Theory and Decision
https://doi.org/10.1007/s11238-021-09864-7

Accountability as a Warrant for Trust: An Experiment on Sanctions and Justifications in a Trust Game

Kaisa Herne[1] · Olli Lappalainen[2] · Maija Setälä[3] · Juha Ylisalo[3]

Accepted: 23 December 2021

Abstract

Accountability is present in many types of social relations; for example, the accountability of elected representatives to voters is the key characteristic of representative democracy. We distinguish between two institutional mechanisms of accountability, i.e., opportunity to punish and requirement of a justification, and examine the separate and combined effects of these mechanisms on individual behavior. For this purpose, we designed a decision-making experiment where subjects engage in a three-player trust game with two senders and one responder. We ask whether holding the responder accountable increases senders' and responders' contributions in a trust game. When restricting the analysis to the first round, the requirement of justification seems to have a positive impact on senders' contributions. When the game is played repeatedly, the experience of previous rounds dominates the results and significant treatment effects are no longer seen. We also find that responders tend to justify their choices in terms of reciprocity, which is in line with observed behavior. Moreover, the treatment combining punishment and justification hinders justifications that appeal to pure self-interest.

Keywords Trust game · Punishment · Altruism · Reciprocity · Accountability

✉ Olli Lappalainen
oljula@utu.fi

Kaisa Herne
kaisa.herne@tuni.fi

Maija Setälä
maiset@utu.fi

Juha Ylisalo
jtylis@utu.fi

[1] Faculty of Management and Business, Tampere University, 33014 Tampere, Finland

[2] Turku School of Economics, University of Turku, 20014 Turun yliopisto, Finland

[3] Faculty of Social Sciences, University of Turku, 20014 Turun yliopisto, Finland

Published online: 31 January 2022

1 Introduction

Accountability is an essential element of various relationships in public and private spheres, including representative politics, public administration, service production, global governance, voluntary associations, and private companies (e.g., Bäckstrand, 2008; Kuyper & Bäckstrand, 2016; Warren, 1996a). Different institutional mechanisms of accountability have been identified in previous studies, and a variety of institutional designs have been formulated to secure accountability in different contexts. This variety also reflects different understandings of accountability in the literature (Bovens, 2010; Bovens et al., 2014).

The common assumption about *public* accountability relations is that accountability solves problems associated with the delegation of powers and responsibility, because it brings the actions of the agents in line with the expectations of the principals (e.g., Manin, et al., 1999). Different conceptualizations of public accountability refer to two fundamental accountability mechanisms involved in different institutional designs: the threat of (material) *sanctions* and the requirement that those who are held accountable should *justify* their actions. Different strands of literature tend to emphasize one of these mechanisms and consider it "the" basis of public accountability, even though the need to combine these mechanisms is often acknowledged. In general terms, rational choice theory and economics-based approaches tend to put emphasis on sanctions (e.g., Besley, 2006), while theories of deliberative democracy (e.g., Gutmann & Thompson, 1996) and social psychology (e.g., Lerner & Tetlock, 1999) tend to emphasize the requirement of a justification. Our aim is to bring these two accountability mechanisms together by studying both the independent and combined effects of sanctions and justifications.

We investigate the two mechanisms of accountability in a laboratory experiment where participants play the trust (investment) game (Berg et al., 1995). In this game, a sender can show trust by allocating resources to a responder who, in turn, can show trustworthiness by returning resources. Sending and returning are mutually beneficial because resources are multiplied if transferred from the sender to the responder. Instead of a standard two-player trust game, we use a modification that involves two senders and one responder. This experimental design shares a key characteristic of mechanisms involved in public accountability relations where public officials make decisions that affect a number of individuals. However, in designing the experiment, we deliberately abstracted away from specific procedures, such as elections, to be able to study the fundamental mechanisms of public accountability, rather than examining specific institutional designs.

The aim of this article is to examine whether the threat of sanctions and the requirement of a justification enhance behavioral trust. We are interested in what Warren (1996b) calls "warranted trust", that is, trust that is shown when specific mechanisms guarantee that those who are trusted behave in a desired manner. We study the influence of each accountability mechanism alone as well as their combined effect. As we will point out below, there are reasons to expect that the requirement of a justification backed with the threat of sanctions provides the most influential form of accountability, because it incentivizes the agents to act in ways

they can publicly justify. In addition, we explore what kinds of justifications are formulated, and how the possibility of facing sanctions affects the content of justifications.

Our main observation is that obliging responders to justify their decisions induces the highest levels of contributions. We also find that responders tend to justify their choices especially in terms of reciprocity, which is in line with their behavior: the more a sender contributes, the larger is the amount a responder tends to return to the specific sender. Moreover, the risk of facing a punishment appears to discourage responders from giving justifications appealing to their self-interest. Our results have implications for the real-world design of institutions of public accountability by especially showing the effects of justification as an accountability mechanism and by showing the tendencies towards reciprocity.

2 Conceptualizations of trust and accountability

Trust and accountability are clearly distinct phenomena. According to Warren (1996b: 4), "[w]hen one trusts, one forgoes the opportunity to influence decision-making, on the assumption that there are shared or convergent interests between truster and trustee". In a similar vein, Berg et al. (1995) define trust in terms of "*belief* in reciprocity" (italics added). In contrast, accountability entails mechanisms intended to ensure that an accountable actor behaves in a manner that the principals require, even in case of divergent or opposing interests. In this respect, mechanisms of accountability can enhance what Warren (1996b: 20) calls "warranted trust". In other words, accountability mechanisms function as "protections and inducements" that help manage the risks involved when placing trust in an actor. For example, in representative democracies, trust in elected representatives is "warranted" because it is secured by specific accountability mechanisms allowing voters to influence or react to representatives' actions when they do not align with their interests (Warren & Gastil 2015).

Different strands of literature focus on different mechanisms of accountability. In political economy and rational choice theory, accountability is primarily understood as a mechanism based on material sanctions or rewards (Fearon, 1999: 55). Besley (2006: 37) defines accountability in terms of the opportunity of the public to punish decision-makers: "A politician is formally accountable if there is some institutional structure that allows the possibility of some action to be taken against him/her (such as being voted out of office) in the event that he/she does poor job." Besley's definition exemplifies a prominent feature of the formally oriented literature on democracy, namely that elected representatives are expected to act in the interests of voters because of the risk of not being re-elected.[1] In addition to material sanctions, various types of immaterial or social sanctions, such as reputational effects, may also motivate those holding public offices (Colombo, 2018; Lerner & Tetlock, 1999).

[1] However, there are theoretical studies showing that voters' ability to use the "electoral weapon" effectively is limited (Barro 1973; Fearon 1999; Ferejohn, 1986).

Although the essence of these approaches to accountability is the risk of utility losses, another approach to accountability is to emphasize the requirement of public justification and reasoning as the key accountability mechanism. According to Philp (2009: 29) "A is accountable with respect to M when some individual, body or institution, Y, can require A to inform and explain/justify his or her conduct with respect to M." According to this view, what defines an accountability relationship is Y's capacity to require A to give an account. In theories of deliberative democracy, the requirement of a public justification is the key feature of accountability. Gutmann and Thompson (1996: 128) define accountability as the requirement of citizens and decision-makers to give justification for their decisions to all those who are bound or significantly affected by them.

Real-world public accountability relations typically involve both mechanisms of accountability. Schedler (1999) argues that those who are accountable are obliged to inform the public about their decisions, as well as to explain and justify their decisions. The public in turn is empowered to monitor decision-makers' conduct and force them to "bear the consequences" in the form of sanctions. Thus, "A is accountable to B when A is obliged to inform B about A's (past or future) actions and decisions, to justify them, and to suffer punishment in the case of eventual misconduct" (Schedler, 1999: 17). Rehfeld (2005: 189–190) argues that a sanctioning dimension is necessary for "any reasonable account of legitimate political representation" and that the discursive mechanism only complements it. Without sanctioning, a representative could act according to his or her self-interest and even justify his or her conduct in these terms.

To sum up, various strands of research give different weights to different mechanisms of accountability and institutional designs supporting them. At present, there is relatively little experimental research on how different aspects of accountability interact with each other and, importantly, how they affect individual action and reasoning. However, we can expect that both sanctions and the requirement of justification enhance trust and thereby cooperation to some extent, but that together they constitute a strong mechanism with the clearest behavioral effects. The existing experimental evidence, to which we will now turn, provides preliminary support for this expectation.

3 Previous experimental studies on sanctions and justifications

The absence, or insufficiency, of accountability mechanisms can lead to collectively sub-optimal outcomes by leaving room for decision-makers' self-serving choices. In experimental studies on electoral accountability, the sanctioning mechanism tends to be reduced to a dichotomous choice between voting for or against an incumbent. Experimental results on electoral accountability provide somewhat ambiguous results. Previous studies suggest that people tend to resort to retrospective voting and punish decision-makers for outcomes they dislike (Landa, 2010; Woon, 2012). Some studies show that electoral mechanisms actually decrease the amounts decision-makers distribute. Arguably, this is due to an effect whereby being elected is perceived as an entitlement to use resources in a self-serving way (Geng et al.,

2011). Others have called into question the result concerning subjective entitlements (Weiss & Wolff, 2013).

In experimental studies on electoral accountability, decision-makers' messages to recipients are often intended to reflect campaign promises, which have been shown to increase the electorate's payoffs (Corazzini et al., 2014; Feltovich & Giovannoni, 2015). Although promises include statements about future action, justifications refer to statements about and reasons for actions that are given simultaneously with an action or retrospectively. Justifications have mainly been understood in terms of blame avoidance and blame management, where the primary interest lies in the conditions under which members of the public are likely to accept explanations that representatives give for damaging, scandalous, or otherwise undesired policy outcomes (McGraw, 1991; McGraw et al., 1993).

Beyond the electoral setting, the influence of costly punishment has been studied in experimental games where individual rationality and collectively optimal action conflict. A number of studies have shown that the possibility of a punishment increases contributions to public goods (e.g., Ostrom et al., 1992; Fehr and Gächter, 2000; Hamman et al., 2011; Lierl, 2016). However, evidence on the trust game suggest that the effect of sanctions is sensitive to the specific design of the experiment, and that sanctions do not always increase trust and trustworthiness (Calabuig et al., 2016; Charness et al., 2008; Fehr & List, 2004; Fehr & Rockenbach, 2003; Houser et al., 2008; Rigdon, 2009). It seems that while sanctions can enhance cooperation in certain contexts, they may also give rise to self-interested choices if players understand them as a price paid for acting selfishly, or if they dampen their intrinsic motivation to cooperate (Houser et al., 2008).

The effects of messages and communication on reasoning and behavior have been studied in a variety of experimental designs. There is a large number of experimental studies on the behavioral consequences of various types of communication in settings where people interact (Bó & Bó 2014; Cason & Gangadharan, 2016; for meta-analyses, see Sally, 1995; Balliet, 2010). Although evidence on the communicative aspect of accountability in social dilemmas is limited, some studies look at the consequences of monitoring and justifications in trust and public good games. Bracht and Feltovich (2009) find that "cheap talk" in the form of messages from the responder before the sender's decision does not have much effect on either player's behavior in a two-person trust game, whereas the opportunity of senders to observe responders' actions in the previous round significantly increases the amounts returned. Experimental evidence on the public good game suggests that an obligation to justify one's choice to other participants increases contributions in particular among those who have larger endowments (De Cremer & van Dijk, 2009). Other studies demonstrate that being required to justify one's choice increases norm-abiding behavior in commons dilemmas and games testing deviations from social norms (de Kwaadsteniet et al., 2007; Xiao 2017).

The key idea of our experimental design is to study the behavioral effects of the requirement of *justification* and *material sanctions*, as well as their interaction. While there are no previous studies with the same experimental design, the relationship between different types of communication (which can vary from an abstract "signal" to a completely free-form communication, cf. Brandts et al. (2019)

for a review), and punishment has been widely studied in the previous literature on social dilemmas. For example, Bochet et al. (2006) found that chat room communication had almost as strong effects on enhancing cooperation as face-to-face communication. They also report that adding a punishment option to a chat room treatment raised contributions only moderately. Dufwenberg et al. (2021) had subjects record a single pre-play message, and report that promises drive the effect of communication on beliefs, and that broken promises lead to higher rates of costly punishment. Furthermore, Ostrom et al. (1992) report that communication increases yields but that overuse of sanctioning and sanctioning without communication reduces net yields in a common pool resource experiment. However, when subjects agree on a joint investment strategy into the common pool and choose their sanction mechanism, they achieve close to optimal results. Likewise, in their experiment on spatial common pool resources, Janssen et al. (2010) found that participants use costly punishment if presented an option to do so, but without communication this does not increase gross payoffs. However, when communication is allowed, performance increases significantly, but it is not sustained if communication is not possible anymore and punishment is available.

Social psychological studies typically focus on the effects of the requirement of justification on individual reasoning. Based on the literature review, Lerner and Tetlock (1999) discuss the conditions under which accountability might amplify biases in reasoning rather than enhance open-mindedness and critical thinking. The authors argue that "experimental work has repeatedly shown that expecting to discuss one's views with an audience whose views are known led participants to strategically shift their attitudes toward that of the audience" (Lerner & Tetlock, 1999: 256). However, when the views of the audience are not known, accountability tends to give rise to what Tetlock (1983) has called "preemptive self-criticism", or increased awareness of their own decision processes and anticipation of potential counter-arguments. More recently, Mercier, et al. (2015) have found support for "the argumentative theory of reasoning", according to which people tend to be lazy when considering their own arguments but critical towards those of others. This highlights the importance of various feedback mechanisms such as dialogue or, as is the case in our experiment, material sanctions.

4 Experimental design and hypotheses

In the standard two-person trust game (Berg et al., 1995), two subjects are randomly assigned into the roles of a sender and a responder and given an endowment of, say, ten units. At the first stage of the game, the sender decides an amount x ($0 \leq x \leq 10$) he or she sends to the responder. The sender keeps $10–x$, and x is tripled by the experimenter to create benefits of cooperation, so that the responder gets $3x$. At the second stage of the game, the responder passes on an amount y ($0 \leq y \leq 3x$) to the sender, and keeps $3x–y$. The sender's choice is understood to model trust, that is, whether the sender believes the responder to reciprocate, and the behavior of the responder is understood to model trustworthiness. Another interpretation of players' behavior is that senders take a risk by sending money.

Both players can earn more money if senders send and responders return money. The standard behavioral result is that trust is frequently observed but that slight variations in the design can produce substantial changes in contributions (Johnson & Mislin, 2011).

We use a three-player variant of the trust game with two senders and one responder, where senders move first and have an opportunity to send money to the responder. The responder moves second and can return money to the senders, possibly returning different amounts to each sender. In the experiment, each participant played the game for *six rounds,* each round consisting of *three stages*. Both senders and responders were first endowed with 12 points (2 points = 1 euro) in the beginning of the first stage of each round.[2] The senders then decided whether and how much to send to the responder. The amount sent was tripled by the experimenter and passed on to the responder. At the second stage of each round, the responder decided whether and how much to return to each sender, that is, how to divide the total amount sent, tripled by the experimenter, plus his or her initial endowment of 12 points between the two senders and him- or herself. Each round ended with a third stage where both senders and responders were given extra 12 points. In the two treatment conditions including the opportunity to punish, the senders could use these extra points to punish the responder. Extra points were also given in treatment conditions without the punishment opportunity to make sure that possible behavioral differences across conditions do not depend on the extra points. The senders made both allocation and punishment decisions individually and simultaneously, so they could not coordinate their choices. The game tree is displayed in Fig. 1.

Our experiment follows a 2 (opportunity to punish) × 2 (requirement of justification) factorial design. In *Baseline*, the three-person trust game was played without punishment possibilities or justification requirements. In *Punishment*, the senders could use the extra points received in the third stage to punish the responder. Punishment was costly, and one point used for punishment decreased the responder's earnings by three points for that round. As a consequence of punishment, a responder's earnings could go down to zero. The costliness of punishments captures an essential feature of most real-world accountability relationships where special efforts are needed in sanctioning the decision-maker. Further, in experimental research, costly punishment is a standard mechanism because it hampers the use of punishment randomly, or for irrelevant reasons.

In *Justification*, responders were asked to write a free-form justification for their decision in an open space. The justification was shown to the senders along with the responder's allocation decision. Finally, in *Justification + Punishment*, the responder was asked to write a justification for his or her decision. The senders had an opportunity to punish the responder after the responder's allocation decision and the

[2] Each point earned paid 50 cents. Amounts sent and returned as well as punishments were constrained to be whole numbers of points.

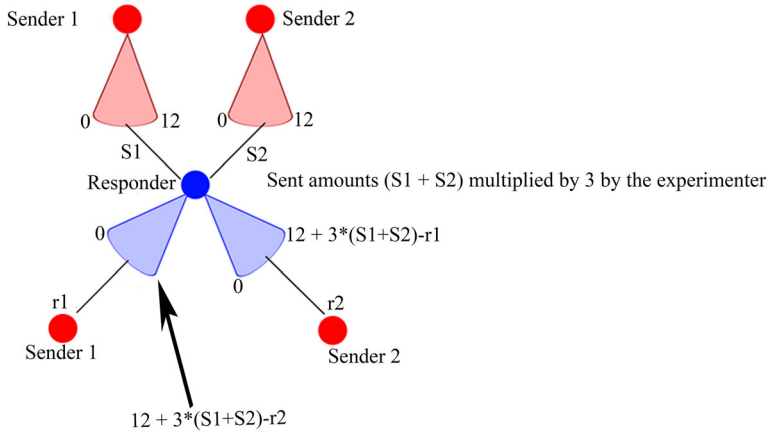

Fig. 1 The structure of the three-player trust game in *Baseline*. In the first round, the amounts sent by the senders (s1 and s2, respectively) are multiplied by three by the experimenter. The responder can then return to sender 1 any amount r1 between 0 and his initial endowment (12) plus the total amount received in the first round (3 × (s1 + s2)), minus the amount returned (r2) to the sender 2

corresponding justification were revealed.[3] The justification was thereby given before the responder knew about the senders' reactions, which allows an analysis of punishments as a feedback mechanism.

In all conditions including punishment, justification, or both, senders were aware of these mechanisms when making their initial decisions implying that senders could anticipate the potential effect of the accountability mechanisms on responder behavior. Table 1 presents the design of the experiment.[4]

We used the three-player variant of the trust game to capture the asymmetries typical in public accountability relations where a small number of decision-makers are entitled to make decisions that affect several individuals so that the decision-makers are able to *discriminate* between these individuals. To be more specific, in a two-player trust game, the responder can be reciprocal toward the sender and return money in proportion to the sender's allocation. In the three-player variant, the responder can likewise be reciprocal by returning money to each sender in proportion to their allocations, but the responder can also return money without reacting to what the senders have done, e.g., split the returned money equally between the senders. It is thereby possible to distinguish between different reaction strategies to sender behavior. We can also investigate what kinds of verbal justifications responders give to their choices, as well as whether the verbal justifications match their choices. To take a real world example, elected representatives may make decisions that benefit either all members of the public

[3] It is noteworthy that another possibility for the *Justification + Punishment* condition would have been to ask responders to formulate their justification before they know about the punishment, but to have senders determine their punishments before they know about the justification (we thank the anonymous referees for pointing this out). However, since we were interested in the influence of the combined behavioral effects of the responder's decision and its justification, we used a condition where senders were aware of the justification when they determined their punishments.

[4] A translation of the experimental instructions is provided in Appendix A.

Table 1 The design of the experiment

	Opportunity to punish	
	No	Yes
Requirement of justification		
No	*Baseline* First stage: All players endowed with 12 points Senders' decisions Amounts tripled and sent to responders Second stage: Responder's decision Third Stage: All players given 12 points	*Punishment* First stage: All players endowed with 12 points Senders' decisions Amounts tripled and sent to responders Second stage: Responder's decision Third Stage: All players given 12 points; Senders given an opportunity to punish the responder
Yes	*Justification* First stage: All players endowed with 12 points; Senders' decisions Amounts tripled and sent to responders Second stage: Responder's decision and justification of the decision Third Stage: All players given 12 points	*Justification + Punishment* First stage: All players endowed with 12 points Senders' decisions Amounts tripled and sent to responders Second stage: Responder's decision and justification of the decision Third Stage: All players given 12 points; Senders given an opportunity to punish the responder

or only a specific subgroup, and they may or may not explain their actions in a manner that corresponds to their actions.

It must be pointed out that in the three-person trust game, senders may be motivated by mutual competition if they anticipate that the responder will return more to a sender who sends more.[5] However, this possibility is reduced by the fact that senders make their choices simultaneously, which makes it impossible for a sender to know the other sender's possible action at the time of making a decision. The senders are also randomly assigned to a new three-person game on each round, which rules out the development of reputational effects regarding sender or responder behavior. It is also noteworthy that as in the regular two-person trust game, the individually rational behavioral strategy also in this setup is to send nothing.

The possibility of punishment and the requirement of justification are operationalizations of the two basic mechanisms of accountability. *Punishment* models a situation where a decision-maker is accountable to the public in the sense that members of the public have an opportunity to sanction the decision-maker. Analogously to the possibility of the public to impose sanctions on decision-makers in real-world accountability relationships, senders in our case have an opportunity to impose monetary fines on responders. Further, senders decide individually whether

[5] We thank an anonymous referee for making this point.

and how much they sanction, which is analogous to the variance in sanctions in real-world accountability relations. *Justification* is intended to capture the requirement that decision-makers provide reasons for their actions. Finally, *Justification + Punishment* models a situation where both accountability mechanisms are in place, that is, the decision-maker is required to justify his or her actions and can also face sanctions.

4.1 Hypotheses

In each treatment condition, the subgame perfect payoff maximizing strategy is for the responder to return nothing and for the senders to send nothing. In the two treatment cells involving costly punishment, not to punish is the dominant strategy for the senders, irrespective of responder behavior in the second stage of the game. However, based on earlier studies on the trust game, we can expect that both senders and responders make contributions. Furthermore, we expect that senders will send more and responders will return more in *Punishment* and in *Justification* compared to *Baseline*.

Punishment gives senders an opportunity to reduce responders' earnings, which is likely to give responders a motivation to return money. Punishment will influence responder behavior if they anticipate that violating the social norm of returning money will lead to punishment and if they care about being punished (De Cremer et al., 2001). Evidence on public good games is rather robust in showing that the opportunity to sanction free riders increases contributions (Ostrom et al., 1992; Fehr and Gächter, 2000; Hamman et al., 2011; Lierl, 2016). Evidence on the effect on punishment in the case of trust games is somewhat more ambiguous (Calabuig et al., 2016; Charness et al., 2008; Fehr & List, 2004; Fehr & Rockenbach, 2003; Houser et al., 2008; Rigdon, 2009), suggesting that the fear of punishment does not automatically affect responder behavior.

The requirement to give a justification does not give senders an opportunity to affect responders' material well-being directly. However, it may still influence responders. Requiring responders to justify their choices may increase the likelihood of following a social norm of behaving in a trustworthy or fair manner (De Cremer et al., 2001). Responders may share more resources when a justification is required because acting fairly is easier to justify and people may care about their ability to give an account for their action (de Kwaadsteniet et al. 2007, De Cremer & van Dijk, 2009). Indeed, evidence shows that being required to justify one's choice increases norm-abiding behavior (de Kwaadsteniet et al., 2007; Xiao 2017). Responders' contributions may also be increased when justifications are required because people tend to care about pleasing one's audience (Lerner & Tetlock, 1999), and returning money accompanied with a justification for that action is likely to please the senders. If senders anticipate that the ability to punish and the requirement of a justification increase responders' likelihood of returning money, senders are likely to feel more confident about investing money. Explicitly stated, our hypotheses regarding sender and responder behavior posit that:

H1: Senders send more money and responders return more money in *Punishment* compared to *Baseline*.
H2: Senders send more money and responders return more money in *Justification* compared to *Baseline*.

Since there appears to be no previous studies comparing the relative effects of material sanctions and verbal justifications, we are not able to make a specific prediction about the difference between *Punishment* and *Justification*. However, we expect that the combination of punishment and justification encourages decision-makers to make decisions that benefit the public more than either accountability mechanism does alone, that is, responders are expected to return the most in *Justification* + *Punishment*. This expectation is based on the role of sanctions as a feedback mechanism between senders and responders, which allows senders to react to responders' decisions and to the corresponding justifications. In other words, when the requirement of a justification is accompanied with a probability of a punishment, responders are likely to anticipate that a discrepancy between their action and justification will not go unpunished. For that reason they are likely to act in a way that can be justified in an acceptable manner, i.e., return money. There is evidence which supports this anticipation by showing that a discrepancy between communication and action prompts sanctions (Dufwenberg et al. 2021). For this reason, we assume a significant interaction effect and predict that senders anticipate the largest returns from responders when both accountability mechanisms are in place. Our third hypothesis is as follows:

H3a: *Justification* induces senders to send more money when they can punish responders in the third stage of the game.
H3b: *Justification* induces responders to return more money when senders can punish responders in the third stage of the game.

In addition to testing these three hypotheses, we conduct an exploratory analysis of the contents of the justifications given by the responders. In particular, we compare justifications given when responders are merely expected to justify their choices to those justifications given when senders have an opportunity to punish. Moreover, we explore the relationship between responder behavior and the justification given for that behavior, as well as the senders' reactions to the combination of responder action and justification.

We conducted an anonymous and computerized experiment (with Z-tree; Fischbacher, 2007). Subjects' allocation to treatments was random. In the beginning of each experimental session, subjects were first randomly assigned to their seats in the decision-making laboratory. Written instructions were then handed out to each participant and read aloud. The experiment began after all participants had successfully completed a practice round. In each session, the game was played for six rounds and subjects were informed about the number of rounds in the initial instructions. The outcome of a round was revealed to the subjects immediately after the round was played, i.e., subjects were aware of the outcome of the previous round before making their decisions on rounds 2–6. Each participant was randomly allocated into the role of a sender or a responder in the beginning of the experiment,

and he or she retained that role throughout the six rounds. Because of random allocation to different roles, entitlement effects should not have impact on our results.

To avoid effects of repeated games, participants were randomly assigned into a new three-player group in the beginning of each round. Each participant was thus in the same role throughout the six rounds but the other group members most likely changed between rounds. At the end of the experiment, each participant was paid the amount he or she had earned from one round, selected by asking the participant to roll a six-sided die.

Each treatment consisted of three experimental sessions of 18 subjects, yielding 54 subjects per treatment, and a total of 216 subjects. The subjects were mostly undergraduate and graduate students of the local university (58 percent female, mean age 27.4, s.d. 6.35). Each subject participated in one experimental session only. Each session took place on a single day at the Decision-Making Laboratory (PCRClab) at University of Turku.

5 Results

We will start with an overview of the points sent and returned, followed by a closer examination of the first round to ensure independent observations. We also examine senders' and responders' behavior across the rounds in order to capture the time dynamics present within each treatment, as well as earnings and sanctions. Finally, we analyze the justifications responders gave in the two treatments where they were required.

5.1 Sender and responder behavior

Aggregated over all four treatments and six rounds, we observed a total of 432 individual plays of the three-player trust game and a total of 864 sender decisions (n = 144, or 36 senders per treatment, each making 6 decisions). Table 2 shows that the average amount of points sent was 5.70 in *Baseline,* 6.20 in *Punishment*, 6.89 in

Table 2 Average points sent and returned, and points used for punishment by treatment, all rounds

	Average sent by senders	Returned by responder to an individual sender	Points used for punishment
Baseline	5.70 (3.98)	8.00 (8.54)	N/A
Punishment	6.20 (4.37)	9.06 (9.04)	1.42 (2.84)
Justification	6.89 (3.89)	9.82 (9.05)	N/A
Justification + Punishment	7.33 (3.84)	11.73 (7.49)	1.23 (2.62)

Standard deviations in parentheses. $N = 6 \times 36$ sender decisions, 6×18 responder decisions per treatment

Justification, and 7.33 in *Justification* + *Punishment*. In accordance with a number of previous studies, senders make strictly positive contributions in our experiment.

In total, our experiment consisted of four (2 × 2) treatment cells, and the single-shot trust game was repeated 6 times in each cell, with stranger matching. We collected data on 432 allocation decision pairs (864 allocation decisions, respectively) by the *responders* (n = 72, or 18 per treatment, each responder making two decisions per round, i.e., 12 decisions). As reported in Table 2, the average amount returned for each sender was 8.00 points in *Baseline*, 9.06 in *Punishment*, 9.82 in *Justification* and 11.73 in *Justification* + *Punishment*. In *Baseline*, the average share returned by responders was 47% of the points received after the multiplication, in *Punishment* the average returned share was 49%, in *Justification* it was 48%, and in *Justification* + *Punishment* 53%.

We test first for an interaction between the treatments, restricting the analysis to the first round to ensure independent choices. The sender behavior was analyzed with a 2 (no punishment vs. punishment) × 2 (no justification vs. justification) ANOVA. The results are reported in Table 3. The ANOVA on senders' first-round choices provides tentative support to our hypothesis H2: As shown in Table 3, the main effect of the justification treatment is almost significant $F(1, 140)$, $p = 0.078$. However, the effect of the *Punishment* manipulation and the interaction effect between the *Justification* and *Punishment* treatments were not significant at the conventional level of 0.05, and thus regarding sender behavior we do not find support for H1 or H3a.

In a similar manner to our analysis of the sender behavior, we restrict the observations to the first round and analyze responders' behavior with a 2 (no punishment vs. punishment) × 2 (no justification vs. justification) ANOVA. As was the case with the senders, the responders seem to return more if treated with *Justification* $F(1, 140)$, $p = 0.051$ as shown in Table 4. However, if we include the amount received in the first stage of the first period as a covariate (Table 5), the main effect vanishes because the first stage sender allocation is a highly significant explanatory variable $F(1, 67)$, $p < 0.00001$ which accounts for most variation in the amounts of points returned. Regarding responder behavior, we therefore find support for H2, if points received are not taken into account, but when they are, the effect is no longer seen. Since *Punishment* or the interaction between *Punishment*

Table 3 Two-way ANOVA of sent allocations, observations restricted to the first round

	Mean				
	df	Sum SQ	SQ	F val	Pr(> F)
Punishment	1	14.7	14.69	1.215	0.2722
Justification	1	38	38.03	3.144	0.0784'
Punishment × Justification	1	0.1	0.11	0.009	0.9238
Residuals	140	1693.2	12.09		

Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, '$p < 0.10$

Table 4 Two-way ANOVA of returned allocations, observations restricted to the first round

	Mean				
	Df	Sum SQ	Mean SQ	*F* val	Pr(> *F*)
Punishment	1	2	2.0	0.017	0.8972
Justification	1	470	470.2	3.957	0.0507’
Punishment × Justification	1	16	16.1	0.135	0.7143
Residuals	68	8080	118.8		

Significance levels: ***$p < 0.001$, ** $p < 0.01$, *$p < 0.05$, ’$p < 0.10$

Table 5 Two-way ANCOVA of returned allocations, observations restricted to the first round, points received in the first stage as a covariate

	Mean				
	Df	Sum SQ	SQ	F val	Pr(> *F*)
Points rec. 1st stage	1	3382	3382	44.943	5.11e−09***
Punishment	1	36	36	0.482	0.49
Justification	1	97	97	1.291	0.26
Punishment × Justification	1	11	11	0.152	0.968
Residuals	67	5042	75		

Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $'p < 0.10$

and *Justification* are not significant, H1 and H3b are not supported in the case of responder behavior.

5.2 Earnings, sanctions, and dynamics of behavior

Overall, we can detect a pattern of reciprocal behavior among the responders: The senders, having no chance to coordinate the amounts they send, were inclined to send an unequal number of points to the responder (in 369 out of 432 times). In return, the responders gave more points to the more generous sender in 80 percent of the time (in 294 out of 369 times). In cases when the responder had received equal allocations from both senders, he or she almost always (95 percent of the time) returned an equal amount to each sender (in 60 out of 63 times).

Sending a larger allocation than the other sender significantly increased the likelihood of getting a larger share in return. In fact, for the more generous sender, the odds of receiving a larger allocation (than the other sender) in return were 78 times higher, (OR: 78.4, 95% CI 29.92–256.92). These results suggest that reciprocity was the dominant motivation for responder behavior (cf. Table B8 in Appendix B).

Table 6 Average gross and net earnings per treatment

	Sender earnings before punishment costs	Sender net earnings	Responder earnings before punishment	Responder net earnings
Baseline	26.31 (6.25)	26.31 (6.25)	42.19 (12.54)	42.19 (12.54)
Punishment	26.85 (8.25)	25.44 (8.48)	43.08 (16.21)	34.56 (13.43)
Justification	26.93 (7.39)	26.93 (7.39)	45.71 (16.13)	45.71 (16.13)
Justification + Punishment	28.40 (6.59)	27.17 (6.91)	44.54 (13.37)	37.18 (12.29)

Standard deviations in parentheses

The average gross and net earnings in different treatments are shown in Table 6. On average, the senders punished the responders quite moderately, and consequently their gross and net earnings are quite close to each other even in treatments where punishments were possible. However, since each point used for punishment decreased the responders' points by three, even this moderate amount of punishment affected responders' payoffs considerably. In the *Punishment* treatment, responders' earnings were on average reduced by 8.52 points, that is, $2 \times 1.42 \times 3$ where 1.42 is the average number of points an individual sender used for punishing the responder. In the *Justification + Punishment* treatment cell, responders' earnings were on average reduced by 7.36 points. In that sense, giving the senders an opportunity to punish the responders did not increase overall efficiency, although this efficiency in terms of points earned increased in the later periods (see Tables 11 and 12 in Appendix B). However, one must keep in mind that this is partly due to the specific parameters of the experiment. The earnings are contingent on the relative cost of punishment or the multiplier for the punishment points that was chosen by experimenters.

We see a growing trend in the average amounts of points sent in each treatment, as shown in Fig. 2. The panel on the left shows the average share of the initial endowment of 12 points senders transferred to responders. Comparing only the first and last period, the average amount sent increased about 50% in each treatment. This growth is most pronounced in *Baseline* where the increase was about 67%. This suggest that there was a form of indirect reciprocity; in other words, positive experiences in previous rounds induced senders to send more, which in turn induced larger returns even though the players did not engage in fixed groups. The panel on the right shows that the average share of points returned remained quite stable in each treatment, whereas the absolute amounts of points grew as the senders' transfers increased. Indeed, as shown in Tables 9 and 10 in Appendix B, the average share returned was always at least 40% and in some occasions even over 50%, making sending money actually a rewarding choice, and in this manner likely contributing to the growing trend of monetary allocations over the six periods in our experiment.

We conducted panel regressions on points sent and returned using the treatment dummies as independent variables. We did not observe significant treatment effects on either amounts sent (Table 13 in Appendix B) or returned (Table 14 in Appendix

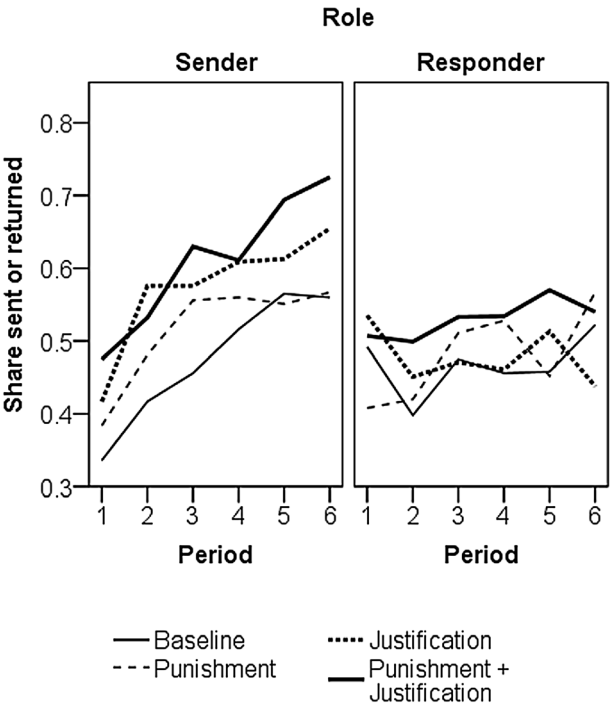

Fig. 2 Average shares of points sent and returned per period

B). However, if we include the amount received in the first stage when explaining responder behavior, or amount received in the previous round when explaining sender behavior as an explanatory variable, the number of points received (first stage/previous round) turns out to be highly significant explanatory variable, which is to be expected. Furthermore, some caution should be used in interpreting the results, given the rather small number of observations per treatment.[6]

To summarize the main results, the average points sent and returned were highest in *Justification* + *Punishment*, lowest in *Baseline,* while *Punishment* and *Justification* fell between these two. When restricting analysis at the first round, the contributions were highest in the treatments with the justification requirement. In accordance with this, responders also returned more money when justifications were required. However, taking into account what responders earned renders the effect of *Justification* insignificant. *Punishment* or the interaction between *Punishment* and *Justification* did not produce statistically significant effects.

[6] As the number of independent units of observation after the first round is rather small (12, or 3 sessions per treatment) any regression analysis would not remain unbiased if observations were taken at face value. To account for this problem, we performed clustered bootstrap estimation as the sessions as clusters (Cameron et al. 2008). However, significant effects of the explanatory variables on amounts given were not observed. Both ordinary and bootsrapped regressions are reported in Appendix B, Tables 6 through 9.

Responders also reacted to senders' behavior by discriminating between the two senders and rewarding them in accordance to their behavior. Furthermore, over the six rounds, the proportional share of points returned remained the same throughout the rounds. However, since senders increased the amount they sent, in absolute terms, retuned amounts also increased. The same pattern was observed independent of the treatment. This virtuous cycle increases *gross efficiency*. However, once the effects of punishments are taken into account, the *net efficiency* in terms of the total payoffs for the whole group of three is actually decreased in treatments with the punishment opportunity.

5.3 Justifications

Since the analysis of treatment effects suggests that responders' obligation to justify their choices had some influence on sender behavior, we will further analyze the content of justifications and their connection to responders' behavior. In the *Justification* and *Justification + Punishment* treatment conditions, responders were asked to write a justification for their decision. What kinds of justifications did they give? Since our hypotheses pertained to participants' *behavior*, the analysis on justifications is exploratory in character. However, insofar as sanctions function as a feedback mechanism, we can expect that the threat of being sanctioned affects the types of justifications responders give. We coded the justifications into five classes: Reciprocity, Equality, Self-Interest, Other, and Empty. The classification is based on our interest in responders' application of different fairness norms. The criteria for these classes as well as examples of each class are given in Table 7.[7] Each justification was placed in one class based on its principal content; the justifications were generally short.

Figure 3 reports the proportional distribution of justifications in *Justification* and *Justification + Punishment* treatments cells. In both cells, messages falling into the Reciprocity class are by far the most common (56% in *Justification* and 52% in *Punishment and Justification*), whereas appeals to the equality norm are much less frequent. In *Justification*, Self-interest ties with Equality, and none of the responders failed to give a justification in this treatment. In *Justification + Punishment*, appeals to self-interest are not observed but two responders failed to justify their choices seven times, and one responder failed to do so once. Although such failures could be considered an analytical nuisance, they can also be substantively meaningful: the failure to give a justification even when explicitly required to do so is still a message to the senders. Note also that the share of Empty in *Justification + Punishment* is almost the same as that of Self-Interest in *Justification.* The distribution of justification types differs between the treatments ($n = 216$, $\chi^2 = 31.761$, $df = 4$,

[7] Justifications were classified by a research group member who did not participate in defining the classification criteria. The reliability of the classification was checked by having two additional persons, who were not involved in the project, classify the justifications independently. Agreement among the coders can be considered substantial (Landis & Koch, 1977) as Cohen's kappa, measuring agreement between pairs of classifications, ranges from 0.693 (external coders A and B) to 0.780 (external coder A and project member) and 0.795 (external coder B and project member). We therefore consider the classification reliable enough for further analysis.

Table 7 Justification classes

Category	Explanation and examples
Reciprocity	The decision-maker announces that the amount of points returned to each sender depends on the number of points they originally sent. The number of points returned is proportional to the number of points sent
	Both get back the same amount they gave
	Points you sent come back doubled
Equality	The decision-maker appeals to the equality of the end distribution among all three participants
	I thought that let's divide the pot we collected or the tripled amounts sent by the senders and my 12 points into three parts among everyone, in which case everyone would get the same amount
	Now total points are about even for three
Self-interest	The decision-maker appeals to his or her own interest
	I maximize own interest. Apologies
	I want all the monies
Other	The justification is too general to be classified, unrelated to the distributive decision or has no substantive content
	I thought this might be a good idea
	I just counted it this way
Empty	A participant assigned to the role of a decision-maker gives no justification

$p < 0.001$). We thereby feel confident to say that the threat of punishment eliminates the propensity to explicitly justify decisions with self-interest.

A question that naturally follows is whether punishments depend on the kinds of justifications responders give or whether the failure to give a justification triggers punishments, which can be investigated in the *Justification + Punishment* treatment cell. To this end, we compared total sanctions (combined sanctions by both senders) directed at responders. Responders seem to face especially harsh punishments if they give no justification, the average number of points used for punishing being 7.46 (s.d. 6.48, $n = 13$) in that case. Recall that the "fine" suffered by the average responder is obtained by multiplying this number by three. In contrast, when the responder evokes the principle of reciprocity, the average sanction is only 1.05 points (s.d. 2.05, $n = 56$). Other justification classes fall in between these two but are closer to Reciprocity than Empty. The Brown-Forsythe test statistic (7.382, $df1 = 3$, $df2 = 28.163$, $p = 0.001$) suggests that average punishments differ among the classes. Furthermore, Tamhane's T2 test shows that there are statistically significant differences between Reciprocity and Empty ($p = 0.023$) as well as between Reciprocity and Other ($p = 0.079$).[8]

[8] Although the number of observations makes it relatively safe to rely on parametric tests, we repeated the analysis using non-parametric tests. The Kruskal–Wallis test also points to differences in average (more specifically, median) sanctions ($\chi^2 = 18.831$, $p < 0.001$). Pairwise Mann–Whitney U-tests show statistically significant differences between Reciprocity and Other ($U = 367.50$, $p = 0.002$), Reciprocity and Empty ($U = 143.00$, $p < 0.001$), Equality and Empty ($U = 61.00$, $p = 0.033$), and Other and Empty ($U = 84.50$, $p = 0.043$). After Bonferroni correction, Reciprocity vs. Other and Reciprocity vs. Empty remain statistically significant.

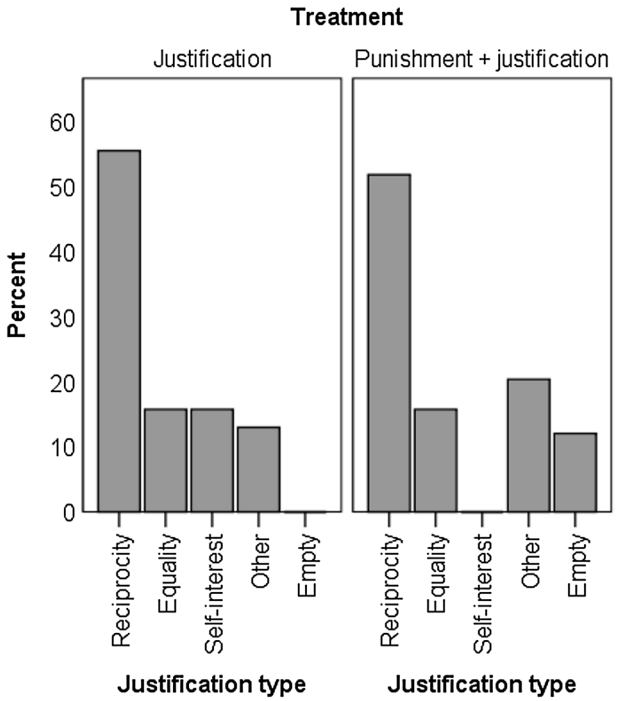

Fig. 3 The distribution of justification classes

Finally, we checked whether responders' behavior was in line with the justifications they gave. As it is not possible to classify choices as unequivocally as justifications, the results are indicative rather than conclusive. When the allocations falling into the category of reciprocity, equality or self-interest are considered, about 61 percent of them match with the associated justification.[9] Justifications matched actual allocations to a somewhat greater degree in the *Justification* treatment cell (67.3 percent) than in the *Justification + Punishment* treatment cell (54.5 percent). According to Pearson's chi square test, the difference between the cells is statistically almost significant ($\chi^2 = 3.599$, $df = 1$, $p = 0.058$, $n = 208$). In the *Justification + Punishment* treatment cell, responders faced on average larger sanctions when the justification did not match the allocation (9.33 points) than when the two matched (3.22 points). The result from a Mann–Whitney U test is statistically significant ($U = 894$, $p = 0.006$, $n = 101$) and indicates that the opportunity to punish was indeed used as a feedback mechanism to sanction

[9] We classified the choice as *reciprocal* if the responder returned more points to the sender who sent more, or returned the same amount of points to both senders after receiving the same amount of points from both of them. The choice was categorized as *equal* if the end distribution among all three players was equal after the responder's choice (and before possible punishments), and *selfish* if the responder returned nothing to either sender after receiving some amount of points from them. The two categories of justifications, Other and Empty, have no clear counterparts in choices.

responders, not just for allocations, but also for inconsistencies between allocations and justifications.

6 Conclusions and discussion

Our experiment was designed to analyze the extent to which different accountability mechanisms increase contributions in a trust game. We designed a three-person trust game to capture asymmetries involved in public accountability relations. The game gauges trust understood as "belief in reciprocity" (Berg et al., 1995: 137), and our experimental treatments make it possible to discern the behavioral effects of two accountability mechanisms, punishment and justification, either separately or combined. We replicate observations from previous studies on the trust game, that is, resources are sent and returned even in the baseline treatment with no accountability mechanisms.

We also see some differences between the treatments. The average sums sent or returned were highest in *Justification + Punishment*, and lowest in *Baseline*. When looking at the first round, the opportunity to punish responders did not increase points sent or returned, whereas the requirement of a justification had an impact on sender behavior. Moreover, responders reacted to sender behavior, they returned more when they were given more, and they also discriminated between the two senders according to how much each sent. The obligation to justify decisions influenced responder behavior, but the effect is not seen when the amount received from senders is taken into account. The analysis of the content of justifications reveals that in the *Justification + Punishment* treatment cell, there were no justifications that appeal to pure self-interest, whereas in the *Justification* cell, appeals to self-interest were presented. Moreover, the failure to give any justification as well as inconsistency between one's behavior and the given justification triggered punishments.

These results suggest that punishment was not an effective way to increase contributions, an observation that is in line with certain other trust game studies (Houser et al., 2008, Clabuig et al. 2016). Regarding justifications, our results give further support to the view that the requirement to justify one's actions increases contributions to others. This may be because people find it easier to give justifications for actions that follow social norms and that are not self-interested. Especially the norm of reciprocity was observed both in responders' actions and in the justifications they gave for their actions. Our results suggest that the difficulty to justify self-interested action may have influenced responders' behavior, because self-interest was not often given as a justification. Furthermore, the fear of punishment prevented appeals to self-interest totally, suggesting that punishment gave senders a tool to give feedback both on decisions and the justifications received from responders. It is worth pointing out, however, that the tendency to follow the norm of reciprocity in terms of both behavior and justifications may also be regarded problematic in public accountability relations where impartiality is expected to be a norm (cf. Rothstein & Teorell, 2008).

Overall, our study offers some support for the importance of justifications as a mechanism of accountability. However, some caution should be used when generalizing from the results because the number of experimental sessions–and consequently, the number independent observations—is somewhat low, and future experiments would therefore be needed to increase the robustness of our findings. Another source of uncertainty arises from the uncertainty regarding sender motivations. While our design was not conducive towards competition between the senders, we cannot conclusively rule out the possibility that some senders might perceive the game as a contest for favors from the responder. Moreover, the frequency of reciprocity as a basis of justifications might lend support to this interpretation of sender behavior.

In our study, experimental subjects were randomly assigned to the sender and responder roles. In this respect, it does not capture one aspect which is present in many public accountability relations, namely the authorization of office-holders. Communication was restricted to responders' justifications for their decisions, whereas future research could address the effects of multilateral communication, including senders' opportunity to publicly discuss and give verbal feedback on responders' behavior and justifications. Moreover, social sanctions such as shaming were precluded since experimental subjects were anonymous, and subjects had no incentives to create reputations since groups were re-shuffled in each round of the experiment. Because our study had an exclusive emphasis on verbal justifications and material sanctions, the effects of social sanctions and reputation building are left for further research.

In our experiment, responders were accountable for one decision at a time, while elected representatives can be accountable for a number of decisions. Responders were accountable to all those individuals who were affected by their decisions, and in treatments involving punishment any (or either) of them could use this opportunity. Elected representatives are typically accountable only to a specific subgroup of the electorate, i.e., constituents, which incentivizes them to act according to the preferences of this particular subgroup (e.g., Chambers, 2004). Future research could also address the behavioral consequences of this restriction.

Appendix A. Instructions

INSTRUCTIONS [Common to all]

Welcome to the decision-making experiment!

All participants will receive a participatory reward of 3 euros. These instructions will explain how you can earn more money from the experiment.

Each participant has been randomly assigned a computer to use in the experiment room. The choices made by the participants are transmitted to other participants via the computer screens. All messaging unrelated to the experiment is forbidden.

You will remain anonymous during the experiment. The results of the experiment will be analyzed on the aggregate level and no choices any participant makes will be

connected back to them. The identities of all participants will be withheld from everyone except the organizers of the experiment.

At the start of the experiment we will read through the instructions together, after which you will get the chance to familiarize yourself with them independently. You will start with three practice tasks. The experiment will begin after all participants have answered correctly to the practice questions.

You are not to discuss with other participants during the experiment. Each participant will make their decisions independently. If you have any questions, please raise your hand. The organizer of the experiment will come see you personally and answer any questions you may have.

Please close your mobile phone for the duration of the experiment.

The experiment [Baseline]

At the start of the experiment each participant will be randomly assigned to either the role of sender or responder. All participants will have a chance to familiarize themselves with both role's instructions. Your role will be randomly assigned during the first round and it will remain the same throughout the experiment. The experiment consists of six rounds. Your reward will be paid based on one randomly selected round. The reward round will be chosen with a roll of a six-sided dice.

In the beginning of each round the participants will be randomly put into groups of three. The groups will change on each round of the experiment, meaning that you will never be in the same group as in earlier rounds.

You will not be given the identity of your group members during or after the experiment.

Each group will consist of three people, two of whom will be senders and one a responder. All rounds of the experiment consist of three stages. The senders will make their decision in the first stage and the responder in the second.

The decisions that you will make during the experiment are about sending points. The points will be converted into rewards so that 2 points = 1 euro.

Stage 1. Only the sender will make a decision

Instructions to the sender

In the first stage of the round each member of the group is given 12 points. It is your job to decide how many points you want to send to the responder. You are free to choose any sum between 0 and 12 points. The other sender in your group will make the same decision.

The points you send to the responder are multiplied by three, meaning that per each point you send the responder will receive three.

Example 1: Sender A sends the responder three (3) points and sender B sends two (2) points. The responder will receive a total of $3 \times 3 + 3 \times 2 = 15$ points.

After the first stage sender A is left with nine (9) points and sender B is left with ten (10) points.

Example 2: Sender A sends the responder four (4) points and sender B send one (1) point. The responder gets a total of $3 \times 4 + 3 \times 1 = 15$ points. After the first stage sender A is left with eight (8) and sender B with eleven (11) points.

Stage 2. Only the responder makes a decision

Instructions to the responder

You have at your disposal the 12 points you were given in the first stage and any points you might have gotten from the senders. In the second stage of the round it is your job to decide how many points you want to send to sender A, how many to sender B and how many you want to keep for yourself.

After you make the decision each group member will be given information about all decisions that were made in stages 1 and 2 by each participant.

Stage 3

All members of the group are given 12 more points in stage three. At the end of stage 3 the total points earned during the whole round will be shown to all group members.

How total points are determined in each round

Both *senders* reward consists of the points the sender kept in the 1st stage, points he or she might have received from the responder in the 2nd stage and the 12 points given in the 3rd stage.

The responder's reward consists of the points he or she kept in the 2nd stage and the points he or she was given in the 3rd stage.

These reward sums will be revealed at the end of each round. To speed up reward payment please write down your rewards (in points and euros) at the end of each round to the sheet provided at your seat.

New round

After the third stage a new round begins, and all participants will be randomly divided into new groups of three. The roles of the participants will remain the same in each round.

The experiment is 6 rounds long. After the experiment the participants will answer a questionnaire about the experiment. After all participants are finished, they will be individually asked to leave the room to claim their reward. The reward is paid based on the result of one, randomly selected round of the experiment. The reward round will be chosen with the roll of a six-sided dice.

The experiment [Punishment]

At the start of the experiment each participant will be randomly assigned to either the role of sender or responder. All participants will have a chance to familiarize themselves with both role's instructions. Your role will be randomly assigned during the first round and it will remain the same throughout the experiment. The experiment consists of six rounds. Your reward will be paid based on one randomly selected round. The reward round will be chosen with a roll of a six-sided dice.

In the beginning of each round the participants will be randomly put into groups of three. The groups will change on each round of the experiment, meaning that you will never be in the same group as in earlier rounds.

You will not be given the identity of your group members during or after the experiment.

Each group will consist of three people, two of whom will be senders and one a responder. All rounds of the experiment consist of three stages. The senders will make their decision in the first stage and the responder in the second.

The decisions that you will make during the experiment are about sending points. The points will be converted into rewards so that 2 points = 1 euro.

Stage 1. Only the sender will make a decision

Instructions to the sender

In the first stage of the round each member of the group is given 12 points. It is your job to decide how many points you want to send to the responder. You are free to choose any sum between 0 and 12 points. The other sender in your group will make the same decision.

The points you send to the responder are multiplied by three, meaning that per each point you send the responder will receive three.

Example 1: Sender A sends the responder three (3) points and sender B sends two (2) points. The responder will receive a total of $3 \times 3 + 3 \times 2 = 15$ points.

After the first stage sender A is left with nine (9) points and sender B is left with ten (10) points.

Example 2: Sender A sends the responder four (4) points and sender B send one (1) point. The responder gets a total of $3 \times 4 + 3 \times 1 = 15$ points. After the first stage sender A is left with eight (8) and sender B with eleven (11) points.

Stage 2. Only the responder makes a decision

Instructions to the responder

At your disposal you have 12 points from the first round and any points the senders might have sent you. In this second stage it is your job to decide how many points you want to send to sender A, how many points to sender B and how many you want to keep for yourself. Justify your decisions to both senders. Please write your justification to the assigned field on your computer screen.

After you make the decision each group member will be given information about all decisions that were made in stages 1 and 2 by each participant and your justifications will be shown to both senders.

Stage 3

All members of the group are given 12 more points in stage three.

Instructions to the sender

If you want, you can now reduce the responder's points. You can choose however many points you would like to reduce from the responder between 0 and 12 points. Any points left unused will be added to your total earnings from the round.

For each point you use three will be reduced from the responder until he or she is down to 0 points. (The responder's points will always total at least 0 points after reductions).

At the end of stage 3 all points collected during the round will be shown to all group members.

How total points are determined in each round

Both *senders* reward consists of the points the sender kept in the 1st stage, points he or she might have received from the responder in the 2nd stage and the 12 points given in the 3rd stage, minus any points the sender used to reduce points from the responder.

The responder's reward consists of the points he or she kept in the 2nd stage and the 12 points he or she was given in the 3rd stage, of which any minus points sent by the sender will be reduced.

These reward sums will be revealed at the end of each round. To speed up reward payment please write down your rewards (in points and euros) at the end of each round to the sheet provided at your seat.

New round

After the third stage a new round begins, and all participants will be randomly divided into new groups of three. The roles of the participants will remain the same in each round.

The experiment is 6 rounds long. After the experiment the participants will answer a questionnaire about the experiment. After all participants are finished, they will be individually asked to leave the room to claim their reward. The reward is paid based on the result of one, randomly selected round of the experiment. The reward round will be chosen with the roll of a six-sided dice.

The experiment [Justification]

At the start of the experiment each participant will be randomly assigned to either the role of sender or responder. All participants will have a chance to familiarize themselves with both role's instructions. Your role will be randomly assigned during the first round and it will remain the same throughout the experiment. The experiment consists of six rounds. Your reward will be paid based on one randomly selected round. The reward round will be chosen with a roll of a six-sided dice.

In the beginning of each round the participants will be randomly put into groups of three. The groups will change on each round of the experiment, meaning that you will never be in the same group as in earlier rounds.

You will not be given the identity of your group members during or after the experiment.

Each group will consist of three people, two of whom will be senders and one a responder. All rounds of the experiment consist of three stages. The senders will make their decision in the first stage and the responder in the second.

The decisions that you will make during the experiment are about sending points. The points will be converted into rewards so that 2 points = 1 euro.

Stage 1. Only the sender will make a decision

Instructions to the sender

In the first stage of the round each member of the group is given 12 points. It is your job to decide how many points you want to send to the responder. You are free to choose any sum between 0 and 12 points. The other sender in your group will make the same decision.

The points you send to the responder are multiplied by three, meaning that per each point you send the responder will receive three.

Example 1: Sender A sends the responder three (3) points and sender B sends two (2) points. The responder will receive a total of $3 \times 3 + 3 \times 2 = 15$ points.

After the first stage sender A is left with nine (9) points and sender B is left with ten (10) points.

Example 2: Sender A sends the responder four (4) points and sender B send one (1) point. The responder gets a total of $3 \times 4 + 3 \times 1 = 15$ points. After the first stage sender A is left with eight (8) and sender B with eleven (11) points.

Stage 2. Only the responder makes a decision

Instructions to the responder

At your disposal you have 12 points from the first round and any points the senders might have sent you. In this second stage it is your job to decide how many points you want to send to sender A, how many points to sender B and how many you want to keep for yourself. Justify your decisions to both senders. Please write your justification to the assigned field on your computer screen.

After you make the decision each group member will be given information about all decisions that were made in stages 1 and 2 by each participant and your justifications will be shown to both senders.

Stage 3

All members of the group are given 12 more points in stage three. At the end of stage 3 the total points earned during the whole round will be shown to all group members.

How total points are determined in each round

Both *senders* reward consists of the points the sender kept in the 1st stage, points he or she might have received from the responder in the 2nd stage and the 12 points given in the 3rd stage.

The responder's reward consists of the points he or she kept in the 2nd stage and the 12 points he or she was given in the 3rd stage.

These reward sums will be revealed at the end of each round. To speed up reward payment please write down your rewards (in points and euros) at the end of each round to the sheet provided at your seat.

New round

After the third stage a new round begins, and all participants will be randomly divided into new groups of three. The roles of the participants will remain the same in each round.

The experiment is 6 rounds long. After the experiment the participants will answer a questionnaire about the experiment. After all participants are finished, they will be individually asked to leave the room to claim their reward. The reward is paid based on the result of one, randomly selected round of the experiment. The reward round will be chosen with the roll of a six-sided dice.

The experiment [Punishment and Justification]

At the start of the experiment each participant will be randomly assigned to either the role of sender or responder. All participants will have a chance to familiarize themselves with both role's instructions. Your role will be randomly assigned during the first round and it will remain the same throughout the experiment. The experiment consists of six rounds. Your reward will be paid based on one randomly selected round. The reward round will be chosen with a roll of a six-sided dice.

In the beginning of each round the participants will be randomly put into groups of three. The groups will change on each round of the experiment, meaning that you will never be in the same group as in earlier rounds.

You will not be given the identity of your group members during or after the experiment.

Each group will consist of three people, two of whom will be senders and one a responder. All rounds of the experiment consist of three stages. The senders will make their decision in the first stage and the responder in the second.

The decisions that you will make during the experiment are about sending points. The points will be converted into rewards so that 2 points = 1 euro.

Stage 1. Only the sender will make a decision

Instructions to the sender

In the first stage of the round each member of the group is given 12 points. It is your job to decide how many points you want to send to the responder. You are free to choose any sum between 0 and 12 points. The other sender in your group will make the same decision.

The points you send to the responder are multiplied by three, meaning that per each point you send the responder will receive three.

Example 1: Sender A sends the responder three (3) points and sender B sends two (2) points. The responder will receive a total of $3 \times 3 + 3 \times 2 = 15$ points.

After the first stage sender A is left with nine (9) points and sender B is left with ten (10) points.

Example 2: Sender A sends the responder four (4) points and sender B send one (1) point. The responder gets a total of $3 \times 4 + 3 \times 1 = 15$ points. After the first stage sender A is left with eight (8) and sender B with eleven (11) points.

Stage 2. Only the responder makes a decision

Instructions to the responder

At your disposal you have 12 points from the first round and any points the senders might have sent you. In this second stage it is your job to decide how many points you want to send to sender A, how many points to sender B and how many you want to keep for yourself. Justify your decisions to both senders. Please write your justification to the assigned field on your computer screen.

After you make the decision each group member will be given information about all decisions that were made in stages 1 and 2 by each participant and your justifications will be shown to both senders.

Stage 3

All members of the group are given 12 more points in stage three.

Instructions to the sender

If you want, you can now reduce the responder's points. You can choose however many points you would like to reduce from the responder between 0 and 12 points. Any points left unused will be added to your total earnings from the round.

For each point you use three will be reduced from the responder until he or she is down to 0 points. (The responder's points will always total at least 0 points after reductions).

At the end of stage 3 all points collected during the round will be shown to all group members.

How total points are determined in each round

Both *senders* reward consists of the points the sender kept in the 1st stage, points he or she might have received from the responder in the 2nd stage and the 12 points given in the 3rd stage, minus any points the sender used to reduce points from the responder.

The responder's reward consists of the points he or she kept in the 2nd stage and the 12 points he or she was given in the 3rd stage, of which any minus points sent by the sender will be reduced.

These reward sums will be revealed at the end of each round. To speed up reward payment please write down your rewards (in points and euros) at the end of each round to the sheet provided at your seat.

New round

After the third stage a new round begins, and all participants will be randomly divided into new groups of three. The roles of the participants will remain the same in each round.

The experiment is 6 rounds long. After the experiment the participants will answer a questionnaire about the experiment. After all participants are finished, they will be individually asked to leave the room to claim their reward. The reward is paid based on the result of one, randomly selected round of the experiment. The reward round will be chosen with the roll of a six-sided dice.

Appendix B

See Tables 8, 9, 10, 11, 12, 13, 14,15, and 16.

Table 8 Reciprocal choices of the responders as a response to amounts received

Senders	Receivers		Total
	Unequal	Equal	
Unequal	294 (79.7%)	75 (20.3%)	369 (100%)
Equal	3 (4.8%)	60 (95.2%)	63 (100%)
Total	297 (68.8%)	135 (31.3%)	432 (100%)

Table 9 Mean return share per treatment: baseline and punishment

Period	Baseline			Punishment		
	Received	Returned	Return-%	Received	Returned	Return-%
1	24.170	11.890	0.492	27.670	11.280	0.408
2	30.000	11.944	0.398	34.670	14.550	0.420
3	32.830	15.610	0.475	40.000	20.450	0.511
4	37.170	16.944	0.456	40.330	21.280	0.528
5	40.670	18.610	0.458	39.670	17.940	0.452
6	40.330	21.056	0.522	40.830	23.170	0.567

Table 10 Mean return share per treatment cell: Justification and Justification + Punishment

Period	Justification			Justification + punishment		
	Received	Returned	Return-%	Received	Returned	Return-%
1	30.000	16.056	0.535	34.170	17.333	0.507
2	41.500	18.722	0.451	38.330	19.111	0.499
3	41.500	19.556	0.471	45.330	24.167	0.533
4	43.830	20.220	0.461	44.000	23.500	0.534
5	44.170	22.667	0.513	50.000	28.500	0.570
6	47.170	20.667	0.438	52.170	28.167	0.540

Table 11 Average net earnings in each treatment cell, senders

Period	Baseline	Punishment	Justification	Justification + punishment
1	25.92	23.39	27.02	25.72
2	24.97	24.02	26.44	25.58
3	26.33	25.67	26.86	27.33
4	26.28	26.89	26.80	26.91
5	26.52	24.92	27.97	29.39
6	27.80	27.72	26.47	28.08

Table 12 Average net earnings in each treatment cell, responders

Period	Baseline	Punishment	Justification	Justification + punishment
1	36.28	30.56	37.94	33.33
2	42.06	35.28	46.79	33.72
3	41.22	32.22	45.94	38.00
4	44.22	36.89	47.61	35.50
5	46.05	37.06	45.50	42.33
6	43.28	35.33	50.50	40.17

Table 13 Variables explaining sender behavior

	I	II
Points returned in previous round		0.243 (0.015) ***
Punishment	0.500 (0.829)	0.205 (0.589)
Justification	1.194 (0.784)	0.733 (0.540)
Justification × Punishment	− 0.06 (1.11)	− 0.214 (0.778)
N	144 × 6 = 864	144 × 5 = 720
Adj. R^2	0.0159	0.283

Standard errors (in parentheses) clustered on individual. Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $'p < 0.10$

Table 14 Variables explaining responder behavior

	III	VI
Points received in the current round		0.494 (0.051) ***
Punishment	2.102 (3.667)	0.618 (2.830)
Justification	3.634 (4.025)	0.095 (3.294)
Justification × Punishment	1.71 (3.397)	1.892 (4.347)
N	72 × 6 = 432	72 × 6 = 432
Adj. R^2	0.034	0.390

Standard errors (in parentheses) clustered on individual. Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $'p < 0.10$

Table 15 Treatments effects on sender behavior, bootstrap estimates of standard errors

Punishment	0.500 (1.71)
Justification	1.194 (1.27)
Justification × Punishment	− 0.060 (2.211)
	0.0159

Bootstrapped standard errors in parentheses, 3,483 replications, 12 clusters (sessions)

Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $'p < 0.10$

Table 16 Treatments effects on responder behavior, bootstrap estimates of standard errors

Punishment	2.102 (5.62)
Justification	3.634 (7.26)
Justification × punishment	1.71 (7.94)
	0.0159

Bootstrapped standard errors in parentheses, 3,483 replications, 12 clusters (sessions)

Significance levels: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $'p < 0.10$

Acknowledgements The authors would like to thank the editor and the two anonymous referees whose comments improved the manuscript considerably. This research has been financially supported by the Academy of Finland project 'Democratic Reasoning' (Decision Number 274305) and Strategic Research Council project "Participation in Long-Term Decision-Making" (Decision Numbers 312671 and 326662)

Funding Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital.

Declarations

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee of University of Turku and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all participants included.

References

Bäckstrand, K. (2008). Accountability of networked climate governance: The rise of transnational climate partnerships. *Global Environmental Politics, 8*(3), 74–102.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution, 54*(1), 39–57.

Barro, R. J. (1973). The control of politicians: An economic model. *Public Choice, 14*(1), 19–42.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior, 10*(1), 122–142.

Besley, T. (2006). *Principled agents? The political economy of good government*. Oxford University Press.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization, 60*(1), 11–26.

Bovens, M. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics, 33*(5), 946–967.

Bovens, M., Goodin, R. E., & Schillemans, T. (2014). *The Oxford handbook of public accountability*. Oxford University Press.

Bracht, J., & Feltovich, N. (2009). Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *Journal of Public Economics, 93*(9–10), 1036–1044.

Brandts, J., Cooper, D. J., & Rott, C. (2019). Communication in laboratory experiments. In *Handbook of research methods and applications in experimental economics*. Edward Elgar Publishing.

Calabuig, V., Fatasb, E., Olcinaa, G., & Rodriguez-Larac, I. (2016). Carry a big stick, or no stick at all: Punishment and endowment heterogeneity in the trust game. *Journal of Economic Psychology, 57*, 153–171.

Cameron, A.C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics, 90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

Cason, T. N., & Gangadharan, L. (2016). Swords without covenants do not lead to self-governance. *Journal of Theoretical Politics, 28*(1), 44–73.
Chambers, S. (2004). Behind closed doors: Publicity, secrecy and the quality of deliberation. *The Journal of Political Philosophy, 12*(4), 389–410.
Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third party intervention. *Journal of Economic Behavior and Organization, 68*(1), 18–28.
Colombo, C. (2018). Hearing the other side?–Debiasing political opinions in the case of the Scottish independence referendum. *Political Studies, 66*(1), 23–42.
Corazzini, L., Kube, S., Maréchal, M. A., & Nicolò, A. (2014). Elections and deceptions: An experimental study on the behavioral effects of democracy. *American Journal of Political Science, 58*(3), 579–592.
Dal Bó, E., & Bó, P. (2014). "Do the right thing": The effects of moral suasion on cooperation. *Journal of Public Economics, 117*, 28–38.
De Cremer, D., Snyder, M., & De Witte, S. (2001). ''The less I trust, the less I contribute (or not)?": The effects of trust, accountability and self-monitoring in social dilemmas. *European Journal of Social Psychology, 31*, 93–107.
De Cremer, D., & van Dijk, E. (2009). Paying for sanctions in social dilemmas: The effects of endowment asymmetry and accountability. *Organizational Behavior and Human Decision Processes, 109*(1), 45–55.
De Kwaadsteniet, E. W., van Dijk, E., Wit, A., De Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin, 33*(12), 1647–1660.
Dufwenberg, M., Li, F., & Smith, A. (2021). Promises and punishment. Available at SSRN 3913750.
Fearon, J. D. (1999). Electoral accountability and the control of politicians: Selecting good types versus sanctioning poor performance. In A. Przeworski, S. C. Stokes, & B. Manin (Eds.), *Democracy, accountability, and representation* (pp. 55–97). Cambridge University Press.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994.
Fehr, E., & List, J. A. (2004). The hidden costs and returns of incentives—Trust and trustworthiness among CEOs. *Journal of the European Economic Association, 2*(5), 743–771.
Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 13*(422), 137–140.
Feltovich, N., & Giovannoni, F. (2015). Selection vs. accountability: An experimental investigation of campaign promises in a moral-hazard environment. *Journal of Public Economics, 126*, 39–51.
Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice, 50*(1/3), 5–25.
Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics, 10*(2), 171–178.
Geng, H., Weiss, A. R., & Wolff, I. (2011). The limited power of voting to limit power. *Journal of Public Economic Theory, 13*(5), 695–719.
Gutmann, A., & Thompson, D. (1996). *Democracy and disagreement*. Harvard.
Hamman, J. R., Weber, R. A., & Woon, J. (2011). An experimental investigation of electoral delegation and the provision of public goods. *American Journal of Political Science, 55*(4), 738–752.
Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior, 62*(2), 509–532.
Janssen, M. A., Holahan, R., Lee, A., & Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science, 328*(5978), 613–617.
Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32*(5), 865–889.
Kuyper, J., & Bäckstrand, K. (2016). Accountability and representation: Non-state actors in UN climate diplomacy. *Global Environmental Politics, 16*(2), 61–81.
Landa, D. (2010). Selection incentives and accountability traps: A laboratory experiment. https://doi.org/10.2139/ssrn.1640033
Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255–275.
Lierl, M. (2016). Social sanctions and informal accountability: Evidence from a laboratory experiment. *Journal of Theoretical Politics, 28*(1), 74–104.

Manin, B., Przeworski, A., & Stokes, S. C. (1999). Introduction. In A. Przeworski, S. C. Stokes, & B. Manin (Eds.), *Democracy, accountability, and representation* (pp. 1–26). Cambridge University Press.

McGraw, K. M. (1991). Managing blame: An experimental test of the effects of political accounts. *The American Political Science Review, 85*(4), 1133–1157.

McGraw, K. M., Timpone, R., & Bruck, G. (1993). Justifying controversial political decisions: Home style in the laboratory. *Political Behavior, 15*(3), 289–308.

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning, 21*(3), 341–355.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review, 86*(2), 404–417.

Philp, M. (2009). Delimiting democratic accountability. *Political Studies, 57*(1), 28–53.

Rehfeld, A. (2005). *The concept of constituency*. Cambridge University Press.

Rigdon, M. (2009). Trust and reciprocity in incentive contracting. *Journal of Economic Behavior and Organization, 70*(1–2), 93–105.

Rothstein, B., & Teorell, J. (2008). What is quality of government? A theory of impartial government institutions. *Governance, 21*(2), 165–190.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society, 7*(1), 58–92.

Schedler, A. (1999). Conceptualizing accountability. In A. Schedler, L. Diamond & M.F. Plattner (Eds), *The self-restraining state: Power and accountability in new democracies*. London: Lynne Rienner Publishers.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74–83.

Warren, M. E. (1996a). Deliberative democracy and authority. *American Political Science Review, 90*(1), 46–60.

Warren, M. E. (1996b). Introduction. In M. E. Warren (Ed.), *Democracy and trust* (pp. 1–21). Cambridge University Press.

Warren, M. E., & Gastil, J. (2015). Can deliberative minipublics address the cognitive challenges of democratic citizenship? *The Journal of Politics*, *77*(2), 562–574.

Weiss, A. R., & Wolff, I. (2013). Does being elected increase subjective entitlements? Evidence from the Laboratory. *Economics Bulletin, 33*(1), 794–796.

Woon, J. (2012). Democratic accountability and retrospective voting: A laboratory experiment. *American Journal of Political Science, 56*(4), 913–930.

Xia, E. (2017). Justification and conformity. *Journal of Economic Behavior and Organization, 136*, 15–28.