.847	0.355	8.578
Coeff: T1 + T1×Inequity = 0	-0.0245	-0.224	-0.0664	0.428
P-value	0.269	0.588	0.248	0.640
Coeff: T2 + T2×Inequity = 0	-0.0164	-0.448	-0.104	-0.703
P-value	0.446	0.303	0.0574	0.411

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on supply chain compliance and production effort vary with participants' demographic characteristics and social preferences. All models include individual covariates (demographics, social preferences, and comprehension-quiz attempts), session fixed effects, and sampling weights. Standard errors (in brackets) are clustered at the Gamma group level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.12 (continued): Supply chain's heterogeneity analysis

Variable	Compliance seconds (share)	Compliance performance	At least min. compl.	Production performance
Fairness (Alpha's request perceived as fair)				
T1	-0.038*	-0.107	-0.048	2.137**
	[0.022]	[0.404]	[0.051]	[0.931]
T2	-0.025	-0.951**	-0.127***	0.990
	[0.021]	[0.407]	[0.044]	[0.917]
Fairness	0.004*	0.441***	0.105***	1.223***
	[0.002]	[0.046]	[0.013]	[0.130]
T1 × Fairness	0.052***	0.029	-0.043*	-2.790***
	[0.005]	[0.111]	[0.023]	[0.222]
T2 × Fairness	0.006	1.661***	0.138***	-1.300***
	[0.007]	[0.184]	[0.036]	[0.295]
Observations	8532	8532	8532	8532
R-squared	0.179	0.219	0.157	0.225
Control mean	0.576	5.837	0.351	8.555
Coeff: T1 + T1×Fairness = 0	0.014	-0.078	-0.0911	-0.653
P-value	0.529	0.856	0.125	0.459
Coeff: T2 + T2×Fairness = 0	-0.0197	0.710	0.011	-0.310
P-value	0.374	0.111	0.873	0.720
Effectiveness (Beta's inspection policy perceived as effective)				
T1	-0.045**	0.039	-0.031	2.667***
	[0.022]	[0.418]	[0.046]	[0.926]
T2	-0.030	-0.349	-0.055	1.605*
	[0.021]	[0.413]	[0.044]	[0.886]
Effectiveness	0.002	1.130***	0.180***	1.934***
	[0.003]	[0.119]	[0.023]	[0.153]
T1 × Effectiveness	0.051***	-0.137	-0.042	-2.801***
	[0.005]	[0.181]	[0.036]	[0.214]
T2 × Effectiveness	0.013***	-0.247*	-0.067**	-1.966***
	[0.004]	[0.132]	[0.026]	[0.152]
Observations	8532	8532	8532	8532
R-squared	0.179	0.218	0.162	0.229
Control mean	0.579	5.567	0.304	8.181
Coeff: T1 + T1×Effectiveness = 0	0.00571	-0.0981	-0.0733	-0.134
P-value	0.790	0.816	0.262	0.881
Coeff: T2 + T2×Effectiveness = 0	-0.0174	-0.596	-0.122	-0.361
P-value	0.422	0.162	0.0341	0.691

Note: This table reports heterogeneity tests examining whether the effects of T1 and T2 on supply chain compliance and production effort vary with participants' demographic characteristics and social preferences. All models include individual covariates (demographics, social preferences, and comprehension-quiz attempts), session fixed effects, and sampling weights. Standard errors (in brackets) are clustered at the Gamma group level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.