WORKING PAPERS

N° 1336

June 2024

“Motivating Public Sector Employees: Public Good Contributions in Addis Ababa Water and Sewerage Authority”

George Joseph, Josepa Miquel-Florensa, Yi Rong Hoo,
Sanjay Pahuja and Tewodros Tebekew

Toulouse
School of
Economics

Motivating Public Sector Employees: Public Good Contributions in Addis Ababa Water and Sewerage Authority [*]

George Joseph
World Bank

Josepa Miquel-Florensa
Toulouse School of Economics

Sanjay Pahuja
World Bank

Yi Rong Hoo
World Bank

Tewodros Tebekew
World Bank

June 3, 2024[†]

We present a lab-in-the-field experiment with employees of the Addis Ababa Water and Sanitation Authority to understand how to improve coordination and collaboration in their daily work. Participants play a series of public good games under different rules: a standard game, a game with a threshold, and a game with a randomly selected anonymous monitor with the power to punish. We show that a common goal, in the form of a threshold to be attained for the group's success, is significantly more effective than a potentially punishing monitor for increasing individual effort and, ultimately, group outcomes (conditional on the threshold being attained). This result advocates for the introduction of team goals as coordination and

[*] We thank Cesar Mantilla, Erin McDonnell and seminar participants at the ASFEE Virtual seminar and SEEDEC for their comments and suggestions, and Guillem Foucalt for excellent research assistance. We thank AAWSA management and employees for their collaboration in the design and implementation of this project. We also acknowledge the support from the World Bank's Global Water Security and Sanitation Partnership, and the Government of Netherland's WaterWorX program. All errors remain our own. The views presented in this paper are those of the authors and do not represent those of the World Bank or its member countries. Josepa Miquel-Florensa acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program). The design was reviewed by the TSE-IAST review board for ethical standards in research (2019-08-001).

[†] George Joseph, World Bank rgjoseph@worldbank.org; Corresponding author: Josepa Miquel-Florensa, Toulouse School of Economics pepita.miquel@tse-fr.eu; Sanjay Pahula spahuja@worldbank.org; Yi Rong Hoo yhoo@worldbank.org; Tewodros Tebekew tadmassu@worldbank.org

motivation devices in settings where tasks are performed by groups and are subject to free-riding and coordination challenges. **Keywords**: Intrinsic motivation, Public utilities, Organizational economics.

JEL codes: J45 (Public Sector Labor Markets), M50 (Personnel Economics), O12 (Microeconomic Analyses of Economic Development).

1. Introduction

Improving the quality of public service delivery is a challenge in many developing countries. Capital investments, institutional reforms, and capacity building are the main forms of intervention aimed at improving the coverage and use of public utilities. However, even when projects succeed in implementing infrastructure and policy reforms, these gains often fail to translate into the expected service delivery improvements for the intended beneficiaries. Physical investments do not necessarily improve service delivery without well-trained and motivated public agency staff. These staff members are responsible for executing the policies and implementing infrastructure improvements. Indeed, they are the "face" of the public utility for consumers and are therefore the crucial last link in ensuring the adoption and use of new services.

Two challenges often faced by public utilities are, first, how to attract public service-motivated individuals to the utilities and, second, how to ensure that the existing employees remain conscientious and motivated to deliver on their assigned tasks (Finan et al., 2017).[1] This paper addresses the second question in the context of the Addis Ababa Water and Sanitation Authority (henceforth AAWSA) - a large water and sanitation service provider in Addis Ababa, the capital and largest city in Ethiopia. The design of effective incentives and organizational schemes is crucial for a public utility to reach its service objectives. An additional challenge faced by these utilities is that tasks are frequently performed in crews, adding a moral hazard and coordination concern to the motivational challenges (Holmstrom, 1982).[2] This is the case in the provision of water and sanitation services,

[1] Social incentives are especially relevant for organizations (see Ashraf and Bandiera (2018) for a survey). This is particularly true in sectors where the quality of outcomes depends primarily on the attitude and behaviors of the last-mile service providers (see Ashraf et al. (2014), Ashraf et al. (2018) and Mbiti and Serra (2018) for examples in the health sector). The evidence on the interaction between economic incentives and social preferences is, however, inconclusive (see Bowles and Polania-Reyes (2012) for a survey).

[2] After the seminal Holmstrom (1982) paper, the literature on monitoring effort in teams is extensive. For example, Gershkov and Winter (2015) study formal versus informal monitoring, Halac et al. (2021) in a theoretical setting look at the difficulty of monitoring heterogeneous

where working crews are formed by four to six employees who each specialize in different roles.

Different instruments have been considered in the provision of incentives for team effort in both the academic literature and in the field. Among them, the difficulty in evaluating individual accountability in teams (Marx and Squintani, 2009), the difficulty in monitoring heterogeneous teams (Halac et al. (2021), Weng and Carlsson (2015)), and the trade-off between formal and informal monitoring schemes (Gershkov and Winter, 2015) have been the most studied options.

In this paper, our focus is from the perspective of the management of a public utility, particularly on the effectiveness of different management policies in increasing crew motivation and hence a team's performance. We ask the following questions: Is the desired result achieved by setting a common goal among employees? Or is it achieved through effective supervision (and potential punishment)? To answer these questions, we perform a lab-in-the-field experiment, where employees of AAWSA play a series of public good games under different rules: a standard game (a Standard Benchmark game), a game with a threshold (a Threshold game), and a game with a randomly selected anonymous monitor with the capacity to punish (a Monitor game).

For the first question, we find that a common goal in the form of a pre-set threshold significantly increases contributions when compared to the standard game. This result is consistent with the literature surveyed in Dannenberg et al. (2015). In relation to the second question, we observe that while players increase their contribution when the monitor figure is introduced, this increase is smaller than that found in the presence of a common goal threshold.[3] However, this result requires further examination when analyzing group outcomes; since contributions are lost when the threshold is not met, the average outcomes at the group level do not differ between the Monitor and the Threshold games.

The identification and empowerment of leading roles improves group cooperation (Kosfeld and Rustagi (2015)) and the voluntary provision of public goods (Jack and Recalde (2015), d'Adda et al. (2017)). Harnessing the power of personal agency and identifying people who can champion change initiatives is vital

teams, and Marx and Squintani (2009) examine individual accountability in teams. Herbst and Mas (2015) summarize the experimental and field literature on workers' output peer effects, showing that it is positive and not statistically significant in the lab versus the field. Hamilton et al. (2003) focuses on team incentives and workers' heterogeneity. Villeval (2020) highlights the importance of norms and institutional dynamics on group performance.

[3] In the econometric analysis, we find that the Monitor game is not significantly different from the Reference game when is played just after the threshold game. We provide further discussion on the ordering effect there.

to achieve the desired outcomes in these public utility sectors. Our experiment is consistent with the literature showing that the monitor figure increases contributions, but significantly less than the introduction of a common goal for the group.

The experiment is accomplished with personnel from AAWSA. There are two facts to be noted about the structuring of work in the water and sanitation branches: (i) there is a vertical distinction across grades and skills of employees, with differences in their training and experience, and (ii) many of the branch activities are performed in teams (or street crews) of four to six employees across different skill levels. This need for collaboration across levels and departments is key to the provision of excellent service to customers. Our experimental results shed light on how to design effective motivational incentive schemes in this setting.

The paper is structured as follows. Section 2 describes AAWSA, the setting in which the experiment took place. Section 3 describes the experimental sample and procedures. Section 4 presents the empirical analysis, and Section 5 presents the results at the individual and group level, together with the monitor choices. Section 6 concludes.

2. Setting: The Addis Ababa Water and Sanitation Authority

The Addis Ababa Water and Sanitation Authority (AAWSA) is a public institution that provides water and sanitation services to Ethiopia's capital city of 3.5 million inhabitants (UNHabitat estimation for 2015). It is structured around eight branches, each with a specific catchment area. In combination, the eight branches cover the entire city. Each of these branches have identical internal organization and governance structures.

Performance measures are tracked at the branch level. Decreasing non-revenue water, improving consumer satisfaction, and ensuring hours of water service provision are the main institutional goals. The success of these goals depends on employee motivation and performance, in a setting where employee retention is challenging, and the labor market is thin. Moreover, the team nature of the work inside the branches plus the rotation of employees amongst the teams (so as to adapt to the specific human needs of the different tasks) makes it unfeasible to establish performance-based incentives at the individual or group level.

Employees work in crews of four to six members with different skills and experience, for example, mechanics, engineers, and a driver. The response speed to issues with the water/sewerage lines and the maintenance of trucks and other materials are critical for timely and effective service delivery. Collaboration within the crew is imperative for a quick reaction, since the efforts of the different members are complementary: as is typical in teamwork, it is very difficult to assign to each crew member a share on the outcome of the intervention. Teamwork, together with the nature of the job (related to external risks such as meteorological), does not allow for output-dependent compensation. Hence, the motivation of these crews, and the improvement of their internal coordination, are the primary concerns for management. Moreover, within each of the water and sewerage institutional departments, which cover 88% of employees, the coordination and information flow between consumer services and street crews is crucial for the timely reaction to the frequent infrastructural issues.

To improve employee motivation, AAWSA has participated in the World Bank Field-Level Leadership (FLL) program.[4] The lab-in-the-field experiment presented in this paper took place between September 2019 and February 2020, on the first day of the Field Level Leadership workshops. These workshops were held at the AAWSA Training facilities located on the outskirts of Addis Ababa. The introductory session of the training workshop started with participants introducing themselves by name and branch, and eating lunch together at the training facility, after which the official training and experimental session began. While interaction between participants cannot be ruled out, the workshop trainers assured the research team that no information on the workshop goals and contents was provided to the participants in the introductory session. The workshops were completed in groups of 25-30 employees from two of the Authority's branches: Arada and Addis Ketema. The two branches are similar in terms of employee characteristics and their distribution across departments (see Table A1). All employees of the two AAWSA branches participated in the training and were invited to volunteer for the experiment, with very high participation rates.[5] The employees were distributed across workshops to ensure diversity across and between departments and skill levels in both branches.

Part A of Table 1 presents the socio-economic characteristics of the participants. The average age is 37 years and employees are predominantly male (over 80%). In terms of education, it is worth noting that most participants have a background of at least some education, leading to higher levels for those with

[4] FLL constitutes a set of interventions aimed at identifying and supporting entrepreneurial and motivated employees in public agencies to lead improvements in performance and service delivery outcomes.

[5] All participants in the workshop (a total of 409) agreed to participate in the experimental session, with only four employees leaving the room after the introduction of the activity.

technical diplomas in infrastructure maintenance tasks. In terms of departments inside AAWSA, Part B of Table 1 shows that the majority of employees (around 88%) work in the main departments of Water and Sewerage. Inside these departments, approximately one-third of employees work on infrastructure tasks (line installation, new connections, etc.), which are those tasks that are more likely to require crews on the street. Coordination between employees working on consumer service tasks within a branch and the street crews is crucial for a rapid reaction to infrastructure issues and for the accurate recording of water and sewerage grid usage and billing.

3. The Public Good Experiment

The literature has considered a variety of experimental games to measure social preferences.[6] In the case of public utility employees, in order to study social preferences, we need to bear in mind that day-to-day work is performed primarily in crews: for example, groups of four to six employees, with the usual concerns on free-riding. This is the case in AAWSA, where crews are generally formed by five employees with different complementary skills. Teams are typically composed of a driver and a team of engineers with different grades and skills, who are frequently re-organized as the work demands. The specificity of the tasks performed in this way makes the public good game an effective instrument to measure social preferences and motivational effort, also enabling the evaluation of interventions to improve team outcomes. We believe that a public good game in teams of five members with a neutral framing can generate insights into how teams, such as the AAWSA crews, can be encouraged to cooperate more effectively.[7] This is true, despite not precisely matching the structure of the workplace, where effort is only partially identifiable and there are no direct economic compensations in terms of the outcome.

The main goal of our experimental design is to test how individual contributions to a standard public good game - and ultimately group outcomes - vary with the introduction of two management strategies: (i) a *common goal* motivation in the form of a step function, and (ii) the inclusion of a *game monitor* who can

[6] See surveys by Charness and Rabin (2002), Camerer and Fehr (2004) and Levitt and List (2007).

[7] While there is no linkage between employees' salaries and their performance or their team's performance, field discussions have highlighted the employees' awareness of the importance of the outcomes of their jobs for their community. It should also be noted that within each department there is considerable awareness of the importance of joint effort in the coordination of tasks and the flow of information.

”punish” participants after observing their contribution amounts.[8]

Our benchmark is the standard public good game. Each of the five participants within a group receives an endowment of 10 tokens[9]; contributions to the pool are doubled and divided equally among the team members. In this setting, the payoff of player i that contributes c_i (while the other four team members contribute $[\sum_{j\neq i} c_j]$) is given by:

$$\pi(c_i, c_{-i}) = [10 - c_i] + \frac{2}{5} * [\sum_{j=1}^{5} c_j] \ ,$$

where the first part of the payment comes from the endowment that is not contributed to the pool, and the second part comes from the tokens distributed from the pool after the amounts contributed by all players have been doubled and then split equally among the five group members.

The variations on the standard public good game played are as follows (see Appendix A for the protocol details). In all games, the contributions of each player were private and anonymous, and no information on group return was revealed until the payment stage (i.e., after all games had been played), so as to avoid hedging between games. All of the games played were paid.

- *Benchmark Public Good game*: Our benchmark is the standard public good game, in groups of five players. The goal of this Benchmark game is to measure the contributions made within a group that is the same size as the employee crews who work together on a daily basis. Tokens contributed to the common pool are doubled and distributed equally among the team members, independently of the total group contribution. A first round was played with anonymous matching: a hypothetical group of five persons in the room, with no information on whether they were sitting at the same table or not. (This first round had the goal of familiarizing the participants with the game and was not used in the analysis). A second round was played

[8] We do not aim to study ”in- vs. out-group behavior” but, rather how behaviors change with the game structure, whilst keeping the composition of the group constant. See Charness et al. (2007), Chen and Li (2009), Charness and Sutter (2012) and Goette et al. (2012) as references on ”minimal group paradigm” and ”group versus individual behavior”. The literature consistently shows how group identification changes individual behavior towards more collaborative decisions. The literature has also looked at how group heterogeneity affects the general result of group identification (see, for example, Weng and Carlsson (2015)). In our setting, there is no heterogeneity in endowment among group participants.

[9] Each token is valued at 1 Ethiopian BIR (0.035 USD at the time of the experiment). The average payment per participant was 66 BIR (2.5 USD), which was approximately the mean salary for two hours of work. It needs to be noted that the experiment took place during the employee’s standard working hours, and hence the payments were adapted accordingly.

with the participants sitting at the same table. The second round, which confirms the table at which the team will be playing in the following rounds, is used as the baseline.

- *Threshold Public Good game*: This game aims to reflect the effort that needs to be exerted to attain success in many of the daily tasks performed by the employee crews. Crew tasks require a minimum total contribution of effort to make any progress toward the job/task completion. Payments in the Threshold game follow a step function: tokens contributed to the pool are doubled when the total contributed amount reaches the threshold of 25 tokens (half of the total tokens of group participants), otherwise all contributions are lost.

- *Public Good game with monitor*[10]: Inspired by the protocol presented in Kosfeld and Rustagi (2015), we add to the standard public good game the figure of the game monitor. This is a member of the group with the power to "punish" participants when observing their anonymous contributions. Monitors are randomly chosen from among session participants by selecting a card from a bag that contains all of the players' codes in front of all participants, without announcing the outcome. Chosen monitors are informed of their role at the end of the session when all games have been played and results have not yet been announced to participants.[11] Once this randomly-assigned chosen monitor sees the contribution of each member of the group, he/she must decide whether to punish one or more of the members of the group (see in Figure F5 the form presented to the monitors).

3.A. Experimental procedures

On arrival to the experiment room, participants are randomly distributed ID tags. When entering the room, participants find the seat that has the number that appears on their tag, with the help of the experiment's assistants. The room

[10] The naming of a player who has the power to punish was the object of discussion with focus groups during the project design. Different suggestions for the name, such as "leader", "monitor", or "supervisor", were considered. The Amharic word used was close to "game leader", and was considered to be the best match in this context. Within the text, we have kept the word "monitor" in order to avoid confusion with leaders not usually being associated with ex-post punishment and with the figure of a leader being associated with a "motivational first mover" in the game.

[11] After all games are played and post-experimental questionnaires have been completed, the randomly selected monitors are discretely informed they had been randomly selected for the monitor role and are asked to make their choices as monitors. See Appendix Protocol.

design places participants in tables of five members, arranged in the room so that all participants can see the experiment director and the projection of the (language-free) instructions. In front of them on the table, participants find four closed receptacles of different colors, one for each of the four games played, and a pen for the post-experimental questionnaire. Each of these receptacles has the participant's ID marked and contains ten white tokens (and some non-valuable black tokens).[12] At the center of the table, participants find a wooden box that acts as a contribution box. This contribution box is closed to maintain privacy and has a hole into which contributions can be slotted. Detailed information on the protocol, as well as the experimental materials, can be found in Appendix A.

The structure of the experimental sessions is presented in Figure 1. The randomly selected groups of five participants seated at the same table are the groups that will play the public good game from the second round on. The first game is the Standard public good game, and is played with partners in the room to allow all participants to become familiar with the structure of the game. Given that participants are not necessarily familiar with abstract reasoning, this first round assists them to become familiar with a public good game. Hence, the order of the first and second games was kept constant across sections and the first game was dropped from the analysis. The last two games, Monitor and Threshold, were played in a different order for the different sessions. In the Monitor game, the random selection of the monitor was achieved by one member per group selecting a card from a box that contained all participants' ID cards. The chosen monitor was only known by the experiment director and not by the group members or the monitor him/herself (thus, the ”leading by example” mechanism is blocked).[13] After all games were played and the post-experimental questionnaire was completed, experiment assistants would call up the randomly chosen monitor privately to inform him/her that their ID had been randomly selected for them to act as monitor and to ask for his/her decision on whether or not to punish any of the group's participants. No information on game outcomes was disclosed until later, at the

[12] Black tokens have no points value. Their role in the individual receptacles and in the common wooden box is to ensure the privacy of the individual contributions. On the one hand, when participants pass the box around the table to make the contributions, the tokens in the common box make noise with movement, ensuring that nobody can guess the amounts contributed and hence added to the box as it goes around the table. On the other hand, the non-value tokens in the individual receptacles also make noise before and after the participant makes the choices, ensuring the privacy of decisions.

[13] By blocking the ”leading by example” mechanism, we want to avoid participants noting the monitor's behavior as a reference, if he/she can be identified. See Eisenkopf and Kölpin (2023) and Billinger and Rosenbaum (2023) for surveys on leading by example and the influence of hierarchies in public good games.

payment stage.

Even if no information is provided during the session about contributions, the fact of playing four consecutive rounds of a public good game under different rules must be considered in relation to the internal validity of the experiment. In addition to including session and group fixed effects in the analysis to control for potential ordering effects, it is pertinent to examine whether participants randomly allocated to the two orders displayed different behaviors. In the results section, we present a discussion on the steps taken to plausibly ensure that the ordering effect is not driving our findings.

Figure 1: **Structure of the Experimental Sessions**

Order 1	Order 2
Test standard PG - Room partners	
Standard PG - Table partners	
Threshold	Monitor
Monitor	Threshold
Post-experimental survey Monitor decisions and payments	

Notes: First and second games (standard PG game with different partners) were not randomized. The order of the Threshold and Monitor games was randomly allocated to the session.

Given that three out of four games are played with the participants sitting at the same table, it is important to question whether or not the participants in each session and at the same table knew each other. Figure A2 in the Appendix shows that around 70% of the participants report knowing at least one other participant sitting at the same table, and that on average in each group, participants knew 2.4 of the table members, (which includes themselves). We control by knowledge of table participants in the empiric analysis. On average, at each table, half of the participants came from the same branch (first quartile 40%, third quartile 60%). For the participants in the sessions, Figure A2 shows that each participant knew between five and 15 other participants. This number is slightly lower than the number of participants in the session they recognized as employees from the same branch.

4. Results

Table 2 presents descriptives of the experimental outcomes. In Part A, we see that at the individual level, the Threshold game is that with the higher average contributions (57% of endowment), followed by the Monitor game with 52% of the endowment. However, given that in the Threshold game group effort is lost when the threshold is not attained (which this is the case in 29.6% of the groups), we see in Part B that the Monitor game is the one that reaches higher average group contributions.

In Part C of Table 2, we see that on average, 72.8% of the game monitors decide to punish at least one member of the group: in these cases, the average number of players punished within a group is 2.13.

4.A. Analysis at the individual level

i) Empirical strategy for within-subject analysis: comparing the behavior of the same individual across the different games

Participants in a given session play four versions of the public good game, as described in Figure 1. The first round is a learning round to familiarize the participants with the experimental procedures, which is not used in the analysis. Hence, we compare the behavior of each subject across his/her decisions in the three rounds of the game, corresponding to the table, threshold, and monitor protocols. In this analysis, we include individual fixed effects controlling for any individual specific characteristics (e.g., pro-sociality, identification with the group, ...) that allow us to identify changes in behavior given by the differences across the games: namely, the introduction of a threshold (Threshold game) and a randomly selected group monitor (Monitor game) compared to the Benchmark standard game with the same table partners.

We estimate:

$$C_{igt} = \alpha + \beta^T * Threshold_g + \beta^L * Monitor_g + \gamma_i + \epsilon_{igt} \tag{1}$$

where C_{igt} denotes the contribution of player i in game g sitting at the table of team t, where $Threshold_g$ and $Monitor_g$ are dummy variables for the threshold and monitor games, respectively, and γ_i are individual player fixed effects. While this is our preferred specification (and the most conservative one), for robustness we run the same estimation with group fixed effects and individual controls for the following: a dummy for player branch of origin, the number of table members

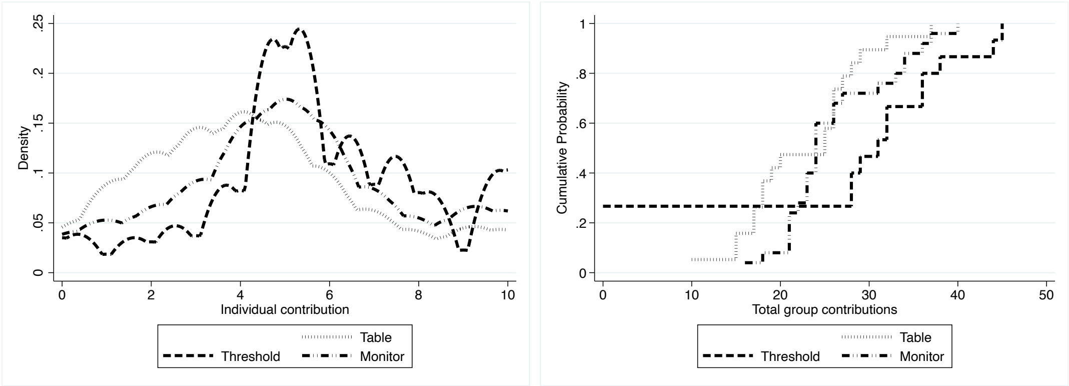

Figure 2: Distribution of public good contributions.

The left figure shows the distribution of individual contributions for the different games. The three distributions are pairwise significantly different (Wilcoxon distribution test, p=0.000).

The right figure shows the cumulative distribution of the group's total contributions. At the group level, the threshold game is not significantly different from the monitor (Wilcoxon p=0.223) or the table games (Wilcoxon p=0.505). Note that the horizontal line on the threshold game at the group level is due to the game structure. All contributions are lost if a minimum of 25 tokens is not reached at the group level.

known by the participant, and whether the sub-department of the employee is more likely to involve team work (installation and connection tasks). We cluster standard errors at the session level.[14]

ii) Results: Comparing public good contributions at the individual level, within–subject analysis

When observing the individual contributions to the public good game, on the one hand, our design allows us to study whether the introduction of a threshold or a randomly assigned game monitor increases the contributions when compared to the benchmark, keeping the game partners constant. On the other hand, we can compare the relative efficiency of these two mechanisms by observing the increase of individual contributions.

Table 2 Panel A shows that both the Threshold and the Monitor games increase the average contribution to the public good game with respect to the Benchmark game with partners at the same table, from 44% to 57%, and 52% of the participant endowment, respectively. Moreover, the three distributions are statistically

[14] Given the relatively small number of clusters (a total of 15 sessions), we report robustness as the p-values obtained using Webb (2013) wild cluster bootstrap. Results are equivalent when the cluster is done at the group level.

different, with the Threshold game maximizing contributions. In the Threshold game, 70% of the participants contribute five or more tokens (25% contribute exactly five), while this share is 50% in the Monitor game. As Figure 2 clearly shows, the Threshold game establishes a reference of contributing half of the endowment and reaching the 25 tokens threshold. This threshold is likely to set a focal point for the players on posterior games, and for that reason, we altered the order of the Threshold and Monitor games, as shown in Figure 1, to be able to control for order effects in our estimations.

The first three columns of Table 3 show the results of estimating equation 1. In all columns, the reference game is that which is played with table partners, with the goal of comparing this game to the two games with partners at the same table, but with a significant "twist" that includes a threshold or a game monitor. Column (1) includes group fixed effects, column (2) adds individual controls, and column (3) is the most conservative estimation with player fixed effects included. All specifications show similar results: the twists in both the Threshold and the Monitor games significantly increase individual contributions, with a greater difference noted in the Benchmark game. At the bottom of Table 3, we show that the coefficients for the two games when compared pairwise are significantly different. We conclude, therefore, that the Threshold game increases contributions by a greater extent when compared with the Benchmark game than the Monitor game.

We present in Figure A3 some heterogeneity analysis of individual contributions with respect to a player's rank in the institution, tenure, and a post-experimental risk-elicitation game. We see that neither of these individual characteristics has a significant impact on contribution level choices in the different games.

iii) Comparing public good contributions at the individual level: accounting for ordering effects

Given the structure of the experimental session, as presented in Figure 1, we present two robustness analyses to alleviate concerns that the results are being driven by ordering effects; that is, between-subject analysis of choices in the same round for different protocols, and interaction of the order with the game dummies in the estimation of Equation 1. In Table A2, we present the between-subject analysis for rounds 3 and 4 of the game. Columns (1) to (4) compare the choices in round 3 of the game between participants who were playing the Monitor game in this round with the choices of the participants playing the Threshold game in the same round. Column (1) presents the raw comparison, controlling only for contributions in round 2. Columns (2) to (4) add a set of individual controls and interact the order of the games with the type of work (e.g., the department more likely to utilise teamwork), and the number of participants known in the session. When comparing participants that played one or the other games in round 3, we see along the columns that the Monitor game contributions were on average one token lower than those in the Threshold game. Columns (5) to (8) present the same exercise for choices in round 4, and show that on average participants playing the Threshold game in round 4 also contributed approximately one token more than participants playing the Monitor game in round 4.

Table A3 presents the estimation of Equation 1, where the game dummies are interacted with the order of the games, and the order control is included. Since order was randomly assigned at the session level, neither session nor group or individual fixed effects can be added. The different columns include individual controls and present clusters of standard errors at the session and the group level. In all specifications, the Threshold game presents higher individual contributions than the Monitor game, and we note that the order of the games leads to smaller contributions in the Monitor game, with the Monitor game becoming not statistically different from the Benchmark game when is played before the Threshold.

4.B. Analysis at the group level

i) Empirical strategy for within-group analysis: comparing outcomes of the same group across the different games

Given the structure of the games proposed, where the common pool contributions are lost when the total of 25 tokens is not reached in the Threshold game, the outcome of the group is not equal to the aggregate of individual contributions in all settings. Hence, a performance indicator of interest is the outcome at the group level. At this level of aggregation, power is limited, since we have one-fifth of the observations at the individual level, with a total of 81 groups.

We estimate:

$$G_{gt} = \alpha + +\beta^{T} * Threshold_{g} + \beta^{L} * Monitor_{g} + \eta_{t} + \epsilon_{gt} \tag{2}$$

where G_{gt} denotes the group outcome for game g of team t, η_t denotes the group fixed effects, and the game dummies are defined as above. Our main specification includes group fixed effects since this is the more conservative approach. However, we present the specifications as robustness with session fixed effects. We cluster standard errors at the session level.

ii) Results: Comparing public good group outcomes The right panel of Figure 2 presents the cumulative distribution of the group aggregate contributions. As expected, we see that the patterns mimic those of the individual contributions (presented in the left panel of Figure 2), with a clear difference in the Threshold game. Specifically, a step appears for groups that contributed less than the 25 points of the threshold, and according to the game rules, all of their contribution is lost, which occurs in 24 out of 81 groups. Table 2 Panel B shows that average contributions are maximized in the Monitor game (26.06 tokens).

The last two columns of Table 3 show the estimation of equation 2. Column (4) includes session fixed effects, and column (5) includes, instead, group fixed effects. We observe that the Monitor game leads to greater group contributions than the Benchmark game that has partners at the table, however this is not the case for the Threshold game. Two differences should be noted here between the individual and the group-level analysis. On the one hand, the Benchmark and Threshold games do not give significantly different group averages, while individual contributions increase in the Threshold game (that is, around one-third of the groups do not reach the threshold, and hence their group outcome is zero). On the other hand, the Monitor game gives significantly greater group contributions than the Threshold game, even if it gives smaller average individual contributions.

4.C. Monitor's behavior

Our experimental design allows us to answer two questions related to the inclusion of a randomly chosen monitor in the game. First, does the randomly chosen monitor generate an increase in the players' contributions, and second, does the monitor use his/her authority to punish other players? Concerning the first question, we see that contributions in the game with a monitor are significantly higher than in the Benchmark game, both at the individual and at the group level (as shown in Table 3). Concerning the second question, Figure 3 (left) shows that monitors do in fact choose to punish other players in approximately 75% of the groups. As a reminder, monitors are randomly chosen, their identity is not revealed to the groups, and their role is not revealed to them until the end of the experimental session, when they are called to make their choices. The choices are made on a form that shows the anonymous contributions of each member of the group where they were randomly allocated as monitors (see Figure F5 in the Appendix). In terms of payoffs, the monitor receives 10 tokens for being chosen for the role, and

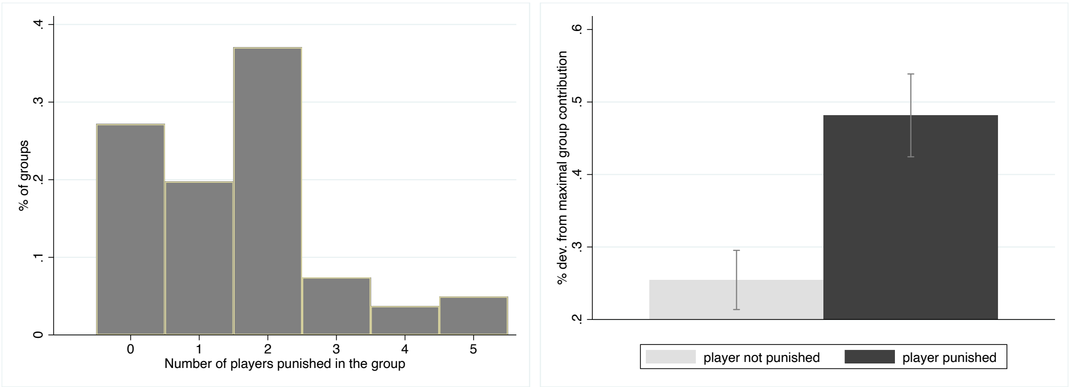

Figure 3: Monitor behavior

The left figure shows the number of players the monitor chooses to punish in the group where he/she is the randomly chosen monitor. The right figure shows how the deviation from the maximum contributor in the group relates to the probability of being punished within the groups where at least one member is punished.

his/her choice is whether to punish a player, which implies a cost of one token for the monitor and three tokens for the player. Hence, punishing a player is a costly decision for the monitor.

We focus on the monitor choice in terms of punishing any one player in the group and on the number of players punished (see columns (1) to (4) in Table 4, where the unit of observation is a monitor allocated to a group). We see that the monitor's discrete decision to punish (i.e., a dummy variable that takes value one if the monitor punishes at least one member of the group) is not significantly correlated to group contributions, nor to the variance of contributions within the group (see columns (1) and (2)). Columns (3) and (4) show that the same is true for the number of players punished inside a group when restricting to groups where at least one player is punished.

We continue our analysis at the individual player level, taking as the unit of observation a game participant, and focusing on players within groups where at least one player was punished. Column (5) of Table 4 shows that the likelihood of a player being punished is significantly negatively correlated with his/her individual contribution, and column (6) shows that this likelihood is positively correlated with the percentage difference of each player's contribution with respect to the highest contribution in the group. Figure 3 (right) plots this relative difference for punished and non-punished players, and shows that high deviations are more likely to be among those who are punished.[15]

[15] A reasonable question at this point is to examine how the monitors themselves behaved in the

5. Conclusions

This lab-in-the-field experiment shows that a common goal, in the form of a threshold to be attained for the group's success, is significantly more effective than the introduction of an anonymous and potentially punishing monitor to increase individual effort. However, at the group level, the threshold is not always attained, and individual contributions are therefore lost in approximately one-quarter of the groups. This result advocates for the introduction of team goals as coordination and motivation devices in settings where tasks are performed by crews or teams, keeping in mind that the goal must be perceived as attainable for the group.

The behavior of the randomly chosen team monitors is also of note when evaluating their impact and in comparing outcomes with the standard public good game. We see that monitors do in fact decide to punish participants who free-ride. This behavior is illustrative of the expectations of participants on others' behavior, especially in this setting, where participants work in crews in their daily schedules and have a clear understanding of the importance of every crew member's effort to complete the tasks.

Two main insights emerge from the experiment to guide reforms in utility authorities, where crews and teams are vital for the provision of services. First, choosing monitors for the teams (randomly or in some setting-dependent acceptable mechanisms) can lead to improvements in individual prosocial behavior under some conditions. And second, setting a common goal for a crew encourages prosocial behavior by the individual employees within the group, and may be more effective than introducing external monitoring. These insights, when used in combination (or in isolation) by the management team, have the potential to improve employee performance standards.

However, the limitations of the extrapolation of lab-in-the-field exercises to actual workplace design need to be acknowledged. First, the (partial) anonymity of the experimental setting abstracts from potential personal incompatibilities among crew and branch members. The history of employees within the institution is likely to play a key role in their reaction to workplace arrangements. Second, the experiment allows for an anonymous monitor, which may not be feasible in reality. Last, in the experimental setting, it is easy to set a threshold for the total contribution, however, the risk faced in service provision makes it difficult

games. Table A4 shows that the randomly selected monitors, who did not know they would be monitors at the time they contributed to the different games, *do not* make significantly different individual contributions to the games than other players in the session. This shows that the monitor's behavior in the games was not significantly different from the other players.

to set thresholds on performance, which is the main argument against piece-rate payments in public utilities.

This experiment does not examine the dynamics of cooperation, since the games are played only once. Moreover, the outcome of the games was not known until the end of the experimental session. We see that when the Monitor game is played first, it generates a smaller increase in contributions than when it is played second. This is in comparison with the Benchmark game, which may be linked to the Threshold game setting a reference. However, we are not able to infer anything about employees' reactions after the threshold has not been reached, or after punishment by the monitor. The analysis of these dynamics is left for future research.

References

Ashraf, N. and Bandiera, O. (2018). Social incentives in organizations. *Annual Review of Economics*, 10(1):439–463.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1 – 17.

Ashraf, N., Bandiera, O., and Lee, S. (2018). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services.

Billinger, S. and Rosenbaum, S. M. (2023). On the limits of hierarchy in public goods games: A survey and meta-analysis on the effects of design variables on cooperation. *Journal of Behavioral and Experimental Economics*, 107:102081.

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

Camerer, C. F. and Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 97:55–95.

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

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352.

Charness, G. and Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, 26(3):157–176.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Dannenberg, A., Löschel, A., Paolacci, G., Reif, C., and Tavoni, A. (2015). On the provision of public goods with probabilistic and ambiguous thresholds. *Environmental and Resource economics*, 61(3):365–383.

d'Adda, G., Darai, D., Pavanini, N., and Weber, R. A. (2017). Do leaders affect ethical conduct? *Journal of the European Economic Association*, 15(6):1177–1213.

Eisenkopf, G. and Kölpin, T. (2023). Leading-by-example: a meta-analysis. *Journal of Business Economics*.

Finan, F., Olken, B., and Pande, R. (2017). Chapter 6 - the personnel economics of the developing state. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, volume 2 of *Handbook of Economic Field Experiments*, pages 467 – 514. North-Holland.

Gershkov, A. and Winter, E. (2015). Formal versus informal monitoring in teams. *American Economic Journal: Microeconomics*, 7(2):27–44.

Goette, L., Huffman, D., and Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, 4(1):101–115.

Halac, M., Kremer, I., and Winter, E. (2021). Monitoring teams.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of political Economy*, 111(3):465–497.

Herbst, D. and Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science*, 350(6260):545–549.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, pages 324–340.

Jack, B. K. and Recalde, M. P. (2015). Leadership and the voluntary provision of public goods: Field evidence from bolivia. *Journal of Public Economics*, 122:80–93.

Kosfeld, M. and Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in ethiopia. *American Economic Review*, 105(2):747–83.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Marx, L. M. and Squintani, F. (2009). Individual accountability in teams. *Journal of Economic Behavior & Organization*, 72(1):260–273.

Mbiti, I. and Serra, D. (2018). Health workers' behavior, patient reporting and reputational concerns: Lab-in-the-field experimental evidence from kenya.

Villeval, M. C. (2020). Public goods, norms and cooperation. In *Handbook of Experimental Game Theory*. Edward Elgar Publishing.

Webb, M. D. (2013). Reworking wild bootstrap based inference for clustered errors. Technical report, Queen's Economics Department Working Paper.

Weng, Q. and Carlsson, F. (2015). Cooperation in teams: The role of identity, punishment, and endowment distribution. *Journal of Public Economics*, 126:25–38.

Table 1: **Descriptives of the experiment participants**

	N. obs.	**Mean**	**sd**
Part A: Socio-Economic characteristics			
Male	348	.816092	.387967
Married	348	.6551724	.4759964
Never married	348	.2873563	.4531808
Primary	348	.1609195	.3679858
Secondary	348	.204023	.4035661
Technical diploma	348	.3850575	.4873096
First degree and above	348	.2413793	.4285359
Language Amharic	348	.7557471	.430262
Lives in Addis	348	.8448276	.3625903
Family breadwinner	348	.8649425	.3422771
Owns dwelling	348	.2442529	.430262
Public rent	348	.1522989	.3598277
Private rent	348	.5316092	.4997184
Part B: Employment – AAWSA departments			
Water services:			
Line installation	348	.1695402	.3757687
Non-revenue water	348	.0632184	.2437058
Water consumer service	348	.25	.4336362
Sewerage:			
Line installation	348	.0775862	.2679045
Sewer connection	348	.0114943	.1067468
Sewerage consumer service	348	.2126437	.4097668
Support department:			
Human resources	348	.0114943	.1067468
Finance	348	.0287356	.1673031
Procurement	348	.0086207	.0925797
General service	348	.0689655	.2537604
Planning and budgeting	348	.0086207	.0925797
Monitoring and evaluation	348	.0057471	.0757005
Non specified:			
Other	348	.0833333	.2767834

Notes: Descriptive statistics are only available for the participants (86%, 348 out of the 405 participants in the experiment) that participated in a parallel employee survey with detailed information on job history.

Table 2: **Summary of experiment results**

Part A: Results at the individual level

	N	**mean**	**sd**
Game Room	405	3.869136	2.562131
Game Table	405	4.424691	2.661121
Game Threshold	405	5.760494	2.528926
Game Monitor	405	5.212346	2.735782

Part B: Results at the group level

	N	**mean**	**sd**
Game Room	81	19.34568	8.293311
Game Table	81	22.12346	8.513787
Game Threshold	81	22.79012	15.43593
Game Monitor	81	26.06173	7.121

Part C: Monitor choices

	N	**mean**	**sd**
Punish at least one player in the group	81	0.7283	0.4475
Number players punished when at least one	59	2.1355	1.09004

Notes: Descriptive statistics of the experimental outcomes. Panel A shows the average contributions in each of the games. Panel B presents the average contributions at the group level, where it is taken into account that in case a group did not reach the threshold all the contributions are lost. Panel C summarizes the behavior of the participants randomly chosen for the Monitor role.

Table 3: **Public good contributions, Individual and group outcomes**

	(1)	(2)	(3)	(4)	(5)
	Individual contribution			Group outcome	
Game Monitor	0.788***	0.825***	0.788***	3.938***	3.938***
	(0.151)	(0.141)	(0.179)	(0.755)	(0.897)
	[0.000]	[0.000]	[0.001]	[0.000]	[0.001]
Game Threshold	1.336***	1.359***	1.336***	0.667	0.667
	(0.208)	(0.207)	(0.246)	(1.530)	(1.818)
	[0.000]	[0.000]	[0.000]	[0.670]	[0.719]
Constant	4.425***	4.692***	4.425***	22.123***	22.123***
	(0.093)	(0.378)	(0.110)	(0.594)	(0.707)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
H0: Leader=0 (WCB p-value)	0.000	0.000	0.000	0.000	0.000
H0: Threshold=0 (WCB p-value)	0.000	0.000	0.000	0.660	0.641
T-test Monitor = Threshold	0.0342	0.0489	0.0674	0.0637	0.1124
Observations	1,215	1,044	1,215	243	243
Number players	405	348	405	–	–
Number groups	–	–	–	81	81
R-squared	0.265	0.292	0.664	0.116	0.616
Session FE	–	–	–	Yes	No
Group FE	Yes	Yes	No	No	Yes
Player FE	No	No	Yes	–	–
Individual controls	No	Yes	No	–	–

Notes: Robust standard errors, cluster session. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. P-values in brackets. Reference in all columns is the game played with table partners. Unit of observation for columns (1) to (3) is the choice of a participant in a given game. Unit of observation for columns (4) and (5) is the group outcome. Given the relatively small number of clusters (15 sessions), we report P-values obtained using a wild cluster bootstrap (WCB) with Webb's Weights for robustness. Individual controls included are branch, number of participants known in the group and sub-department with technical street tasks. Column (2) that includes individual controls has a smaller number of observations due to lack of controls information for 14% of participants.

Table 4: **Monitor behavior**

	(1)	(2)	(3)	(4)	(5)	(6)
	Group Level				**Player Level**	
	Any punishment all groups		Number punished if number > 1		Player punished inside punished group	
Group contribution	-0.005413	-0.005741	-0.01340	-0.01075		
	(0.0126)	(0.0123)	(0.0293)	(0.0293)		
Std. dev. group contributions		0.07856		0.2031		
		(0.0717)		(0.1649)		
Individual contributions					-0.07456 **	
					(0.0256)	
% deviation from max group contributor						0.6189 **
						(0.2210)
Constant	0.8695 **	0.6864	2.4850 ***	1.8990 **	0.8161 ***	0.2096 **
	(0.3286)	(0.4375)	(0.7652)	(0.7633)	(0.1338)	(0.0777)
Observations	81	81	59	59	295	295
R^2	0.1485	0.1709	0.2868	0.3070	0.3217	0.3 214
Session FE	Yes	Yes	Yes	Yes	–	–
Group FE	–	–	–	–	Yes	Yes

Notes: Robust standard errors, cluster session. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. For columns (1) to (4) the unit of observation is a group, the dependent variable is the behavior of the monitor randomly and anonymously allocated to this group. For columns (5) and (6) the unit of observation is a player, and the sample is the players in groups where at least one member was punished by the monitor.

Table A1: **Descriptives of the experiment participants by branch**

	Addis Ketema		**Arada**		T-test Difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
Water services					
Line installation	178	0.169	170	0.171	-0.002
		(0.028)		(0.029)	
Non revenue water	178	0.084	170	0.041	0.043*
		(0.021)		(0.015)	
Water Costumer Service	178	0.275	170	0.224	0.052
		(0.034)		(0.032)	
Sewerage					
Line installation	178	0.073	170	0.082	-0.009
		(0.020)		(0.021)	
Sewer connection	178	0.000	170	0.024	-0.024**
		(0.000)		(0.012)	
Sewerage Costumer Service	178	0.185	170	0.241	-0.056
		(0.029)		(0.033)	
Support department					
Human resources	178	0.011	170	0.012	-0.001
		(0.008)		(0.008)	
Finance	178	0.022	170	0.035	-0.013
		(0.011)		(0.014)	
Procurement	178	0.006	170	0.012	-0.006
		(0.006)		(0.008)	
General service	178	0.062	170	0.076	-0.015
		(0.018)		(0.020)	
Planing and Budgeting	178	0.017	170	0.000	0.017*
		(0.010)		(0.000)	
Monitoring and Evaluation	178	0.006	170	0.006	-0.000
		(0.006)		(0.006)	
Non-specified					
Other	178	0.090	170	0.076	0.013
		(0.021)		(0.020)	

Notes: Descriptive statistics are presented for the participants (87.3%) that participated in a parallel employee survey with detailed information on job history. The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A2: Between-subjects analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Round 3 individual choices				Round 4 individual choices			
Monitor first	-1.0086***	-1.1446**	-1.0221**	-1.0612**	0.7904***	0.7120***	0.9209***	0.9873**
	(0.2668)	(0.4031)	(0.4248)	(0.4412)	(0.2926)	(0.1975)	(0.2466)	(0.3546)
Number participants Know in group		-0.1241	-0.1235	-0.1150		0.08878	0.08976	0.1187
		(0.0916)	(0.0922)	(0.1117)		(0.0730)	(0.0714)	(0.0891)
Sub-Department team work		-0.3473	-0.2469	-0.3476		-0.6347**	-0.4635	-0.6355**
		(0.2429)	(0.2743)	(0.2429)		(0.2607)	(0.2897)	(0.2602)
Arada branch		-0.02343	-0.01936	-0.02296		0.2698	0.2767	0.2713
		(0.2093)	(0.2179)	(0.2115)		(0.3014)	(0.3012)	(0.2998)
Order × Team Department			-0.6549				-1.1162**	
			(0.4776)				(0.4569)	
Order × participants know				-0.03686				-0.1216
				(0.1801)				(0.1340)
Constant	3.5856***	3.9487***	3.8946***	3.9254***	3.2477***	3.1700***	3.0778***	3.0932***
	(0.2709)	(0.4262)	(0.4544)	(0.4921)	(0.2289)	(0.3440)	(0.3517)	(0.3710)
Observations	405	348	348	348	405	348	348	348
R^2	0.2560	0.2923	0.2942	0.2924	0.2457	0.2495	0.2540	0.2500
Individual controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes

Notes: Robust standard errors, cluster session. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Unit of observation is the choice of a participant in a given game: columns 1 to 4 have as dependent variable the choice in round 3, and columns 5 to 8 the choice in round 4. Individual controls included are branch, how many participants in the group the participant knows, and whether the employee works in a department where team work is more prominent. Choice in round 2 of the game (game played with table partners) is included as control in all specifications.

Table A3: **Public good contributions, Individual outcomes**

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var: Individual contributions						
Threshold (β^T)	1.359***	1.359***	1.193***	1.193***	1.183***	1.183***
	(0.199)	(0.176)	(0.208)	(0.189)	(0.209)	(0.194)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Threshold * Monitor first (β^{TO})			0.550	0.550	0.712*	0.712
			(0.477)	(0.416)	(0.398)	(0.429)
			[0.268]	[0.191]	[0.095]	[0.101]
Monitor (β^M)	0.825***	0.825***	1.027***	1.027***	1.023***	1.023***
	(0.136)	(0.148)	(0.127)	(0.152)	(0.118)	(0.174)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Monitor * Monitor first (β^{MO})			-0.922***	-0.922***	-0.802***	-0.802***
			(0.170)	(0.272)	(0.200)	(0.279)
			[0.000]	[0.001]	[0.001]	[0.005]
Order: Monitor first	-0.261	-0.261	0.147	0.147	-0.231	-0.231
	(0.449)	(0.332)	(0.534)	(0.430)	(0.531)	(0.428)
	[0.570]	[0.434]	[0.788]	[0.734]	[0.670]	[0.591]
Constant	4.480***	4.480***	4.387***	4.387***	4.472***	4.472***
	(0.387)	(0.329)	(0.183)	(0.221)	(0.394)	(0.336)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
T-test: Threshold = Monitor	0.0416	0.0025	–	–	–	–
F-test $\beta^T + \beta^{TO} = 0$	–	–	0.0012	0.0000	0.0001	0.0000
F-test $\beta^M + \beta^{MO} = 0$	–	–	0.3691	0.6434	0.1925	0.3145
Observations	1,044	1,044	1,215	1,215	1,044	1,044
R-squared	0.054	0.054	0.051	0.051	0.064	0.064
Ind. controls	Yes	Yes	No	No	Yes	Yes
Cluster	Session	Group	Session	Group	Session	Group

Notes: Robust standard errors ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. P-values in brackets. Reference in all columns is the game played with table partners. Unit of observation is the choice of a participant in a given game. Individual controls included are Branch, number of participants known in the session and belonging to a sub-department where team work is more relevant. F-tests of joint significance of (β^T and β^{TO}) and (β^M and β^{MO}) in pairs, and the four together give p-values of 0.0000

Table A4: **Monitors contributions in the games**

	(1)	(2)	(3)	(4)
	Room	Table	Threshold	Monitor
Monitor	0.5844 *	0.7537 *	0.07139	0.04946
	(0.3200)	(0.3800)	(0.3649)	(0.4437)
Constant	3.7508 ***	4.2721 ***	5.1979 ***	5.7505 ***
	(0.0648)	(0.0769)	(0.0739)	(0.0898)
Session FE	Yes	Yes	Yes	Yes
Observations	405	405	405	405
R^2	0.0776	0.0884	0.0697	0.0523

Notes: Robust standard errors. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The unit of observation is a player.

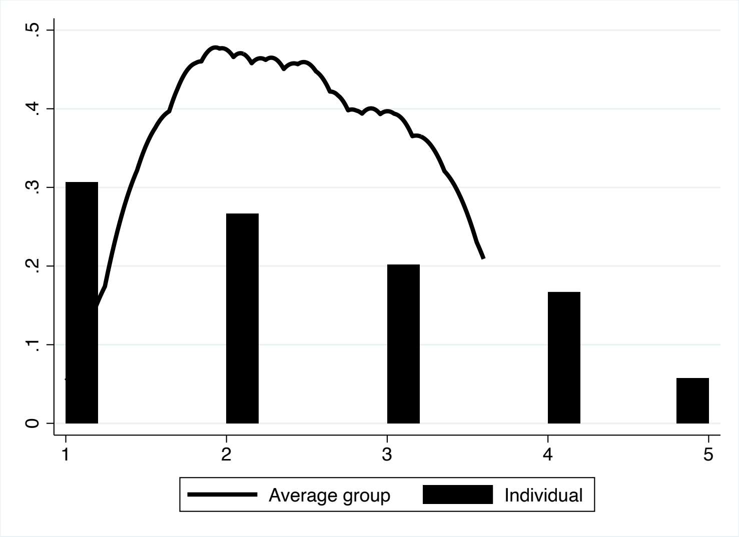

Figure A1: Knowledge of other participants in the group.

The histogram shows the answer of the participants to the post-experimental survey question *"Including yourself, how many participants in your group do you know?"*. The line represents the kernel density of the average of this question by group. On average participants know 2.5 of the members of their 5 person game group.

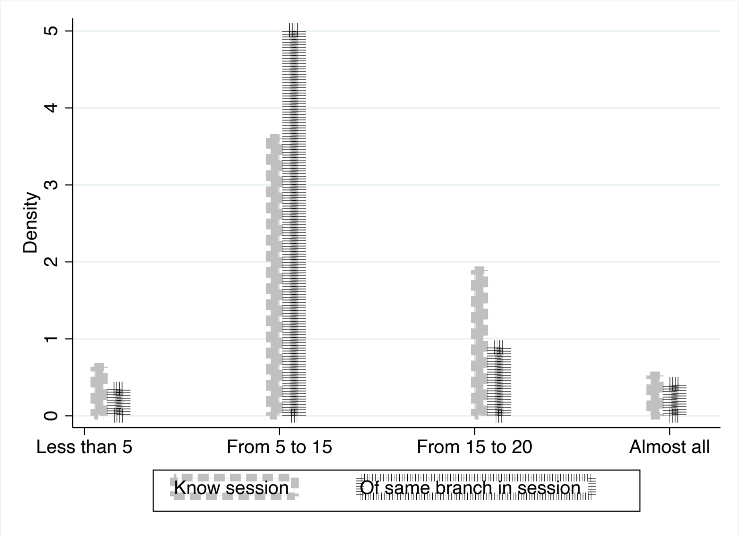

Figure A2: Knowledge of other participants in the session.

The histogram shows the answer of the participants to the post-experimental survey question *"How many of the participants in the session do you know?"* (lighter bar) and *How many participants in the session work at the same AAWSA branch?"* (darker bar).

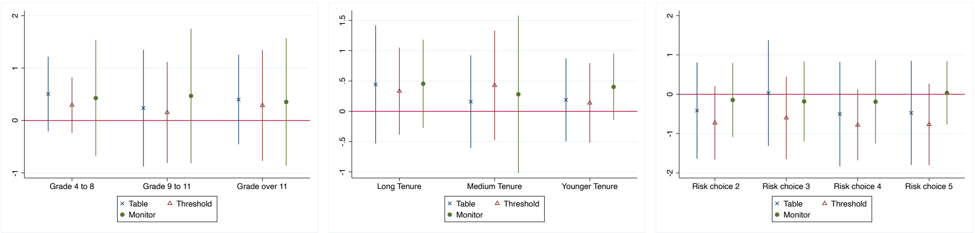

Figure A3: Heterogeneity of individual contributions.

The figures show coefficients of estimating $C_{igt} = \alpha + \beta_i * \sum_i Variable_i + \gamma_t + \epsilon_{igt}$, where $Variable_i$ is a set of dummies on the individual characteristic of interest, and C_{igt} is contribution of player i in group t in game g. Standard errors clustered at the session level. Tenure dummies are defined as per tenure quartiles being the reference the shortest tenure quartile. Risk choices come from post-experimental risk-elicitation game, being the reference the safest alternative offered (choice 1), and with alternatives increasing in risk. Results from the restricted sample (348 out of 405 participants) for which there is information on parallel employee survey.