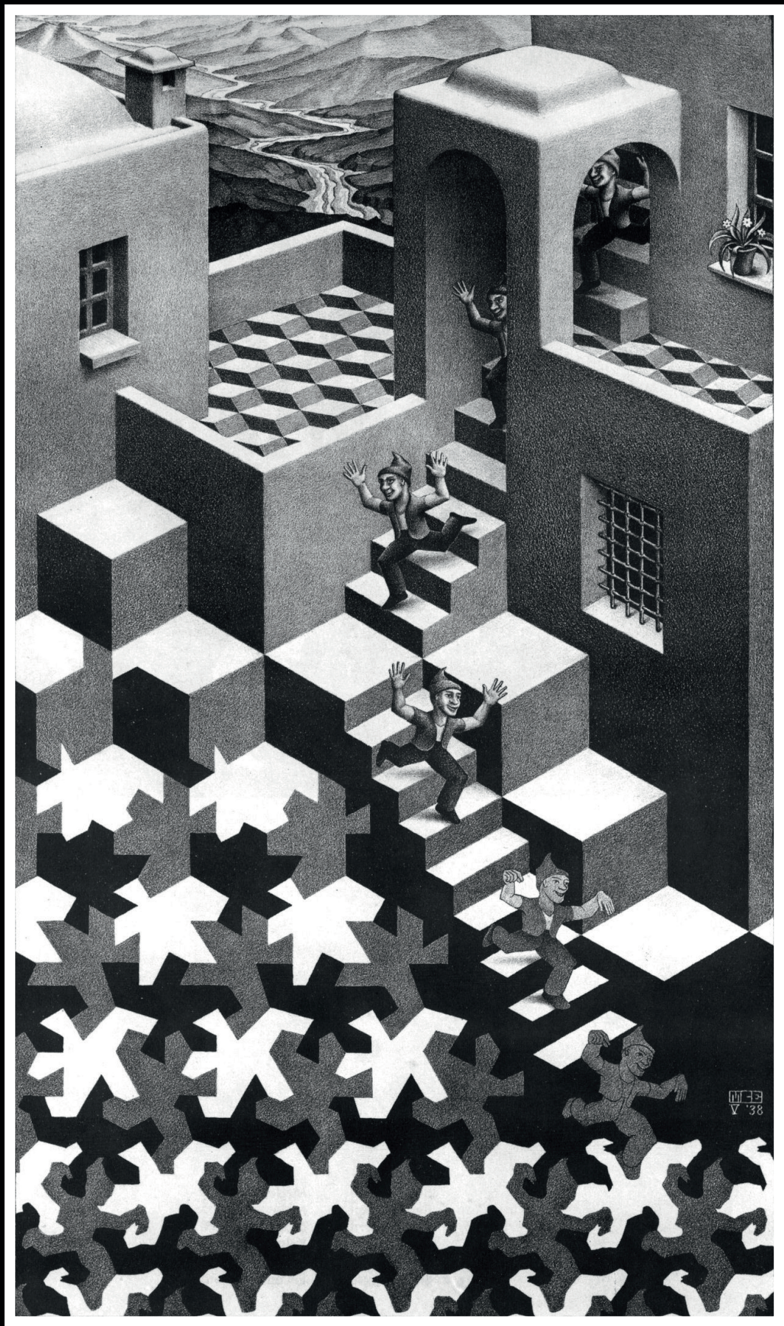


Cooperation in changing groups

How newcomers and norms shape public good provision in the lab, online games, and the field

Kasper Otten

Cooperation in changing groups

How newcomers and norms shape public good provision
in the lab, online games, and the field

Kasper Otten

Cooperation in changing groups: How newcomers and norms shape public good provision in the lab, online games, and the field
— Kasper Otten

Cover: M.C. Escher's “Cycle” © 2023 The M.C. Escher Company, The Netherlands.
All rights reserved. www.mcescher.com
Printing: Ridderprint BV, ridderprint.nl
ISBN: 978-90-393-7573-0

Cooperation in changing groups

How newcomers and norms shape public good provision
in the lab, online games, and the field

Coöperatie in veranderende groepen
Hoe nieuwkomers en normen publieke goederen beïnvloeden
in het lab, online spellen, en het veld
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 8 september 2023 des middags te 12.15 uur

door

Kasper Daniël Otten

geboren op 9 oktober 1993
te Enschede

Promotoren:
Prof. dr. ir. V.W. Buskens
Prof. dr. N. Ellemers

Copromotor:
Dr. W. Przepiorka

Beoordelingscommissie:
Prof. dr. D. Baldassarri
Prof. dr. A. Flache
Prof. dr. ir. A.G. van der Lippe
Prof. dr. M. Lubbers
Prof. dr. D.T. Scheepers

Dit proefschrift is onderdeel van het onderzoeksprogramma Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). Dit proefschrift werd mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en het Ministerie van Onderwijs, Cultuur, en Wetenschap (OCW) in het kader van een in 2017 toegekende Zwaartekracht subsidie aan SCOOP (grant number 024.003.025). Dit proefschrift is afgedrukt met financiële steun van de J.E. Jurriaanse Stichting.

Table of contents

List of Figures

List of Tables

Chapter 1. Synthesis

1.1. Background

1.1.1. *Public good provision*

Public good provision has been a crucial activity throughout human history. Hunter-gatherers contributed to public goods when they engaged in large-scale communal hunts, participated in camp-wide food sharing, and worked on collective facilities like hunting nets, drivelines, and irrigation (Boyd & Richerson, 2022; Hawkes et al., 1993; Hill, 2002). People contribute to crucial public goods when risking their lives in war to defend their countries (Mathew & Boyd, 2011), when risking their jobs to strike for better collective working conditions (Naylor, 1989), and when joining protests and revolutions against dictatorial regimes (Muller & Opp, 1986). In everyday life, people contribute to public goods when they volunteer for community events or local clubs, clean the dishes at shared living spaces, do their share of team tasks, or take measures to reduce their carbon footprint (Van Vugt et al., 2000). And in modern societies, people contribute via taxes to national public goods such as public transport, public education, law enforcement, and social security (Kallhoff, 2014).

Although all these activities differ in many ways, they share a common feature – they require individually costly contributions that benefit the group as a whole, including members who did not contribute to the public good themselves. Indeed, the two characteristics often used to define public goods are that (1) members cannot be excluded from benefiting of the good (non-excludability) and (2) the benefits reaped by any particular member do not come at the expense of other members (non-rivalry). Non-excludability and non-rivalness are matters of degree rather than binary conditions, so a good is rarely perfectly non-excludable and non-rival (Malkin & Wildavsky, 1991). As long as people can benefit from the good without contributing to it themselves, scholars commonly regard it as a public good (Cornes & Sandler, 1994, 1996). This feature of public goods means that the group as a whole is better off if the public good is provided, but individual members have a temptation to free-ride on the contributions of others. Once a public good is provided, its maintenance often requires sustained contributions from future generations, meaning that temptations to free-ride can be a persisting threat over time. Whether and how people cooperate for public goods despite the temptation to free-ride continues to receive societal and academic interest.

Much previous research has focused on the conditions that promote cooperation for public goods, but research is still scarce on how to sustain cooperation when conditions change (Lazega et al., 2022; Olsson et al., 2015). In particular, we still have a limited understanding of whether and how cooperation can be sustained when groups change due to the arrival of newcomers whose views and needs differ from those of the incumbents (i.e., members who were already in the group). In this thesis, we study whether and how cooperation for public goods is sustained in changing groups consisting of newcomers and incumbents. Before turning to theory about how newcomer entry may affect cooperation, the problem of public good provision is first outlined more generally.

The temptation to free-ride in public good provision has led some to argue that people will not voluntarily contribute to public goods, an argument often referred to as the 'zero contribution thesis' (Olson, 1965; Ostrom, 2000; Samuelson, 1954). In *The Logic of Collective*

Action, Olson argued that "even if all of the individuals in a large group are rational and self-interested, and would gain if, as a group, they acted to achieve their common interest or objective, they will still not voluntarily act to achieve that common interest or objective" (Olson, 1965, p. 2). Olson suggests that some form of central or external authority is necessary to achieve successful public good provision when individuals are tempted to free-ride. Similarly, in *The Tragedy of the Commons*, Hardin suggests that "we institute and (grumblingly) support taxes and other coercive devices to escape the horror of the commons" (Hardin, 1968, p. 1247). These predictions echo earlier philosophical works by Hobbes and Hume. In *Leviathan,* Hobbes argued that life without a sovereign is "solitary, poor, nasty, brutish, and short" (Hobbes, 1968 [1651], p. 186). In *A Treatise of Human Nature,* Hume suggests that public good provision is unlikely in large groups because "each seeks a pretext to free himself of the trouble and expense, and would lay the whole burden on others." (Hume, 1952 [1739], p. 239). Although these scholars were mostly theorizing about public goods that rule out any other circumstances beyond the free-rider dilemma, their work has led to a common view that public goods will be under-provisioned without coercion from a central authority (Malkin & Wildavsky, 1991).

In modern societies, we indeed find that several national public goods such as public transport, public education, and social security, are centrally organized via the government and regulated by law enforcement. It is hard to deny that law enforcement from a central authority is one way to achieve public good provision. However, law enforcement in itself is a public good that can only function if citizens agree on the rules that need to be enforced (Cowen, 1992; Tullock, 1971). Indeed, a democratic institution to effectively govern laws is often considered one of today's most important national public goods, and can be a prerequisite for participating in international cooperation arrangements such as the European Union. A democratic institution depends on voluntary contributions from citizens in the form of voting, communicating with legislators, and being politically informed (Birch, 2018; Downs, 1957). What is more, national public goods are only one category of public goods. There are many situations in which free-rider problems have to be addressed without an institutionalized authority, ranging from small-scale cooperation problems such as team work to international problems such as combatting the climate crisis. A large body of evidence shows that such voluntary public good provision is possible. Humans survived for thousands of years by participating in communal hunting and food sharing (Hawkes et al., 1993), workers have dramatically improved their working conditions by participating in collective strikes (Conell & Cohn, 1995), and citizens have risked their lives to join protests that contributed to the fall of the Berlin Wall (Tietzel & Weber, 1994) and overthrew corrupt regimes during the Arab Spring (Steinert-Threlkeld, 2017).

A few mechanisms have been suggested that may explain such cooperation without a central authority. First, individuals may cooperate because they are family. From an evolutionary perspective, cooperating with relatives benefits one's own genes because people share part of their genes with their relatives (this is also referred to as kin altruism, see Hamilton, 1964). However, many of the above examples of public good provision occur among unrelated individuals. A mechanism promoting cooperation among unrelated individuals is direct reciprocity. With direct reciprocity, individuals may behave cooperatively toward others because this increases the chance that these others will "return the favor", allowing for a mutually beneficial pattern of cooperation (Axelrod, 1984; Trivers, 1971). For example, hunter-gatherers who caught more prey than they can consume may share their surplus with others in the hope that these others will do the same when they catch prey in the future. Direct reciprocity thus relies on repeated encounters between the same individuals. Indirect reciprocity can promote cooperation when repeated encounters occur between different people (Nowak & Sigmund, 1998). An individual may act cooperatively toward one person because this improves their reputation and hence the chance that other persons will act cooperatively toward them (Buskens & Weesie, 2000; Granovetter, 1985; Raub & Weesie, 1990). For example, persons may lend a neighbor some tools not because they expect this neighbor to return the favor, but rather because it improves their reputation in the neighborhood and thereby allows them to tap into a wider exchange network.

Kin altruism, direct reciprocity, and indirect reciprocity are important mechanisms for cooperation, not only among humans but also other non-human animals (Bshary & Grutter, 2006; Carter, 2014). However, the mechanism often suggested to be most important for explaining the unique scale and diversity of human cooperation consists of social norms (Bicchieri, 2006; Elster, 1989; Fehr & Schurtenberger, 2018; House et al., 2019). Social norms are social standards of behavior prescribing how people should and should not behave (Elster, 1989; Fehr & Fischbacher, 2004a; Tomasello & Vaish, 2013; Young, 2015). For example, social norms tell us to wait in line at store counters, throw our trash in the bin rather than on the ground, and to respect others' personal space. In most conceptualizations of social norms, social norms indicate how people *ought* to behave (injunctive norms) rather than how most people actually behave (descriptive norms), although what *ought* and *is* may in practice overlap (Bicchieri, 2006; Cialdini et al., 1990). Humans have an evolved psychology for learning and conforming to social norms (Boyd et al., 2011; Henrich, 2015; House et al., 2019). When norms can be socially enforced by punishing norm violators and rewarding norm conformists, cooperation can be achieved even in situations where gains from reputation are minimal or absent (Fehr & Gächter, 2002). Indeed, the ability to develop and enforce social norms has been suggested to be one of the distinguishing characteristics of the human species (Fehr & Fischbacher, 2004b; Quervain et al., 2004). Of course, the different mechanisms promoting cooperation are not completely independent of each other and may interact. For example, social norms may prescribe direct reciprocity (Flache & Macy, 2006; Kimbrough & Vostroknutov, 2016) and signal which behaviors lead to good and bad reputations (Ohtsuki & Iwasa, 2006).

Altogether, previous research has identified several mechanisms to bring about cooperation for public goods in different sets of circumstances – from interactions within families to large-scale cooperation among unrelated individuals. However, we know less about whether and how cooperation can be sustained when circumstances change. Societal developments such as changes in the environment, population, and work arrangements can alter the conditions on which cooperation was established and render existing forms of cooperation ineffective. For example, cooperation during the industrial revolution led to mass production using fossil fuels that brought prosperity but eventually caused climate change. Despite knowing for decades that this form of cooperation is no longer sustainable, the use of fossil fuels continues to rise (Jackson et al., 2019). Next to rendering existing forms of cooperation ineffective, changing circumstances can also lead to a general decay in cooperation, for example in the form of decreasing voter turnouts (Hill, 2006), deterioration of public places

(Cialdini et al., 1990), and gradual collaboration failures in work organizations (De Dreu & Gelfand, 2008).

The SCOOP research program 'Sustainable Cooperation: Roadmaps to a Resilient Society' aims to study how cooperation can be sustained despite changing circumstances (SCOOP, 2019). By studying this question, the program hopes to contribute to resilient societies – societies that are able to maintain high levels of cooperation despite the sustainability challenges posed by changing circumstances. One of the main sustainability challenges identified by the SCOOP program is the arrival of newcomers into communities and organizations. Newcomers run the risk of being seen as free-riders, who benefit from public goods without contributing at first. They may initially not contribute because they are unaware of the prevailing social norms that regulate public goods or contribute differently because they have different needs and views for the public goods. These aspects of newcomer entry may also reduce incumbents' willingness to contribute to public goods, thereby threatening group cooperation altogether. At the same time, newcomers are necessary to maintain public goods over periods of time that span more than one generation. Incumbents will have to depart the group at some point, and the public goods they created will disappear if there is no one to replace them. Hence, sustainable public good provision requires cooperation between newcomers and incumbents. As part of the SCOOP program, this thesis studies whether and how changing groups consisting of newcomers and incumbents can cooperate for public goods.

In what follows, we first elaborate on the role of social norms in public good provision. We then describe how the arrival of newcomers can challenge social norms and thereby public goods. Afterward, we outline how we examine this problem in this thesis.

1.1.2. Social norms

Just like public good provision, social norms have been observed throughout all known human societies, from hunter-gatherer societies to modern nations (Fehr & Williams, 2018). Norms of egalitarian food sharing are found in hunter-gatherer societies (Wiessner, 2005), norms against 'crossing the picket line' are observed during strikes by trade unions (Naylor, 1989), norms requiring young men to join the army were observed during both world wars (Simkins, 1988), and norms of tax compliance uphold public goods that are nationally funded (Wenzel, 2004). Norms are so ubiquitous and influential in human societies and their public goods that they have been called 'the grammar of society' (Bicchieri, 2006) and 'the cement of society' (Elster, 1989). By indicating what the group considers the right and wrong thing to do, norms constitute an important mechanism of behavior regulation in groups (Ellemers et al., 2013). Conforming to group norms demonstrates that one is a "good" group member who is willing to forego selfish temptations for the group's benefit. Conversely, violating the group norms can cause other members to distance themselves from the norm violators or even exclude them from the group (Ellemers & Van der Toorn, 2015).

The enforcement of social norms typically occurs via informal sanctioning, i.e., the rewarding of norm-conforming behaviors and punishing of norm-violating behaviors. These sanctions can be material (e.g., withholding resources such as food or money) or social (e.g., social exclusion or withholding social approval), and both can be effective means of norm enforcement (Masclet et al., 2003). Rewards for norm conformity may consist of high status, good reputation, or material benefits (Balliet et al., 2011; Gallus, 2017). Punishment for norm violations may take the form of ostracism, ridicule, gossip, and verbal reproach (Boehm, 1999; Guala, 2012). For example, observations of hunter-gatherer societies show that members who free-ride in hunting and food sharing are ostracized by having to leave the camp (Gurven, 2004; Marlowe, 2010). In mining communities, strikebreakers were socially ostracized from community life by being excluded from football teams, bands, choirs, and other social activities (Francis, 1985). And in World War 1, young men who did not voluntarily join the British army were publicly shamed by community members who attached big red patches to the young men's front doors at night (Fehr & Williams, 2018; Simkins, 1988).

Lab experiments corroborate that a wide variety of norm enforcement mechanisms can promote public good provision. These include monetary punishments and rewards (Balliet et al., 2011; Van Miltenburg et al., 2014), ostracism (Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010), symbolic punishment and rewards such as peer (dis)approval (Dugar, 2013; Gallus, 2017; Masclet et al., 2003), and negotiation via verbal communication (Dawes et al., 1977; Isaac & Walker, 1988a). The enforcement of social norms not only helps to change free-riders' behaviors, but also serves as a signal to would-be free-riders that such behavior is not accepted and will be punished (Fehr & Gächter, 2000; Xiao & Houser, 2011). Some suggest this can make norms self-enforcing, i.e., the threat of norm enforcement can be enough to maintain conformity without the need for actual enforcement (Voss, 2001). This creates a stable pattern of behavior that is upheld by the collective belief in the norm and the perceived consequences of breaking it. An important precondition for this positive effect of norm enforcement is the existence of shared views on what constitutes the norm (Herrmann et al., 2008). In groups that lack shared views of what behavior is considered appropriate, norm enforcement via punishment can lead to counter-punishment and conflict (Nikiforakis et al., 2012; Rauhut & Winter, 2017).

Norms are thus regarded as one of the most important factors promoting cooperation for public goods. To be sure, there are also norms that can harm, rather than improve, the wellbeing of the group or some of its members (sometimes also referred to as public bads). Examples are norms of 'honor killings' (Kulczycki & Windle, 2011), norms of foot binding (Mackie, 1996), and norms of genital mutilation (Efferson et al., 2020). However, it is more common that norms promote cooperation for public goods (Cubitt et al., 2011; Curry et al., 2019; Kimbrough & Vostroknutov, 2016; Reuben & Riedl, 2013). In fact, one popular view is that social norms are solutions to cooperation problems, i.e., they prescribe actions that have positive externalities (cooperation) and proscribe actions that have negative externalities (free-riding) (Coleman, 1990; Ullmann-Margalit, 1977). This is not to say that cooperation for public goods is objectively good for everybody. For example, a country providing the public good of a national army is good for the country itself but bad for other countries that it is at war with. When we say that social norms promote public goods, we mean that they promote goods that bring (non-excludable) benefits for the group producing the good.

There are different views on how social norms emerge (Brennan et al., 2013), but many scholars agree that repeated interaction across the same set of members facilitates the emergence of social norms (Baldassarri & Abascal, 2020; Bendor & Swistak, 2001; Coleman, 1988; Ostrom, 2000; Przepiorka et al., 2022; Titlestad et al., 2019; Voss, 2001). That is, repeated interaction with the same group members facilitates a learning process that allows members to reach a shared understanding of what is appropriate or "good" behavior (Duffy & Ochs, 2009). This in turn helps group members to know what to expect from others and hold one another accountable for uncooperative behavior (Brennan et al., 2013). Hence, stability in group composition is an important factor in promoting social norms and cooperation for public goods. But what happens when groups change?

1.1.3. Group changes

Even though stability in group composition is important for social norms and public good provision, virtually all groups will encounter old members leaving and new members arriving at some point. Countries, cities, and neighborhoods change in composition due to migration, work organizations hire new workers and let go of existing workers who retire or move to other organizations, volunteer organizations and cooperatives attract new members and see other members leave, and so on. What is more, group changes are becoming increasingly common due to recent patterns of migration, globalization, and technological change (Tannenbaum et al., 2012). Immigration to Western Europe and North America has been growing in recent decades, which has spurred debates about whether and how immigrant groups and native groups can cooperate to provide public goods (Baldassarri & Abascal, 2020). The demise of the 'job for life' has led to increased employee turnover in work organizations, leading to questions about how work teams can maintain successful collaboration among new and old employees (Berton & Garibaldi, 2012; Cahuc et al., 2016). Group changes are said to be the "new normal" (Huckman et al., 2009) and the "rule rather than exception" (Grund et al., 2018). The importance of stable group compositions for social norms and public good provision suggests that such *changes* in group composition may threaten public goods.

Indeed, the notion that public goods and the social norms that support them are threatened by changes in group composition is common. For example, Elinor Ostrom writes "Out-migration may change the economic viability of a regime due to loss of those who contribute needed resources. In-migration may bring new participants who do not trust others and do not rapidly learn social norms that have been established over a long period of time." (2000, p. 153). Group members who expect to leave the group soon may be more tempted to not contribute to public goods because they lack a "shadow of the future". That is, they are less affected by potential future sanctions for not contributing and have less invested interest in the maintenance of the public good (Przepiorka & Diekmann, 2021). In turn, newcomers may not contribute to public goods because they lack a "shadow of the past". That is, they still have to learn the social norms in the group or they are accustomed to different social norms. Indeed, normative differences between newcomers and incumbents (i.e., members who were already in the group) are often argued to be a major challenge to cooperation for public good provision (Collier, 2013; Habyarimana et al., 2009; Hainmueller & Hopkins, 2014).

Because newcomers may be unaware of the social norms or are accustomed to different norms, they may initially not conform to the prevailing contribution norm in the group and be perceived as free-riders. This threatens cooperation, as people are often only willing to contribute to public goods if they trust that others will also contribute. Indeed, people have a strong aversion to being taken advantage of, and will often reduce their own contribution when they believe others do not contribute enough (Gächter, 2007; Thöni & Volk, 2018). This conditional cooperation implies that cooperation is fragile. Supposed free-riding by some members can spark free-riding among other members and set in motion a downward spiral leading to norm erosion and deterioration of public goods. As mentioned, norm enforcement can help to overcome free-rider problems, but it is only expected to help if people agree on the norm that is to be enforced (Herrmann et al., 2008). Accordingly, norm enforcement is often argued to be ineffective in culturally heterogeneous groups such as groups consisting of immigrants and native-majority members (Habyarimana et al., 2009; Winter & Zhang, 2018). Altogether, cooperation for public goods is thought to be threatened when newcomers do not contribute to the prevailing norm because they are unaware of it or because they support different norms.

A related threat is that the prevailing norm and form of public good provision among incumbents may be tailored to the incumbents' needs and wishes, and not to those of the newcomers. Norms often favor the ingroup over the outgroup (Bernhard et al., 2006; Romano et al., 2021), and such ingroup biases can lead to a disadvantaged position for newcomers. For example, public holidays in a group may be centered around the incumbents' religion while largely ignoring the newcomers' religion (e.g., paid leave during Christmas and not during Eid al-Fitr). Research suggests that newcomers are initially not seen as full ingroup members by the incumbents (Rink et al., 2013). Only after a socialization process, during which incumbents attempt to get the newcomers to conform to the incumbents' norms, do incumbents and newcomers perceive each other as full ingroup members (Levine & Moreland, 1994; Pratsinakis, 2018). Thus, only over time, newcomers are said to make a transition from 'outsiders to being insiders' (Bauer et al., 2007). Research on natives' perceptions toward immigrant groups suggests that this process can take decades, and even then certain immigrant groups are still not considered to be fully part of the native ingroup (Coenders & Scheepers, 2008; De Coninck et al., 2021; Lee & Fiske, 2006; Storm et al., 2017). As long as newcomers do not belong to the ingroup, incumbents' ingroup bias can prevent an inclusive form of public good provision that benefits all members. In sum, changing group compositions can not only threaten cooperation for public goods, but also lead to a form of public good provision that disadvantages newcomers.

1.1.4. Research aim and focus

Aims. There has been little empirical research on the relationship between group changes and public good provision, let alone research that tests the supposed underlying mechanisms of normative differences and ingroup favoritism between incumbents and newcomers. Most research on public good provision focuses either on stable groups ('partner design') or groups of strangers who only meet once ('stranger design'). While such research has greatly improved our understanding of public good provision, it leaves out a type of group that occurs often in real life – groups with a gradually changing composition due to the arrival of newcomers and departure of incumbents. There are only a few studies that relate changes in group composition to public good provision, and they report mixed results so far. One study found that the entry of 1-2 temporary newcomers into a group of four has a negative influence on cooperation for public goods because these newcomers contribute less than incumbents (Grund et al., 2015). Two other studies report that the entry of a single newcomer in a group of four does have a positive influence on cooperation if the newcomer has prospects of staying longer in the group (Duffy & Lafky, 2016; Sonnemans et al., 1999). Finally, one study examined the entry of four newcomers into a group of two incumbents and shows that cooperation is higher if the four newcomers enter at separate moments instead of jointly at the same time (Ranehill et al., 2014). While these studies are informative, they present only a small subset of the possible ways in which a group can change.

We have little evidence on the underlying mechanism of normative differences, and most of it comes from observational and case studies, which makes drawing causal inferences especially difficult (Collier, 2013; Habyarimana et al., 2009). Indeed, newcomers often differ not only from incumbents in their normative views, but also in their resources, social status, social opportunities, and several other factors. When these differences co-occur, it is difficult to pinpoint which of them is (or are) responsible for cooperation difficulties between newcomers and incumbents. Indeed, a recent review suggests that negative estimates for the effect of normative differences between immigrant groups and native groups on public good provision are confounded with other aspects of interethnic groups such as poverty and political instability (Baldassarri & Abascal, 2020). Longitudinal and experimental research is vital for a causal understanding of public good provision in changing groups.

In this thesis, we aim to contribute to our understanding of cooperation for public goods in changing groups. The thesis contains five empirical chapters addressing the following three main questions:

1) What is the relationship between changes in group composition and public good provision?
2) What is the effect of normative disagreement between newcomers and incumbents on public good provision?
3) To what extent does public good provision in changing groups involve ingroup bias?

Focus. We mentioned that this thesis is part of the interdisciplinary SCOOP program on the sustainability of cooperation under changing circumstances. Cooperation is considered sustainable if it is stable and valuable (SCOOP, 2019). Stability refers to ongoing cooperation, while value refers to the benefits that cooperation brings to the group members (and the wider social context). Applied to public good provision, sustainable cooperation thus means that (1) the *level* of public good provision is maintained and (2) that the *form* of public good provision brings value to the group members. In studying whether group changes affect the level of public good provision (research questions 1 & 2) and whether the form of public good provision involves ingroup biases (research question 3), this thesis focuses on both aspects of sustainable cooperation under changing circumstances.

The thesis incorporates insights from multiple disciplines, including sociology, social psychology, and behavioral economics. With regard to sociology, the thesis touches upon the three main sociological themes of culture, social relations, and inequality (Van Tubergen, 2020). It relates to (1) culture in the form of norms, (2) social relations in the form of changing social networks and their implications for cooperation, and (3) inequality in the form of unequal benefits derived from public goods between newcomers and incumbents. With regards to social psychology, the thesis studies what the arrival of newcomers does for (1) feelings of group identity among incumbents and newcomers, (2) the development of shared group goals and norms and commitment toward them, and (3) social dynamics of influence and conformity between newcomers and incumbents. Finally, the thesis incorporates aspects of behavioral economics via (1) the focus on public good provision as the central outcome, (2) the use of behavioral games, and (3) the inclusion of psychological insights into economic decision-making.

The boundaries between these disciplines are not clear-cut and are becoming increasingly blurred (Klein, 2017). The factors suggested to be relevant in one discipline do not work independently of the factors in other disciplines. For example, insights from psychology on how newcomers and incumbents identify with their group are important to explain sociologically whether members conform to norms and forego their economic incentive to free-ride on public goods. Thus, instead of studying psychological, sociological, and economic factors as distinct influences in separate disciplinary chapters, we study them as interrelated factors in each chapter. In doing so, we aim to capture the interplay between individual behavior, group norms, and collective action for public goods that cuts across disciplinary boundaries. Indeed, the SCOOP research program argues that complex societal problems can only be fully understood when insights from different disciplines and methods are combined.

One of the main sustainability threats identified by the SCOOP program is the arrival of newcomers as an external shock that tests the ability of existing cooperation arrangements. External in this case means that the newcomer entry is not initiated by the existing members of the group. This means that we do not focus on the related but separate literature on *endogenous* changes in groups, in which members themselves have control over who enters and who leaves (Guido et al., 2019; Gürerk et al., 2014). Experimental research suggests that endogenous changes in group composition allow groups to solve the free-rider problem by expelling free-riders and allowing only cooperative actors to join the group (Charness & Yang, 2014). Similarly, there is research on dynamic networks in which individuals can decide with whom they want to interact, allowing them to create and break ties with others based on others' cooperative behavior. Most economic experiments suggest that such dynamic networks can help to solve cooperation problems because they allow people to break with free-riders and connect with cooperative actors (Fehl et al., 2011; Melamed et al., 2018; Rand et al., 2011; Riedl & Ule, 2002; Wang et al., 2012), although some find mixed effects (Corten et al., 2020).

This type of *endogenous* group change – group change under the control of its members – can thus be a solution to free-rider temptations rather than a problem. While such research is certainly informative for real-life situations in which people can decide with whom they want to interact, there are also many real-life situations where incumbents have little say in how the composition of their group changes. In such cases, the group changes are largely *exogenous* (external) for the incumbents. For example, residents in neighborhoods or communities often have little control over who enters or leaves, and employees in many work organizations have to collaborate with colleagues that are chosen not by themselves but by their employers. In such situations, newcomers who deviate from the incumbents' norm or free-ride cannot simply be excluded, and may therefore threaten public good provision. Indeed, the essential characteristic of public goods is that members cannot readily be excluded from its benefits. In this thesis, we therefore focus on public goods in changing groups where excluding free-riders is unfeasible. This means that newcomers and incumbents have to find a way to sustainably cooperate for public goods despite their potential normative differences.

1.2. Methods

This thesis follows a multi-method approach. In particular, we combine three different methods: (1) lab experiments using public goods games, (2) analysis of longitudinal digital trace data from an online multiplayer game, and (3) a field experiment on a real-life public good. Each of these methods has its advantages and disadvantages, and they all involve different levels of abstraction. Simple public goods games are at one end of the spectrum of abstraction and real-life public goods at the other. Although each method can bring important insights, arguably a better understanding comes from combining different methods at different levels of abstraction. If results are consistent across the different approaches, we can have more confidence that they are not purely driven by the limitations of any particular approach (Buskens & Raub, 2013; Jackson & Cox, 2013). In what follows, we discuss each of the three methods and their advantages and disadvantages.

1.2.1. Lab experiments and the public goods game

Lab experiments are a widely used method in the natural sciences. With the exception of psychology, lab experiments are still relatively uncommon in the social sciences, although they are becoming more popular over time (Falk & Heckman, 2009; Jackson & Cox, 2013). Lab experiments are uniquely suited to test causal statements, as they provide a level of controlled variation that is not possible in observational studies. When finding that variable A predicts variable B in observational data, it is difficult to determine whether A causes B, or B causes A (reverse causality), or another omitted variable C causes both A and B (contextual confounding). Researchers can attempt to draw causal conclusions from observational data by statistically controlling for various confounders. However, unless all relevant control variables are captured and these control variables are measured without any error, this will only partially adjust for confounding (Westfall & Yarkoni, 2016). What is more, unless the observational data perfectly matches the model specifications, adding control variables does not prevent confounding (Miller & Chapman, 2001; Westfall & Yarkoni, 2016). Longitudinal observational data helps to reduce the chances of reverse causality and contextual confounding, but unfortunately does not eliminate them altogether (Leszczensky & Wolbring, 2022; VanderWeele & An, 2013). Lab experiments can prevent both the issues of reverse causality and contextual confounding by randomly assigning participants to different conditions. As we will see, lab experiments also have limitations, in particular with regard to realism and generalizability. However, they can serve as a useful complement to observational methods, in particular when the goal is to test causal relationships and mechanisms that are difficult to observe at a large scale.

Approaches to conducting lab experiments in the social sciences differ in the extent to which they abstract away from real-life contexts (Ariely & Norton, 2007; Camerer, 1996). One approach is to entirely strip the real-life context from the experiment, creating an abstract and simple setting. This may help to reduce unwanted influences of real-life complexities on behavior in the lab, and thereby allows for a potentially stricter control of the variable(s) of interest. A different approach is to import a richer context into the lab, for example by devising elaborate cover stories. This may help to engage the participants and to have them behave as they would in real life. In this thesis, the lab experiments mostly serve to abstract away from real-life contexts because the other methods of the thesis do incorporate richer contexts (the online multiplayer game and the field experiment on a real-life public good).

In particular, our lab experiments follow the paradigm of the public goods game. This game is a common tool in interdisciplinary research on public good provision, with contributions from scientific disciplines such as economics, sociology, psychology, biology, philosophy, anthropology, physics, political science, and computer science (Chaudhuri, 2011; Kollock, 1998; Rand & Nowak, 2013; Spadaro et al., 2022; Thielmann et al., 2021; Van Dijk & De Dreu, 2021). Initially, only game theorists used games to model human behavior, but soon the use of games gained popularity across the abovementioned disciplines (Deutsch, 1958; Thielmann et al., 2021). By now, more than a thousand studies have been conducted using the public goods game (Spadaro et al., 2022), leading to important insights into the motives, institutions, and dynamics related to public good provision under controlled settings (Balliet et al., 2011; Chaudhuri, 2011; Guido et al., 2019; Ledyard, 1994; Van Lange et al., 2013). It is rare that scientists from such a broad range of disciplines share a standardized tool like the public goods game. This has facilitated the comparison of results found across disciplines and societies, and contributed to a transdisciplinary understanding of public good provision. Because of the importance of the public goods game in the study of public good provision – and because we use it in multiple chapters of this thesis – it is worth discussing the game and its advantages and disadvantages in some more detail. Later, we also discuss contextualized games and field experiments.

The public goods game is sometimes also referred to as the voluntary contribution mechanism (VCM) or *n*-person prisoner's dilemma. It is designed to capture the dilemma between self-interested free-riding and contributing to the public good. The common version of the public goods game is presented in Figure 1.1. At the beginning of the game, each participant receives some monetary endowment (step 1). Each participant then has to decide how much of their endowment to contribute to a public good (step 2). The experimenter multiplies all contributions by a number above 1 and smaller than the number of participants in the game (step 3). Finally, the multiplied contributions are equally distributed among all members, including those who did not contribute to the public good (step 4). Because contributions to the public good are multiplied by a factor larger than 1, contributions increase the total payoff for the group as a whole, and the group earns the most money if everybody contributes their entire endowment. However, because the multiplier is smaller than the number of participants, each participant receives less from the public good than they contributed themselves, so each participant earns the most money by free-riding on the contributions of others.

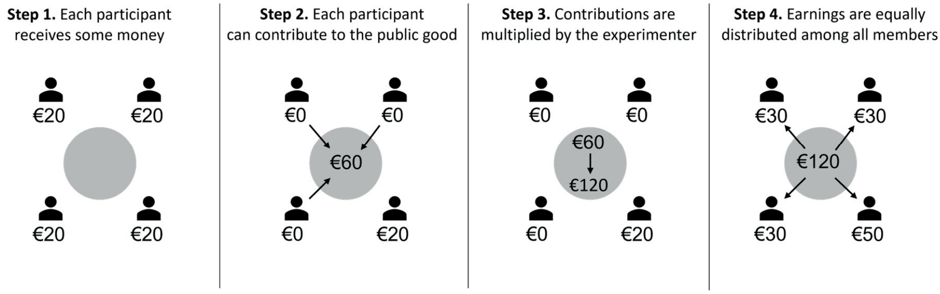


Figure 1.1. The public goods game.
Note: Each participant receives some monetary endowment, in this example 20 euros (step 1). Each participant then has to decide how much of their endowment to contribute to a public good (step 2). In this example, three members contribute all of their 20 euros, while one participant keeps the 20 euros. The experimenter then multiplies all contributions by some number, in this example by 2 (step 3). Finally, the multiplied contributions are equally distributed among all members (30 euros for each in this example), including those who did not contribute to the public good (step 4).

In the example of Figure 1.1, contributions are multiplied by 2 and then equally divided among four participants. This means that if a participant contributes 20 euros, this contribution becomes worth 40 euros. These 40 euros are then equally divided among the four participants, giving the participant who contributed only 10 euros, which is less than the 20 euros the participant initially had. In other words, you benefit others by contributing to the public good, but you yourself earn the most money by not contributing. The public goods game recreates the essence of the cooperation problem in public good provision; the group earns the most if everybody contributes to the public good, but each individual member earns the most by free-riding on others' contributions. This property makes it a social dilemma – a situation in which people have to choose between individual and collective interests. The public goods game is often regarded as a prisoner's dilemma with more than two people and a classic laboratory paradigm for studying collective action problems (Kube et al., 2015; Rand et al., 2009; Rand & Nowak, 2013).

People can have different motives or social preferences in the public goods game. Those with egoistical motives may attempt to free-ride on others' contributions. Those with prosocial motives may contribute to the public good. Still others may wish to minimize unequal outcomes or to follow what others are doing. Indeed, a wealth of evidence shows that people vary widely in their motives in the game and that these motives depend on the social context (Alexander & Christia, 2011; Fehr & Gintis, 2007; Van Lange et al., 2013). Using the public goods game by no means implies assuming that people behave only with the motive to maximize their own monetary payoff, sometimes referred to as the *homo economicus* model. On the contrary, the public goods game has been used to show the myriad of ways in which human behavior *deviates* from the homo economicus model (Henrich et al., 2001). Insights from psychology and sociology have enriched our understanding of how people behave in public goods games, often showing a large role for social norms guiding people's contributions (Fehr & Fischbacher, 2004a; Kimbrough & Vostroknutov, 2016; Ostrom, 2000).

1.2.2. Validity of the public goods game

Real-life public good provision is much more complex than the public goods game, so how can such an abstract game nevertheless help us to understand real-life public good provision? Indeed, the problem of how to achieve cooperation in the presence of free-rider temptations occurs in many different situations with many different complications – ranging from small-scale public goods such as cleaning your dishes at home or doing your share of team tasks at the office to larger problems such as providing national public goods and solving international pandemics and climate problems. To make progress in this wide variety of situations that have the property of free-rider temptations, we can benefit from a model that represents what is common in these situations without getting lost in the unique complexities of each particular situation. This is what makes the public goods game useful; it abstracts away from the complexities and focuses on the core free-rider problem. Once we understand behavior in the simple model, we can expand it to incorporate moderating variables that capture common subsets of real-life situations, following a method of decreasing abstraction (Lindenberg, 1992). In sum, it is the very complexity of reality that makes the use of abstract models and games helpful to improve understanding (Axelrod, 1984; Binmore, 2006; Van Dijk & De Dreu, 2021). Care needs to be taken that the level of abstraction does not become so extreme that it no longer helps to understand real-life problems. However, as we will see, the mechanisms and patterns studied with public goods games map onto several real-life observations.

Despite public goods games trading off realism for tractability, studies find that behavior in the game is related to behavior toward real-life public goods (Barr et al., 2014; Fehr & Leibbrandt, 2011; Gelcich et al., 2013; Gneezy et al., 2016; Hergueux et al., 2015; Laury & Taylor, 2008; Rustagi et al., 2010). Although there are also some studies providing counter-evidence (Galizzi & Navarro-Martinez, 2019; Levitt & List, 2007), most studies find support for a relationship between behavior in the public goods game and in real-life public goods. One interesting example comes from a study on Brazilian fishers (Fehr & Leibbrandt, 2011). When catching shrimp, Brazilian fishers use bucket-like traps in which holes are cut to allow immature shrimp to escape. The fishers can cut holes of any size, and this presents a public good problem. Large holes are better for the group of fishers as a whole because these allow immature shrimp to escape, thereby preserving the stock for future shrimping. Small holes are better for the fisher himself, as it increases the fisher's personal catch. A similar public goods problem arises when the fishers use fishnets to catch fish. Fishnets differ according to their mesh size, with smaller mesh sizes being more likely to catch immature fish, which leads to a larger catch for the fisher himself but depletes the future stock for all fishers. The study found that fishers who cooperate more in a public goods game also cooperate more in the field by cutting larger holes in their shrimping buckets and by using fishnets with larger mesh sizes. In another study, scholars conducted public goods games in 15 different small-scale societies in developing countries (Henrich et al., 2005). Not only did they find that behavior in the public goods game was related to the way social life was organized, the participants themselves also recognized similarities between the public goods game and everyday public good provision (Henrich et al., 2005, p. 811):

"The fact that group-level measures of economic and social structure statistically explain much of the between-group variance in experimental play suggests that there may be a relationship between game

behavior and patterns of daily life in these places. In several cases the parallels are striking, and in some cases our subjects readily discerned the similarity and were able to articulate it. The Orma [a pastoralist society living in Savanna-woodlands], for example, immediately recognized that the PGG [public goods game] was similar to the harambee, a locally initiated contribution that Orma households make when their community decides to pursue a public good, such as constructing a road or school. They dubbed the experiment "the harambee game" and contributed generously (mean 58% with 25% full contributors)."

Similarly, a study on Ugandan farmers found that real-life cooperation is related to cooperation in public goods games when the experimental treatment resembles the real-life environment (Grossman & Baldassarri, 2012). Other studies find that citizens who contribute more in public goods games are more likely to participate in local and national accountability institutions (Barr et al., 2014) and that community-forest users who contribute more in public goods games also engage more in collective action behaviors that support common forests (Bluffstone et al., 2020). In sum, while public goods games lack some of the realism from real-life public goods, behavior in these games seems to be predictive of behavior toward real-life public goods.

A common concern is that the incentives or 'stakes' in public goods games are not high enough for the participants to take the game seriously and invest cognitive effort to understand the implications of their decisions. Typically, the stakes are such that participants on average earn the local hourly wage. These stakes might be high enough if we want to understand everyday contributions to public goods, but may not reflect very costly contributions to public goods such as foregoing one's weekly or monthly wage when joining a strike. However, empirical research suggests that increasing the stakes does not appreciably change contribution and punishment behavior in public goods games (Kocher et al., 2008). In the prisoner's dilemma, people playing with stakes averaging over $20.000 were even found to cooperate at qualitatively similar levels to people in prisoner's dilemmas with regular stakes (Van Den Assem et al., 2012). Finally, the aforementioned significant effects of symbolic (non-monetary) punishments and rewards show that monetary stakes are not the only possible incentive to take social situations in the lab seriously. Indeed, as we mentioned, people may have several motives that drive their behavior in the game, both monetary and non-monetary.

Many studies on public goods games rely on student participant pools, which understandably leads to concerns about generalizability. One common response to this concern is that public goods experiments are meant to test theories whose predictions are often independent of assumptions concerning participant pools (Falk & Heckman, 2009). However, we can also turn to empirical evidence to assess how results depend on participant pools. Research suggests that students contribute less than non-students in public goods games, the difference is practically small but significant and consistent (Belot et al., 2015; Carpenter & Seki, 2011; Gächter et al., 2004; Stoop et al., 2012). However, most public goods experiments are not designed to examine absolute contribution levels, but rather the *effect* of some manipulated variable(s) on contribution levels. Hence, of particular interest is whether observed effects in a particular sample generalize to other samples.

One informative study in this regard comes from Herrmann et al. (2008), who conducted public goods games in 16 participant pools that were spread out all over the world and differed strongly in cultural and economic backgrounds. The authors found that levels of absolute contributions and punishment differed significantly between participant pools. However, the manipulated variable in each participant pool was whether norm enforcement via punishment was possible. Without punishment, a declining trend in contribution levels was visible in nearly each participant pool. The study found that punishment had a stabilizing effect on contributions in almost all participant pools. The level at which contribution levels stabilized differed between participant pools, and depended on the culture and norms in the country of the participant pool (Gächter et al., 2010; Herrmann et al., 2008). This suggests that a similar mechanism of using punishment to enforce norms applies across the different participant pools, but because the norms differ between participant pools, they end up with different contribution levels. Hence, observed mechanisms may generalize better than absolute contribution levels. This is in line with other research showing that punishment effects on cooperation observed in a nationally representative Dutch participant pool are qualitatively similar to effects found in student participant pools (Egas & Riedl, 2008). What is more, a recent meta-analysis of 1506 studies on prisoner's dilemmas and public goods games reassuringly found that cooperation did not differ much between regions, societies, and cultures (Spadaro et al., 2022). Chapters 2, 3, and 5 of this thesis involve lab experiments using public goods games. Before we turn to a summary of these chapters, we first discuss other methods used in the thesis.

1.2.3. Contextualized public goods in an online multiplayer game

The second method used in this thesis involves contextualized public goods games – these are games that bring in some real-life context while keeping the well-constrained and standardized environment of games. As such, they form a middle ground between pure public goods games and real-life public goods. For example, the public goods game may be contextualized as a group of employees working together in a team (Chen et al., 2009; Van Gerwen et al., 2018) or a group of people stranded on an island that need to build a raft to leave the island (Baptista et al., 2013). Such scenarios may help participants to engage with the problem of public good provision and pay more attention to it, thereby increasing ecological validity. However, the risk is that their behavior will be driven by particular features of the scenario rather than the general public good problem. This is why a combination of different approaches may give the most complete answer.

Even when results are consistent across contextualized and abstract public goods games, they can still be impacted by limitations of lab experiments more generally. Some of these were already mentioned in the 'validity' section of public goods games (e.g., unrepresentative participant pools and low stakes) but there are two more limitations that deserve to be mentioned. First, participants in lab experiments know that they are being observed, which might affect their behavior, for example through experimenter demand effects (Zizzo, 2010). Second, lab experiments can typically only run for a couple of hours at most, so only short-term public good provision can be studied. Both of these limitations can be addressed by turning to another type of game – online multiplayer games. These are games played by people for fun, often over long periods of time and from the comfort of one's home. In the scientific literature, these games are sometimes also referred to as virtual worlds. For example, the game EverQuest II has been used to analyze networks and personality (Burt, 2012), the game

Runescape has been used to test economic theories (De Sousa & Munro, 2012), and games such as Second Life and World of Warcraft also offer opportunities for innovative scientific research (Bainbridge, 2007; Chesney et al., 2009; Innocenti, 2017).

Because players of these games are not aware of taking part in a scientific study and because gameplay occurs over much longer periods of time than typically possible in a lab study, they largely overcome the aforementioned limitations of contextualized public goods games in the lab. Although some players may realize their behavior can be seen by the game creators, this is less likely to affect them than being observed by experimenters. What is more, data from these virtual worlds often span large populations of players with near-full information on a standardized environment, a combination that is difficult to achieve both in the lab and in the field. Of course, there are also some particular disadvantages to this method. For example, the group of gamers may not be representative of the wider population, people may behave differently in contextualized games than in real-life, and data access is at the discretion of the company running the games. In Chapter 4, we use large-scale and long-term data from the online multiplayer game Ikariam, which has public goods games deliberately built in by the game designers. The game is set in an ancient Greek archipelago, and each island of the archipelago involves a public good. The public good is contextualized as a sawmill that produces wood at a rate depending on contributions by all players on the island.

1.2.4. Field experiments

Finally, we turn to field studies involving real-life public goods. We discussed before that real-life public goods are complex, often involving many interrelated factors that together influence how people contribute to the public good. This makes isolating the effect of any particular factor difficult in observational studies of public goods, and is why lab experiments using (contextualized) public goods games are useful. Still, if we want these lab studies to inform real-life cases of public good provision, it seems prudent to test whether results from the lab translate to real-life cases. In other words, we also want to know whether theoretical predictions are supported by data from real-life public good provision. Some examples of studies on real-life public goods were already mentioned in the 'validity' section of this chapter. However, compared to the large body of abstract public goods game studies, field studies are still rare. What is more, field studies are often observational, which carries the aforementioned risks of confounding by unmeasured variables. A small but growing research area involves the use of field *experiments* on cooperation and real-life public good provision. This combination of the experimental method and real-world situations allows for causal inferences in naturally occurring contexts.

The newcomer-incumbent relation most often studied in field experiments is that between immigrants and native-majority members. Field experiments have shown that native-majority members are more likely to cooperate with native-majority members than members with an immigration background (Aidenberger & Doehne, 2021; Choi et al., 2019, 2021a, 2021b; Zhang et al., 2019). What is more, these immigrant members are more likely to be sanctioned for violating norms than native-majority members (Aidenberger & Doehne, 2021; Mujcic & Frijters, 2020; Winter & Zhang, 2018). These studies suggest that the immigrant-native relation is a salient newcomer-incumbent relation in real-life cases of cooperation and norm enforcement. In Chapter 6, we report on a field experiment to study the free-rider problem in public transport. In particular, we examine whether people display ethnic group biases in facilitating or sanctioning free-riding in public transport.

Field experiments also have their limitations. Typically, it is more difficult to manipulate a real-life social environment than a laboratory environment. This means that field experiments are constrained in the type of variables that they can examine, and often focus on rather small-scale interventions such as nudges instead of systemic interventions that affect structural elements of public good provision (e.g., societal income inequality). What is more, field experiments are often limited to recording the behavior of participants, while lab experiments can more readily tap into the participants' attitudes and experiences, and can even include physiological or neurological measures (Scheepers & Derks, 2016). Virtual worlds often have richer data on the social environment and long-term patterns of behavior than possible in field experiments, and allow for more control of the social environment (Bainbridge, 2007). Lab experiments, contextualized games, and field experiments complement each other. Each method has its own advantages and disadvantages, and by combining them we can attempt to overcome the limitations of any particular method. Table 1.1 shortly summarizes the main strengths and weaknesses of our different methods.

The replicability of scientific research has been called into question in recent years (Errington et al., 2021; Open Science Collaboration, 2015). Across our different methods, we take measures to increase the replicability of our research. In particular, replication chances have been suggested to increase with high statistical power, preregistration, and methodological transparency (Protzko et al., 2020). Therefore, we provide power calculations for each study (in the chapter itself or in the online pre-registration), pre-registered all studies for which we collected the data ourselves, and made the data and code underlying our own data collections openly available (the links are provided in the first page of each chapter).

Table 1.1. The strengths and weaknesses of our different methods.

	Lab experiments	Contextualized game	Field experiment
Internal validity • experimental control • tractability	+	+/-	+/-
External validity • realism • representativeness	-	+/-	+
Tracking long-term behavior • repeated observations • duration and frequency	+/-	+	-

Note: + strength, - weakness, +/- neutral.

1.3. Chapter summaries

The five empirical chapters in this thesis address different sub-questions related to public good provision in changing groups. Table 1.2 gives a broad overview of the sub-questions and methods per chapter and how these relate to our three main research questions. We present the chapters in the chronological order they were written. We start in Chapters 2 and 3 by investigating a key mechanism that has been suggested to determine whether newcomers and incumbents will be able to cooperate – their (dis)agreement on social norms. These chapters use a lab experiment to test the causal effect of normative disagreement on public good provision, both in newly formed groups and in groups consisting of newcomers and incumbents. In Chapter 2, we report on the part of the experiment related to newly formed groups, which lays the groundwork for Chapter 3, in which we report on the groups consisting of newcomers and incumbents. In Chapter 4, we zoom out to examine the overall relationship between changes in group composition and public good provision. Because there are many ways in which groups can change, a systematic investigation into this relationship requires large-scale data involving a broad range of group changes. Additionally, groups need to be tracked for long periods of time to observe how cooperation develops after group change. We make use of an online multiplayer game that fits both of these requirements.

In Chapters 5 and 6, we move our focus from the *level* of cooperation to the *form* of cooperation. In particular, we examine to what extent cooperation in changing groups involves ingroup bias. In Chapter 5, we use a lab experiment to study how norms of ingroup cooperation can impede collective cooperation for public good provision, both in newly formed groups and in groups consisting of incumbents and newcomers. This requires a type of public goods game in which people face not only a dilemma between free-riding and contributing to a public good, but also between contributing to a subgroup public good and a collective public good. This is captured in the multilevel public goods game, in which different subgroups are nested within a larger collective group and participants have to decide at which level to cooperate (if any). In Chapter 6, we use a field experiment to study another type of ingroup bias in public good provision – this time on how people respond to others who free-ride in public transport. In particular, we examine if members with an immigration background are less likely to be allowed to free-ride than native-majority members and more likely to face norm enforcement when attempting to free-ride. In what follows, we summarize the main results of each chapter.

Table 1.2. Overview of chapters.

Main research question	Chapter	Sub-questions	Method
What is the effect of normative disagreement on public good provision in changing groups?	2	a) What are the normative views and expectations about public goods with heterogeneous returns and to what extent do these vary across people and time? b) What is the effect of normative disagreement on public good provision in newly formed groups?	Lab experiment using a public goods game with heterogeneous returns
	3	a) What is the effect of normative disagreement on public good provision in groups consisting of newcomers and incumbents? b) What is the effect of normative disagreement on newcomer-incumbent relations in terms of group identification, the emergence of a social norm, and costly punishment?	
What is the relationship between changes in group composition and public good provision?	4	a) What is the relationship between changes in group composition and cooperation in terms of contributions to a public good? b) Can this relationship be attributed to the contributions of the newcomers, the incumbents, or both? c) How do newcomers' contributions develop over time?	Longitudinal analysis of data from an online multiplayer game
To what extent does public good provision in changing groups involve ingroup bias?	5	a) What is the effect of punishment on subgroup and collective cooperation in multilevel public good problems? b) Do these effects depend on whether the multilevel structure is present from the start or arises due to migration across subgroups?	Lab experiment using a multilevel public goods game
	6	a) Is there ingroup bias in helping behaviors toward free-riders? b) Is there ingroup bias in norm enforcement toward free-riders?	Field experiment on free-riders in public transport

1.3.1. Chapter 2

Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels.

We test the causal effect of normative disagreement on contributions in a public goods game using a lab experiment with 192 participants. The lab experiment is divided into two parts; the first involves newly formed groups and the second involves groups consisting of newcomers and incumbents. This chapter is based on the first part and thus focuses on newly formed groups. Previous experiments only tested the effect of normative disagreements indirectly, by comparing situations with more and less normative disagreement without manipulating the level of normative disagreement. To directly test the influence of normative disagreement, we first elicit participants' views on the appropriate way to contribute to a public good. We then use this information to sort people into groups. In one condition, we sort participants with similar normative views together, while we sort participants with dissimilar normative views together in the other condition. Groups in both conditions subsequently make several contribution decisions within their groups, allowing us to test whether contributions are lower in the condition with normative disagreement. We make use of a public goods game with heterogeneous returns and peer punishment because prior research suggests that the potential for normative disagreement is larger in this type of public goods game (Reuben & Riedl, 2013).

We find that participants in the condition with normative disagreement do not contribute significantly differently from participants in the condition without normative disagreement. Hence, we do not find a negative effect of normative disagreement on cooperation for public goods in newly formed groups. Instead, groups seem to reach a compromise between the different normative views of their members; the group-average normative view correlates highly with the group-average contribution. The results suggest that norms can sustain cooperation even in situations of normative disagreement in newly formed groups, and lay the groundwork for the next chapter on normative disagreements between newcomers and incumbents.

1.3.2. Chapter 3

Cooperation between newcomers and incumbents: The role of normative disagreements.

We test the causal effect of normative disagreement between newcomers and incumbents on contributions in a public goods game using 192 participants. We make use of the same lab experiment and manipulation as in Chapter 2, but this time using the data from the second part of the experiment. This second part uses a group change manipulation that creates groups consisting of both newcomers and incumbents. We test not only whether normative disagreement between newcomers and incumbents affects contributions to the public good, but also newcomer-incumbent relations in terms of group identification, the emergence of a social contribution norm, and costly punishment. While we find that normative disagreement does not affect contributions to the public good, we do find that it negatively affects the emergence of a shared social norm and lowers feelings of group identification. We furthermore find that incumbents enforce their own norm on the newcomers by punishing them when they deviate from it, and they punish free-riding newcomers more so than free-riding incumbents. Consequently, the way public goods are provided is tailored mostly to the incumbents' views.

The main finding – that public good provision is possible despite normative disagreements between newcomers and incumbents – is in line with a recent literature review suggesting that prior negative effects of normative differences on cooperation are due to unaccounted confounding factors in observational studies (Baldassarri & Abascal, 2020). This underlines the importance of experimental research for causal inference. While the finding that normative disagreements between newcomers and incumbents need not harm public good provision is reassuring, our results also imply that focusing only on whether a public good is provided may not give the whole picture. The results on group identification and norm enforcement reveal that the way the public good is provided may not always be equally valuable for everyone. Newcomers start by contributing according to their own normative view, but this leads to punishment that induces them to conform to the incumbents' view. Public goods may thus be provided despite normative disagreements, but they are based mostly on the incumbents' standards and are accompanied by lower group identification.

1.3.3. Chapter 4

Human cooperation in changing groups in a large-scale public goods game.

We analyze data from the online multiplayer game Ikariam, which has public goods games deliberately built in by the game designers and involves a broad range of group changes as players progress through the game. Gameplay in Ikariam is context-rich, occurs over a time span of many months, is free of observer bias, and involves a diverse participant pool. We analyze longitudinal data on about 1.5 million contribution decisions for public goods by about 135 thousand players in about 11.3 thousand groups. We examine the relationship between these contribution decisions and a total of about 234 thousand changes in group composition. We find robust evidence that changes in group composition relate negatively to contributions to the public good. This negative relationship holds for groups of different sizes, different magnitudes of newcomers, and newcomer entry at different time periods.

Although incumbents slightly reduce their contributions when newcomers enter the group, the negative relationship is mainly driven by players contributing less as newcomers than as incumbents. However, as players spend more time in their new group, they increase their contributions to the incumbents' level. This suggests that, in the process of moving from newcomer to incumbent status, individuals contribute more and conform to the incumbents' contribution norm. These findings are in line with the theoretical mechanism that newcomers need time to become accustomed to the prevailing contribution norm in the group. We also find that the newcomer-incumbent difference in contributions is smaller when there is less inequality between newcomers and incumbents in terms of contribution capacities and benefits of the public good. We conclude that, rather than marginalizing the short-term low contributions of newcomers, it is better to realize their long-term contribution potential by giving them the time to adjust and put them in a better position to contribute to public goods.

1.3.4. Chapter 5

Cooperation, punishment, and group change in multilevel public goods experiments.

The standard public goods game involves a single group, and people thus have to choose between not contributing to benefit themselves and contributing to benefit their group. However, in many real-life instances of public good provision, there are multiple groups nested within a larger collective group. Such public goods have been labeled multilevel public goods. For example, multilevel public good problems may arise in multiethnic societies consisting of the native-majority group and immigrant groups. In multilevel public good problems, individuals have to decide between not contributing, contributing to subgroup public goods, and contributing to collective public goods that also benefit other subgroups. Collective public good provision is the best outcome for everybody combined, but ingroup biases or self-interest can lead to subgroup public good provision or no public good provision at all.

Because there are multiple public goods that one can contribute to in the multilevel public goods problem, there is the risk that people disagree about which public good to provide. What is more, ingroup biases can lead to subgroup public good provision at the expense of collective public good provision. Norm enforcement via peer punishment is known to promote public good provision in single-group scenarios (Chaudhuri, 2011), but the heightened risks of disagreement and ingroup bias in multilevel public good problems may limit effective norm enforcement for (collective) public good provision. We conduct a lab experiment on the multilevel public goods game with 220 participants to examine the effect of punishment on subgroup and collective public good provision.

We find that punishment has limited effectiveness in multilevel public good problems. In particular, punishment only promotes cooperation for collective public goods if people have prior experience with solving single-group public good problems. Groups that do not have this prior experience fail to achieve high levels of collective public good provision even when punishment is possible, partly because they contribute to subgroup public goods at the expense of collective public goods. The results refine the boundary conditions for the effectiveness of norm enforcement via punishment and suggest that 'starting small' is important for successful collective public good provision. These findings may be indicative of how cooperation can be sustained in modern societies that become increasingly multicultural through group changes and immigration.

1.3.5. Chapter 6

Double standards in facilitating norm violations: A field experiment on attempted free-riding in public transport.

In the last chapter, we look at another form of ingroup bias in public good provision – whether people behave differently toward free-riders based on the free-riders' group membership. We conduct a field experiment in which confederates behave as potential free-riders nearby Dutch train stations. The confederates approach travelers who are about to go through check-in gates and request to follow them without checking in themselves, i.e., they request assistance with free-riding in public transport. We observe whether travelers enforce the public goods norm by rejecting this request or violate the public goods norm by helping the confederate to free-ride. We hired five confederates with a native-Dutch background and five with an immigration background, which together approached 801 travelers at three train stations.

We find robust evidence for ingroup bias in the treatment of free-riders. Confederates with a native-majority background are more likely to receive help with free-riding than confederates with an immigration background across all three train stations, travelers of all age groups, and all experimental sessions. The verbal answers that travelers gave to the confederates were subsequently rated by 57 independent coders. Verbal answers received by free-riders with an immigration background were rated as more disapproving and less helpful. These results suggest that members with an immigration background are more likely to face norm enforcement when free-riding in the public good of public transport. This is in line with the lab experiment in Chapter 3 showing that newcomers face more norm enforcement when free-riding in the public goods game.

Structure of chapters

The chapters were written such that they can be read independently from each other and in any order. This also means there is some unavoidable overlap between chapters, mainly in the chapters' Introduction sections. What is more, most chapters are written for interdisciplinary journals whose article structure deviates somewhat from the conventional structure in sociology journals (Chapter 3 being the exception). In these chapters, the Methods section appears at the end, and the theory is integrated into the Introduction section instead of being a separate section.

1.4. Conclusion and discussion

1.4.1. Conclusions

Public good provision is crucial for the functioning of prosperous societies. While a large body of research shows how to get cooperation going for public goods in stable groups, the dynamics of public good provision under changing group compositions have received less attention. We studied cooperation for public goods in changing groups, with a particular focus on the norm dynamics that arise when newcomers enter a group. Newcomers run the risk of being seen as free-riders when they are unaware of the prevailing contribution norm or accustomed to different norms. The potential friction this causes in the group has led several scholars to question whether changing groups can sustain cooperation (Collier, 2013; Habyarimana et al., 2009; Ostrom, 2000). For instance, can sports clubs that rely on volunteering parents (e.g. for transportation or bar duty) continue to offer sports facilities for children whose parents are not aware that they should contribute or disagree to contribute in a particular way? And if so, does the form of volunteering take into account the needs of newly joining parents or are old ways maintained that may no longer be efficient? We examined such issues in contexts where group changes are beyond the control of incumbents. This means that incumbents could not simply solve the problem by allowing only cooperative newcomers or excluding newcomers who do not contribute according to incumbents' standards. In doing so, we studied whether and how newcomers and incumbents can sustainably cooperate despite their potential normative differences. We organized this research problem into three main research questions, which the five empirical chapters aimed to answer.

Our first research question – what is the relationship between changes in group composition and public good provision – was answered with large-scale data from an online multiplayer game. The results suggest that the relationship is mostly negative in the short-term; people contribute less as newcomers than as incumbents, meaning that groups consisting of more newcomers obtain lower contributions. Because individuals in the game occupy both the roles of newcomer and incumbent, we could compare within individuals how contribution behavior changes depending on their role in the group. Hence, we could largely exclude individual dispositions as explanations for differences between newcomers and incumbents. That we find a contribution difference between newcomers and incumbents while ruling out individual dispositions suggests that just having a 'newcomer status' can already reduce people's tendency to contribute. However, the long-term contribution potential of newcomers is as high as those of incumbents. Newcomers increase their contributions over time and eventually contribute more in line with the incumbents' contribution norm. These findings are consistent with models of group socialization (Levine & Moreland, 1994). For example, the model of organizational socialization argues that new employees in work organizations undergo a process of learning and adapting in order to assume a role that aligns with the organization's needs and one's own needs (Bauer et al., 2007).

The results suggest that changes in group composition are indeed a relevant factor in public good provision and can be a challenge for sustainable cooperation. However, this does not mean that changes in group composition are bad or should be avoided. In fact, one sure way to let public goods deteriorate and eventually disappear is to never allow any newcomers. Incumbents will have to leave the group at some point, and if there is nobody to replace them, the public goods they provided will cease to exist. For example, pension systems and work organizations can only be maintained over time if departing incumbents are replaced by newcomers. Newcomers are thus necessary to uphold public goods over periods of time that span more than one generation. As we have seen, newcomers may initially contribute less than incumbents, but have high long-term contribution potential. Groups can benefit from this long-term potential if they are patient and resist the temptation to exclude or expel newcomers who initially do not contribute. Knowing that newcomers will contribute to the public good after a period of adjustment may help incumbents to be patient and avoid newcomer-incumbent tensions driven by anxieties that the public good will disappear altogether.

Allowing newcomers the time to integrate and putting them in a position to contribute effectively helps in this regard. For example, new residents may be more likely to volunteer at neighborhood events once they are settled and have become part of the community (Ghimire & Skinner, 2019), and immigrants may be more likely to contribute to charitable organizations once they have secured a job and are integrated in the host country (Bekkers & Wiepking, 2007; Hainmueller et al., 2016). Only by incorporating newcomers, can cooperation for public good provision be sustainable in the long-term. To further understand the conditions for sustainable cooperation in changing groups, we can turn to the results on the underlying norm dynamics between newcomers and incumbents. This brings us to the second research question.

The second research question – what is the effect of normative disagreement between newcomers and incumbents on public good provision – was addressed mostly with lab experiments. Previous observational research suggested normative differences between incumbents and newcomers to be an important factor impeding cooperation (Collier, 2013; Habyarimana et al., 2009; Hainmueller & Hopkins, 2014). However, because newcomers and incumbents usually differ on several additional aspects (e.g., status, income, religion), it was so far unclear whether normative differences are responsible for cooperation problems or some other mechanism. Our experiments suggest that normative disagreement does not lower public good provision. This implies that normative disagreement is not the key mechanism driving cooperation problems in changing groups. Instead, it mostly seems that newcomers over time learn the incumbents' norm, and the time this takes depends on the complexity of the public good provision.

The public goods game forms a very simple environment, which makes it relatively easy for newcomers to learn the incumbents' contribution norm. Accordingly, we found that newcomers very quickly learned and conformed to the incumbents' norm in the public goods game (Chapter 3). The online multiplayer game Ikariam forms a more complex environment, in which public good provision happens in a context-rich setting, over a longer period of time, and across a broader range of players with different resources and needs. These features make it more difficult to learn the incumbents' contribution norm, and we indeed saw that it took newcomers considerably more time in Ikariam to contribute in line with the incumbents (Chapter 4). This suggests that newcomers adapt to the incumbents' contribution norm at a speed that depends on the complexity of the public good and the norm supporting it. Thus, next to being patient, one potential way to facilitate cooperation in changing groups is to provide newcomers with clear and complete information on the prevailing contribution norm.

However, we also found that incumbents impose their own norms of public good provision on the newcomers, that newcomers were punished harsher for free-riding than incumbents, and that normative disagreement led to lower feelings of group identity, in particular among newcomers. These findings suggest that the way public good provision is organized in situations of normative disagreement is not always equally valuable for everyone and can have unwanted side-effects in terms of group identity. These latter findings provide some tentative answers to our third research question – to what extent does public good provision in changing groups involve ingroup bias? We employed additional lab and field experiments to delve further into this question. We focused in particular on ingroup biases in (1) social sanctions for free-riding on public good provision and (2) the form of public good provision. We find evidence for ingroup biases in both these domains. In the first domain, we find that newcomers receive more social sanctions when free-riding in abstract lab settings and in real-life public good settings. In the second domain, we find that when people have to decide between (exclusive) subgroup goods and (inclusive) collective goods that also benefit (new) members from outgroups, a significant share opts for subgroup goods. However, this bias can be overcome if – rather than requiring people to cooperate with outgroup members immediately – they can gradually expand their circle of cooperation. For example, it may help if people can first establish cooperative relations with familiar neighbors (e.g., in terms of borrowing and lending tools) before they are required to cooperate with new and unfamiliar neighbors.

Thus, our research also revealed that it may not always be desirable for newcomers to adapt to the incumbents. This one-way adaptation process neglects the needs of newcomers and can lead to a form of public good provision that is only valuable for the incumbents. This threat is more difficult to detect; it is missed if we focus only on the *level* of public good provision and not on the *form* of public good provision. A public good may be maintained at a high level, but if it only provides benefits for a subset of its members, it may not be optimal or sustainable. Sustainable cooperation involves adjustment to changing conditions; retaining public goods despite a continuing influx of newcomers with new needs may require the nature and content of public goods to be adjusted over time (SCOOP, 2019). To some this may feel like a 'loss' of public goods that raises tensions, but in the long-term it is necessary for changing societies to survive. This can be compared to 'creative destruction' in organizations – survival in dynamic and competitive industries requires organizations to continuously replace older and less efficient activities with innovative and more efficient activities (Igami, 2017; Schumpeter, 1942). This can involve a sense of loss of identity and discontinuity that leads to resistance, but organizations that are able to successfully manage these changes can emerge stronger in the long run.

Our multi-methods approach allows us to examine the robustness of findings across different methods and levels of abstraction. If comparable results are obtained using different methods, they are arguably more robust. The first relatively robust result is that newcomers initially contribute *differently* from incumbents. Newcomers initially contribute according to different normative views than incumbents in the standard public goods game (Chapters 2-3), they initially contribute less than incumbents in the contextualized public goods game (Chapter 4), and individuals with an immigration background differed from native-majority members in how they uphold the norm of paying for public transport (Chapter 6). Second, the finding that newcomers conform to the incumbents' views over time was present both in the standard and contextualized public goods game, although the time to conform differed between studies (Chapters 3 and 4). Third, that the resulting form of public good provision involves ingroup bias was observable in the standard public goods game (Chapter 3), the multilevel public goods game (Chapter 5), and the field experiment on public transport (Chapter 6). Fourth, that free-riding is met with norm enforcement more so for newcomers than incumbents was visible both in the standard public goods game and the field experiment in public transport (Chapters 3 and 6).

Finally, we find across all studies that the *role* of being a newcomer or incumbent in the group is important for how people themselves behave and how they are treated by others. Some theoretical accounts suggest that differences between newcomers and incumbents in public good provision are explained by differences between them in their underlying dispositions to cooperate (Collier, 2013; Gereke et al., 2021). According to this view, groups should mainly be concerned with avoiding the entry of "bad apples that spoil the barrel" (Grund et al., 2018). Our thesis shows that differences between newcomers and incumbents can occur without any differences in their underlying dispositions. In all our studies, newcomers and incumbents came from the same participant pool and thus did not differ in their underlying attributes. In the online multiplayer game, players had to undergo both the roles of newcomers and incumbents. We found that the same players contributed differently depending on whether they were a newcomer or incumbent. In the lab experiments, the roles of newcomers and incumbents were randomly assigned, so there are on average no differences between them in any underlying dispositions. Still, we found that newcomers were treated differently than incumbents (e.g. in normative influence and received punishment for free-riding). In the field experiment, all confederates made the request to free-ride using the same neutrally phrased sentence. Yet, people responded differently toward confederates with an immigration background than confederates with a native-majority background. Hence, in none of the chapters can the observed incumbent-newcomer differences be attributed to differences in their underlying dispositions to cooperate. Instead, just the role of being a newcomer seems to already importantly affect both newcomers' own behavior and how they are treated by incumbents.

1.4.2. Limitations and future research

As with any research project, there are several limitations that should be taken into account. First, recent research suggests that 30-60% of findings disappear or weaken when independent research teams attempt to replicate them (Camerer et al., 2016, 2018; Errington et al., 2021; Nosek et al., 2022; Open Science Collaboration, 2015). Sometimes such 'replication failures' may be attributed to alterations in the experimental design or statistical analyses, but other times findings fail to replicate even under nearly identical designs and contexts (Fabrigar et al., 2020). To some extent, we were able to assess replicability for the online multiplayer game across game servers from different countries and for the field experiment across different locations within the Netherlands. However, the lab studies were carried out at one location and would therefore particularly benefit from replication efforts. Although we took steps to increase the chance of replication (see Methods), results have to be treated with some degree of uncertainty before they are independently replicated, including ours. Future research can reduce such uncertainty from the outset by moving toward collaborative knowledge creation (Ellemers, 2021) and big-team science (Coles et al., 2022), e.g., by designing multi-lab studies in which the same hypothesis is tested across multiple sites and countries.

There are many aspects of changing groups that can affect cooperation, and we could not study them all. We discuss a few aspects of changing groups that could be further examined in future research. First, we focused mostly on *normative* differences between newcomers and incumbents. While this is a theoretically important cultural factor, there are also economic, religious, and ethnic differences between newcomers and incumbents that can be of importance for public good provision. For example, incumbents often have more resources and status than newcomers. Such inequality in resources and status can create tensions and impede cooperation (Hauser et al., 2019; Zhang et al., 2019), thereby providing alternative mechanisms through which group changes can affect public good provision. Our research suggests that the norm dynamics between newcomers and incumbents matter for cooperation, but this does not imply that it is the only important factor or even the most important one. What is more, there is often an interplay between the different newcomer-incumbent aspects and their effects on cooperation. For example, if newcomers have more resources and status, they might also have more normative influence (Andersen & Moynihan, 2018; Lefkowitz et al., 1955). This can potentially lead to a form of public good provision that serves the needs of both incumbents and newcomers, or even the adaptation of newcomer norms by incumbents (e.g., gentrification affecting neighborhood norms).

We mainly studied the relationship *between* the two categories of incumbents and newcomers. However, often there is also considerable variation *within* these two categories. In society, there are strong differences across immigrant groups in terms of status, integration, and resources. For example, immigrants from the Middle East, Africa, and Latin America face more discrimination in the Netherlands than immigrants from Europe, America, and Asia (Thijssen et al., 2020), and these origins matter for whether immigrants are considered part of the ingroup by native populations (Lee & Fiske, 2006). The societal context within which any particular newcomer-incumbent relation is embedded matters for how newcomers and incumbents perceive and treat each other, and can thereby impact their willingness to cooperate for public goods. Much of this thesis deliberately abstracted away from such real-life complexities and instead studied public goods using minimal groups in a standardized laboratory environment. However, there are ways to incorporate *natural* groups into the laboratory. For example, the public goods game can be played using participant pools that include both native-majority and ethnic-minority members (Gereke et al., 2021). As an extension of this thesis, the team of researchers involved in Chapter 5 is currently undertaking a replication in Israel which involves natural groups of Jewish and Arabic ethnicity.

Another important natural group of newcomers that this thesis did not explicitly study is refugees. Of particular relevance are the recent large-scale influxes of refugees from Syria, Ukraine, and several other countries to European societies, which have led to scientific and political debates on cooperation issues between host communities and refugees (Agustín & Jørgensen, 2019; Wike et al., 2016). One approach to incorporate refugees in cooperation research is to invite them to the lab and have them cooperate with natives (Drouvelis et al., 2021). Another approach – similar to how we applied the online multiplayer game to study entire populations consisting of incumbents and newcomers – is to use register data covering entire national populations consisting of natives and refugees (Van der Laan et al., 2022). Although these register data often do not explicitly include individuals' contributions to public goods, they can be enriched with representative survey samples to tap into self-reported contribution intentions or behaviors. More broadly, the newcomers and incumbents studied in this thesis mostly come from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010). Although some of the chapters incorporate data that go somewhat beyond WEIRD societies (e.g., Turkish and Greek servers in the online multiplayer game and confederates with Turkish and Moroccan backgrounds in the field experiment), future research is needed to establish to what extent conclusions generalize to non-WEIRD societies.

An aspect of group changes that we did not focus on is situations in which newcomer entry increases the group size. In society, a concern voiced by some citizens and politicians is that national public goods can only support the native population and would become insufficient under large waves of immigration (Ruist, 2015). We did not study this threat, but empirical evidence suggests that population growth is not a major threat to public good provision. In empirical research on public goods, the effect of group size is often mixed and non-linear (Carpenter, 2007; Isaac & Walker, 1988b; Nosenzo et al., 2015; Yang et al., 2013). A meta-analysis suggests that group size has no significant effect on contributions in public goods games (Zelmer, 2003), and studies with group sizes of 100 members (Weimann et al., 2019) or even 1000 members (Pereda et al., 2019) suggest that public good provision in large groups is qualitatively similar to small groups. A related threat that is sometimes argued to play a role in newcomer-incumbent tensions is resource competition. For example, newcomer-incumbent tensions may be driven by conflicts over scarce resources in the labor and housing market (Lubbers et al., 2006). However, public goods are typically non-rival, meaning that the use of one person does not go at the expense of another person. Thus, our focus on cooperation for public goods mostly precludes tensions over competing resources. In other areas of cooperation – for example in common-pool situations – perceived resource competition may be important for newcomer-incumbent tensions.

Because research on the relationship between group changes and public good provision has been scarce, this thesis mostly focused on understanding this relationship and the potential threats that it entails for newcomers and incumbents. This means that the thesis did not yet test the effectiveness of potential interventions to overcome the identified threats. Indeed, before we can design effective interventions to solve a problem, we first need to understand the problem (IJzerman et al., 2020). Our findings provide a few potential directions for interventions that can be tested in future research. Recall that we found that newcomers contribute differently from incumbents not because they want to take advantage of incumbents or disagree with them, but rather because they need time to learn the incumbents' norm before they can conform to it. This suggests that interventions that provide newcomers with information on the prevailing contribution norm may help. Incumbents, in turn, need to look beyond initially low contributions and realize the high long-term contribution potential. Research shows that, when possible, incumbents tend to exclude or expel newcomers who initially do not contribute enough (Ahn et al., 2008; Maier-Rigaud et al., 2010). Simulations suggest such exclusions can lead to a loss of contribution potential (de Matos Fernandes et al., 2022), and our empirical evidence corroborates this. Providing information to incumbents on the long-term contribution potential of newcomers may help them to be more patient and put effort into helping newcomers integrate.

We further found that the difference in contributions between newcomers and incumbents is smaller when there is less inequality between them. Providing newcomers with enough resources can be another way to speed up the process of newcomers contributing to public goods. To ensure that the form of public good provision is not only tailored to incumbents' needs, interventions that address ingroup biases may help. We observed that one way of overcoming ingroup bias is to make sure groups have experience with ingroup public good provision before turning to the more complex multigroup public good provision (Chapter 5). Other research shows that newcomers signaling commitment to the group (Choi et al., 2019, 2021) and signaling good intentions (Rösler, 2022) also help to mitigate ingroup bias. Reducing the salience of ingroup-outgroup divisions in incumbent-newcomer relations may further help to prevent ingroup bias, for example by emphasizing superordinate identities and goals (Gaertner & Dovidio, 2000; Sherif, 1966).

Changing group compositions bring unique challenges to cooperation for public goods. The sustainability of public good provision depends on the group's ability to overcome these challenges and cooperate in a way that brings value to all its members. In an increasingly dynamic world, sustaining cooperation in the face of group changes becomes ever more important. Future research on cooperation for public goods in changing groups can help to understand and overcome these challenges.

Chapter 2. Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels[1]

Abstract

Norms can promote human cooperation to provide public goods. Yet, the potential of norms to promote cooperation may be limited to homogeneous groups in which all members benefit equally from the public good. Individual heterogeneity in the benefits of public good provision is commonly conjectured to bring about normative disagreements that harm cooperation. However, the role of these normative disagreements remains unclear because they are rarely directly measured or manipulated. In a laboratory experiment, we first measure participants' views on the appropriate way to contribute to a public good with heterogeneous returns. We then use this information to sort people into groups that either agree or disagree on these views, thereby manipulating group-level disagreement on normative views. Participants subsequently make several incentivized contribution decisions in a public goods game with peer punishment. We find that although there are considerable disagreements about individual contribution levels in heterogeneous groups, these disagreements do not impede cooperation. While cooperation is maintained because low contributors are punished, participants do not use punishment to impose their normative views on others. The contribution levels at which groups cooperate strongly relate to the average normative views of these groups.

[1] A slightly different version of this chapter has been published as Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2020). Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels. *Scientific Reports*, *10*(1), 1-12. Otten wrote the manuscript, developed and executed the experiment, and did the analysis. All authors contributed to experiment development and manuscript writing. The study is preregistered at doi.org/10.17605/OSF.IO/GY8ST. All data and code are openly available at doi.org/10.24416/UU01-87KATL.

2.1. Introduction

Norms indicate social standards for individual behavior, and are considered to be one of the main factors sustaining cooperation among humans (Bicchieri, 2006; Coleman, 1990; Curry et al., 2019; Fehr & Schurtenberger, 2018; Gavrilets & Richerson, 2017; Tomasello & Vaish, 2013; Young, 2015). Cooperative goals often require people to bear an individual cost to benefit the group as a whole. This can create a social dilemma where each member is individually best off by free-riding on contributions of others, while the group as a whole is best off if everybody contributes their share. Prosocial contribution norms can solve this social dilemma by prescribing sufficient contributions for the public good. Experiments using linear public goods games (PGGs) show that groups often sustain high contribution levels if they have the opportunity to enforce prosocial contribution norms through peer punishment, whereas they largely fail to do so without options for norm enforcement (Chaudhuri, 2011; Fehr & Gächter, 2000).

However, research suggests that the potential of norms to promote cooperation for public good provision may be limited to homogeneous groups in which individuals can contribute similar amounts and benefit equally from each other's contributions. Heterogeneous groups in which individuals differ in how much they can contribute to and benefit from the public good often achieve lower levels of cooperation, even if there is an opportunity to enforce contribution norms (Gangadharan et al., 2017; Kingsley, 2016; Nikiforakis et al., 2012; Reuben & Riedl, 2013). These lower levels of cooperation are commonly attributed to normative disagreement within heterogeneous groups (Kingsley, 2016; Nikiforakis et al., 2012; Rauhut & Winter, 2017; Winter et al., 2012). While homogeneous groups largely agree upon a single appropriate level of contributions, heterogeneity within groups brings about a plurality of different and conflicting views about how much group members should contribute (Reuben & Riedl, 2013). This normative disagreement has been conjectured to harm cooperation. If group members are dissatisfied with the contributions of others, they may react with lower contributions themselves, leading to outcomes that are worse for everyone (Fischbacher & Gächter, 2010). Here we study experimentally whether normative disagreement harms cooperation in heterogeneous groups.

In everyday life, individual heterogeneity is the rule rather than the exception (Scheffer et al., 2017). The importance of incorporating heterogeneity in experiments is increasingly recognized (Fischbacher et al., 2014; Hauser et al., 2019; Kube et al., 2015; Przepiorka & Diekmann, 2018; Robbett, 2016). One ubiquitous type of heterogeneity concerns the returns from the public good. For example, the costs of public facilities such as dams and parks are shared by all taxpayers, even if these often provide different benefits to individuals depending on their distance to or frequency of enjoying the facility. In workplaces, employees may differ in their returns from contributing to teamwork (Hamilton et al., 2003), such as researchers benefitting differently from joint publications at different career stages. Countries have different interests and costs in jointly addressing global problems such as climate change (Lange, 2006) or the refugee crisis (Thielemann, 2018).

In all of these examples, there are at least two common and conflicting views about how people ought to behave (Fischbacher et al., 2014; Nikiforakis et al., 2012; Reuben & Riedl, 2013). One view is that all actors should contribute equally to the public good, which implies that those obtaining higher returns from the public good also end up earning more. The other view is that the actors with higher returns from the public good contribute more than the others, such that earnings are equalized. For example, some may argue that citizens living in areas with higher risks of flooding should contribute more to the construction of dams, whereas others may argue that all people should contribute equally regardless of where they live. Thus, there is a potential for normative disagreement (sometimes referred to as normative conflict) in heterogeneous groups between the views of equal-contributions and equal-earnings.

Lab experiments can manipulate the level of normative disagreement in a controlled decision environment, which helps to isolate its impact on cooperation from potential confounders that exist in real-life contexts (Falk & Heckman, 2009). The PGG is a classic laboratory paradigm for studying cooperation problems in groups. In previous experimental research on the PGG, the existence of normative disagreement and its influence on cooperation were mainly inferred indirectly from the way people contributed to the public good and sanctioned each other (Fehr & Fischbacher, 2004b, 2004a; Henrich et al., 2006; Kingsley, 2016; Nikiforakis et al., 2012; Rauhut & Winter, 2017; Reuben & Riedl, 2013; Winter et al., 2012). Normative views were rarely measured in the PGG (see for exceptions; Cubitt et al., 2011; Hauge, 2015; Lindström et al., 2018; Reuben et al., 2015; Reuben & Riedl, 2013; Spiller et al., 2016). This means that to date we have limited information on what people actually think constitutes appropriate contribution behavior, or how this might influence the choices they make. What is more, to our knowledge, group disagreement on these normative views has not been manipulated experimentally.

The role of normative disagreement in explaining (the lack of) cooperation in heterogeneous groups therefore remains ambiguous, and alternative explanations cannot be ruled out. For example, heterogeneity in groups may also be related to confusion about the cooperation problem, difficulty with coordinating behavior among group members, or the existence of inefficient norms, all of which could negatively affect cooperation (Gangadharan et al., 2017; Ramalingam et al., 2018; Reuben et al., 2015). Hence, without explicitly measuring and manipulating normative disagreement, our understanding of its causal influence on cooperation remains limited. This research makes two main contributions. First, we explicitly measure normative views in the PGG and analyze their content. Second, we manipulate whether groups agree or disagree on these normative views and examine the causal influence of this (dis)agreement on cooperation in terms of contributions to the public good.

An overview of our experiment is shown in Figure 2.1. Participants (N = 192) play the PGG in fixed groups of three for 10 rounds. In each round, participants obtain 20 monetary units (MU) and choose how many of these MUs to contribute to a group project and how many to keep for themselves. The contributions of all group members are added up, multiplied by a factor larger than 1, and then distributed among the group members based on individual returns. While the multiplication factor is larger than 1, the individual return of each member is smaller than 1. This constitutes a social dilemma because each member individually is best off by not contributing (retaining their own MUs), while the group as a whole is best off if all members fully contribute. Participants can punish individuals whose contribution decisions they do not approve. They can do so by deducting MUs from group members after each contribution decision (Fehr & Gächter, 2000; Reuben & Riedl, 2013). As mentioned, normative disagreement can ensue when individuals differ in the benefits they obtain from cooperating.

To introduce this possibility, we assign heterogeneous individual returns of the public good. Per group, one member receives a fifty percent higher return than the two other members (0.75 vs 0.50). A prior study suggests that with this level of heterogeneity, there is considerable variation among people in whether they support the rule of equal-contributions or equal-earnings (Reuben & Riedl, 2013).

Before communicating which participants obtain the low and high returns, we ask participants to report their normative views on the appropriate contributions that the high-return and the low-return members should make. We also ask them to report their normative expectations, i.e., what they expect others to consider appropriate contributions. Participants' own normative views are used to position them on the spectrum of equal-contributions to equal-earnings. Supporters of equal-contributions would answer that both types of players should contribute equally to the public good, whereas supporters of equal-earnings would answer that high-return types should contribute twice as much as low-return types. Participants who support a balance between both rules would answer that high-return types should contribute more than low-return types, but not twice as much. It turns out that almost all of our participants fall within one of these three categories and are rather evenly distributed across these three categories.

Group-level normative disagreement is manipulated by sorting participants into groups with either similar or dissimilar views (two conditions) on the spectrum of equal-contributions to equal-earnings. That is, in the normative agreement condition, we sort participants from the same side of the spectrum together, whereas in the normative disagreement condition we sort participants from different sides of the spectrum together (see Figure 2.1, and for more details Methods). To keep our design comparable to related research (Gangadharan et al., 2017; Kingsley, 2016; Nikiforakis et al., 2012; Reuben & Riedl, 2013), we neither inform participants about the normative views of their group members, nor about the method of group formation (see Methods). In both conditions, participants play the PGG within their group for 10 rounds. As we will show, participants' normative views remain largely stable over the 10 rounds. This implies that the difference in normative disagreement between conditions remains largely stable over time as well. By comparing the average contribution levels between the two conditions, we can assess to what extent normative disagreement negatively influences cooperation in terms of contributions to the public good.

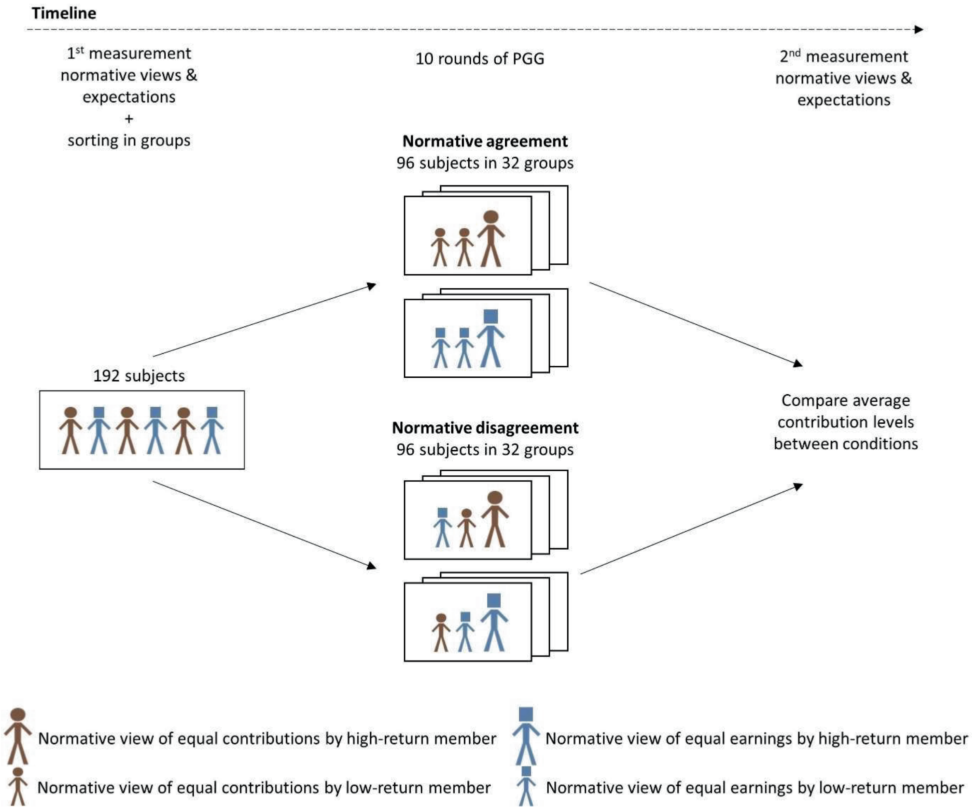


Figure 2.1. Experimental setup.

Note: We ran 8 sessions with 24 participants each, leading to a total sample of 192 participants. Randomization took place at the session-level. Within each session, participants reported their normative views before and after playing the game and were sorted on a continuum from equal-contributions to equal-earnings. This is denoted by the red (circle-headed) and blue (square-headed) figures supporting equal-contributions and equal-earnings respectively. The sorting differs by condition. In the condition for normative agreement, participants were sorted with similar others. In the condition for normative disagreement, participants were sorted such that one participant per group disagrees with the two other group members (see Methods). One member per group has a higher return from the public good than the other two, as is illustrated here by differences in body size.

The prevailing theoretical prediction is that normative disagreement has a negative effect on public good provision (Gangadharan et al., 2017; Kingsley, 2016; Nikiforakis et al., 2012; Rauhut & Winter, 2017; Reuben & Riedl, 2013; Winter et al., 2012). Many people are conditional cooperators, who contribute only if others are also contributing their share (Frey & Meier, 2004; Rustagi et al., 2010). If people contribute according to different normative views while observing others' contributions, they can discover that their own view is not adhered to by others. The expected consequence is that conditional cooperators who think others are not contributing enough will reduce their own contribution, causing a downward trend in contribution levels. Research on PGGs without punishment finds that contributions decline significantly already when one actor contributes less than others in groups of three (De Oliveira et al., 2009), four (Kurzban & Houser, 2005), and six (Boosey, 2017).

While peer-punishment is known to facilitate cooperation in contexts without normative disagreements, it is not expected to help much in contexts with normative disagreements, and can make matters even worse (Kölle, 2015; Rauhut & Winter, 2017). Previous research shows that, without punishment, contributions steadily decline toward free-riding in all groups, regardless of differences in norms and treatments. However, clear differences in cooperation occur when punishment is possible (Reuben & Riedl, 2013). Actors who receive punishment while behaving according to their own normative view may deem the punishment unjustified, and retaliate by further reducing their contributions or by counter-punishment (Herrmann et al., 2008). We therefore hypothesize that normative disagreements have a negative influence on public good provision in contexts with punishment possibilities. To test this hypothesis, we focus on public good games with punishment possibilities and do not consider contexts without punishment possibilities.

We start the results section with an overview of the type and distribution of normative views and expectations among our participants. Moreover, we examine the temporal stability of participants' normative views. Then, we statistically test our hypothesis. Our paper concludes with a brief exploratory analysis and a general discussion of our findings.

2.2. Results

2.2.1. Describing normative views

To elicit normative views, we showed the participants a hypothetical group of three members, two of which obtain a low-return and one of which obtains a high-return, the exact same composition of returns as used in the actual contribution rounds of the experiment. We subsequently asked: "According to you, what is the appropriate amount that each member should contribute to the group account". The participants could then indicate a contribution for each of the three members between 0 and 20 (see also Figure A.1 in the supplementary material). Because there is no right or wrong normative view, we did not incentivize these decisions. We subsequently elicit participants' normative expectations by letting them guess what their two other group members' answers were. The elicitation of normative expectations is incentivized in line with prior studies (Krupka & Weber, 2013) to motivate participants to seriously put themselves in the shoes of the other participants (see Methods).

In Figure 2.2a, we plot participants' normative views regarding the appropriate contribution of group members with a low return rate (x-axis) against these participants' normative views regarding the appropriate contribution of group members with a high return rate (y-axis). Figure 2.2a shows that there is considerable heterogeneity in normative views between participants. Yet, as anticipated, virtually all observations fall within the range between equal contributions (the $y = x$ line) and equal earnings (the $y = 2x$ line). Three normative views are especially prevalent: (1) equal and full contributions (i.e., collective efficiency), achieved by all group members contributing their full endowment; (2) equal earnings (i.e., equality), achieved by the high-return members contributing their full endowment and the low-return members contributing half of it, and (3) a mix between equal-contributions and equal-earnings, achieved by the high-return members contributing three-quarters of their endowment, 50% more than the low-return members. Each of these normative rules has about the same number of advocates among our participants (about 20% each). We thus find that participants' normative views map well along the dimensions of equal-contributions and equal-earnings, and that there is substantial heterogeneity among participants on these views.

To examine whether participants expect others to hold the same normative views as themselves, we plot participants' normative expectations (*x*-axis) against their normative views (*y*-axis) in Figure 2.2b. Observations on the diagonal represent participants for which normative views and expectations fully overlap. As can be seen, most participants are on the diagonal; they expect others to hold the same normative views as they do. On average, the correlation between views and expectations is .64 ($p < .001$). The extent of overlap between normative views and expectations is largely the same for answers on high- and low-return members. In both cases, about 60% of participants expect others to hold exactly the same views as they do, 20% support higher contributions than they expect others to support, and 20% support lower contributions than they expect others to support.

We elicited participants' normative views at the start and at the end of the ten rounds of the PGG. This allows us to assess the temporal stability of participants' normative views. In Figure 2.2c, we plot participants' initial normative views (*x*-axis) against their normative views at the end of the ten PGG rounds (*y*-axis). Most observations lie on the diagonal, representing participants with stable normative views. The average correlation between views before and after the ten PGG rounds is .60 ($p < .001$). Participants whose view matched neither the equal-contributions nor the equal-earnings rule from the start are considerably more likely to change it. Out of the 101 participants who matched one of the two rules at first measurement, only 33 changed them (33%), whereas out of the 91 participants who did not match one of the two rules, 74 changed them (81%) (Mann-Whitney ranksum test, $p = <.001$). About half of the participants who initially did not uphold one of the two focal normative views and then changed their views, switch to one of the two focal rules (34 out of 74 participants). Thus, the two normative rules of equal-contributions and equal-earnings have a substantial number of supporters from the start, and these supporters are unlikely to change their views over time. Participants whose views do not match one of the two focal rules from the start are more likely to change them, and if they do, they often end up supporting one of the two focal rules.

We find little support for self-serving normative views (high-return participants supporting the equal contributions view and low-return participants the equal earnings view because that is in their best interest). Table 2.1 shows that both high-return and low-return participants report roughly the same normative views, both before and after the ten PGG rounds. We also find no indication that participants are more likely to change their view if it receives minority instead of majority support. Both the absolute and relative change in normative views between the ten PGG rounds are similar for participants holding a minority and majority view (see supplementary material, Table A.1). In sum, our descriptive findings suggest that participants hold well-defined normative views about contributions to the public good under return heterogeneity, there is substantial between-participant variation in these views, these normative views largely overlap with normative expectations, and they are mostly stable over time.

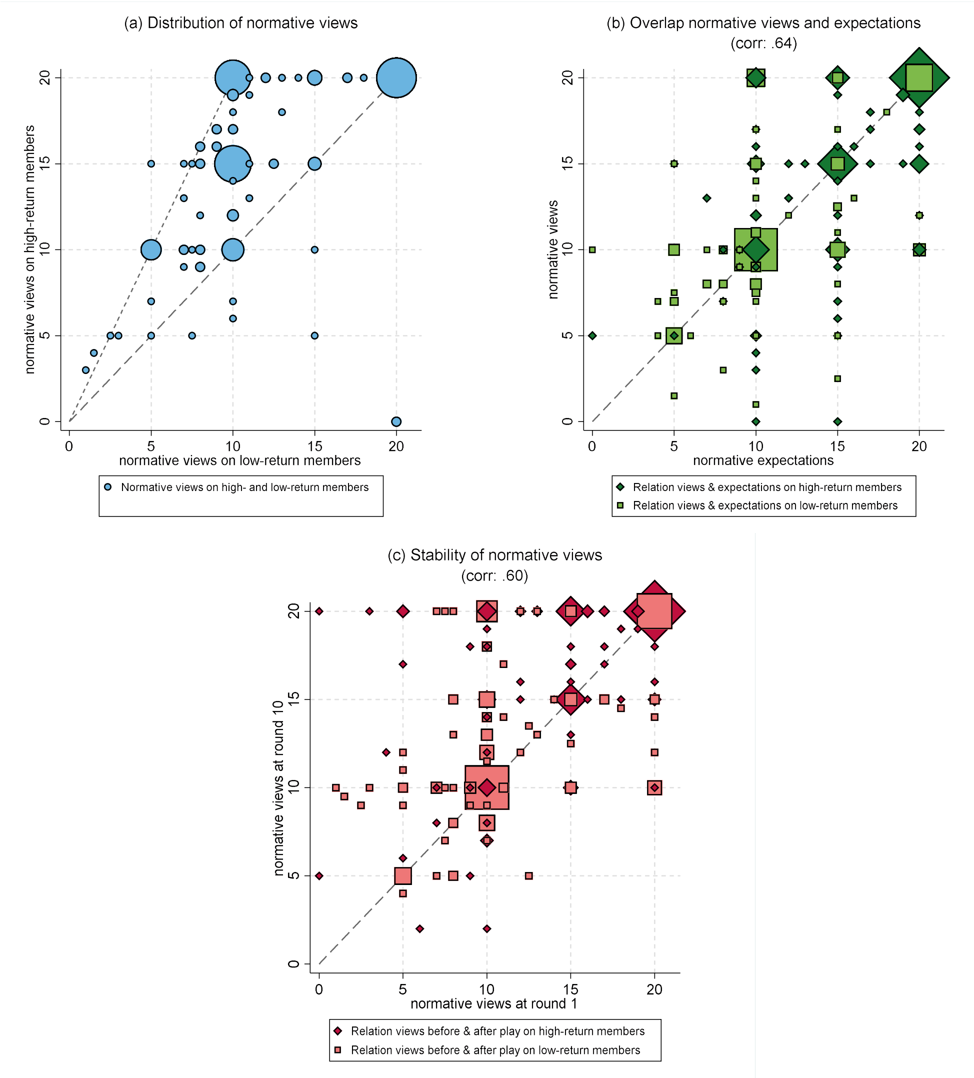


Figure 2.2. Normative views and expectations: distribution, overlap, and temporal stability.

Note: Number of observations in panel (a) is 192, and 384 in panels (b) and (c). Marker size is weighted by the number of observations. The marker size in the legends display the size for single participants. Normative views on contributions of high- and low-return members are separated by axis in panel (a), and by marker symbol in panel (b) and (c) (diamond for views on high-return members, squares for views on low-return members). Although participants were asked to provide their normative view and expectation on the appropriate contribution for each of the two low-return members separately, more than 90% provided the same answer for both low-return members. We therefore average the answers on the two low-return members.

Table 2.1. Descriptive statistics on contributions, punishments, and normative views.

	Average	Return of participant		Mann-Whitney test of difference by return-type	
		low	high	*z*-statistic	*p*-value
Contribution	13.87	12.93	15.74	-10.78	<.001
	(5.72)	(5.67)	(5.36)		
Punishment assigned	0.72	0.78	0.58	1.67	.09
	(1.87)	(2.03)	(1.49)		
Punishment received	0.72	0.61	0.92	-2.32	.02
	(1.87)	(1.60)	(2.31)		
Normative views before game					
on high-return members	15.76	15.66	15.97	-0.07	.94
	(4.77)	(5.04)	(4.20)		
on low-return members	11.92	11.95	11.85	0.47	.64
	(4.89)	(5.05)	(4.60)		
Normative views after game					
on high-return members	17.40	17.75	16.70	1.41	.16
	(4.10)	(3.80)	(4.60)		
on low-return members	13.08	12.96	13.31	-0.62	.54
	(4.82)	(4.80)	(4.88)		

Note: There are two low-return members and one high-return member per group. Standard deviations are in parentheses.

2.2.2. Normative disagreement and public good provision

The heterogeneity in normative views allowed for a successful manipulation of disagreement at the group-level. To assess how our sorting procedure affected group-level disagreement, we rank participants within each group based on how much more they think high-return members should contribute than low-return members. Group-level disagreement is measured by comparing this difference between the highest-ranked and lowest-ranked participant of each group. In the disagreement condition, the difference supported by the highest-ranked participant was on average 7.70 contribution points larger than that of the lowest-ranked participant. In the agreement condition, the difference supported by the highest-ranked participant was on average only 1.23 contribution points higher than that of the lowest-ranked participant (Mann-Whitney ranksum test, $p = <.001$).

In Figure 2.3, we show the average contribution (left *y*-axis) and punishment level (right *y*-axis) for each condition over the course of the ten PGG rounds. In line with prior research on groups with heterogeneous returns, we find that participants contribute about two-thirds of their endowment on average (Reuben & Riedl, 2013). This is also in line with the conclusion that public good provision is lower in heterogeneous groups than in homogeneous groups, where the contribution level is typically closer to full contributions when peer-punishment is possible (see also Figure A.2 in SI) (Chaudhuri, 2011). However, Figure 2.3 suggests that the lower levels of public good provision found in heterogeneous groups are not due to normative disagreement. The contribution levels are similar in both the agreement and disagreement condition, suggesting no support for the hypothesis that normative disagreement harms public good provision.

We use population-averaged regression models, which account for repeated measures obtained from the same participant or group, to statistically test this hypothesis, with the contribution decision as the dependent variable and the experimental condition as the predictive factor. Across six models we vary whether the outcome variable is on the individual-level or group-level, and whether we include all rounds, only the first round (1), or only the last round (10) as observations. Table 2.2 shows that, regardless of which model is used, we find no significant difference in contribution levels between conditions according to conventional standards ($p < .05$, with Bonferroni adjustment for multiple comparisons). Non-parametric (Mann-Whitney) tests lead to the same conclusion (see Table A.2 in SI). The hypothesis that normative disagreement harms public good provision is therefore not supported. The between-condition variance in contribution levels is negligible (< 1%), implying that virtually all variation is within conditions. Similarly, Figure 2.3 suggests no substantial differences in punishment levels between conditions, which is corroborated by non-parametric (Mann-Whitney) tests (see Table A.2 in SI). In both conditions, the average punishment allocated falls mostly between 0.5 and 1 whereas the possible range is from 0 to 10. We find no significant difference in punishment allocation between high-return and low-return participants (see Table 2.1). When subdividing contribution and punishment levels by the participants' return rate, we also find no difference between conditions (see supplementary material, Figure A.3). In exploratory analyses presented in the supplementary material, we also do not find a significant effect of normative disagreement on contribution levels under alternative conceptualizations of normative disagreement and model specifications (Figure A.6 and Tables A.3-A.5).

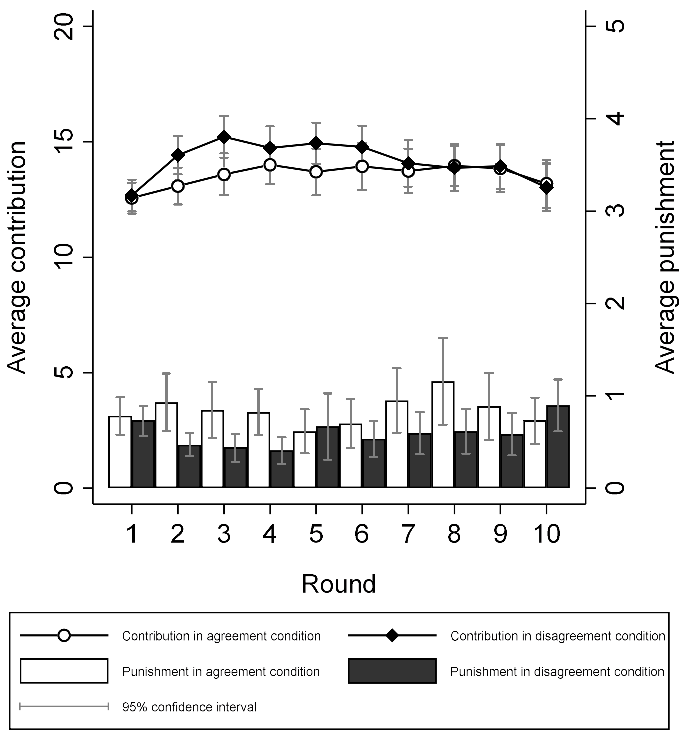


Figure 2.3. Average contribution and punishment per round and condition.

Table 2.2. Population-averaged model tests of hypothesis.

	(1)	(2)	(3)	(4)	(5)	(6)
treatment (disagreement)	.608	.115	-.156	.608	.115	-.156
	(.672)	(.764)	(.929)	(1.014)	(.831)	(1.274)
intercept	13.561***	12.563***	13.188***	13.561***	12.563***	13.187***
	(.475)	(.540)	(.657)	(.717)	(.588)	(.901)
N observations	1920	192	192	640	64	64
N participants	192	192	192			
N groups	64	64	64	64	64	64

Note: We use population-averaged regression models to statistically examine the hypothesis that the contribution level is higher in the normative agreement condition than in the normative disagreement condition. The contribution decision is the dependent variable and the experimental condition is the predictive factor. Across six models we vary whether the outcome variable is on the individual-level (models 1-3) or group-level (models 4-6), and whether we include all rounds (models 1 & 4), only the first round (models 2 & 5), or only the end round (models 3 & 6) as observations. Regardless of which model is used, we find no significant difference in contribution levels between conditions according to conventional standards: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Bonferroni-adjusted $p/6$, two-tailed tests). Standard errors in parentheses. In model 1 and 4 there are repeated measures on individuals/groups. We take these repeated measures into account with the exchangeable working correlation matrix.

The exploratory analyses show that, instead of normative disagreement, the average normative view of the group strongly relates to the average contribution level that it actually achieves. In Figure 2.4, we present the bivariate correlation between group-mean normative views and group-mean contributions per round. We find that in the initial rounds of the game, group-mean normative views almost perfectly predict group-mean contribution levels in both conditions ($r > .8$). Although the influence of normative views decreases somewhat over time, it remains substantial throughout the entire 10 rounds (average $r = .66$). The normative views participants have before they start interacting with each other thus strongly predict how they behave in the subsequent interactions. In the supplementary material, we show that the association between normative views and contributions is also present when subdividing by return-type, condition, and different normative views (Figure A.3-A.5). We conclude that the contribution level reached within groups is strongly related to the mean normative views within the group. Low contribution levels are thus not related to the existence of normative disagreement, but rather to the existence of group-mean normative views that support these low levels.

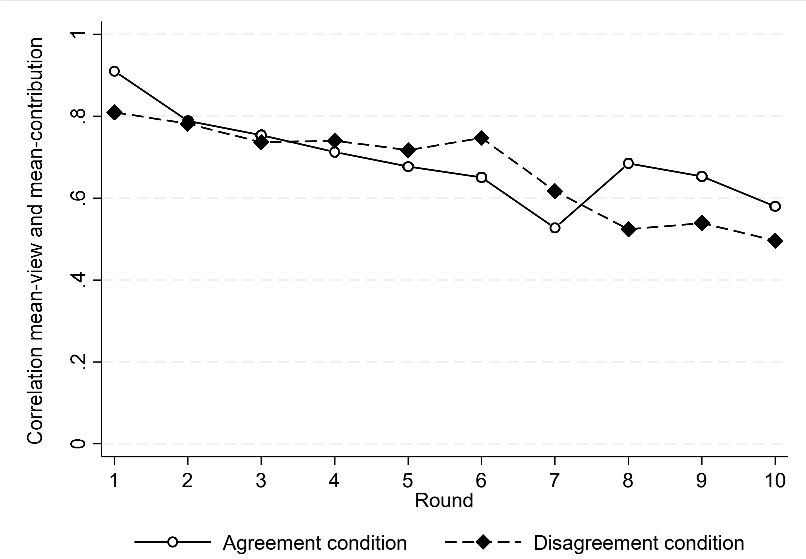


Figure 2.4. Correlation between group-mean normative views and contributions.

There are multiple reasons why normative disagreement does not influence public good provision while the group-mean view does. We consider three potential explanations: (1) people quickly adjust their normative views in line with the contributions of their group members, (2) people retain their normative views but reach a contribution level that compromises between the different views, and (3) people do not impose their own normative views on others. Figure 2c suggests that the first explanation is unlikely: most participants hold the same normative views before and after the game (~60%). What is more, this likelihood of holding stable normative views does not differ between conditions (Mann-Whitney ranksum test for stability on equal-contributions versus equal-earnings spectrum, $p = .19$). Our data also provides little support for the second explanation. If the contribution level is a compromise between group members' views, the absolute gap between what participants themselves think they should contribute and what they actually contribute should be bigger in the disagreement condition (i.e., where compromise is necessary) than in the agreement condition (i.e., where compromise is not/less necessary). However, this difference is small: the gap is on average 3.19 in the agreement condition and 3.68 in the disagreement condition (Mann-Whitney ranksum test, $p = .04$).

To assess the third explanation, that participants do not impose their normative views on others, we turn to punishment behavior. In Figure 2.5, we plot the relationship between contributions (x-axis) and the received punishment (y-axis) for high- and low-return participants. We see that for low-return participants, the received punishment is very low as long as they contribute 10 or more. Recall from Figure 2.2a that the most common way to achieve the equal-earnings rule is for low-return members to contribute 10 while high-return members contribute 20, and the most common way to achieve the equal-contribution norm is for all members to contribute 20. Thus, as long as low-return participants do not contribute below a level required to fall between one of the two prosocial contribution norms, they are hardly punished. Yet, when asking about contribution norms, only a minority of participants reports that equal-earnings is the appropriate norm (the majority of participants balance equal-contributions and equal-earnings, see Figure 2.2a). This suggests that participants do not impose their own normative views on their group members, as long as the group members' contributions fall between one of the two prosocial contribution norms. Contributions of 10 by high-return members fit neither of the two prosocial rules (see Figure 2.2a) and therefore do lead to punishment. Similarly, free-riding fits neither of the two prosocial rules and is punished regardless of return-type. Post-experiment measurements on normative views of punishment also support the conclusion that participants do not think it appropriate to punish contributions that do not conform to their own normative views, as long as these contributions conform to one of the prosocial views (see SI, Figure A.7).

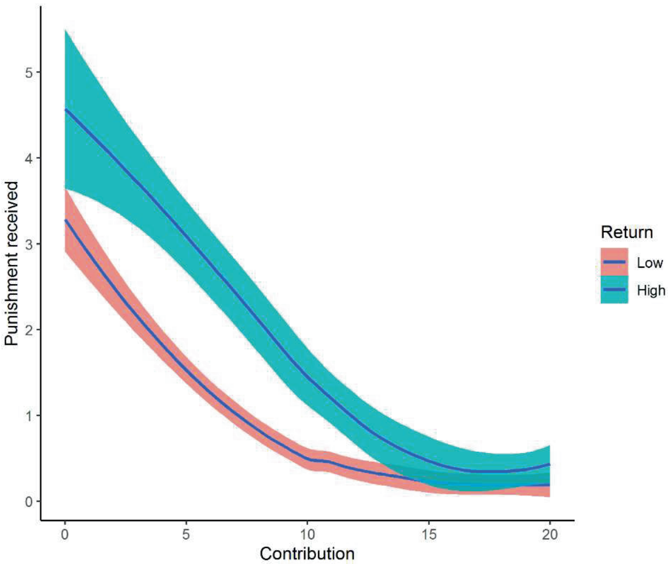


Figure 2.5. Received punishment as a locally estimated (LOESS) function of contributions with 95% confidence intervals.

2.3. Discussion

It is commonly conjectured that normative disagreements in heterogeneous groups negatively affect cooperation. To our knowledge, we are the first to directly test this conjecture. In a lab experiment, we measure each participant's view regarding the appropriate way to contribute to a public good with heterogeneous returns and use this information to manipulate whether members of a group agree or disagree on these normative views. Our results show that participants vary considerably in their normative views, but in a predictable way. Virtually all participants subscribe to a norm of equal-contributions, equal-earnings, or a balance between these two. However, disagreement between these normative views does not negatively affect public good provision. Group composition in terms of normative views does have a strong relationship with the level of public good provision. However, it is the group-mean, rather than the group-disagreement, that matters.

Joint payoffs are maximized when all members contribute fully to the public good, i.e., under the efficient equal-contributions rule. We find, however, that many participants support deviations from the equal-contributions rule to achieve equal-earnings or a balance between the two rules. Notably, in our experiment, equal-earnings could only be achieved by lowering the earnings of high-return members and not by increasing the earnings of the low-return members. Still, many participants supported lowering the joint payoffs to achieve more equalized earnings. Furthermore, participants who do support the equal-contributions rule nevertheless largely refrain from enforcing others to contribute alike. We therefore suggest that there is an alternative reason than normative disagreement for why public goods with heterogeneous returns are underprovided in terms of joint payoffs. Most people support this under-provision to achieve more equalized earnings, and if they do not support it, they refrain from punishing it.

Normative disagreements have been put forth to explain the lack of cooperation in several pressing real-life issues that require parties to invest in a shared public good, such as the climate crisis (Lange et al., 2010), the European debt crisis (Rathbun et al., 2019), or the refugee crisis (Thielemann, 2018). We used a lab experiment to isolate the effect of normative disagreement from other complexities that arise in such real-life situations. Our results suggest that normative disagreement alone is not sufficient to harm cooperation.

It is possible that normative disagreements would have been harmful under alternative design choices. For example, had we made the differences in returns from the PGG larger or told participants about the (conflicting) normative views of the members they were grouped with, we may have observed a negative effect of disagreement on contribution levels. However, the prior studies that led to the conjecture that normative disagreement causes lower contribution levels used similar levels of heterogeneity and also did not provide participants with information about each other's normative views (Nikiforakis et al., 2012; Reuben & Riedl, 2013). We kept our design as close as possible to these experimental designs. Our results thus suggest that under these standard experimental conditions, normative disagreement is not the explanation for lower contribution levels.

Future research can further chart the boundary conditions under which normative disagreements affect cooperation. A promising direction involves manipulating what participants know about others' views and how they can express their own. We list three suggestions: First, one could manipulate whether participants have direct information on other's normative views and the level of disagreement. Norm disagreements may become more salient when participants have this information, leading to potentially more harmful effects, e.g., when disagreement triggers attempts to convince other group members that one's own views are superior. Second, one could on top of revealing other's norms manipulate whether contributions are private or public information. This would allow one to better disentangle the effects of descriptive and injunctive norms on contribution levels (Cialdini et al., 1990). Third, one could vary how participants can react to each other. In our experiment, participants could react through contribution and punishment decisions. Future experiments could simplify our design by removing punishment for example, or enrich it by allowing for additional reactions through direct communication or rewards for example (Andrighetto et al., 2013; Bicchieri et al., 2021).

Our study results are also relevant for research on PGGs in general. Participants entered the laboratory with their normative views already well-defined. Participants' views mapped well along the spectrum between equal-contributions and equal-earnings, were largely stable throughout the experiment even when confronted with the conflicting behavior of others, and guided these participants' behavior. The minds of participants entering the lab are thus not blank slates. Some scholars conjecture that participants take the social norms they adhere to in everyday life with them to the lab, and that ignoring these norms when looking at these participants' behavior leads to substantial misinterpretations of lab experimental results (Binmore, 2010; Kimbrough & Vostroknutov, 2016). Our study results corroborate this conjecture. The normative views participants have before they start interacting with each other strongly relate to how they behave in the subsequent interactions. Therefore, not measuring these views runs the risk of missing a large part of the variation in within-condition behavior.

One concern is that merely asking participants about norms might encourage norm compliance. However, the available evidence suggests that norm elicitation as such does not affect behavior (D'Adda et al., 2016). For our study, we can directly compare the results with the results of an experiment that used the exact same game-parameter values but did not elicit norms (Reuben & Riedl, 2013). The behavioral patterns are similar, suggesting that our norm elicitation did not substantially affect behavior (see Figure A.2 in SI).

Norms are considered to be an important element in explaining human cooperative behavior across numerous disciplines (Bicchieri, 2006; Coleman, 1990; Curry et al., 2019; Fehr & Schurtenberger, 2018; Gavrilets & Richerson, 2017; Tomasello & Vaish, 2013; Young, 2015). Recently, it has been conjectured that when groups are heterogeneous rather than homogeneous, norms can also harm cooperation because they cause disagreement. We find that although there is considerable disagreement about normative views in heterogeneous groups, this disagreement does not impede cooperation. Groups cooperate despite normative disagreement, at different contribution levels depending on the average level supported by their members. Our results suggest that norms can sustain cooperation even in situations of normative disagreement.

2.4. Methods

We conducted the computerized experiment in the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University during October-November 2019. The experiment was programmed with z-Tree software (Fischbacher, 2007). We recruited participants among students at Utrecht University using the internet recruitment system ORSEE (Greiner, 2015). We ran 8 sessions with 24 participants each, leading to a total of 192 participants. Each session lasted about 75 minutes. Payment depended on behavior in the game, participants earned on average 15 euros (min = 5, max = 22). Participants were on average 24 years old, 127 (66%) were female, 62 male, and 3 other. Almost all participants were students at Utrecht University, 87 were Dutch and 105 from various other countries.

Participants were randomly placed in an individual cubicle, so they could not see, or communicate with, each other. They were informed about the experiment through written instructions (provided in the SI). There were two parts of the experiment. In the first part, participants report on their normative views and expectations, play 10 rounds of the PGG with peer punishment, and report on their normative views and expectations again. This part is designed to examine the influence of normative disagreement on public good provision. In the second part, participants are switched between groups, they play another 10 rounds of the PGG within the newly formed groups consisting of new and old members, and afterward answer questions on their normative views, expectations, meta-norms, social preferences, and background characteristics. This second part is designed to examine the influence of newcomer entry on public good provision. Participants were informed at the beginning of the experiment that there would be two parts of the experiment, but that they would only receive the information about the second part of the experiment after the first part was finished. Because we only study the influence of normative disagreement on public good provision in this paper, we only analyze the game data of the first part of the experiment. The effect of newcomer entry will be analyzed in another paper.

Each round of the PGG with peer punishment has two stages. In the first stage, each individual i in a group composed of 3 members receives an endowment of 20 MUs and must decide how much of this endowment to contribute to a public good, c_i, where $c_i \in \{0, 1, \ldots, 20\}$. The part of the endowment that is not contributed to the public good is kept for the individual. The public good consists of the sum of the contributions made by all individuals. Each individual receives a return per contributed point (sometimes also referred to as marginal per capita return) to the public good ($m_i < 1$). The sum of these returns makes up the total multiplication factor of the public good 1.75. After all individuals in a group have made their contribution decision, the contributions and payoffs of each player are communicated to all and the first stage is finished.

In the second stage, each individual is given the opportunity to assign punishment points $p_{ij} \in \{0, 1, \ldots, 10\}$ to each group member j. Each punishment point costs 1 point to the punisher, and reduces the payoff of the punished player by 3 points. The individual payoff after one round of this two-stage is given by:

$$\pi_i = 20 - c_i + m_i \sum_j c_j - \sum_{j \neq i} p_{ij} - 3 \sum_{j \neq i} p_{ji}$$

Individuals do not see who punished them (to prevent confounding of normative behavior with revenge motives), and repeatedly play rounds of this two-stage game within the same group. We assign heterogeneous returns of the public good per group: one participant with a high return $m_i = .75$, and two participants with a lower return $m_i = .50$. For comparability to previous research, all other parameter values are set to follow the typical form of the PGG with peer punishment (Fehr & Gächter, 2000).

Prior to sorting participants into groups and assigning them their returns, we present them with the game and elicit their normative view by letting the participants answer what they consider to be the appropriate contribution decisions for another hypothetical group. Each of the participants is asked to indicate the appropriate contribution for each of three group members, one with return $m_i = .75$, and two with $m_i = .50$. The participants can try out different combinations of contributions, and see how it affects the earnings of each group member (see instructions and screenshots in the SI, Figure A.1). After the participants reported their personal normative view, we tell them that their group members were also asked to indicate appropriate contributions for three members in the PGG. Each of the participants is then asked to guess the answers submitted by their group members. To incentivize the guess, the participants are told that we will randomly pick one of their guesses, and give an additional payment of 100 MU (~€1.40) when it matches the answer of at least one of the group members. Only at the end of the experiment are participants informed of whether they were correct in the guess we randomly chose. This measure is inspired by the following earlier studies (Bicchieri et al., 2014; Krupka & Weber, 2013; Reuben et al., 2015).

We sort 192 participants into 64 groups of 3 members each based on their normative views. Within each session of 24 participants, we assign each of the participants a ranking in terms of how much they support the equal-contributions rule versus the equal-earnings rule compared to the other participants in the session. The precise score used to assign ranks is: $c_H - \bar{c}_L + .02\bar{c} + .0001R$, where c_H is the participant's view on the appropriate contribution for the high-return member, $\bar{c}_L$ is the participant's view on the appropriate contribution of the two low-return members on average, $\bar{c}$ is the mean appropriate contribution over all three members, and R is a random number between 0 and 1. The addition of $.02\,\bar{c}$ makes sure that participants who assign a contribution of 20 to all members obtain slightly higher scores than participants who assign a contribution of 0 to all members. This helps to differentiate between different absolute levels of achieving the equal-contributions rule in the sorting method. The number .02 is chosen such that whether contributions are relative to returns or not always has dominance in the sorting mechanism over the absolute level of contributions. The addition of $.0001R$ is to avoid tied scores. The method of sorting within the two conditions based on these scores is described in Figure 2.6.

Sample size was determined based on sample sizes in comparable studies (see for example Nikiforakis et al., 2012; Reuben & Riedl, 2013). Indeed if the hypothesized effect would have been substantial, we would most likely have found confirmation for it. Namely, a Mann-Whitney ranksum test with individuals as unit of analyses has a high power (.96) to detect medium-sized effects (.5) given our sample size. With groups as unit of analyses, we have relatively low power (.61) to detect medium-sized effects (.5), but high power (.92) to detect large effects (.8). These estimates are conservative as they are based on one observation per individual/group while we have ten (correlated) observations per individual/group (one observation per round). The effects of normative disagreement that we find do not reach statistical significance neither at the individual nor at the group level and are in terms of effect size small.

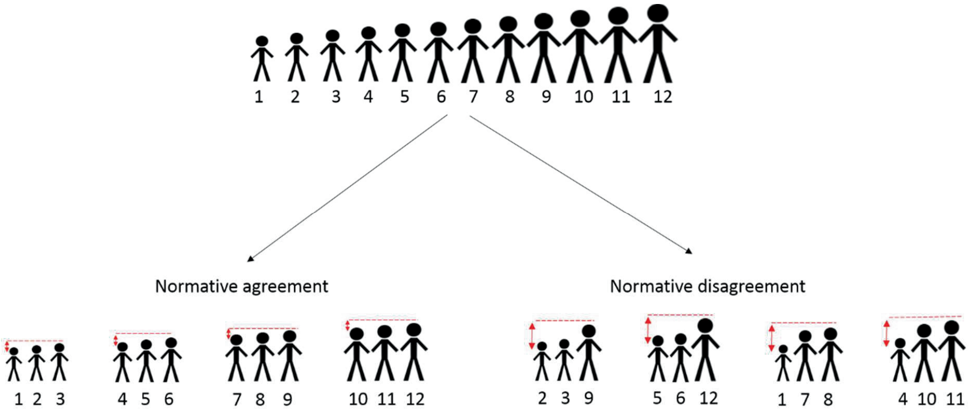


Figure 2.6. Example for method of sorting participants.
Note: At the beginning of the experiment, participants are ranked in terms of their normative views for contributions relative to returns. In the example presented here, there are 12 participants sorted into 4 groups. When sorting for normative agreement, we first form a group of the three highest-ranked participants (1-3), then of the remaining participants we again form a group of the three highest-ranked participants (4-6), and so on until all participants are grouped. Compared to the condition with normative agreement, in the condition with normative disagreement we select the highest-ranked low-return participant from the first group in the first half of the groups (ordered in terms of support for contributions relative to returns), and replace it with the lowest-ranked low-return participant from the first group in the second half of the groups, and repeat this procedure with the remaining groups. In this way, the extent of normative disagreement (in terms of rank-differences) is equal for all groups.

We decided not to inform participants on the method of group formation for three main reasons. First, we modeled our design on prior studies that gave rise to the conjecture that normative disagreements harm cooperation (Reuben & Riedl, 2013) (see for example the comparison in the supplementary material, Figure A.2). The main difference we introduced was our addition of the norm elicitation and the associated sorting based on this norm elicitation, both of which do not directly affect the participants' beliefs. Had we also told participants about the method of group formation, we would likely have altered the participants’ beliefs on how cooperative their group members are, which makes comparison with other studies more difficult. Second, by telling participants that their normative views will be used for group formation, we might create experimenter demand effects on the importance of these views for behavior. That is, it might lead the participants to believe that the experimenter judges these normative views to be important for behavior in the game. Participants who want to comply with the experimenter’s belief may as a result act more in line with their normative views once they know about the sorting procedure. Third, revealing that the normative views will be used for group formation may provide participants with an incentive to misrepresent their normative views. For example, in the normative agreement condition, if participants know that they will be grouped according to their normative views with similar others, they might report their normative views to be more prosocial than they actually are, so as to be grouped with others that hold prosocial views. Note that although we did not reveal how grouping was done, we did not offer untruthful information to the participants about the group formation, e.g., we did not say the formation was random. Instead, we told the participants when the formation had happened (directly after the norm elicitation), but not how it had happened.

After reading the instructions, participants were given six questions to test their understanding of the game. Upon completion, they were shown which questions they had answered correctly and incorrectly. At the first try, 158 participants answered 5-6 questions correctly, 26 participants answered 3-4 questions correctly, and 8 participants answered less than 3 questions correctly. Participants had to redo all questions that they answered incorrectly until all answers were correct. At the end of the experiment, participants were asked to rate whether their understanding of the experiment was (1) bad, (2) not bad, not good, or (3) good. 169 participants reported a good understanding, 22 reported not bad, not good, and 1 reported a bad understanding. These figures give us confidence that the experiment was adequately understood. Throughout the experiment, participants had the opportunity to ask questions to the lab official, less than a handful did so. The experimental data is openly available at doi.org/10.24416/UU01-87KATL. The experiment was preregistered before data collection at Open Science Framework, osf.io/gy8st. Informed consent was obtained from all participants and the experimental procedures were approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University. All research was in line with relevant regulations.

Chapter 3. Cooperation between newcomers and incumbents: The role of normative disagreements[1]

Abstract

Cooperation in groups often requires individual members to make costly contributions that benefit the group as a whole. Prior research suggests that shared norms can help to support ingroup cooperation by prescribing common standards of how much to contribute. These common standards may be disrupted when groups undergo membership change, i.e., when members from outgroups enter the ingroup. When newcomers and incumbents have different notions about how much to contribute, a normative disagreement ensues that could undermine cooperation and the extent to which individuals identify with the group. In a laboratory experiment, we manipulate whether newcomers and incumbents disagree about how much to contribute in a public goods game with peer punishment. We examine whether normative disagreement between newcomers and incumbents affects newcomer-incumbent relations in terms of group identification, the emergence of a social norm, and costly punishment. The main goal is to test whether normative disagreement and the resulting newcomer-incumbent relations harm cooperation in terms of contributions to the common good. We find that normative disagreement between newcomers and incumbents negatively affects the emergence of a shared social norm and lowers feelings of group identification. Contrary to expectations, normative disagreement does not affect cooperation negatively. Instead, participants adjust their behavior to each other's standards, using punishment for norm enforcement. This punishment is especially directed at low-contributing newcomers, leading them to conform to the incumbents' higher contribution standards.

[1] A slightly different version of this chapter has been published as Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2021). Cooperation between newcomers and incumbents: the role of normative disagreements. *Journal of Economic Psychology*, *87*, 102448. Otten wrote the manuscript, developed and executed the experiment, and did the analysis. All authors contributed to experiment development and manuscript writing. The study is preregistered at doi.org/10.17605/OSF.IO/GY8ST. All data and code are openly available at doi.org/10.24416/UU01-J5M2LE.

3.1. Introduction

Group cooperation often requires individual members to make costly contributions that benefit the group as a whole. For example, countries provide public and social security based on citizens' tax payments, neighborhoods maintain clean and safe parks if residents abstain from littering and keep watch, and work organizations survive and grow as a result of collaboration between workers (Dur & Sol, 2010; Sanders, 2009; Van Gerwen et al., 2018; Wageman, 1995). The composition of such groups changes frequently due to the arrival of new members and departure of old members. For example, work organizations hire new workers and let go of existing workers who retire or move to other organizations; countries, cities, and neighborhoods change in composition due to migration, and volunteer organizations and cooperatives attract new members and see other members leave. Sustainable group cooperation thus requires that contributions to the common good continue, regardless of the turnover in group members. However, this cooperation is often theorized to be impeded by newcomers and incumbents having different notions about how much should be contributed to the common good (Collier, 2013; Habyarimana et al., 2009; Ostrom, 2000), a situation which we will refer to as normative disagreement. In this study, we examine experimentally whether normative disagreement between incumbents and newcomers harms cooperation in terms of contributions to the common good. We also explore how perceptions of the contribution norm and feelings of ingroup identification shift due to the arrival of newcomers in the group.

There is a vast body of research that studies cooperation experimentally in the lab using social dilemma games (Chaudhuri, 2011). In social dilemmas, individual and collective interests are at odds. Cooperation in social dilemmas thus requires that people forego their individual interest to act in line with the collective interest. For example, taking measures to reduce one's carbon footprint are often individually costly but benefit society. When one's individual interest is aligned with cooperation, no social dilemma exists, and cooperation is more easily achieved. However, we focus on social dilemmas, which means that cooperation is arguably more fragile and vulnerable to normative disagreement. Most social dilemma research is directed either at intragroup cooperation (i.e., cooperation between members of the same group) or intergroup cooperation (i.e., cooperation between members of different groups). A main finding is that intragroup cooperation can be sustained via the development and enforcement of contribution norms, i.e., norms on how much to contribute to the collective good (Fehr & Schurtenberger, 2018). Although these norms promote intragroup cooperation, they can impede intergroup cooperation (De Dreu et al., 2020). Norms are typically group-specific and favor the ingroup over the outgroup (Bernhard et al., 2006; Titlestad et al., 2019). This can lead to conflict when different groups have to cooperate together but each group wants to stick to their own norm (Jetten et al., 1996). Indeed, there is ample research showing that intergroup relations are often characterized by conflict rather than cooperation (Balliet et al., 2014; Böhm et al., 2020).

While research using social dilemma games has advanced our knowledge about intragroup and intergroup cooperation, we know little about cooperation in groups where newcomers enter and have to cooperate with incumbents to contribute to the common good. We regard newcomer-incumbent relations as an intermediate case between intragroup and intergroup relations. Studying these intermediate cases is important because newcomers, although situated within the incumbent-group, are often seen as outsiders by the incumbents (Rink et al., 2013) and also themselves do not readily identify with the incumbents (Moreland, 1985). The change in group composition brought about by newcomer entry and the resulting newcomer-incumbent relations may have consequences for contributions to the common good.

Stability in group composition is commonly theorized to promote contributions to the common good via shared social norms (Fehr & Schurtenberger, 2018). Repeated interaction with the same group members facilitates the reaching of a social norm, i.e., a common understanding of what is an appropriate contribution level (Duffy & Ochs, 2009). This in turn helps group members to know what to expect from others and hold one another accountable for uncooperative behavior. Because norms are often group-specific, different groups develop different norms on how much to contribute (Bernhard et al., 2006; Gangadharan et al., 2017; Henrich et al., 2001). When members migrate between groups with different and incompatible norms, normative disagreement ensues. When members instead migrate between groups that hold similar norms, the arrival of newcomers does not increase normative disagreement.

According to normative conflict theory (Rauhut & Winter, 2017; Winter et al., 2012), groups are able to cooperate by contributing high amounts for the collective good if their members agree on normative views and can enforce these views, for example by the punishment of norm violations. However, if group members disagree on normative views, conflict is expressed in both low contribution levels and high punishment levels. When members differ in their normative views and contribute according to their own view, they observe that their view is not adhered to by others. The expected consequence is that members who feel others are not contributing enough reduce their own contribution and/or punish those they disapprove of. While punishment typically helps to promote cooperation if members agree on their normative views, it is predicted to be harmful when members disagree on normative views. Members who receive punishment while behaving according to their own normative view deem the punishment unjustified, and retaliate by reducing their contribution and by counter-punishment, leading to lower contribution levels and higher punishment levels.

Hence, if a newcomer brings in a normative view that is at odds with the normative views of the incumbents, a normative disagreement ensues that is predicted to lead to conflict in terms of cooperation failure. If the newcomer's normative view does not conflict with those of the incumbents, there is normative agreement, and cooperation is expected to be sustained. The few studies that have examined the influence of newcomers on cooperation in social dilemma games mainly looked at overall effects, i.e., whether newcomer entry has a positive or negative effect on contributions to the common good. Some of these studies suggest a positive effect of newcomers on contributions to the common good (e.g., Duffy & Lafky, 2016; Sonnemans et al., 1999), whereas other studies report a negative effect (e.g., Grund et al., 2015; Ranehill et al., 2014). This prior literature has not yet examined the role of normative disagreement between newcomers and incumbents in explaining the level of cooperation. Using a public goods game, we sort participants into groups of three and let them make contribution decisions in two sets of rounds. After the first set, we replace one member per group for a member from another group, such that each group consists of one newcomer and two incumbents. We manipulate whether newcomers and incumbents agree or disagree on normative views and examine if this influences cooperation in the second set of rounds.

Although newcomers do not always come from a salient outgroup, there are many situations in which they do. For example, work organizations may hire employees that previously worked at a competitor to obtain inside information from that competitor, players in professional sports are regularly switched between competing teams, and immigration also often involves ingroup-outgroup considerations in social interactions and conflict (Hainmueller & Hopkins, 2014). We focus on a situation where a newcomer comes from a relatively salient outgroup.[2]

Before we test the impact of normative disagreement on contribution levels, we examine how normative disagreement affects the development of newcomer-incumbent relations. The influence that normative disagreement has on cooperation may depend on the type of newcomer-incumbent relations that develop under normative disagreement. We examine three dimensions of the newcomers-incumbent relations. We examine whether normative disagreement between newcomers and incumbents affects (1) group identification, (2) costly peer-punishment, and (3) the emergence of a social norm (i.e., convergent normative expectations on how much should be contributed to the common good).

3.2. Previous research

3.2.1. Membership changes in public good games

One of the earliest experiments on membership changes in the public goods game was conducted by Sonnemans et al. (1999). The study examined 4-player groups in which one randomly selected member is replaced by another after a prespecified number of rounds. Group composition and the schedule of group changes are common knowledge from the start. The authors found that participants decrease their contribution just before they leave the group, but in the new group substantially increase their contribution. As a result, contribution levels increase considerably just after each membership change. Similarly, Duffy and Lafky (2016) found that periodically replacing old members by new members helps 4-player groups to sustain contributions to the public good over a longer period of time compared to groups of stable composition.

However, other studies report negative effects of membership changes on contribution levels. Grund et al. (2015) examined contribution levels in 4-player groups, where some group members stay together for all rounds (partners) and other members switch groups every round (strangers). There are four conditions that differ in whether the groups consist of (1) all partners, (2) three partners and one stranger, (3) two partners and two strangers, or (4) all strangers. The authors found that the contribution level is lower in groups with more strangers (i.e., more membership changes). Ranehill et al. (2014) studied how the rate of newcomer entry affects contribution levels in growing groups. They found that a higher rate of newcomer entry (i.e., more newcomers at once vs each newcomer entering in separate rounds) negatively affects contribution levels. Salmon and Weber (2017) also find that higher rates of newcomer entry

[2] Before the game starts, participants are told that each group receives a color: half the groups a blue color and half the groups an orange color. When switching members across groups, the incumbents are explicitly told that the newcomer comes from a group with a different color, and the newcomer is also explicitly told that the incumbents have a different color from the newcomer's prior group. Such minimal groupings have been shown to make group identities salient and lead to intergroup biases (Lane, 2016).

negatively affect contribution levels. In addition, they find that restrictions via entry quotas that limit the number of newcomers or via entry quizzes in which newcomers need to demonstrate sufficient competence can help to reduce this negative effect. Finally, McCarter and Sheremeta (2013) found that newcomers have a negative effect on the effort devoted to group cooperation.[3]

The discussed studies generally looked at overall effects, i.e., whether newcomers affect contributions to the public good, and produced mixed results. Potential mechanisms that drive newcomer effects have received much less attention. Our study focuses on the role of normative disagreements between incumbents and newcomers. In the next section, we turn to literature that suggests how newcomer-incumbent relations can be regarded as a special case positioned between intragroup and intergroup relations and how this depends on normative disagreement between newcomers and incumbents.

3.2.2. Intergroup differentiation in newcomer-incumbent relations

When examining the relevance of intergroup research for newcomer-incumbent relations, a first question that arises is whether incumbents categorize newcomers as ingroup or outgroup and vice versa. According to the group socialization model (Levine & Moreland, 1994), newcomers are initially not seen as full ingroup members by the incumbents. Instead, newcomers occupy a position between non-members and full members. Only after newcomers have experienced a socialization process, during which incumbents attempt to get the newcomers to act in line with the group's goals and norms, do the newcomers become full members. Thus, only over time newcomers are said to make a transition from 'outsiders to being insiders' (Bauer et al., 2007). In support of this group socialization model, most empirical evidence suggests that newcomers are initially regarded as outsiders by incumbents (Rink et al., 2013). Similarly, newcomers do often not immediately identify with the incumbent-group (Moreland, 1985). Newcomers' identification with the incumbent group may be especially low when the membership change is not initiated by the newcomers themselves but rather by an external decision-maker, such as the experimenter as in our study (Arrow & McGrath, 1993).

Self-categorization theory explains when and why people consider themselves as members of a particular group (Turner et al., 1987), and can therefore be used to predict when people are likely to identify with a group despite the presence of newcomers. According to this theory, one of the aspects that makes a set of people be seen as a self-relevant group is normative fit. When assessing normative fit, people compare a potential member's attributes and behaviors with one's expectations about dimensions that should distinguish between members of different groups in a particular situation. Thus, whether incumbents and newcomers contribute according to similar or different normative views will matter for whether they see themselves as one group or two groups. If newcomers act according to different

[3] There is a related literature where individuals can form groups and thus also have the option to enter and leave groups (endogenous group formation). However, these studies are designed to examine group formation processes rather than effects of membership changes. Because the role of newcomer and incumbent are assumed endogenously in these studies, causal effects of newcomer entry and newcomer status are more difficult to identify. We therefore refrain from reviewing this literature. However, a recent overview can be found in Guido et al. (2019).

normative views than incumbents, normative fit will be lower. This, in turn, will lead to a lower chance that newcomers are categorized as ingroup by incumbents.

Is categorization into different (sub)groups between incumbents and newcomers able to lead to conflict in terms of cooperation failure? According to social identity theory, ingroup favoring biases that impede collective cooperation may emerge once people are categorized into different groups (Tajfel & Turner, 1986). This has been demonstrated in several situations and expressions of bias, even where categorization into ingroups and outgroups is based on arbitrary criteria (see for a meta-analysis of economic experiments: Lane, 2016). Nevertheless, cooperation failure is especially likely when newcomers threaten the group's norms and goals (Böhm et al., 2020; Thravalou et al., 2020) Newcomers that contribute in line with a normative view that conflicts with the normative view of incumbents bring a larger threat to both the group's norm and the group's goal in terms of realizing the common good. Consequently, normative disagreements between incumbents and newcomers may not only matter for how newcomers are categorized, but also for the likelihood of cooperation failure.

In short, newcomers and incumbents do not readily identify with each other, and instead need a socialization process before they feel they belong to the same group. When newcomers are not regarded as part of the ingroup by the incumbents and vice versa, cooperation failure becomes more likely, especially when newcomers and incumbents are in normative disagreement. We next briefly review how the effect of normative disagreement on cooperation in public goods games has been studied so far.

3.2.3. Normative disagreement in public good games

Prior experiments on normative disagreement in public good games typically study disagreement between the commonly supported norms of equal-contributions and equal-earnings (Gangadharan et al., 2017; Nikiforakis et al., 2012; Reuben & Riedl, 2013). The tension between these norms arises when there are heterogeneous returns of the public good, i.e., when some group members benefit more from the public good than others. The equal-contributions norm prescribes that all members contribute equally, which implies that those who obtain a higher return from the public good also end up earning more. The equal-earnings norm prescribes that members who obtain a higher return from the public good also contribute more such that earnings are equalized. Several studies suggest that groups disagreeing on these two norms are more likely to fail to cooperate in terms of contributing to the public good (Kingsley, 2016; Nikiforakis et al., 2012; Rauhut & Winter, 2017; Winter et al., 2012). In these studies, participants do not have direct information on each other's normative views, but they do have information on each other's contributions. The effect of normative disagreement is hypothesized to work via observing each other's contributions and judging whether these contributions match one's own normative view. If people differ in their normative views and contribute according to their own view, they will observe that their view is not adhered to by others. The expected consequence is that people who feel others are not contributing enough will reduce their own contribution, causing a downward trend in contribution levels.

While punishment of low contributors mostly promotes cooperation if members agree on their normative views, it is predicted to be ineffective or harmful when members disagree on normative views. If members feel that their contribution is sufficient and nevertheless are punished, they may refuse to increase their contribution and retaliate against the other members with counter-punishment. Groups with normative disagreement are thus expected to obtain lower contribution levels than groups with normative agreement when punishment is possible. Prior research suggests that without punishment opportunities most groups show a trend to free-riding regardless of any between-group differences in normative disagreement (Reuben & Riedl, 2013).

The first study to manipulate normative disagreement experimentally found no evidence for a negative effect on contribution levels in groups where all members are new (Otten et al., 2020). However, such groups are substantially different from groups consisting of incumbents and newcomers with both having prior group affiliations and experiences. In groups with all new members, these members are initially unaware of what others deem appropriate contribution behavior. In the course of interacting with each other, a common standard of behavior emerges and turns into a group-specific norm that members become accustomed and attached to (Diekmann & Przepiorka, 2016; Titlestad et al., 2019). Once in place, norms are not easily changed. Many experiments show that participants often keep conforming to a norm even if the incentive structure changes in favor of norm transgression or movements to a new norm (Andreoni, Nikiforakis, & Siegenthaler, 2021; Duffy & Lafky, 2021; Guala & Mittone, 2010; Smerdon, Offerman, & Gneezy, 2019). Furthermore, participants take the norms they have learned in prior interactions with them when entering new social situations (Engl et al., 2021; Peysakhovich & Rand, 2013; Stagnaro et al., 2017), although not indefinitely (Arechar et al., 2018; Duffy & Ochs, 2009). Hence, newcomers are expected to be attached to the norm of their prior group, which may impede cooperation if this norm is different from that of the incumbents. We will test the following hypothesis:

Hypothesis. *Normative disagreement between newcomers and incumbents harms cooperation in terms of contributions to the public good.*

3.3. Methods

We conducted a computerized experiment in the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University with 192 participants sorted into 64 groups of 3 members each. Sample size was determined based on sample sizes in comparable studies (see for example Nikiforakis, Noussair, & Wilkening, 2012; Reuben & Riedl, 2013) and desired power. A Mann–Whitney ranksum test with individuals as the unit of analysis has a high power (0.96) to detect medium-sized effects (Cohen's d=0.5) given our sample size. We deem a medium-sized effect of Cohen's d=0.5 reasonable, as this is close to the average effect size found in the social dilemma literature (average effect size in the Cooperation Databank is 0.49; Spadaro et al., 2020). We recruited participants among students at Utrecht University using the internet recruitment system ORSEE (Greiner, 2015) during October-November 2019. We conducted 8 sessions. Each session included 24 participants and lasted about 75 minutes. Payment depended on behavior in the game. On average, participants earned 15 euros (min = 5, max = 22). The average age of participants was 24 years, 127 (66%) were female, 62 male, and 3 other. Almost all participants were attending courses at Utrecht University, 87 were of Dutch nationality, and 105 from various other countries. Participants were randomly placed in an individual cubicle and informed about the experiment through written instructions (provided in the supplementary material). The main part of the experiment consists of repeated rounds of a version of the public goods game with peer punishment (Fehr & Gächter, 2000).

3.3.1. Game

Each game round has two stages. First, each individual i receives an endowment of 20 monetary units (MU) and decides how much to contribute to a public good, c_i, where $c_i \in \{0, 1, \ldots, 20\}$. The part of the endowment that i does not contribute is kept for oneself. The public good benefits each group member j and consists of the sum of the contributions made by all members of the group $\sum_j c_j$. We use a group size of $N = 3$. Each member receives a return m_i per MU contributed to the public good, with $m_i < 1$. The sum of these returns is the multiplication factor of the public good M, with $N > M > 1$. Because $m_i < 1$, it is most profitable for the individual group member to contribute nothing in each round. However, since $M > 1$, it is most profitable for the group if every group member contributes their full endowment. These two aspects together constitute the social dilemma, i.e., the conflict between individual and collective interests, of public good provision. After all group members made their contribution decisions, the contributions and payoffs of each member are communicated to all group members.

Second, each group member is given the opportunity to assign punishment points $p_{ij} \in \{0, 1, \ldots, 10\}$ to each group member $j \neq i$. Each assigned punishment point p_{ij} costs 1 MU to the punisher and each received punishment point p_{ji} reduces the payoff of the punished group member by 3 MU. This provides participants with an informal instrument for norm enforcement (Fehr & Schurtenberger, 2018; Reuben & Riedl, 2013). The individual payoff (π_i) after one round of this two-stage game is calculated as follows:

$$\pi_i = 20 - c_i + m_i \sum_j c_j - \sum_{j \neq i} p_{ij} - 3 \sum_{j \neq i} p_{ji}$$

As is common in public good games with peer punishment, we do not let participants see who punished whom. This curbs punishment driven by revenge motives instead of dissatisfaction with others' contributions and thereby helps to analyze punishment as an instrument for norm enforcement.

As described in the review section, there is more variation in normative views between participants when there is heterogeneity in the returns of the public good. As a consequence, there is a higher likelihood that different groups develop different contribution norms (Gangadharan et al., 2017), which increases the chance of normative disagreements when members are switched between groups. Per group of three members, we randomly assign two members a low return of $m_i = .50$ from the public good and one member a higher return of $m_i = .75$ (this makes the joint multiplication factor $M = 1.75$). Participants know which members have the low-return and which the high-return. A prior study suggests that with this level of heterogeneity, people vary considerably in their normative views (Reuben & Riedl, 2013).

3.3.2. Normative views and expectations

Before playing the game and assigning the individual returns, we measured participants' normative views. To do so, we showed participants a hypothetical group of three members, two of which obtain a low-return ($m_i = 0.50$) and one of which obtains a high-return ($m_i = 0.75$)

from the public good. The exact same composition of returns is used in the actual contribution rounds of the experiment. We subsequently asked: "According to you, what is the appropriate amount that each member should contribute to the group account". Participants could then indicate a contribution for each of the three members between 0 and 20. Participants can try out different combinations of contributions, and see how it affects the earnings of each group member (see instructions and screenshots in the supplementary material). The normative views are measured again after the first 10 rounds of the game (before announcing the membership change), and also one final time after the second 10 rounds of the game. Before this second and third measurement, we explicitly remind participants that they need not be consistent between the different measurement moments.

Every time after participants report their personal normative views, we tell them that their group members were also asked to indicate appropriate contributions for three members in the public goods game. Each participant is then asked to guess the answers submitted by their group members, i.e., to report their normative expectations. To incentivize the normative expectation, participants are informed that one of their guesses will be picked randomly and yield an additional payment of 100 MU (~€1.40) if it matches the actual answer of at least one of the group members. Only at the end of the experiment, participants are informed of whether they were correct in the guess we randomly chose. This measure is inspired by earlier studies (Krupka & Weber, 2013; Reuben et al., 2015).

Participants' normative views are used to position these participants on a spectrum from equal-contributions to equal-earnings. Supporters of equal-contributions would answer that both types of players should contribute equally to the public good (ratio of contributions by high-return members to low-return members = 1), whereas supporters of equal-earnings would answer that high-return types should contribute twice as much as low-return types (ratio of contributions by high-return members to low-return members = 2). Participants who support a balance between both rules would answer that high-return types should contribute more than low-return types, but not twice as much. As we will show, almost all of our participants fall within one of these three categories and are rather evenly distributed across the three categories.[4]

We regard normative views as attitudes on what is appropriate. Unlike preferences, attitudes cannot be incentivized without running the risk of introducing confounds. Although it is possible to link a participant's normative view to monetary consequences for other unrelated participants (similar to spectator methods), this still runs the risk of confounding by other-regarding and distributional preferences (for example, a competitive spectator may wish to allocate few points to others, even though the spectator does think it is appropriate that others receive more points). For this reason, normative views are often elicited in a non-incentivized

[4] The precise score used to assign ranks is: $c_H - \overline{c_L} + 0.02\,\bar{c} + 0.0001R$, where c_H is the participant's view on the appropriate contribution for the high-return member, $\overline{c_L}$ is the participant's view on the appropriate contribution of the two low-return members on average, $\bar{c}$ is the mean appropriate contribution over all three members, and R is a random number between 0 and 1. The addition of $0.02\,\bar{c}$ makes sure that participants who assign a contribution of 20 to all members obtain slightly higher scores than participants who assign a contribution of 0 to all members. This helps to differentiate between different absolute levels of achieving the equal-contributions rule in the sorting method. The number 0.02 is chosen such that whether contributions are relative to returns or not always has dominance in the sorting mechanism over the absolute level of contributions. The addition of $0.0001R$ is to avoid tied scores.

way (Bicchieri, 2006; Cubitt et al., 2011; Hauge, 2015). However, as it turns out, normative views correlate highly with normative expectations in our experiment (correlation of .71, $p < .001$). Hence, in practice, it would not have made much of a difference if we used the incentivized normative expectations for group sorting or the unincentivized normative views. What is more, participants' normative views are also strongly correlated with their incentivized contribution behavior after the measurement. A participant's normative view measured after round 10 correlates at .72 with the participant's contribution decision in round 11. This gives additional indication that participants did take the measurement of normative views seriously.

3.3.3. Procedure

We implement two conditions, that differ only in the method of sorting and resorting participants in groups based on their normative views on the spectrum of equal-contributions vs equal-earnings, as shown in Figure 3.1. In the first condition (upper part of Figure 3.1), we initially sort participants from the same side of the spectrum together, whereas in the second condition (lower part of Figure 3.1) we initially sort participants from different sides of the spectrum together. After this sorting, each group receives a color (orange or blue) and the participants play an initial 10 rounds of the public goods game within their group. Participants are told that there will be a second part of the experiment after these 10 rounds and that they will receive information about this second part just before it starts. The first 10 rounds of the experiment are analyzed in another study (Otten et al., 2020) and we do not find a difference between conditions in outcomes (contributions, group identification, social norms) at the end of these first 10 rounds (also shown in supplementary material, section B.2). In the current study, we focus on the second part of the experiment.

After the first part of the experiment, we exchange one member per group for a member from another group with another color in both conditions, such that each group consists of one newcomer and two incumbents. We inform the participants of this membership change, and communicate to the incumbents that the newcomer comes from a group with a different color. Likewise, we communicate to the newcomer that the entered group is of a different color than that of the newcomer's prior group. We then let the reshaped groups play a second set of 10 rounds. How members are exchanged differs by condition. In the first condition, we replace a member that had a similar normative view to the two other members by a newcomer that has a dissimilar normative view to the two other members (upper part of Figure 3.1). Here, the newcomer is thus in normative disagreement with two incumbents. In the second condition, we replace a member that had a dissimilar normative view to the two other members by a newcomer that has a similar normative view to the two other members (lower part of Figure 3.1). Here, the newcomer is thus in normative agreement with the two incumbents. By comparing the contribution levels between both conditions during the second set of 10 rounds, we can test our hypothesis of the negative influence of normative disagreement among newcomers and incumbents on cooperation. To prevent experimenter demand effects, we did not inform participants about the method of group (re)sorting. We did not use deception, i.e., we did not offer untruthful information to the participants and they were aware that they did not have information on how group sorting happened.

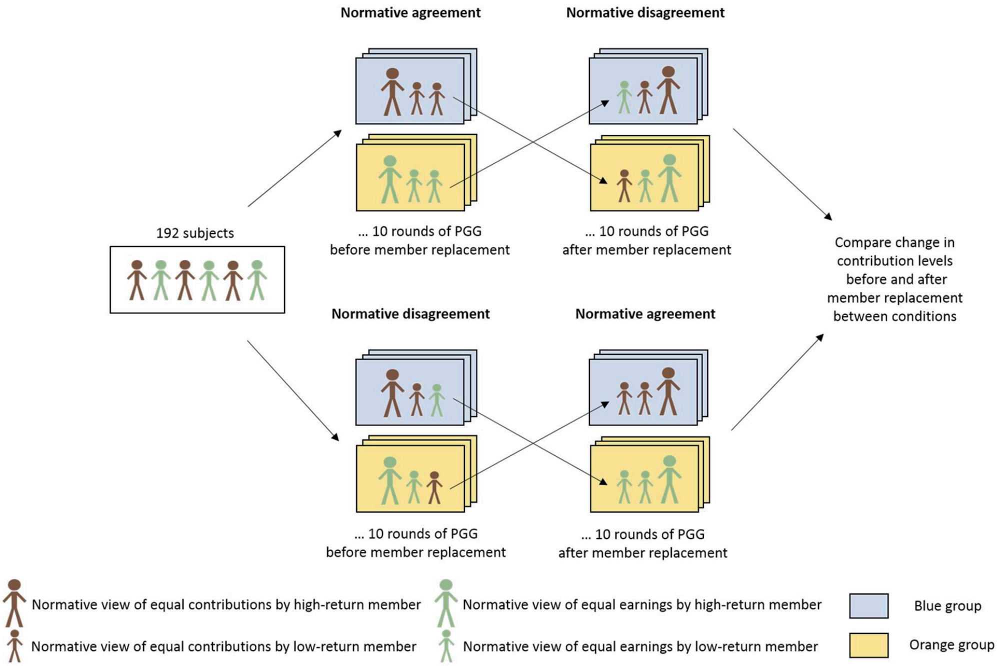


Figure 3.1. Experimental design.

Because there are two low-return members and one high-return member per group, we switch low-return members between groups. This allows us to compare the behavior of the low-return newcomer with the low-return incumbent when doing individual-level analyses, i.e., to prevent confounding of newcomer-incumbent differences with return-rate differences. The exact method of sorting and resorting is outlined in Figure 3.2.

Note that the variation in normative views among participants is used *within* conditions to sort them into groups. However, the manipulation *between* conditions (sorting for agreement vs disagreement) adheres to the strict experimental method. Participants are randomly assigned between these conditions, and hence the normative views (and other characteristics) of participants are also randomly distributed across conditions. As is common in studies on normative disagreement in public good games, we did not show participants each other's normative views. This helps to isolate the impact of normative differences from potential confounders such as expectations on what one's group members will contribute. Participants are thus initially not aware of how much they agree or disagree with their group members, and instead infer this indirectly from their contribution and punishment decisions. Participants are generally quite good at guessing other's normative views (as per the elicitation of these participants' normative expectations); 56% guess their group members' normative views correctly on average before the experiment, and 71% guess correctly on average after the experiment.

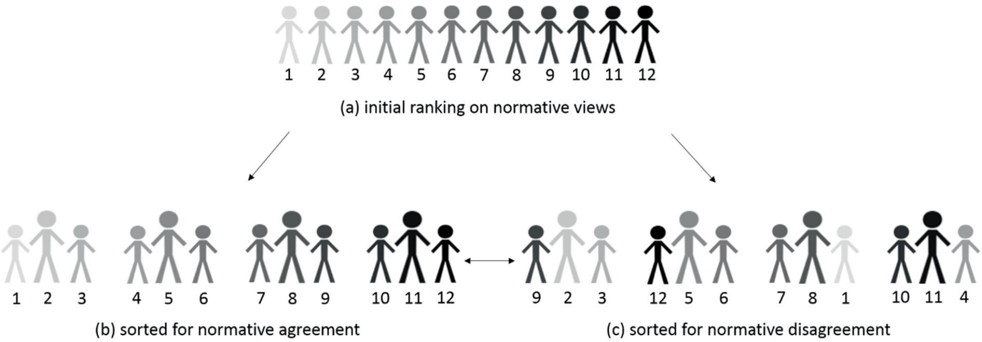


Figure 3.2. Example for method of sorting and resorting participants.
Note: At the beginning of the experiment (a), participants are ranked in terms of their normative views on the spectrum of equal-contributions vs equal-earnings (indicated with numbered grey shading). In the example presented here, there are 12 participants sorted into 4 groups. When sorting for normative agreement (b), we first form a group of the three highest-ranked participants (1-3), then of the remaining participants we again form a group of the three highest-ranked participants (4-6), and so on until all participants are grouped. When (re)sorting such that groups become dissimilar (c), we select the highest-ranked low-return participant from the first group in the first half of the groups (ordered in terms of support for equal-earnings over equal-contributions) and replace the participant with the lowest-ranked low-return participant from the first group in the second half of the groups, and repeat this procedure with the remaining groups. In this way, the extent of normative disagreement (in terms of rank-differences) is equal for all groups. In one condition participants start in groups sorted on similar normative views and then members are exchanged such that groups are sorted on dissimilar normative views (b → c), and in the other condition we reverse this order (c → b). In both conditions, there is one member per group that obtains a higher return from the public good than the other two members, as indicated by the size of the figures.

3.3.4. Social norms

Many definitions of social norms share the view that social norms involve shared expectations between group members on what actions are considered appropriate (Fehr & Gächter, 2000; Ostrom, 2000). Accordingly, mutually consistent normative expectations in a group are commonly used as an indicator of the existence of a social norm (Bicchieri, 2006; Bicchieri et al., 2014; Krupka & Weber, 2013). Therefore, to assess the presence of social norms in the experiment, we examine the overlap in normative expectations between members within groups. As mentioned, all participants were asked to guess what their group members deemed the appropriate contribution for each of three hypothetical members. A participant thus makes three guesses, and each of these guesses may be the same as, or different from, the guesses made by the other two group members. We examine the proportion of these three guesses that were exactly the same between all three members of a group (so the possible values per group are 0, 1/3, 2/3, and 1). The higher this proportion, the more the members share their normative expectations and hence share a social norm. As with the measure of normative views, normative expectations are elicited three times: once before the first round, once after the first 10 rounds (before the membership change), and once after the last 10 rounds. Just after membership change, we additionally asked participants to report their normative expectations toward their new members, but because these correlated highly with normative expectations toward prior group members (correlation = .72, $p < .001$), we do not analyze them separately.

We focus on the similarity in normative views and expectations after the last 10 rounds, but we give the similarity at all measurement moments in the supplementary material (Figure B.2).

3.3.5. Group identification

Both after the first and second set of 10 rounds, we ask participants to indicate their agreement on a Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree) to the following six statements:

1. I identify with other members of this group
2. I feel strong ties to this group
3. I am like other members of this group
4. This group is an important reflection of who I am
5. I feel proud to be a member of this group
6. I would like to continue working with this group

These items are commonly used to measure group identification in experiments (Leach et al., 2008; Ouwerkerk et al., 1999). We take the average across the six items as the group identification score. This score has high reliability, as indicated by a Cronbach's alpha of .93. We focus on the group identification score after the last 10 rounds, but we give the group identification score after the first 10 rounds in the supplementary material (Figure B.3).

3.3.6. Post-experiment measures and other information

After the experiment, participants were asked to provide information on background characteristics such as age, sex, and nationality, as well as some other measures such as their social value orientation. These post-experiment measures are not analyzed in this paper. They are described in detail in the pre-registration of the experiment: osf.io/gy8st. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons et al., 2012). All data and code are available at https://doi.org/10.24416/UU01-J5M2LE.

3.4. Results

We first examine whether the level of normative disagreement differs by experimental treatment as intended, i.e., whether our manipulation was successful. Recall that all participants provided their normative view on the appropriate contribution that high-return members and low-return members should make. To measure group-level disagreement, we rank participants within each group based on their ratio of appropriate contributions by high-return members to low-return members. Recall that a ratio of 1 means support for the equal-contributions norm and a ratio of 2 means support for the equal-earnings norm. The level of disagreement of each group is measured by subtracting the ratio supported by the lowest-ranked participant from the ratio supported by the highest-ranked participant. Figure 3.3a shows that groups in the condition where newcomers and incumbents are sorted for disagreement indeed end up with considerably more disagreement than groups in the condition where newcomers and incumbents are sorted for agreement (Mann-Whitney ranksum test, $p < .001$). This can also be seen in the supplementary material, Figure B.1 and B.2, where we show the normative views and disagreement for each group separately and how the disagreement developed over time throughout the experiment. Figure 3.3b shows that groups in the disagreement condition also end up with weaker social norms, as indicated by more dissimilarity in normative expectations among group members (Mann-Whitney ranksum test, $p < .001$). Normative disagreement between newcomers and incumbents thus also impedes social norm emergence.

However, Figure 3.3c shows that the extent of normative disagreement between newcomers and incumbents does not harm cooperation: the average contributions to the common good are very similar in both conditions. We find no significant differences in the contribution levels between the disagreement and the agreement conditions, neither before membership change (proportion of endowment contributed of .68 vs .71, OLS regression with cluster-robust standard errors, $p = .37$) nor after membership change (proportion of endowment contributed of .69 vs .75, OLS regression with cluster-robust standard errors, $p = .08$). Also when using population-averaged regression models, which account for repeated measures obtained from the same participant or group, we find that the change in contribution levels before and after membership change does not differ significantly between conditions, see supplementary material Table B.1. The contribution levels per round can be found in supplementary material Figure B.4 and also indicate no difference between conditions before and after membership change.

We additionally test the effect of normative disagreement between newcomers and incumbents conditional on contribution levels before membership change. The conditional test suggests that the change in normative disagreement brought about by the membership change does impact groups that had low contribution levels before membership change. Previously low-contributing groups do not change their contribution level after membership change in the disagreement condition, whereas previously low-contributing groups do increase their contribution level in the agreement condition (supplementary material Table B.2). However, for groups that had high contribution levels before membership change, the change in contribution level after membership change did not depend on the experimental condition. Finally, we run regressions with group-level contributions as the outcome variable and the beforementioned group-level normative disagreement as the predictor (while controlling for other factors known to influence contributions). These models indicate no significant relationship between disagreement and cooperation, also not in conditional tests (supplementary material Table B.2-B.3). Altogether, we find no support for the hypothesis that normative disagreements between incumbents and newcomers harm cooperation in terms of contributions to the common good.

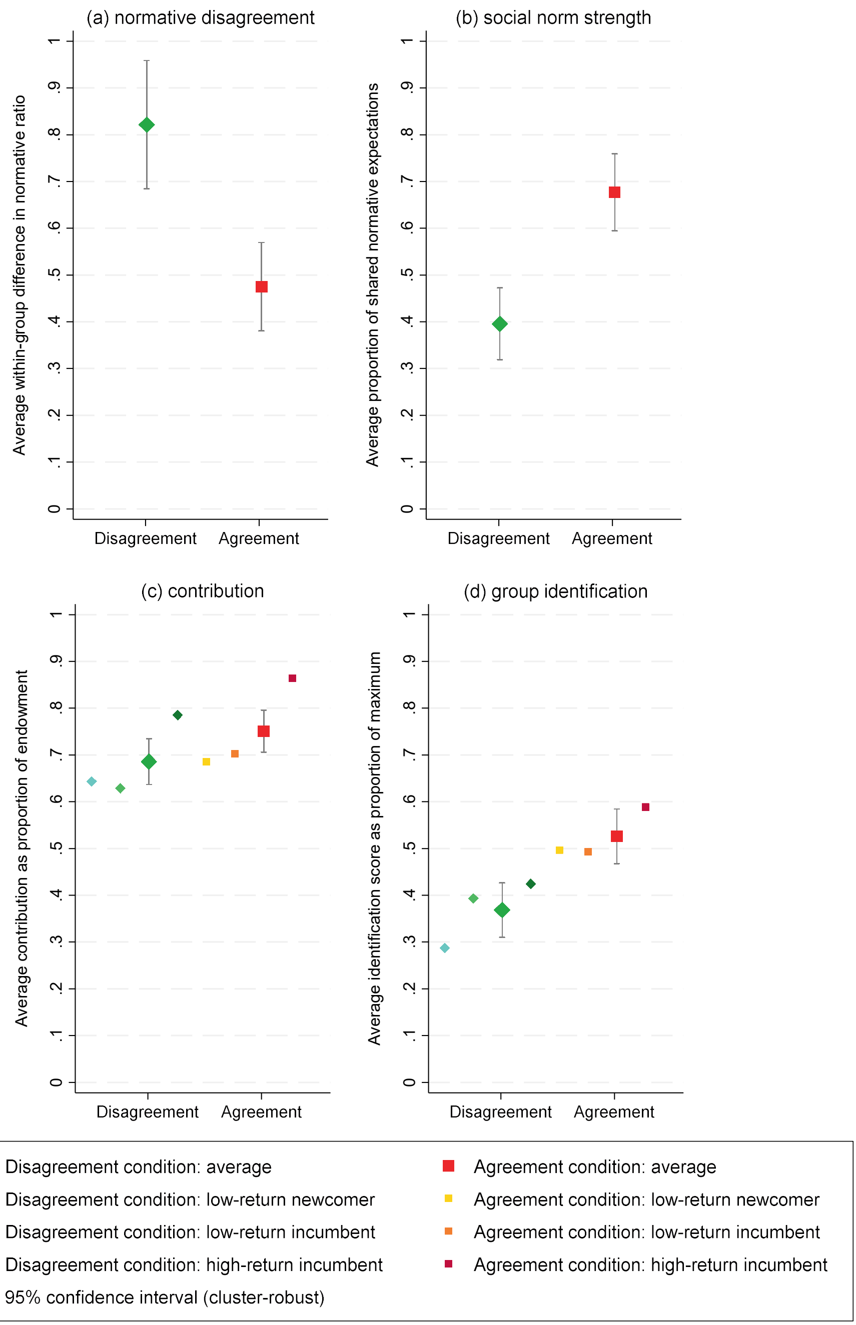


Figure 3.3. Disagreement, social norms, contributions, and group identification by condition.

Note: Disagreement refers to the condition where newcomers and incumbents are in normative disagreement. Agreement refers to the condition where newcomers and incumbents are in normative agreement. Normative disagreement (a), social norm strength (b), and group identification (d) values are based on the measurement after round 20. Contributions (c) were measured every round and we take the average of rounds 11-20.

Figure 3.3d shows that normative disagreement between newcomers and incumbents does harm group identification: group identification is about 40% higher in the condition where newcomers and incumbents agree instead of disagree (Mann-Whitney ranksum test, $p <. 001$). The differences between experimental conditions appear for both newcomers and incumbents, but newcomers report on average about 20% less group identification than incumbents (Mann-Whitney ranksum test, $p = .04$). Regression models with group identification as the outcome variable and group-level normative disagreement as the predictor (while controlling for other factors related to group identification) also indicate a significant negative association between normative disagreement and group identification (see supplementary material, Table B.4).

We next examine if normative disagreement between incumbents and newcomers affects the amount of punishment received, and whether punishment differs between incumbents and newcomers. Figure 3.4a indicates that punishment levels are considerably higher in the early interactions between newcomers and incumbents in the condition with normative disagreement. In the first three rounds after membership change, the punishment level is about twice as high when newcomers and incumbents disagree instead of agree about how much to contribute (0.61 vs 0.30 punishment points, Mann-Whitney ranksum test, $p < .001$, full details in Table B.5). However, the punishment level decreases over time in the condition with disagreement and thereby also the difference between the conditions. Regressions with group-level punishment as the outcome variable and group-level disagreement as the predictor (while controlling for other factors related to punishment) do not suggest a significant association between these two variables, also not in the first rounds after membership change (see Table B.6). The finding that normative disagreement between newcomers and incumbents leads to more punishment in the early interactions is thus not robust to alternative analysis. There is an outlier in terms of punishment in the last round (the tenth round after membership change), which is related to the so-called endgame effect: contributions tend to drop in the last round of the experiment, leading to higher levels of punishment in both conditions.

In Figure 3.4b, we show the average punishment points received by incumbents and newcomers for different levels of contributions. We see that there are no significant differences between incumbents and newcomers when they contribute medium to large amounts. In this case, punishment is low for both newcomers and incumbents. However, we find that newcomers are more strongly punished for low contributions. That is, newcomers receive about double the number of punishment points when they contribute ≤ 25% of their endowment compared to incumbents making these contributions (1.58 vs .76 punishment points received, Mann-Whitney ranksum test, $p < .001$, full details in Table B.5). Subdividing by condition suggests that this difference is mostly a result of normative disagreements between incumbents and newcomers. We find that this difference is significant when newcomers and incumbents are in disagreement (Mann-Whitney ranksum test, $p < .001$, Table B.5), but not when they are in normative agreement (Mann-Whitney ranksum test, $p = .17$, Table B.5). The differences in punishment points assigned to newcomers and incumbents also appear when examining not their absolute contributions, but how much they deviate from the average contribution in the group (controlling for return-rate differences). Newcomers are more strongly punished than incumbents for deviating from the average contribution, see supplementary material Figure B.6.

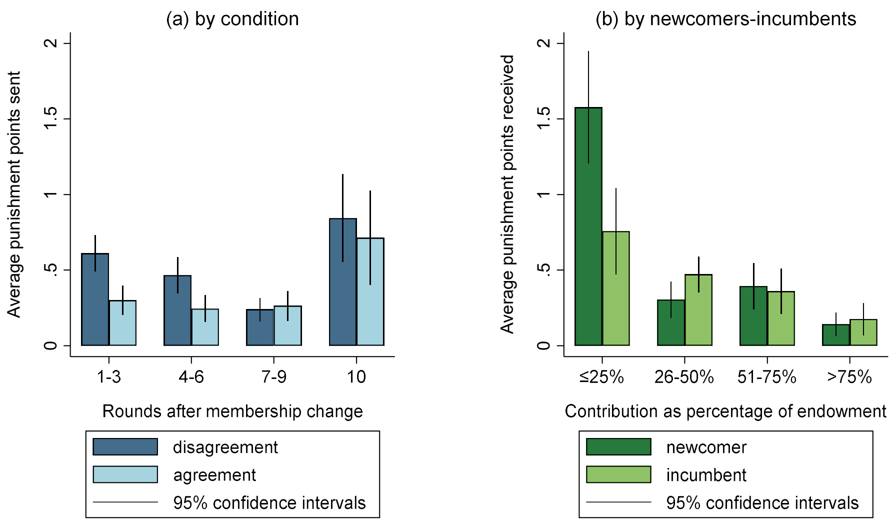


Figure 3.4. Punishment.

Note: we show the number of punishment points sent by condition and round in panel (a) and the number of punishment points received by newcomer-incumbent division and contribution amount in panel (b). Agreement refers to the condition where newcomers and incumbents are in normative agreement. Disagreement refers to the condition where newcomers and incumbents are in normative disagreement. The values in both panels are based on dyads, i.e., punishment sent to, and received from, individual group members. Because each participant has two co-members, the punishment levels will be twice as large in practice. The number of observations totals to 3840 in panel (a): 192 participants × 10 rounds × 2 co-members to punish or receive punishment from. To keep the return rate constant when comparing newcomers and incumbents in panel (b), we only compare (low-return) newcomers to low-return incumbents (which make up two-thirds of the participants), giving 128 participants × 10 rounds × 2 co-members = 2560 observations. When subdividing by contribution level, the number of observations is 326 for contributions ≤ 25%, 792 for contributions 26-50%, 516 for contributions 51-75%, and 926 for contributions > 75%.

3.4.1. Predicting contribution levels

To examine why contribution levels do not change after membership change and what alternatively predicts contribution levels, we conduct further exploratory analyses. These full analyses can be found in supplementary material B.6. Here we explain the main findings. We find that the change in the group-average normative view that is brought about by the membership change is an important factor in predicting whether the membership change increases or decreases the group-average contribution (i.e., whether the group-average contribution in the second 10 rounds is higher or lower than in the first 10 rounds). A participant's average normative view measures what a participant deems to be the appropriate average contribution in the group. For example, if a participant thinks the high-return member should contribute 20 and the two low-return members should contribute 10, the participant's average normative view would be (20+10+10)/3 = 13.33. The group-average normative view is then the average of its three participants' average normative views. We find that if newcomer entry leads to a higher group-average normative view, then the group-average contribution increases, and the opposite holds when newcomer entry reduces the group-average normative view. This is not just a function of the normative view that the newcomer brings in, it also depends on the normative view of the incumbent that the newcomer replaces. The group-average normative view increases if the newcomer's view is higher than the view of the incumbent that is replaced. The change in the group-average normative view correlates at .53 with the change in the group-average contribution (OLS regression on 64 groups, $p <. 001$, also when controlling for other related factors, Table B.3). The group-average normative view is an important reason for the null-effect of membership change on contributions. When averaging over all groups (also within conditions), newcomers do not change the group-average normative view; in some groups they increase the group-average whereas in other groups they decrease the group-average, which balances out when averaging over all groups.

When predicting how contribution levels develop after membership change on an individual level, we find that a participant's own normative view and the contribution of the participant's group members explain about half of the variation in contribution levels. However, the relative influence of these two variables is different for newcomers and incumbents. Compared to incumbents, newcomers contribute more in line with the contributions of their group members and less in line with their own views. We furthermore find that the influence of others' contributions is driven to a large extent by punishment, which holds both for incumbents and newcomers. When moving to the next decision round, participants tend to contribute the same amount if they were not punished in the prior round. But if they were punished, they increase their contribution. Moreover, the stronger the punishment, the more participants move away from their prior contribution behavior in favor of higher contributions. With high levels of received punishment (about 10 punishment points), a participant's prior contribution is no longer related to the participant's subsequent contribution. However, if participants are punished when contributing large amounts (about more than 15 MU), they do not significantly change their behavior (see Table B.7 in the supplementary material).

Because newcomers act more in line with the contribution of incumbents, and the contribution of incumbents is related to the incumbents' normative views, newcomers end up conforming more to the incumbents' normative views than their own. The relative influence of the newcomer's own normative views and that of the incumbents is best shown in the condition with disagreement between newcomers and incumbents, because that is where they have to decide what view to conform to. In Figure 3.5, we show for newcomers (a) and incumbents (b) how their contributions in each round correlate with their own normative view and the normative view of their group members. Figure 3.5a shows that while newcomers start by contributing according to their own normative view in the first round after membership change, already in the second round after membership change they contribute according to their group members' normative view. In contrast, Figure 3.5b shows that incumbents contribute more in line with their own normative view throughout all rounds. These findings suggest that when there is disagreement between incumbents and newcomers on how to contribute to the public good, newcomers largely concede to incumbents. That this occurs despite our finding that normative disagreement harms social norm emergence (Figure 3.3b) suggests that the contribution norms in these groups are descriptive rather than injunctive. Further analyses indicate that the finding of newcomers conceding to the incumbents cannot be attributed to newcomers changing their normative view or being a minority (see supplementary material, Figure B.7-B.8).

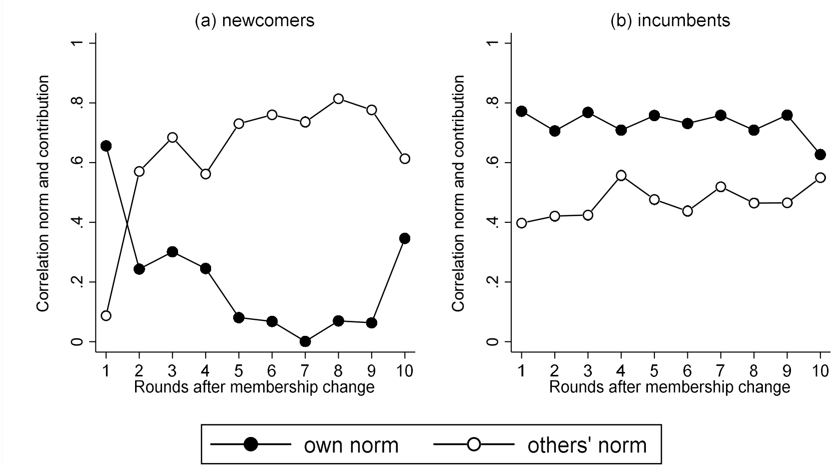


Figure 3.5. Correlation between contribution and own or others' normative views.

Note: All participants provided their normative view on the appropriate contribution that a high-return member should make and the appropriate contribution that each of the low-return members should make. We examine if participants contribute in line with their normative view on how much they themselves should contribute (i.e., if they have a low-return, how much they think low-return members should contribute; if they have a high-return, how much high-return members should contribute) or in line with how much their group members think they should contribute. The results are separated by newcomers and incumbents and shown for the condition where the newcomers and incumbents hold different normative views and thus have to decide what view to conform to. We use the normative views as measured just before the membership change (round 10) and correlate it with the subsequent contribution decisions.

3.5. Discussion

Groups often rely on shared norms to achieve cooperation. These shared norms may be challenged when new members enter the group and old members leave. If newcomers hold normative views that are incompatible with the normative views of incumbents, a normative disagreement ensues that may harm cooperation. Using the public goods game, we studied experimentally whether the impact of membership change on cooperation depends on the level of normative disagreement between incumbents and newcomers. While we find that normative disagreement negatively affects newcomer-incumbent relations in terms of group identification and social norm emergence, we find that contributions to the common good are not impacted. Newcomers and incumbents can thus cooperate for the provision of common goods even in the presence of normative disagreement and the associated lower levels of group identification and social norm emergence. Our findings suggest that norm enforcement plays an important role in sustaining cooperation in groups consisting of newcomers and incumbents. Norm enforcement was possible in our experiment via punishment. Low-contributing participants respond strongly to punishment by increasing their subsequent contributions, and newcomers in particular are punished strongly for making low contributions that deviate from the incumbents' contribution norm. The result is that newcomers mostly adapt to the incumbents' norm. That this occurs despite newcomers sticking to their normative views and expectations suggests that the contribution norms are descriptive rather than injunctive (Cialdini et al., 1990).

These results paint a nuanced picture of the effects of normative disagreements between newcomers and incumbents. While it may be reassuring that contribution levels are unaffected, the newcomer-incumbent relations, as measured by group identification and social norm emergence, are negatively impacted by normative disagreements. And while cooperation is achieved, it is mostly at a level supported by the incumbents and not the newcomers. There is thus a newcomer-incumbent inequality in the extent to which normative views are realized. This also speaks to the ongoing debate on whether the integration of migrants is a one-sided or two-sided process (Klarenbeek, 2021), with our experiment providing support for the former. Similarly, the punishment results indicate a newcomer-incumbent inequality in terms of how strongly low contributors are punished. Altogether, we can thus not unequivocally conclude whether normative disagreements are harmful; it depends on which outcome measures are considered. Focusing only on the achieved contribution level as outcome may lead to a misrepresentation of how group members are doing in terms of subjective experiences.

The prediction that cooperation for public good provision is harmed when newcomers and incumbents do not share the same norms is relatively common, especially when considering cooperation between immigrants and natives (Collier, 2013; Habyarimana et al., 2009; Ostrom, 2000). It might therefore come as a surprise that we find no support for it in our experiment. However, a recent literature review on cooperation between natives and immigrants suggests that prior estimates for the effect of normative differences between immigrants and natives on public good provision are often confounded with other aspects of interethnic groups, such as poverty and political instability (Baldassarri & Abascal, 2020). To get a causal estimate on the effect of normative differences between newcomers and incumbents on group cooperation, experimental manipulation such as in our study is helpful to isolate the effect of normative differences from potential confounders. Our study suggests that normative disagreements alone do not predict whether newcomers and incumbents cooperate toward public good provision.

Indeed, an emerging literature challenges the commonly held view that the perceived threats of immigration are of an economic nature, and instead suggests that the perceived threats are of a cultural nature (Hainmueller & Hopkins, 2014). Our findings that disagreements between newcomers and incumbents do not harm contributions to the public good but do harm group identification and impede social norm emergence seem to corroborate this view. That is, we found no economic impacts in terms of contribution to the public good, but did find impacts in terms of group identification and social norm emergence. However, we did find some evidence that norm enforcement via punishment is initially higher under normative disagreement. Since punishment is costly, this does suggest some temporary harmful economic impacts of normative disagreement. This norm enforcement is especially directed at newcomers, who are punished twice as much as incumbents for contributing low amounts and hence incur extra costs. A recent field experiment on norm enforcement in Germany found a similar pattern; natives impose norms on immigrants considerably more so than the other way around (Winter & Zhang, 2018). These authors suggest that social control therefore will increase as more newcomers enter and communities become increasingly ethnically diverse. The norm examined in this field experiment is different from our study, namely the anti-littering norm, and the setting is very different as well (field vs lab). That we nevertheless find the same pattern of newcomers being punished more strongly than incumbents for norm violations suggests that this finding may generalize to other norms and settings.

As with any experimental study, our conclusions might depend on design choices. We tried to stay close to the prior studies on normative disagreements in public goods games in our design. Future research can assess to what extent our findings also hold under different designs. We provide a few suggestions. We studied normative disagreements between the norms of equal contributions and equal earnings, both of which are common and important norms in heterogeneous groups (Nikiforakis et al., 2012; Reuben & Riedl, 2013). Future research can study whether disagreements differently impact group relations when other norms are involved. For example, group relations between native-majority members and immigrants are often argued to be harmed by disagreements on religion or work-related norms (Hainmueller & Hopkins, 2014). Our finding that disagreement about contribution norms lowers group identification might therefore extend to these other forms of normative disagreement, which future research can test. Moreover, we studied groups of three in which one incumbent is replaced by one newcomer. Normative disagreements might be more harmful when the number and share of newcomers are larger, as this increases the chance for separate ingroup-outgroup subdivision and the newcomers' power to influence norms (Pettigrew, 1991). Prior research suggests that if more than one newcomer enters a group, the newcomers tend to identify with each other instead of with the incumbents (Moreland, 1985), which may further lower a common sense of group identification among incumbents and newcomers. Finally, future research could examine newcomer-incumbent cooperation using natural rather than minimal groups, for example by letting immigrants and natives play the public goods game together (Drouvelis et al., 2021). Although natural identities are conflated with potential confounders (e.g., differences in status and income), they can enhance the external validity and thereby complement research using minimal groups.

We conclude that the absence of conflict in terms of cooperation failure does not imply that newcomer-incumbent relations are harmonious. Our results suggest that a fuller understanding of newcomer-incumbent relations is achieved when multiple dimensions are measured simultaneously, e.g., behavior (in our case contributions and punishment), subjective experiences (in our case group identification), and normative perceptions (in our case normative views and expectations). As we showed, the results between these different dimensions need not be in line with each other. While this makes it more difficult to draw clear conclusions, we think it is important to allow for a nuanced view on intergroup relations. High levels of contributions to the common good are not always unequivocally good or bad; it may depend on how group members think about it. We showed that normative disagreements between newcomers and incumbents do not lead to conflict in terms of cooperation failure. Yet, they do seem to negatively impact group identification and lead to contribution levels that predominantly reflect the normative views of incumbents rather than newcomers.

Chapter 4. Human cooperation in changing groups in a large-scale public goods game[1]

Abstract

How people cooperate to provide public goods is an important scientific question and relates to many societal problems. Previous research studied how people cooperate in stable groups in repeated or one-time-only encounters. However, most real-world public good problems occur in groups with a gradually changing composition due to old members leaving and new members arriving. How group changes are related to cooperation in public good provision is not well understood. To address this issue, we analyze a dataset from an online public goods game comprising 1.5 million contribution decisions made by 135 thousand players in 11.3 thousand groups with 234 thousand changes in group composition. We find that changes in group composition negatively relate to cooperation. Our results suggest that this is related to individuals contributing less in the role of newcomers than in the role of incumbents. During the process of moving from newcomer status to incumbent status, individuals cooperate more and more in line with incumbents.

[1] A slightly different version of this chapter has been published as Otten, K., Frey, U. J., Buskens, V., Przepiorka, W., & Ellemers, N. (2022). Human cooperation in changing groups in a large-scale public goods game. *Nature Communications*, *13*(1), 6399. Otten carried out the investigation, wrote the original draft, and made the visualizations. Otten and Frey conceptualized the research and conducted the formal analysis. All authors were responsible for the methodology and review & editing of the writing. For this study, we used existing data owned by Gameforge. Gameforge allowed the authors to use the raw data for academic purposes, but did not allow the raw data to be openly shared. We provide an aggregated dataset at doi.org/10.17605/OSF.IO/3WYE9.

4.1. Introduction

Cooperation to provide public goods has been a central human activity throughout history. In our ancestral past, communal hunting, food sharing, and warfare produced public goods (Bowles & Gintis, 2013; Hawkes et al., 1993; Hill, 2002). In contemporary life, important public goods include law enforcement, public education, social security, public transport, voting, and tackling climate change (Kallhoff, 2014). Despite the ubiquity of public goods, their provision is rarely trivial. Because every member of a group benefits from the public good, including those who do not contribute to its provision, individuals have an incentive to free-ride. However, if too many individuals free-ride, the public good is not provided and nobody benefits (Olson, 1965). Public good provision is thus a social dilemma because individual and collective interests are not aligned. How groups can cooperate to overcome this social dilemma is a central scientific question that is still not fully answered.

Researchers have predominantly used economic game experiments to study the factors involved in cooperation in public good provision (Van Dijk & De Dreu, 2021). In typical public goods games, participants receive a monetary endowment and are asked to decide how much of it to keep for themselves and how much to contribute to a group project that also benefits other participants. Collective returns are maximized if everybody contributes their full endowment to the group project, but individual returns are maximized by not contributing anything, irrespective of what others do. More than a thousand studies have been conducted using public goods games (Spadaro et al., 2022), leading to many important insights into the motives (e.g., self-interest, reciprocity, fairness), institutions (sanctioning and reputation systems, social norms), and dynamics (e.g., conditional cooperation) related to human cooperation (Balliet et al., 2011; Chaudhuri, 2011; Guido et al., 2019; Ledyard, 1994; Van Lange et al., 2013). Many studies also suggest that cooperative behavior in public goods games is predictive of cooperative behavior in natural settings (Barr et al., 2014; Fehr & Leibbrandt, 2011; Gneezy et al., 2016; Hergueux et al., 2015; Laury & Taylor, 2008; Rustagi et al., 2010), although there is also some counter-evidence (Galizzi & Navarro-Martinez, 2019; Levitt & List, 2007).

Public goods games are typically studied in one of two contexts: in repeated interactions among the same group members (also known as partner matching) and in interactions among members who change randomly after each interaction (also known as stranger matching) (Chaudhuri, 2011; Fehr & Fischbacher, 2003; Rand et al., 2014). Research shows that contribution levels are considerably higher with partner matching than with stranger matching (Botelho et al., 2009; Fehr & Gächter, 2000; Ghidoni et al., 2019; Keser & Winden, 2000; Zelmer, 2003). However, in real life, most public good problems occur in groups that lie between these two extremes, where group composition gradually changes over time due to old members leaving and new members arriving (Collier, 2013; Trainer et al., 2020). In such contexts, groups that produce public goods typically consist of a mix of incumbents and newcomers. Common examples are immigrants entering new countries, residents entering new neighborhoods, and employees entering new organizations and work teams. Despite their relevance for real-life groups, relations between ongoing changes in group composition and cooperation have not been studied systematically.

One reason for this may be high data requirements. Because groups are the units of analysis, sample sizes must be considerably larger than in experiments studying individual behavior. What is more, there are many ways in which group compositions can change over time, and the number of possibilities increases considerably with group size. This further amplifies the requirement for large sample sizes. Additionally, groups need to be tracked for longer periods of time to observe compositional changes and their effects. So far, the few studies that relate changes in group composition to cooperation in public goods games have had to rely on relatively small sample sizes (between 100-300 participants, in groups of 2-6 members) and were only able to observe a limited set of group composition changes in a short time span (Duffy & Lafky, 2016; Grund et al., 2015; Otten et al., 2021; Ranehill et al., 2014; Sonnemans et al., 1999).

There is a related literature on cooperation in dynamic networks. In dynamic networks, actors have some control over whom they interact with, allowing them to form and break ties with others based on others' cooperation decisions. Evolutionary models show that such strategic tie formation and dissolution can promote cooperation (Cavaliere et al., 2012; Santos et al., 2006). In particular, cooperation is expected to be higher if actors can frequently break with defectors and link with cooperative actors (Fu et al., 2008). Behavioral experiments generally support these predictions; cooperation is higher in dynamic networks than in static networks and leads to clusters of cooperation (Fehl et al., 2011; Melamed et al., 2018; Rand et al., 2011; Wang et al., 2012). However, this literature leaves largely unaddressed what happens in situations where individuals have little say in how the composition of their group changes and hence cannot easily break with defectors. For example, residents in a neighborhood typically do not get to choose who enters or leaves and employees in work organizations frequently have to accept with whom they have to collaborate based on the decision of their employers. What is more, exit costs are typically substantial in these situations, meaning that incumbents have little option to leave if they are dissatisfied with the newcomers. In sum, more research is needed on how group changes are related to cooperation when avoiding free-riders is not feasible.

In this paper, we analyze large-scale data from the multiplayer online game Ikariam, in which public goods games are deliberately built in by the designers and are central to players' success. The public goods games are played over a time span of multiple months in groups with a broad range of compositional changes. Moreover, options to leave the group or exclude free-riders are limited. The data are ideal to shed light on our two main research questions. First, what is the relationship between changes in group composition and cooperation in terms of contributions to a public good? Second, can this relationship be attributed to the contributions of the newcomers, the incumbents, or both? Two prior studies have shown that Ikariam players use contribution strategies that can be categorized as free-riding, conditional cooperation, and high cooperation (Frey, 2017, 2019), but how group changes relate to cooperation and whether newcomers and incumbents contribute differently has not yet been examined.

The theoretical answers to these questions can broadly be categorized into two opposing arguments. The first argument posits that newcomers will initially contribute less than incumbents because they lack a shared history with the incumbents. Newcomers may therefore have a lower concern for the group's welfare and a lower awareness of the norms prescribing contributions or a lower willingness to conform to them (Collier, 2013; Ostrom, 2000; Rand et al., 2014). Newcomers are then expected to increase their contribution to the incumbents' level with more time spent in the group, as this increases the shared history they have with incumbents and allows them to get accustomed and socialized to the prevailing contribution norm in the group. The second argument posits that newcomers will initially contribute more than incumbents because they are under special scrutiny when entering the group and need to show their worth to the incumbents (Barclay & Willer, 2007; Cimino, 2011). Hence, newcomers are expected to decrease their contribution to the incumbents' level with more time spent in the group, as their position will have been earned over time and the scrutiny decreases. Although both these arguments mainly suggest a role for newcomers' contributions in the relationship between changes in group composition and group cooperation, it is also possible that incumbents condition their contributions on changes in group composition. Indeed, some studies suggest that incumbents anticipate lower contributions by newcomers and other incumbents and will therefore reduce their contribution if newcomers enter (Grund et al., 2015; McCarter & Sheremeta, 2013).

In the Ikariam game, each player starts as the ruler of a town on an ancient Greek island with up to 16 other players on the island. On an island, individuals accumulate resources in real-time (i.e., the game continues after a player logs out) and have to make strategic decisions on how to use these resources. The resources can be invested in private goods, such as constructing and upgrading different types of buildings in one's town, e.g., a town hall, trading post, museum, or tavern. The resources can also be contributed to public goods. The main public good is a sawmill that provides wood. Wood is a crucial resource needed to develop one's town and hence to advance in the game. The sawmill is the main way for players to obtain wood. The rate at which individual players can extract wood from the sawmill depends on how many units of wood have been collectively contributed to the sawmill by all players on the island. This means that individuals who do not contribute nevertheless benefit from the contribution of other players on the island (non-excludability). The rate at which a player can extract wood from the sawmill does not depend on the rate at which other players extract wood (non-rivalry). The non-excludability and non-rivalry of the island's resource extraction make the game a pure public good analogous to public goods games (Frey, 2017, 2019). More details on the public good dynamics in Ikariam are available in the Methods section.

An individual can enter additional groups by building a town on additional islands. Entering additional islands is an essential part of progressing in the game, as individuals eventually will need more resources than produced on their first island(s). When players newly join an island, they are able to extract resources from the island at a rate that depends on how much the incumbents of that island have contributed so far. Entering an island thus means entering a new group with a specific state of public good provision. An individual can become part of up to 12 groups, which means that an individual can become a newcomer several times. This also means that the same individual will sometimes be an incumbent in one group and a newcomer in another. Incumbents have no say in who enters their group and cannot exclude members.

We analyze longitudinal data on about 1.5 million contribution decisions of about 135 thousand players located in about 11.3 thousand groups. We examine the relationship between these contribution decisions and a total of about 234 thousand changes in group composition that occur over a time span of about a year. The results suggest a robust negative relationship between group changes and cooperation in public good provision. Newcomers contribute less to public goods than incumbents and thereby lower groups' average contributions. However, as newcomers spend more time in a group, they increase their contributions to the public good and contribute more in line with the group. That is, in the process of moving from newcomer to incumbent status, individuals' cooperation increases to the average level of the other group members.

4.2. Results

The data are systematically structured in 28 biweekly intervals, i.e., we have one observation per two weeks per player-group combination. We refer to each biweekly interval as a time period, so we have 28 time periods (analogous to rounds of public goods games in lab experiments). Players' contributed resources are divided by their available resources to obtain their contribution percentage at each time period. We first examine the relationship between the average contribution percentage and the number of newcomers within each bi-weekly period. We regard an individual as a newcomer if the individual was not present in the group before the current period. We regard an individual as an incumbent if the individual was present in the group before the current period. Figure 4.1 shows how the group's average contribution percentage relates to the number of newcomers, group size, time periods, and the combination of these three variables. Note first that, for groups with zero newcomers, contribution patterns resemble those of aforementioned lab experiments with no newcomers in several aspects: (1) the initial contribution percentage lies between 40 and 60 percent, (2) the contribution percentage decreases over time, and (3), contribution percentages are higher in smaller groups. However, of particular interest to us is how the contribution percentage relates to the number of newcomers.

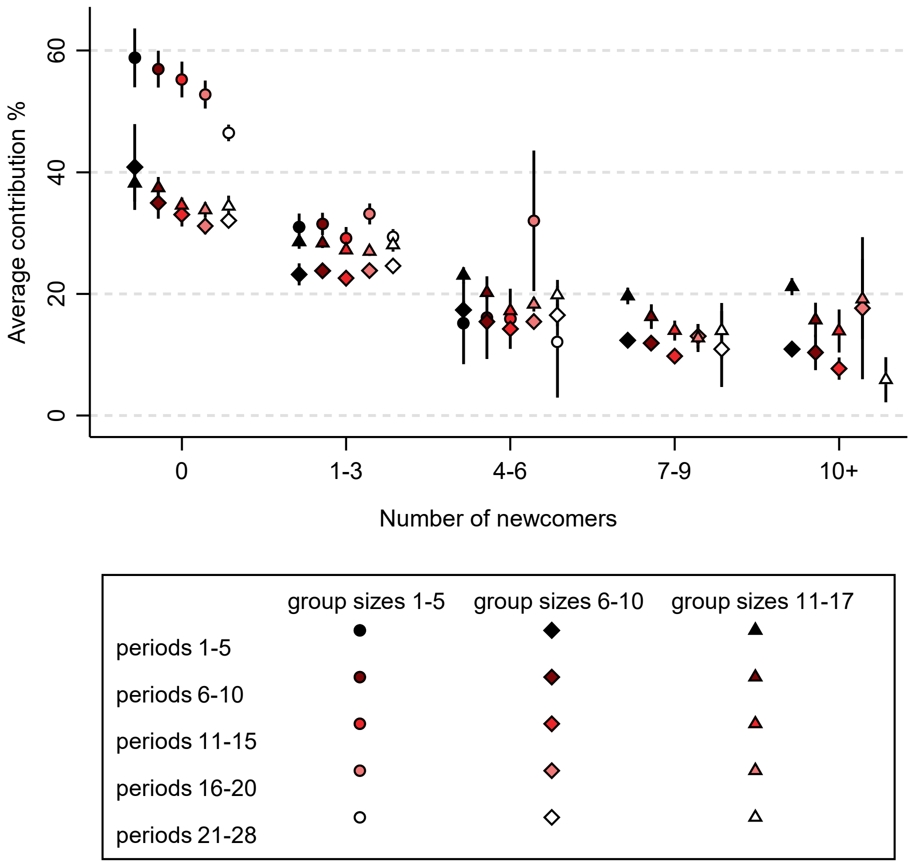


Figure 4.1. Groups' average contribution percentages by the number of newcomers, group size, and time periods.

Note: Data are presented as mean values and group cluster-robust 95% confidence intervals are provided via vertical spikes. The data are discretized in this figure for visualization purposes, the non-discretized analyses can be found in Table 4.1. The combination of a large number of newcomers and a small group size is not possible because a large number of newcomers implies a large group size. Therefore, no markers are shown for the combination of the group size category of 1-5 members and the upper two categories of the number of newcomers (7-9 and 10+ newcomers). Groups' average contribution percentage negatively relates to the number of newcomers. This holds for groups of different sizes, but more so for small groups (1-5 members) and also holds regardless of the time periods in which the newcomers enter. Results include 11348 groups, with groups existing on average for 17-18 periods, giving a total number of observations of 199530 group-period combinations.

We see that a group's average contribution percentage negatively relates to the number of newcomers. The negative relationship is strongest when moving from no newcomers to a moderate number of newcomers (4-6) and is somewhat smaller when moving from a moderate to a large number of newcomers (+10). The contribution percentage decreases from about 50% to about 10% when moving from the smallest to the highest number of newcomers. The negative relationship holds for groups of different sizes, although it appears stronger for small groups (1-5 members), and also holds regardless of the time periods in which the newcomers enter. Although the variables are discretized in Figure 4.1 for visualization purposes, we do not discretize data in any of the statistical analyses and we find the same patterns there. All tests are two-tailed. Statistical tests using fixed effects regressions to account for between-group confounders reported in Table 4.1 confirm that the negative relationship between the number of newcomers and the average contribution percentage is significant and stronger for smaller groups. We also find that the negative relationship is significantly stronger at later time periods, although the size of this interaction is small. Out of the group size, time period, and number of newcomers, it is the number of newcomers that relates most strongly to the contribution percentage.

Table 4.1. Regression model of average contribution percentages with group fixed effects.

	Model 1	Model 2
Number of newcomers	-2.72***	-4.87***
	(0.04)	(0.06)
Group size	-0.32***	0.04
	(0.03)	(0.03)
Period	-0.27***	-0.33***
	(0.01)	(0.01)
Number of newcomers × group size		0.48***
		(0.01)
Number of newcomers × period		-0.05***
		(0.01)
Intercept	33.75***	33.75***
	(0.07)	(0.07)
R^2 (overall)	0.04	0.05
Rho	0.58	0.58

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients of independent variables and intercept are marginal effects. Standard errors are provided in parentheses. Statistical significance is calculated using two-sided t-tests. Results include 11348 groups, with groups existing on average for 17-18 periods, giving a total number of observations of 199530 group-period combinations.

We next examine whether the relationship between the average contribution percentage and the number of newcomers also holds over the entire duration of the game instead of within time periods. Figure 4.2 shows the relationship between a group's average contribution percentage over all 28 time periods and the total number of newcomers that entered during this time. We once again see that there is a negative relationship between the number of newcomers and the average contribution percentage. Groups with very low total numbers of newcomers obtain contribution percentages of about 40-50% whereas groups with very large total numbers of newcomers obtain contribution percentages of about 10%. The negative relationship between a group's average contribution percentage and the total number of newcomers across all time periods is significant and explains about 5 percent of the variation in contribution percentages (see Supplementary Table C.3). The bivariate correlation between the number of newcomers and the contribution percentage is -.18 within periods and -.23 across periods, which are regarded as small to moderate effect sizes in related research (Hemphill, 2003).

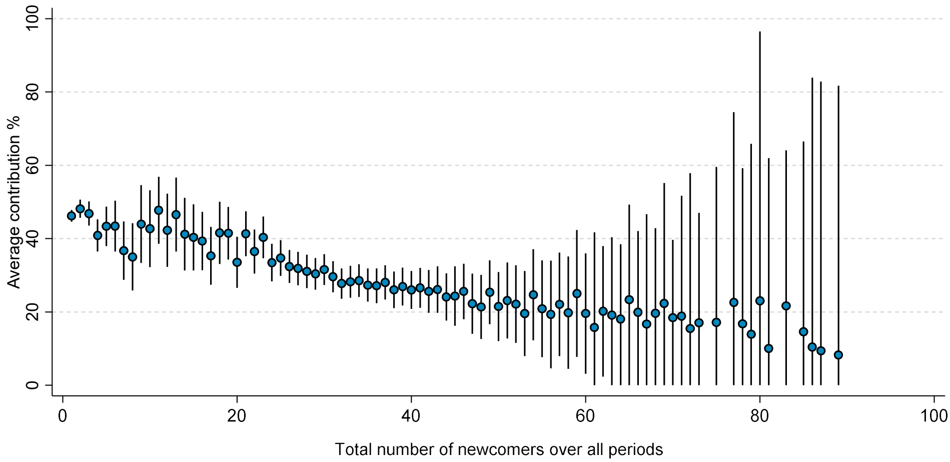


Figure 4.2. Average contribution percentages by the total number of newcomers over all periods.

Note: Data are presented as mean values and 95% confidence intervals are provided via vertical spikes. In contrast to Figure 4.1, we look at a group's average contribution percentage over all 28 periods, which means there is only 1 observation per group. Results include 11353 groups. At any given period, a group can only consist of up to 17 players and hence only up to 17 newcomers can enter. However, groups may also experience incumbents leaving in some periods, opening up new spaces for newcomers. So the total number of newcomers across all time periods can be higher than 17. Because there are few groups with a very high total number of newcomers, the contribution percentages for these groups have larger confidence intervals. We cut off confidence intervals below 0 because contribution percentages below 0 are not possible. We see a negative relationship between the group's average contribution percentage across all time periods and the total number of newcomers that have entered during this time. In Table C.3 of the Supplementary Information, we show that this relationship is significant, also when controlling for the group's average group size and time period.

Since there is a maximum group size of 17 members, a total number of newcomers above 17 is only possible if there were also individuals leaving the group. The leavers thus make way for the newcomers. In fact, the number of newcomers and leavers are closely related in Ikariam; the total number of newcomers entering a group over all 28 periods correlates at .97 with the total number of leavers. Within periods, the correlation between the number of leavers and the subsequent number of newcomers is .49, which gives us some room to examine whether the number of leavers also independently relates to the contribution percentage. In Figure C.2 and Table C.15-16 of the Supplementary Information, we show that when including both the number of newcomers and the number of leavers as predictors of the contribution percentage, it is mostly the number of newcomers that is associated with lower contribution percentages. The negative relationship between the number of newcomers and the contribution percentage is robust to different operationalizations of the incumbent/newcomer and contribution variables, analyses excluding outliers, analyses per game server, analyses incorporating crossed fixed effects that control simultaneously for group and player characteristics, analyses controlling for the public good level, and other model specifications (see Supplementary Tables C.4-C.14). We did not adjust p-values for multiple comparisons, but the relationships between group changes and cooperation would remain significant also when adjusting the cut-off p-values substantially downwards (e.g., when using $p < .001$ as the cut-off for significance instead of the conventional $p < .05$).

4.2.1. Differences in contribution behavior between incumbents and newcomers

We next address what role incumbents and newcomers play in the negative relationship between the number of newcomers and incumbents, i.e., do newcomers contribute less than incumbents and/or do incumbents condition their contributions on the number of newcomers? To do so, we first look at the difference in contribution behavior between newcomers and incumbents. Since players in Ikariam are part of multiple groups and will occupy both the roles of incumbent and newcomers, we can perform a within-player analysis to assess whether being a newcomer is indeed associated with lower contributions. Such an analysis reduces the number of confounding factors due to between-player differences. Figure 4.3a shows that players contribute considerably less as newcomers than as incumbents. Whereas players contribute 32% on average as incumbents, they only contribute about 14% as newcomers. Even when we control simultaneously for player and group characteristics in a crossed fixed effects model, we find that incumbents contribute more than newcomers (Supplementary Table C.13). Figure 4.3b shows that players increase their contribution percentage in the process of moving from newcomer to incumbent status. Whereas the contribution percentage is low when first entering a group, it increases steadily with more periods spent in the group. After ten periods in the group, the contribution percentage is 40% and remains mostly stable afterward. Thus, in the progress of moving from newcomer to incumbent status, players increase their contribution percentage.

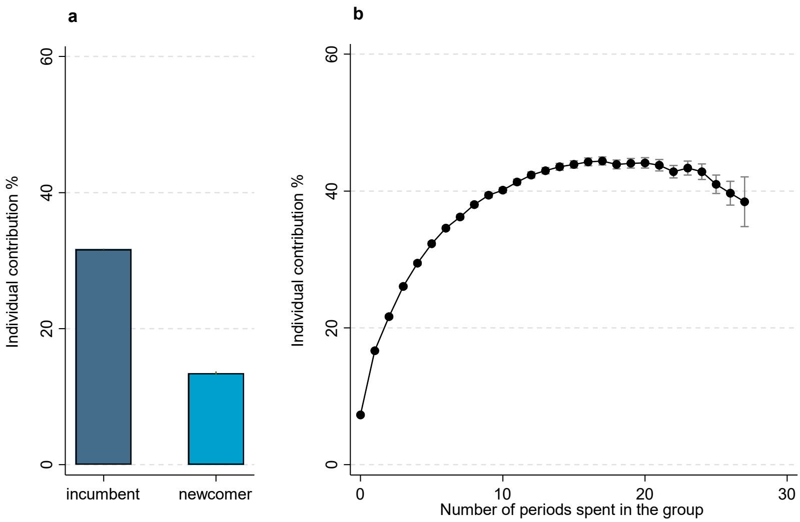


Figure 4.3. Individual contribution percentage by newcomer status (a) and time spent in the group (b).
Note: For (a), results are based on a player fixed effects regression with the individual's contribution percentage as the dependent variable and a factor on whether the individual is a newcomer or incumbent as the independent variable. We control for the time period and group size. For (b), results are based on a player fixed effects regression with the individual's contribution percentage as the dependent variable and a factor on the number of periods spent in the group as the independent variable. We control for the time period and group size. Data are presented as mean values and 95% confidence intervals are included via vertical spikes. Results for both panels include 1572734 contribution decisions made by 134631 players.

That players contribute considerably less as newcomers than as incumbents suggests that newcomers' contribution behaviors play a role in the negative relationship between the number of newcomers and the average contribution percentage. We can further assess the role of newcomers in this negative relationship by examining whether the relationship remains when subtracting newcomers' contributions in calculating the group-average contribution. That is, we examine the relationship between the number of newcomers and the incumbents' contribution percentage. Table 4.2 shows that the negative relationship is indeed strongly reduced. Whereas originally each additional newcomer was associated with a 2.72 lower average contribution percentage (Model 1 in Table 4.1), each additional newcomer is only associated with a 0.33 lower contribution percentage among incumbents (Table 4.2). Although this association is still significant, it is only a small fraction of its original size, which suggests that most of the negative relationship can be linked to the newcomers' (lack of) contributions.

Table 4.2. Regression model of incumbent contribution percentages with group fixed effects.

	Incumbents' contribution
Number of newcomers	-0.33***
	(0.05)
Group size	-0.72***
	(0.03)
Period	-0.43***
	(0.01)
Intercept	37.35***
	(0.08)
R^2 (overall)	0.03
Rho	0.61
N	187758

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients are marginal effects with standard errors provided in the parentheses. Statistical significance is calculated using two-sided t-tests. Results include 10940 groups (groups with no incumbents are excluded), with these groups existing on average for ~17 periods, giving a total number of observations of 187758 group-period combinations.

4.2.2. Mechanisms behind the newcomer-incumbent difference

The finding that players contribute less as newcomers than as incumbents but do contribute more the more time they spend in the group is in line with the theoretical mechanism that newcomers need time to become accustomed to and socialized with the prevailing contribution norm in the group. To further delve into this norm-based mechanism, we examine the extent to which a player's contribution percentage relates to the group's average contribution percentage (excluding the player's own contribution percentage) and how this develops with time spent in the group. An incumbent's contribution percentage correlates at .21 with the group's average contribution percentage. A newcomer's contribution percentage correlates at .13 with the group's average contribution percentage, significantly lower than the correlation for incumbents (Supplementary Table C.17). This suggests that newcomers indeed contribute less in accordance with the group contribution norm than incumbents. What is more, the relationship between a player's contribution percentage and the group's contribution percentage increases with time spent in the group (Supplementary Table C.17 and Figure C.4). That is, as players spend more time in the group, they contribute more in line with the group contribution norm. This is further in line with the norm-based mechanism specifying that newcomers need time to get accustomed to the prevailing contribution norm in the group.

We find little evidence that having a shared history with the incumbents in and of itself (i.e., without socialization to the group's contribution norm) is related to higher contributions. To assess this, we examine whether newcomers who already know the incumbents from their other groups contribute more than newcomers who do not know the incumbents from other groups. Although newcomers' contribution percentage is slightly higher if they know more incumbents from prior groups, the effect size is very small (0.7 percentage points, Supplementary Table C.18). The difference between newcomers and incumbents is also present when comparing players who are part of only one group and hence do not have to share their attention across groups (Supplementary Table C.19). We also find little evidence that newcomers use the contribution norm of their other groups to inform their contribution decision in new groups (Supplementary Table C.20). Instead, newcomers seem to start with low contribution percentages in their new groups and over time contribute more in line with the group's norm.

We further examine to what extent the difference between newcomers' and incumbents' contribution percentages is related to inequality between them in how much they can contribute to, and benefit from, the public good. The role of inequality in resources between incumbents and newcomers is largely prevented by examining not absolute contributions but instead the percentage of one's resources contributed. Newcomers can thus always achieve as high contribution percentages as incumbents, and not doing so is a choice. Still, having more resources might motivate one to contribute higher percentages, as having more resources increases the efficacy of one's contribution. Similarly, lower benefits might also lead newcomers to choose to contribute lower percentages.

We examine the role of inequality in resources and benefits in two ways. First, incumbents generally have higher-level private goods than newcomers. A player's private good level on an island is captured by the player's town hall level. Every increase in a player's town hall level increases the maximum number of citizens allowed in the player's town. Because citizens produce resources, an increase in the maximum number of citizens generally means an increase in resources. Hence, a higher town hall level means more resources. Because incumbents generally have higher private good (town hall) levels than newcomers, they have more resources to contribute to the public good and also benefit more from contributing because they are more in need of strong public goods to support their higher private good levels. Without controlling for private good level, we estimated the incumbents' contribution to be on average 18 percentage points higher than the newcomers' contribution (see Figure 4.3A). Controlling for the private good level decreases the difference in the contribution percentage between newcomers and incumbents from 18% to 9% (Supplementary Table C.21). Hence, while the newcomer-incumbent difference is halved when controlling for private good levels, it remains substantial and significant.

A second way to examine the role of inequality in the efficacy and benefits of contributing is by examining the difference between newcomers' and incumbents' contribution percentages depending on the public good level. If public good levels are low, it takes little resources to increase the public good level and the returns of increasing the public good level are high (see also Supplementary Information, Table C.2). In this situation, both newcomers and incumbents are in a position to effectively contribute to the public good and benefit from doing so. If public good levels are high, it takes more resources to increase the public good level and returns are lower. In this situation, incumbents are in a better position than newcomers to effectively contribute due to their higher amount of resources and benefit from doing so. Hence, if inequality in benefits and efficacy of contributing matters, we would expect the difference in contribution percentages between newcomers and incumbents to be lower with lower public good levels. Indeed, we find that the difference in the contribution percentage is lower with lower public good levels (Supplementary Table C.22). Whereas the average newcomer-incumbent difference in contribution percentage is about 18% (Figure 4.3A), the difference is about 9% at the lowest public good level (Supplementary Table C.22). Again, this shows that the newcomer-incumbent difference is smaller when incentives to contribute are similar for newcomers and incumbents, but also that it remains substantial and significant.

4.3. Discussion

Using large-scale data from a multiplayer online game that incorporates public good dilemmas, we find that changes in group composition relate negatively to contributions to the public good. This negative relationship holds both when looking at short-term group changes and when looking at the total number of group changes across the entire observed duration of the game. Although incumbents slightly decrease their contributions if newcomers enter, the negative relationship between group changes and contributions is linked mostly to players who enter the group as newcomers. That is, newcomers contribute considerably less than incumbents and thereby lower groups' average contribution percentages. However, as players spend more time in their new group, they increase their contributions to the public good.

The results are consistent with group socialization models suggesting that newcomers gradually increase their tendency to act in line with the group's welfare and norms when making the transition from outsider to insider (Bauer et al., 2007; Levine & Moreland, 1994). Indeed, individuals do not only contribute more if they are longer in the group, their contribution behavior also starts to resemble that of their group members more. Hence, individuals act more in line with the prevailing contribution norm as their time in the group increases. This norm-based mechanism seems to be more important than shared history explanations; newcomers who already know the incumbents from other groups hardly contribute more than newcomers who do not know the incumbents. Our findings also suggest that part of the newcomer-incumbent difference in contribution behavior is related to newcomers being in a disadvantaged position in terms of how effectively they can contribute to the public good and benefit from it.

The dynamic networks literature suggests that group changes can promote cooperation when they allow individuals to create ties with cooperative actors and break or avoid ties with uncooperative actors. However, not all situations allow individuals to determine how the composition of their group changes. For example, residents usually do not get to select who enters or leaves their neighborhood, and employees in many work organizations often do not get to form their own teams. Because exit costs are also typically substantial in these situations, incumbents cannot easily leave their group if they are dissatisfied with the newcomers. Similarly, incumbents in Ikariam have no say in who enters their group and, since exit costs are high, leaving the group is usually not an option. Hence, in contrast to most of the dynamic networks literature, strategically linking with cooperative actors and breaking with defectors is not a solution to cooperation in our study context. This may explain we find a negative relationship between group changes and cooperation instead of positive.

Our finding that newcomers' contributions are initially lower than those of incumbents but do increase over time resonates with field studies on newcomer contributions. Studies suggest that residents are more likely to volunteer at community events if they are longer part of the community (Ghimire & Skinner, 2019), that workers' output in organizations is higher with higher tenure (Shaw & Lazear, 2008), and that immigrants contribute more to charitable organizations with more time spent in the country (Bekkers & Wiepking, 2007). Typically, these studies only observe individuals at one point in time, so changes over time within individuals as they switch roles are not accounted for. In our study, the same individuals take both the roles of newcomer and incumbent, allowing us to compare within individuals how contribution behavior changes depending on one's role in the group. This is important for current debates on newcomer contributions, in which a common argument is that group changes have negative effects because newcomers have lower dispositions to contribute and are therefore expected to contribute structurally different from incumbents (Gereke et al., 2021). For example, immigrants are sometimes said to contribute less than native populations because they come from different nations with different levels of public good provision (Collier, 2013). Our study allows us to rule out individual dispositions and assess whether individuals condition their contributions on their role in the group. That we find a contribution difference between newcomers and incumbents while ruling out individual dispositions suggests that one's role in the group is a meaningful element in the difference between newcomers' and incumbents' contributions. Individuals' contribution percentages show a clear increase when roles switch from newcomer to incumbent.

What is more, the initially lower contributions by newcomers do not seem to be related to a motivation to take advantage of the incumbents, as is sometimes feared (Collier, 2013). Rather, newcomers initially have lower benefits and efficacy of contributing to public goods. With more time spent in the group, newcomers' benefits and efficacy of contributing increase and they increase their contributions to the incumbents' level. This finding on the role of inequality in benefits/efficacy of contributing is in line with prior theory (Hauser et al., 2019; Otten et al., 2020) and empirical research (Cherry et al., 2005; Hargreaves Heap et al., 2016) suggesting that inequality can hamper cooperation. Similar processes may play a role in explaining immigrant contributions to public goods. For example, cross-sectional research suggests that part of immigrants' lower contributions to public goods may be attributed to their lower education and income level (Bekkers & Wiepking, 2007). If immigrants have higher education and income levels, they are in a better position to contribute to public goods and also do so. The link between newcomers' disadvantaged position and their lower contribution implies that, rather than marginalizing newcomers' contributions or avoiding group changes, it is better to give newcomers the time to adjust and put them in a better position to effectively contribute to public goods (Frey & Rusch, 2012; Gächter et al., 2008; Hainmueller et al., 2016). Similarly, our finding of a negative relationship between changes in group composition and contributions to the public good should not be interpreted as evidence that change is bad. Changes in group compositions are unavoidable; incumbents will leave their groups at some point. Initial low contributions by a newcomer are better than no contributions at all if incumbents are not replaced. This becomes especially apparent if one considers the long-term contribution potential of newcomers whenever they are given the time to transition to an incumbent role. Our results simply suggest that to understand a group's current contribution to its public goods, it is informative to know its composition in terms of newcomers and incumbents.

Compared to typical research using public good games, our sample is broader and more heterogeneous. The inclusion of players from Germany, the United Kingdom, France, Greece, and Turkey means our sample goes slightly beyond typical WEIRD samples (Western, Educated, Industrialized, Rich, and Democratic) (Henrich et al., 2010). A survey reporting the average age of Ikariam players to be around 31 years (Frey, 2017) suggests that our sample more closely reflects the global median age than most other public good game studies which predominantly recruit younger university undergraduates (Ritchie & Roser, 2019). Whereas social dilemma studies are typically somewhat overrepresented by women (Spadaro et al., 2022), Ikariam is largely overrepresented by men (~80% men) as is common for computer games. We do not have access to data on the income or education of Ikariam players, so cannot establish representativeness in these aspects. Altogether, our sample presents an improvement in terms of representativeness in some areas (e.g., global coverage and age), but still has limited representativeness in other areas (e.g., sex).

Virtual worlds such as Ikariam present an exciting and growing opportunity to study cooperation in context-rich settings over longer periods of time and broader ranges of group compositions. These virtual worlds offer opportunities to unobtrusively track behavior of all individuals in an entire population within a constrained and well-understood environment. Lab experiments are still needed to draw causal inferences and field studies to bring external validity. However, virtual worlds provide an insightful addition to these more traditional research methods to together provide a fuller understanding of human cooperation.

4.4. Methods

4.4.1. Game context

Ikariam is a free online browser-based strategy game that has been played by more than 50 million individuals so far. Results from a survey suggest that about 80 percent of the players are male, with an average age of 31 years (Frey, 2017). The game is financed by some players paying real money to unlock in-game advantages such as obtaining more resources. Although players do not have a monetary incentive to act selfishly or cooperatively in the game, there are real incentives in terms of time investment and game progression. Because the game is played in real-time (i.e., the game continues after a player logs out), players typically have to log in multiple times per day to be sure that their towns keep running well and to respond to unforeseen events such as running out of resources. How quickly people progress in the game during this time is dependent, among other factors, on how selfishly or cooperatively they behave. Competition via leaderboards further incentivizes individuals to perform well.

The game is set in an ancient Greek archipelago and each island is regarded as a separate group tasked with producing public goods. The composition of the island will change over time as newcomers enter the island and incumbents leave the island, similar to how the composition of real-life groups producing public goods, such as countries, neighborhoods, and organizations, changes over time. A survey of Ikariam players confirms that they are well aware that contributing to the island is a cooperative act and that there are incentives to free-ride on others' contributions (Frey, 2017). The presence of the public goods problem in Ikariam is further indicated by the language used in the community. Ikariam players have a specific term for free-riding, namely leeching, in their online community pages. They can read about leeching on the wiki (Ikariam fandom, 2020), and can even find user-built tools to detect leechers (free-riders) in their group based on different contribution rules (*Ikariam Leecher Checker*, 2011).

The context of public good provision in Ikariam falls between the constrained setting of the lab and the unconstrained setting of the field. Compared to the lab, public good provision in Ikariam is context-rich, long-term, free of observer bias from experimenters' presence, and observed among a more diverse pool of individuals. These features make the context of Ikariam arguably more similar to field settings of public good provision (Henrich et al., 2010; Levitt & List, 2007; Zizzo, 2010). There is a growing body of research that examines whether cooperation patterns found in the lab also translate to field settings of public good provision (Kraft-Todd et al., 2015; Przepiorka & Berger, 2016; Smith et al., 2018). For example, the peer-production of Wikipedia has been analyzed and compared to public good provision in lab experiments (Gallus, 2017; Hergueux et al., 2015; Piskorski & Gorbatai, 2017). Such studies of public good provision in the field generally improve external validity, while lab studies remain important to provide a constrained environment that reduces the possibility of confounding variables. Compared to the complex environment of public good provision in the field, Ikariam consists of a highly standardized environment of which every aspect is recorded. Thus, in contrast to most field studies, we have information on the entire context in which public good provision takes place. All groups face the exact same public good problem and contribution behavior is therefore directly comparable across groups and individuals. Altogether, Ikariam provides a middle ground between a constrained lab environment and an unconstrained field setting.

4.4.2. Public good dynamics

There are two public goods per island, a sawmill producing wood and an island-specific good producing wine, marble, crystal glass, or sulfur. Both public goods work exactly the same, it is only the produced good that differs. For simplicity, we use the sawmill as an example when explaining the public good mechanism. In doing so, we refer to several existing variants of public goods games that bear resemblance to public good provision in Ikariam. Four specific characteristics of the public good in Ikariam are (a) it provides returns in real-time, (b) the returns increase step-wise by the total contribution of all group members, (c) the public good is durable, and (d) individuals differ in how much they can absolutely contribute.

The sawmill produces wood for each group member at a certain rate per hour, with the rate depending on how much wood has been contributed to the sawmill in total by all group members (players on the island). At the start of the game, when nobody has contributed yet, the sawmill produces 30 units of wood per hour for each group member. Hence, over 10 hours, a player would receive 300 units of wood. The hourly production rate can be increased if group members contribute wood to the public good. However, wood contributed to improve the sawmill cannot be used to develop one's town, giving an incentive to free-ride on the contribution of others. The public good increases step-wise as a function of the total contributions made to it by all group members. For example, the rate of 30 units of wood per hour can be increased to 38 units of wood per hour if all members combined contribute 394 units of wood to the public good. The public good is then said to have increased from level 1 (return of 30 units of wood per hour) to level 2 (return of 38 units of wood per hour). In total, there are 50 steps of improvement in succession. The step-wise increase in the benefits of a public good after total contributions surpass a threshold is commonly studied in lab experiments with step-level public goods games (Croson & Marks, 2000; Diekmann & Przepiorka, 2016). We provide the thresholds and the step-returns associated with these thresholds in the Supplementary Information (Supplementary Table C.2). The continuous-time flow of benefits from the public good is akin to continuous-time public goods games (Friedman & Oprea, 2012; Oprea et al., 2014).

The public good is durable: if a certain production rate per hour has been reached, it will never drop back to a lower rate. Relatedly, contributions are cumulative and not rebated. For example, if an individual contributes 500 out of 1000 units of wood required to move to the next public good level, the 500 remains in the sawmill even though the public good level (and hence the hourly production rate of wood) is unaffected. Only if another 500 units of wood are contributed to surpass the threshold, the hourly production rate of wood increases. See for prior empirical work on durable public goods games (Battaglini et al., 2016; Duffy et al., 2007). Finally, individuals that have accumulated more resources in the game can also contribute more. This is similar to the dynamic public goods game, where the wealth accumulated in prior rounds determines the endowment that can be contributed to the public good in subsequent rounds (Gächter et al., 2017). To compare contributions across individuals with different endowments, we examine not the individuals' absolute wood contributed, but instead the percentage of wood contributed out of the total wood that they had available on the island at the time of measurement.

Players can see the contributions of other players on the island at any time (an example is provided in Supplementary Table C.1). This observability of the contributions of all members is an important element that allows for cooperation norms to be at play via reciprocity, where individuals can condition their contributions on their group members' contributions (Ohtsuki & Iwasa, 2006; Santos et al., 2018). That contributions on each island are observable to all members allows us to examine whether players indeed contribute more when their group members also contribute more. A contribution of our study is that it also allows us to examine to what extent a player's tendency to contribute in line with the group differs between newcomers and incumbents, and whether newcomers contribute more in line with the group as they spend more time in the group.

Once an individual has entered a new group, it is generally not possible to leave, with two exceptions. The first is if an individual quits the game altogether. The second is if an individual spends real money to be able to move their town from one group to another, which is very rare. An individual can choose any group to enter, as long as the group has not reached the maximum group size of 17 yet. Incumbents thus have no say in who enters their group and cannot exclude members. When choosing which group to enter next, individuals have information on the island-specific resource that is produced, the location and size of the group, the incumbents in the group, and the current level of the public goods in the group. Generally, individuals will prefer groups that produce the island-specific resource that individuals are most in need of, groups that are located close to their current group(s), and groups with high public good levels.

There is the option to attack other players in Ikariam. Attacks are not publicly seen by others and can serve multiple purposes. For example, they can be used to take away resources from other players or to sanction those who do not contribute enough. However, attacks are very costly to both the attacker and the player defending the attack because both attacking and defending require armies that take up large amounts of resources. Attacks therefore only occur infrequently and are not central to the gameplay of most players. Since we do not have data on attacks and because they are not central to the gameplay, we do not focus on attacks in this study.

4.4.3. Data collection and analysis

The data were collected and provided by Gameforge, the creator of the game. Players of the Ikariam game provide consent for (third-party) analyses of their non-personal data when signing up for the game. We did not have any access to personal data and obtained ethical approval for the study protocol from the Faculty Ethics Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University. We have data from five servers, each from a different country: Germany, the United Kingdom, France, Greece, and Turkey. Each server contains a fixed number of 5351 islands, but the number of players differs per server. Data collection is identical and synchronous for each country. Given that there are no large differences by country (Frey, 2019), we pool the data across countries. The data is structured in 28 biweekly snapshots between April 2013 and February 2014 (the snapshots are biweekly on average; they start out weekly in April, get biweekly in August, and still later it is a four-week interval). The first snapshot coincides with the start of new game servers, so we begin our observation at the actual beginning of a game. The data were analysed using Stata MP 15.1. The raw data is the property of Gameforge. Gameforge allowed the authors to use the raw data for academic purposes, but did not allow the raw data to be openly shared. We provide an aggregated dataset at https://doi.org/10.17605/OSF.IO/3WYE9.

Since group size ranges from 1 to 17 players per group, and each player can either be a newcomer or an incumbent at each time period, we can observe many different combinations of the number of incumbents and newcomers. This allows giving a comprehensive answer to our first research question on how changes in group composition relate to contributions to the public good. We take advantage of the longitudinal nature of the data by examining this relationship within groups, which reduces potential confounding by between-group differences.

Furthermore, because players in Ikariam are part of multiple groups and experience both roles of newcomer and incumbent, we can examine differences in contribution behavior between newcomers and incumbents within players. By analyzing whether the same player contributes differently depending on whether the player is an incumbent or newcomer in the group, we can exclude selection effects based on different personal characteristics and assess whether just one's role in the group already relates to contribution behavior. This allows us to rule out individual disposition when answering the second research question concerning differences in contribution behavior between newcomers and incumbents, which is typically not possible in related prior research. For example, when finding differences in contributions to national public goods between migrants and native populations, it is typically not possible to pinpoint whether these differences arise from the role that one has in the group (newcomer vs incumbent) versus selection effects (migrants having different individual dispositions than native populations) (Baldassarri & Abascal, 2020).

Recall that there are two public goods that individuals can contribute to in each group (the sawmill and the island-specific good). Resources that an individual contributes to one of the two public goods cannot be contributed to the other public good. The contributions to the two public goods are added up and divided by the total resources available at the time of measurement to obtain an individual’s contribution percentage. Analyses that examine each public good separately are provided in the Supplementary Information and show no substantial differences between the two (Supplementary Table C.7-9).

Because individuals can move resources between their groups, it can happen that they contribute more to the public good of a group than the total resources they had available in that group, i.e., individuals can end up with contribution percentages above 100 percent. Likewise, because we only have snapshots of an individual’s available resources instead of a continuous-time overview of an individual’s resources, it is possible that an individual had more (or fewer) resources available than we see at the snapshot, which can also lead to contribution percentages above 100 percent. This happens in 6.5% of our analyzed cases. In the Supplementary Information (Table C.11), we show that the negative relationship between the contribution percentage and the number of newcomers remains significant when leaving out these cases or setting them to 100 percent.

Chapter 5. Cooperation, punishment, and group change in multilevel public goods experiments[1]

Abstract
Peer punishment is regarded as an important element in sustaining human cooperation for public good provision. Many behavioral experiments have shown that public good provision is higher if cooperation norms can be enforced by peer punishment. However, these experiments predominantly focus on single-group public goods, in which people have to choose between their private interests and the interests of their group. In many societal problems, people are involved in multilevel public goods problems, where multiple local groups are nested within a larger global group. We study experimentally how punishment affects cooperation and norms in multilevel public goods games. In our games, two local groups are nested within a larger global group. Participants have to choose between not contributing, contributing locally, and contributing globally. Local contributions would lead to a polarized outcome where two separate local public goods are provided, whereas global contributions would lead to a unified global good that benefits all. Moreover, we study whether cooperation and punishment patterns depend on the type of public good participants are initially exposed to: single-group or multilevel. Participants either begin in a single-group public goods game and then shift to a multilevel public goods game or vice versa. We find that punishment is less effective in multilevel public goods games than in single-group public goods games. In particular, punishment only promotes cooperation in multilevel public goods games if people have prior experience with solving single-group public goods games. Our results refine the boundary conditions for the effectiveness of punishment and suggest that 'starting small' by first solving single-group public goods problems is necessary for successful multilevel public good provision.

[1] This chapter is based on a paper written by Kasper Otten, Vincent Buskens, Wojtek Przepiorka, Boaz Cherki, and Salomon Israel. The paper has been submitted to an international peer-reviewed journal. Otten wrote the manuscript, developed and executed the experiment, and did the analysis. All authors contributed to experiment development and manuscript writing. The study is preregistered at doi.org/10.17605/OSF.IO/5KAWT. All data and code are openly available at doi.org/10.17605/OSF.IO/2DKGP.

5.1. Introduction

Human cooperation to provide public goods is key for the success of social groups ranging from families and work teams to nations and international organizations. Cooperation for public good provision often presents a social dilemma – contributing is individually costly but brings benefits for the group. This means that there are incentives to free-ride on others' contributions, but the public good is not provided if everybody free-rides (Olson, 1965). Our ancestors had to overcome such free-rider incentives when they hunted large animals, shared food, or engaged in warfare (Bowles & Gintis, 2013; Hawkes et al., 1993; Hill, 2002). Contemporary humans face free-rider incentives in solving team production tasks, when paying taxes, or when taking measures to reduce their carbon footprint (Kallhoff, 2014). How people cooperate despite the incentives to free-ride is still a major scientific question. Peer punishment is often suggested as a solution to achieving cooperation for public good provision (Balafoutas & Nikiforakis, 2012; Balafoutas et al., 2014; Fehr & Gächter, 2002; Fehr & Schurtenberger, 2018; Henrich et al., 2006). If cooperation norms can be enforced through punishment, free-riding is discouraged and cooperation becomes more likely. A large number of behavioral experiments show that giving people the opportunity for peer punishment indeed promotes cooperation (Balliet et al., 2011; Chaudhuri, 2011) and group welfare in the long run (Gächter et al., 2008).

Previous experiments on the effect of punishment on cooperation typically focus on single-group public goods problems, where people have to choose between not cooperating or cooperating with their own group. However, in many real-life instances of public good provision, there are multiple local groups nested within a larger global group. Such public goods have been labeled multilevel public goods (sometimes also referred to as nested social dilemmas) (Aaldering & Böhm, 2020; Aaldering et al., 2018; Blackwell & McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner & Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009; Polzer et al., 1999; Wit & Kerr, 2002). In multilevel public goods problems, individuals have to choose to what extent they act in their private interests, cooperate with their own local group (local cooperation), and cooperate with the larger global group (global cooperation). For example, employees are clustered in teams that are themselves clustered in departments or organizations and have to choose their effort for each level. Because multiple group memberships are increasingly the rule rather than the exception, interest in multilevel public goods has been growing (Aaldering et al., 2018; Aaldering & Böhm, 2020; Buchan et al., 2009, 2011; Espinosa et al., 2019). However, we do not yet know whether and how peer punishment promotes cooperation in multilevel public goods problems. The first aim of this study is to experimentally test the effect of peer punishment on cooperation in multilevel public goods problems.

Examples of multilevel public goods problems can be found in several domains. In charitable giving, people need to decide how much to donate and whether to donate to local charities or global charities. For national public goods such as social benefits, native populations need to decide whether and how they grant access to other groups such as immigrants (Degen et al., 2019). Scholars have also suggested that pressing international issues such as tackling the climate crisis and the COVID-19 pandemic involve elements of multilevel public goods problems (Buchan et al., 2011, 2009; Romano et al., 2021; Tavoni et al., 2011).

In these examples, global cooperation can be impeded by individuals' tendency to pursue local interests or private interests. In work organizations, merging failures have been attributed to difficulties that employees face in collaborating across departments or divisions (Weber & Camerer, 2003). In charitable giving, people generally donate more to domestic causes than global causes, even when donations to global causes can do a lot more good (Grimson et al., 2020). In two-party political systems, parties sometimes pursue forms of public good provision that benefit their own supporters instead of the general public, leading to polarization and suboptimal policies (Dimant, 2022; Schultz, 1996). During the COVID-19 pandemic, several countries bought up a disproportionally large share of vaccines for their own populations, even though a more equitable global distribution would have been more effective to limit virus mutations that evade the vaccines (Ye et al., 2022). Finally, in combatting climate change, countries regularly take actions that reduce the negative environmental consequences for their local population but not for the global population (e.g., toxic waste trading) (Cotta, 2020).

Prior experiments on multilevel public goods problems show that also in the lab, global cooperation is often impeded by a tendency for local cooperation, even when global cooperation is collectively more beneficial (Aaldering & Böhm, 2020; Aaldering et al., 2018; Blackwell & McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner & Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit & Kerr, 2002). These experiments employ the multilevel public goods game (sometimes also referred to as a nested social dilemma), in which usually two local groups are nested within a larger global group. Participants then have to decide whether they want to not contribute at all, contribute to their local group, or contribute to a global good that benefits all. The multilevel public goods game is considered a central paradigm to study behavioral group polarization (Van Dijk & De Dreu, 2021). If people contribute locally instead of globally, a polarized outcome ensues in which two separate local public goods are provided instead of a unified global good. The extent to which people prefer to cooperate locally or globally depends on the returns to both types of cooperation (Blackwell & McKee, 2003; Böhm et al., 2014; Fellner & Lünser, 2014; Gallier et al., 2019), the observability of both types of cooperation (Fellner & Lünser, 2014), resource inequality (Lange et al., 2022), framing (Polzer et al., 1999), and social categorization (Wit & Kerr, 2002). Similar to single-group public goods experiments, cooperation decays over time in multilevel public goods experiments, and the decay is especially noticeable for global cooperation (Blackwell & McKee, 2003; Fellner & Lünser, 2014). The well-known solution to prevent cooperation decay in single-group public goods experiments is peer punishment (Balliet et al., 2011; Chaudhuri, 2011), but this solution has not yet been tested for multilevel public goods experiments.

Theoretically, punishment may be less effective in promoting cooperation in multilevel public goods problems than in single-group public goods problems. The reason is that the potential for normative disagreement is larger in multilevel public goods problems. In single-group public goods problems, the only way to cooperate is to contribute to the group and most agree that this is the appropriate thing to do (Cubitt et al., 2011; Kimbrough & Vostroknutov, 2016; Reuben & Riedl, 2013). Hence, the cooperation norm is clear, which facilitates the use of peer punishment to effectively enforce norms. Indeed, previous experiments suggest that punishment is only socially beneficial if complemented by strong cooperation norms (Bicchieri et al., 2021; Herrmann et al., 2008). In multilevel public goods problems, there are several plausible cooperation norms; local cooperation, global cooperation, or a mix between the two (Catola et al., 2021). Consequently, there is more ambiguity about the cooperation norm, creating the potential for disagreement (Nikiforakis et al., 2012; Otten et al., 2020; Rauhut & Winter, 2017; Winter et al., 2012; Winter et al., 2018; Wit & Kerr, 2002). Individuals who do not cooperate and are subsequently punished may not always know whether the punishment is meant to induce them to cooperate locally, globally, or a mix between the two. As a result, they may not react with the type of cooperation that the punisher intended to instigate. Further punishment or a decay in cooperation in general may be the result. Individuals who cooperate at a certain level (e.g., locally or globally) because they believe this to be appropriate may be reluctant to change their behavior when being punished (Rauhut & Winter, 2017). In sum, punishment may be less effective in enforcing cooperation norms when the norm that is to be enforced is more ambiguous or disputed.

Even if punishment promotes cooperation in multilevel public goods problems, there is no guarantee that it leads to a collectively efficient form of cooperation. In multilevel public goods problems as strictly defined, global cooperation is more efficient than local cooperation, which in turn is more efficient than no cooperation at all (Blackwell & McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Gallier et al., 2019; Lange et al., 2022; Wit & Kerr, 2002). Hence, if punishment promotes local cooperation over global cooperation, it leads to a collectively inefficient and polarized form of cooperation (in which two separate local goods are provided instead of a unified global good). As mentioned, previous experiments suggest that people indeed tend to cooperate more with ingroup members than outgroup members (Aaldering & Böhm, 2020; Aaldering et al., 2018; Balliet et al., 2014; Blackwell & McKee, 2003; Böhm et al., 2014; Buchan et al., 2009; Fellner & Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit & Kerr, 2002). Moreover, evidence from intergroup experiments suggests that punishment benefits ingroup members more than outgroup members (Bernhard et al., 2006). Based on prior research showing that punishment promotes cooperation (Balliet et al., 2011; Chaudhuri, 2011), we preregistered the hypothesis that total cooperation (combining local and global cooperation) in multilevel public goods problems is higher when punishment is possible. Exploratively, we examine what type of cooperation, if any, is increased by punishment (local versus global). We also assess whether the effect of punishment in multilevel public goods problems is different from that in single-group public goods problems.

Previous multilevel public goods experiments impose the multilevel structure from the start (Aaldering & Böhm, 2020; Aaldering et al., 2018; Blackwell & McKee, 2003; Böhm et al., 2014; Buchan et al., 2011, 2009; Espinosa et al., 2019; Fellner & Lünser, 2014; Gallier et al., 2019; Israel et al., 2012; Lange et al., 2022; Polzer et al., 2009, 1999; Wit & Kerr, 2002). However, in society, multilevel structures often arise over time from membership changes between local groups. That is, people start in local groups with single-group public goods problems but over time end up in a multilevel public goods problem when members from other local groups enter the situation. For example, many nations have become more ethnically diverse through immigration over the past several decades, increasing the need for collective multicultural cooperation. Work organizations may also become more multilevel over time through changes in their hierarchical structure and mergers. Whether the multilevel public goods problem is present from the start or arises after an initial single-group public goods problem may be crucial for the cooperation norms that emerge and hence the norms that are enforced through peer punishment. Groups that were initially in a single-group public goods problem may have developed norms of local public goods provision (Bernhard et al., 2006; Choi et al., 2019, 2021; Otten et al., 2022; Titlestad et al., 2019). The presence of these norms of local public good provision may hamper global public good provision when the multilevel public goods problem arises. That is, individuals may stick to their norms of local public good provision even when they enter a multilevel public goods problem in which global cooperation is collectively more beneficial. Previous experiments indeed suggest that norms are sticky and spill over to new settings (Andreoni et al., 2021; Duffy & Lafky, 2021; Efferson & Vogt, 2018; Engl et al., 2021; Guala & Mittone, 2010; Otten et al., 2021; Peysakhovich & Rand, 2013; Przepiorka et al., 2022; Smerdon et al., 2019). In contrast, people who are in a multilevel public goods problem from the start may not be constrained by prior norms of local public good provision and therefore more easily coordinate on global cooperation. We therefore preregistered the hypothesis that global cooperation will be higher when the multilevel public goods problem is present from the start than when it arises after an initial single-group public goods problem.

5.1.1. Experiment

We conduct an incentivized experiment with 220 participants. Using a repeated public goods game (PGG) (Fehr & Gächter, 2000), we vary whether participants can use peer punishment to enforce norms and whether they move from a single-group PGG to a multilevel PGG or vice versa. The experimental design is presented in Figure 5.1. At the start of the experiment, each participant is randomly assigned a specific local group. Participants are then placed into groups of four members facing either a single-group or a multilevel PGG. In single-group PGGs, all four members belong to the same local group. In multilevel PGGs, the group of four consists of two local groups with two members each. In each round of the game, participants receive an endowment of 20 monetary units (MU) and have to allocate this endowment between a private account that benefits only themselves, a local account that benefits only members of their own local group (local cooperation), and a global account that benefits all four members regardless of which local group they belong to (global cooperation). Individual marginal returns are highest for the private account (1), followed by the local account (0.7), and then the global account (0.5).

In the single-group PGG (where all four members belong to the same local group), it is best for the individual to not contribute at all whereas it is best for the collective (all four members combined) to contribute to the local account. This can be seen from a numerical example. For every participant who keeps the 20 MU, they get 20 MU themselves, while all other members of the group receive nothing (collective payoff = 20 MU). For every participant who contributes the 20 MU to the local good, they get $20 \times 0.7 = 14$ MU themselves, and the three other members of the local group also each receive 14 MU (collective payoff = 56 MU). For every participant who contributes the 20 MU to the global account, they get $20 \times 0.5 = 10$ MU themselves, and the three other members of the local group also each receive 10 MU (collective payoff = 40 MU). Hence, not contributing is best for each individual participant (20 MU vs 14 when contributing locally vs 10 when contributing globally), but contributing locally is the best for the collective (56 MU vs 40 MU when contributing globally vs 20 MU when not contributing at all).

In the single-group PGG, the dilemma is thus between not contributing and contributing to the local account. Despite global contributions being neither individually nor collectively efficient in single-group PGGs, we still provide participants the option to contribute globally. This way, we do not create artificial differences between the single-group and multilevel PGGs in terms of the complexity of the decision situation. However, as we will see, virtually nobody contributes globally in single-group PGGs. The tension between not contributing, contributing locally, *and* contributing globally only arises when multiple local groups are nested within a global group. This is what we capture in the multilevel PGG, where the group of four members consists of two local groups with 2 members each.

In the multilevel PGG, it is best for the individual to not contribute at all whereas it is best for the collective (all four members combined) to contribute to the global account. This can again be seen from a numerical example. For every participant who keeps the 20 MU, they get 20 MU themselves, while all other members of the group receive nothing (collective payoff = 20 MU). For every participant who contributes the 20 MU to the local account, they get 20 × 0.7 = 14 MU themselves, the other member of the same local group also receives 14 MU, but the two members from the other local group receive nothing (collective payoff = 28 MU). For every participant who contributes the 20 MU to the global account, they get 20 × 0.5 = 10 MU themselves, and all three other members of the group (both from the same local group as from the other local group) also each receive 10 MU (collective payoff = 40 MU). So the individual payoff is highest when not contributing, followed by contributing locally, followed by contributing globally (20 vs 14 vs 10 MU). The collective payoffs follow the exact opposite order; they are highest when contributing globally, followed by contributing locally, followed by not contributing (40 MU vs 28 MU vs 20 MU). Finally, the payoffs of one's local group member are highest when contributing locally followed by contributing globally and not contributing (14 MU vs 10 MU vs 0 MU).

Participants first play 10 rounds of either a single-group PGG (all four members of the same local group, Figures 5.1A and 5.1B) or a multilevel PGG (two local groups with two members each, Figures 5.1C and 5.1D) within their group. After these 10 rounds, we replace two members per group with two members of another group in such a way that single-group PGGs become multilevel PGGs and vice versa. After this group change, the reshaped groups play another set of 10 rounds. Because we vary whether groups start with a single-group or multilevel PGG and whether punishment is possible, we have 2 × 2 = 4 conditions. We refer to the conditions in which people move from single-group PGGs to multilevel level PGGs as *single-group to multilevel* (Figures 5.1A and 5.1B), and the conditions with the opposite order as *multilevel to single-group* (Figures 5.1C and 5.1D). In the punishment conditions (Figures 5.1B and 5.1D), participants can assign each other punishment points after each contribution decision. In line with most previous research on punishment (Fehr & Gächter, 2000; Reuben & Riedl, 2013), each punishment point costs the punisher 1 MU and the punished person 3 MU, i.e., punishment is costly for the punisher but even more so for the punished person. For example, if a participant assigns another member two punishment points, the participant loses 2 MU herself and the punished group member loses 2 × 3 = 6 MU. Because our theory and hypotheses rest on the idea that punishment is a means of norm enforcement, we also measure normative expectations in round 10 of the single-group and multilevel PGG. This allows us to examine to what extent cooperation patterns are reflected in norms and to what extent norms and normative disagreement differ between single-group and multilevel PGGs. More details on the experiment are available in the Methods section.

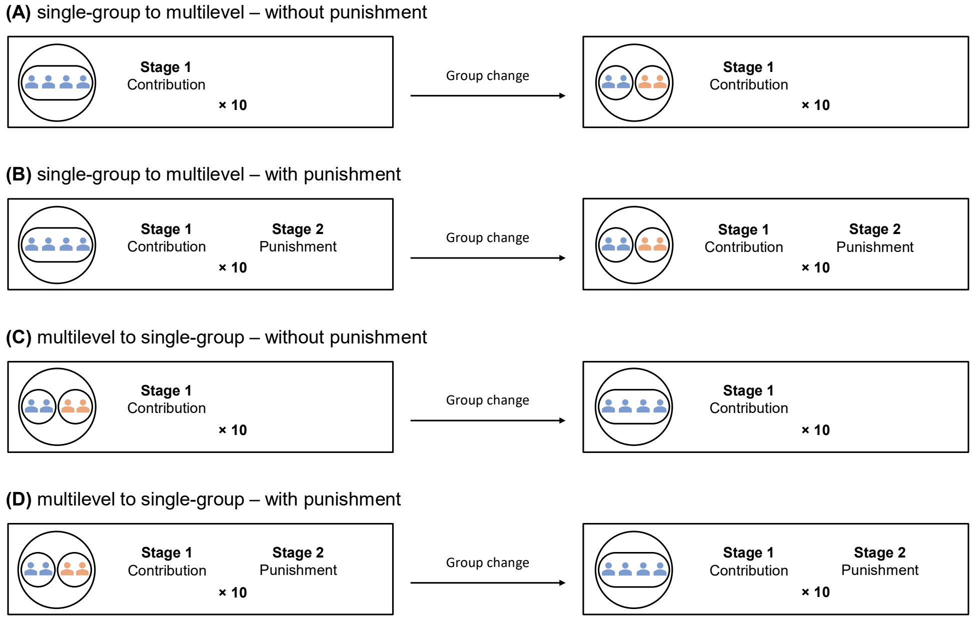


Figure 5.1. Experimental design.
Note: Participants either start with a single-group public goods game (PGG) and after group change move to a multilevel PGG (single-group to multilevel) or vice versa (multilevel to single-group). We additionally manipulate whether peer punishment is possible. In both the single-group and multilevel PGG, participants play 10 rounds within their group of four members. In each round, participants receive an endowment of 20 monetary units and have to allocate this endowment between a private account that benefits only themselves, a local account that benefits only members of their own local group, and a global account that benefits all members regardless of which local group they belong to. Individual marginal returns are highest for the private account (1), followed by the local account (0.7), and then the global account (0.5). The accounts and conversion rates are available at all times.

5.2. Results

5.2.1. Cooperation in single-group public goods games

Figure 5.2 shows the cooperation levels in single-group PGGs for all four conditions (where all four members belong to the same local group). As mentioned, the collectively efficient form of cooperation in single-group PGGs is local cooperation, and we see that virtually all cooperation is indeed local cooperation. This suggests that participants have little difficulty coordinating on the efficient way of cooperating in single-group PGGs. Consistent with previous research, we find that cooperation in single-group PGGs is higher when peer punishment is possible (3.40 MU higher in condition *single-group to multilevel*, $p < .001$, Figure 5.2B vs 5.2A; 3.75 MU higher in condition *multilevel to single-group*, $p < .001$, Figure 5.2D vs 5.2C; see Supplementary Material Table D.1).

Further in line with previous research, we find that there is a decaying trend in cooperation without punishment (Figures 5.2A and 5.2C, Supplementary Material Table D.4). An exception to this trend is the last round. Norms were elicited just before this round, and prior research shows that this can make norms temporarily more salient (D'Adda et al., 2016; Krupka & Weber, 2009), thereby promoting cooperation in the last round. If punishment is possible, cooperation is sustained at a high level throughout all rounds (Figures 5.2B and 5.2D). The effect of punishment on cooperation in single-group PGGs does not depend on the order of the problem (*single-group to multilevel* vs. *multilevel to single-group*; the difference in effect is 0.35 MU, $p = .78$; derived from Table D.1). In sum, punishment promotes cooperation in single-group PGGs regardless of whether the single-group PGG comes before or after the multilevel PGG.

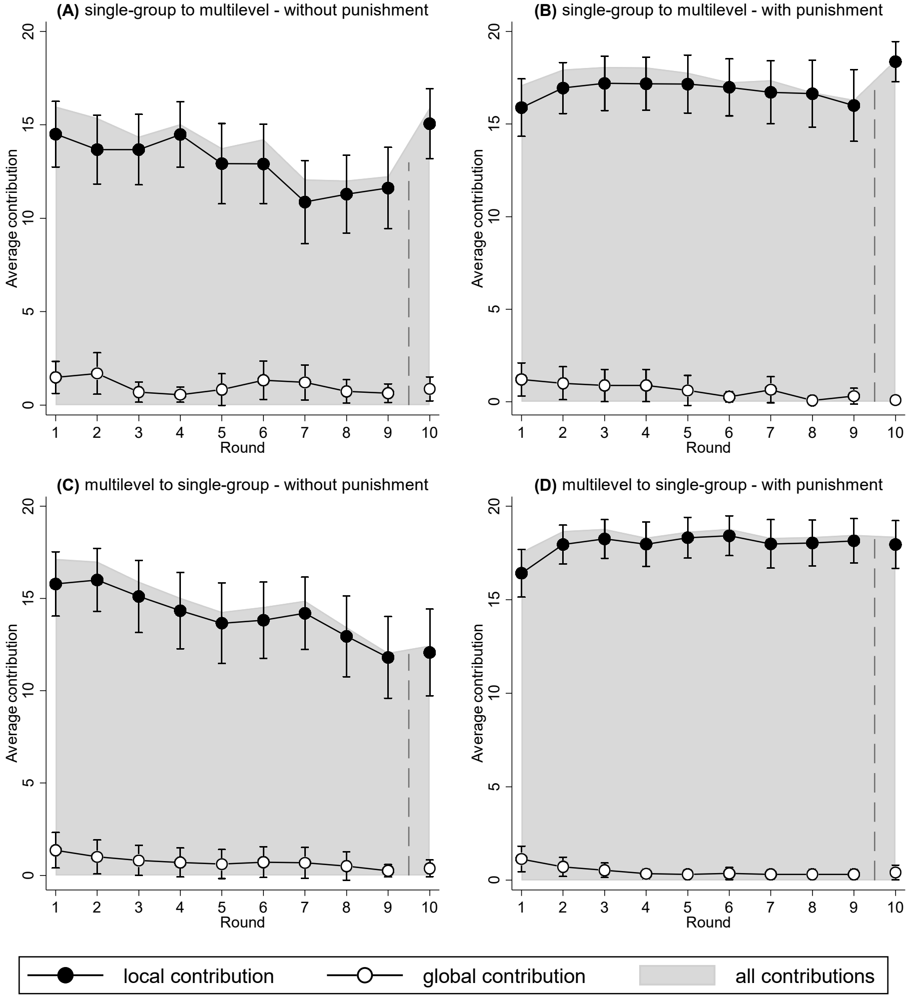


Figure 5.2. Cooperation in single-group public goods games.

Note: We present the 10 contribution decisions in the single-group public goods game for each condition. Mean contributions are indicated via markers and 95% confidence intervals via capped spikes. The vertical dashed line between rounds 9 and 10 indicates that norm measurements occurred between these rounds. We find that virtually all cooperation is local cooperation and that local cooperation is higher if peer punishment is possible (panels B vs A, and D vs C).

5.2.2. Cooperation in multilevel public goods games

Figure 5.3 shows the cooperation levels in multilevel PGGs for all four conditions (where groups of four members consist of two local groups with two members each). In the conditions where the multilevel PGG comes before the single-group PGG (Figures 5.3C and 5.3D), we find a roughly equal mix of local and global cooperation. Punishment does not promote cooperation in these conditions; the levels of local and global cooperation are not significantly different between the non-punishment and punishment conditions (average local contributions of 7.20 MU vs. 8.43 MU, $p = .17$, Table D.2; average global contributions of 8.25 MU vs. 8.45 MU, $p = .85$, Table D.3; average total contributions of 15.45 MU vs. 16.88, $p = .05$, Table D.1). This means that we do not find support for the hypothesized positive effect of punishment on total cooperation in multilevel PGGs that come before the single-group PGG.

The pattern is strikingly different in the conditions where a single-group PGG was encountered before the multilevel PGG (Figures 5.3A and 5.3B). Here, we see that global cooperation is higher than local cooperation. Although this difference between global cooperation and local cooperation is also present without punishment (8.31 MU vs. 4.94 MU, $p = .001$, Table D.7), it is especially remarkable when punishment is possible (14.71 MU vs. 3.38 MU, $p < .001$, Table D.7). Punishment strongly promotes global cooperation in these conditions (6.41 MU, $p < .001$, Table D.3), and only slightly lowers local cooperation (-1.56 MU, $p = .04$, Table D.2). Consequently, total cooperation is promoted by punishment (4.85 MU, $p < .001$, Table D.1). This means that we do find support for the hypothesized positive effect of punishment on total cooperation in multilevel PGGs that come after a single-group PGG.

When punishment is not possible, we do not find a difference in global cooperation between the conditions where a single-group PGG comes before the multilevel PGG and conditions with the opposite order (8.31 MU vs 8.25 MU, $p = .96$, Table D.7, and see Figures 5.3A vs 5.3C). When punishment is possible, global cooperation is much higher in the condition where a single-group PGG comes before the multilevel PGG than in the condition with the opposite order (14.71 vs. 8.45 MU, $p < .001$, Table D.7, and see Figures 5.3B vs 5.3D). This runs counter to our hypothesis that global cooperation will be higher when the multilevel public goods problem is present from the start than when it arises after an initial single-group public goods problem.

With regard to time trends, we see that in conditions *multilevel to single-group* (with and without punishment, Figures 5.3C and 5.3D and Tables D.5 and D.6), global contributions are being replaced by local contributions over time. Local contributions significantly increase over time (Table D.5) while global contributions significantly decrease over time (Table D.6). This means that contributions are increasingly separated into two subgroup-specific goods instead of a unified global good, indicating a trend of group polarization. In the condition *single-group to multilevel* without punishment, local cooperation is stable over time (Table D.5 and Figure 5.3A) but global cooperation still decreases over time (Table D.6 and Figure 5.3A). Only in the condition *single-group to multilevel* with punishment is global cooperation stable and high (Table D.6 and Figure 5.3B). This suggests that starting with single-group PGGs before turning to multilevel PGGs in combination with punishment possibilities helps to avoid group polarization. Note that, again, we find an exception to these trends in the last round because norms were elicited just before this round (see Methods for details).

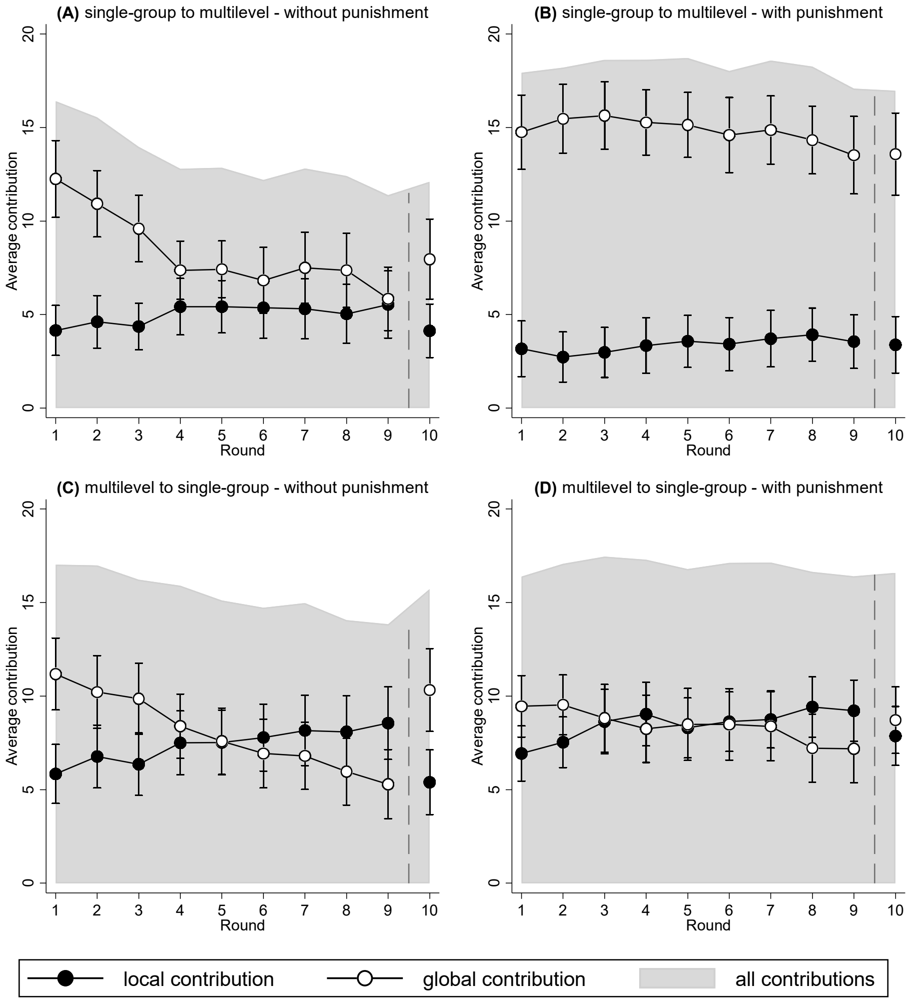


Figure 5.3. Cooperation in multilevel public goods games.
Note: We present the 10 contribution decisions in the multilevel public goods game for each condition. Mean contributions are indicated via markers and 95% confidence intervals via capped spikes. The vertical dashed line between rounds 9 and 10 indicates that norm measurements occurred between these rounds. We find that punishment increases global cooperation when participants encountered a single-group public goods game before the multilevel public goods game (panel B vs A). Punishment does not increase global cooperation when participants encountered the multilevel public goods game before the single-group public goods game (panel D vs C).

5.2.3. Punishment patterns

Recall that we suggested that punishment may be less effective in promoting cooperation in multilevel public goods problems than in single-group public goods problems because the potential for normative disagreement is higher in multilevel public goods problems. Prior research suggests that normative disagreement is reflected in more punishment (Nikiforakis et al., 2012; Rauhut & Winter, 2017; Winter et al., 2012, 2018). Hence, if normative disagreement is higher in multilevel public goods problems, we would expect more punishment in multilevel PGGs. Indeed, we find that punishment occurs more often in multilevel PGGs than in single-group PGGs (participants in multilevel PGGs punish at least one group member 29.5% of the time compared to 16.0% in single-group PGGs, $p < .001$, Table D.8). However, this difference is only significant in the condition where the multilevel PGG comes before the single-group PGG (36.8% vs 16.3%, $p < .001$, Table D.8). In the condition where a single-group PGG comes before the multilevel PGG, punishment is not significantly higher in the multilevel PGG than in the single-group PGG (21.0% vs 15.6%, $p = .19$, Table D.8). This suggests that initial experience in single-group PGGs may help to prevent normative disagreement in subsequent multilevel PGGs and reduce the extent of punishment necessary to maintain cooperation.

Recall further that we expected groups that start with a single-group PGG to develop norms of local contributions, which then potentially carry over to subsequent multilevel PGGs and thereby lead to local instead of global contributions. Groups that start with a multilevel PGG are not constrained by these norms of local contributions that developed in the single-group PGG and may therefore more easily achieve global contributions. Because punishment is regarded as a means of norm enforcement, we can examine which contribution behaviors are punished to get a first indication of which norms are present in single-group and multilevel PGGs and whether these depend on the order of the problem. In Figure 5.4, we show what type of contribution behavior is punished in single-group and multilevel PGGs. In single-group PGGs (Figures 5.4A and 5.4B), punishment is directed at low contributions to local goods and high contributions to global goods. Participants contributing nothing to local goods receive 2.11 punishment points whereas those contributing fully to local goods receive 0.14 punishment points (difference is 1.97 punishment points, $p < .001$, derived from Table D.9). Participants contributing nothing to global goods receive 0.33 punishment points whereas those contributing fully to global goods receive 2.14 punishment points (the difference is 1.81 punishment points, $p < .001$, derived from Table D.9). These punishment patterns do not differ significantly between the condition where the single-group PGG comes before the multilevel PGG (Figure 5.4A) and the condition with the opposite order (Figure 5.4B; Table D.9). The punishment patterns are in line with a norm of contributing to local goods in single-group PGGs.

In multilevel PGGs (Figures 5.4C and 5.4D), punishment is directed at high contributions to local goods and low contributions to global goods, but mostly in the condition where a single-group PGG comes before the multilevel PGG (Figure 5.4C; Table D.9). In this condition, participants contributing nothing to local goods receive 0.40 punishment points whereas those contributing fully to local goods receive 1.53 punishment points (the difference is 1.13 punishment points, $p < .001$, derived from Table D.9). Participants contributing nothing to global goods receive 2.36 punishment points whereas those contributing fully to global goods receive no punishment points (the difference is significant at $p < .001$, derived from

Table D.9). In the condition with the opposite order (Figure 5.4D), punishment is relatively independent of how people contribute, but a tendency for punishment of low contributions to global goods is also discernable. In this condition, contributions to local goods are not significantly related to received punishment (null-contributors receive 0.99 punishment points compared to 1.04 punishment points for full contributors, p = .85, derived from Table D.9). Contributing to global goods is still associated with lower received punishment, but the association is significantly smaller than in the condition where a single-group PGG comes before the multilevel PGG (Table D.9). These punishment patterns are in line with a norm of contributing to global goods in multilevel PGGs, in particular when a single-group PGG was encountered before the multilevel PGG. Hence, in contrast to our expectations, groups with prior experience in single-group PGGs seem to more easily achieve a norm of global cooperation in multilevel PGGs than groups that start with the multilevel PGG.

In Supplementary Material Figure D.1, we show how people respond to received punishment depending on their contribution behavior. In single-group PGGs, low local contributors typically respond with higher local contributions, in particular in the condition where a multilevel PGG comes before the single-group PGG. High local contributors typically do not change their behavior when being punished, and neither do the few low or high global contributors (global contributions in single-group PGGs may reflect a lack of understanding the game, which may also explain why global contributors do not respond to punishment). In multilevel PGGs, reactions to punishment strongly depend on whether a single-group PGG comes before the multilevel PGG or vice versa. In the condition where a single-group PGG comes before the multilevel PGG, low global contributors react with higher global contributions, high local contributors react with lower local contributions, and high global contributors or low local contributors do not react to punishment. These reactions are in line with a norm of global cooperation in multilevel PGGs. In the condition where the multilevel PGG comes before the single-group PGG, low global contributors also react with higher global contributions, but low and high local contributors do not react to punishment, and high global contributors respond to punishment with lower global contributions. These reactions are less clearly in line with the norm of global cooperation.

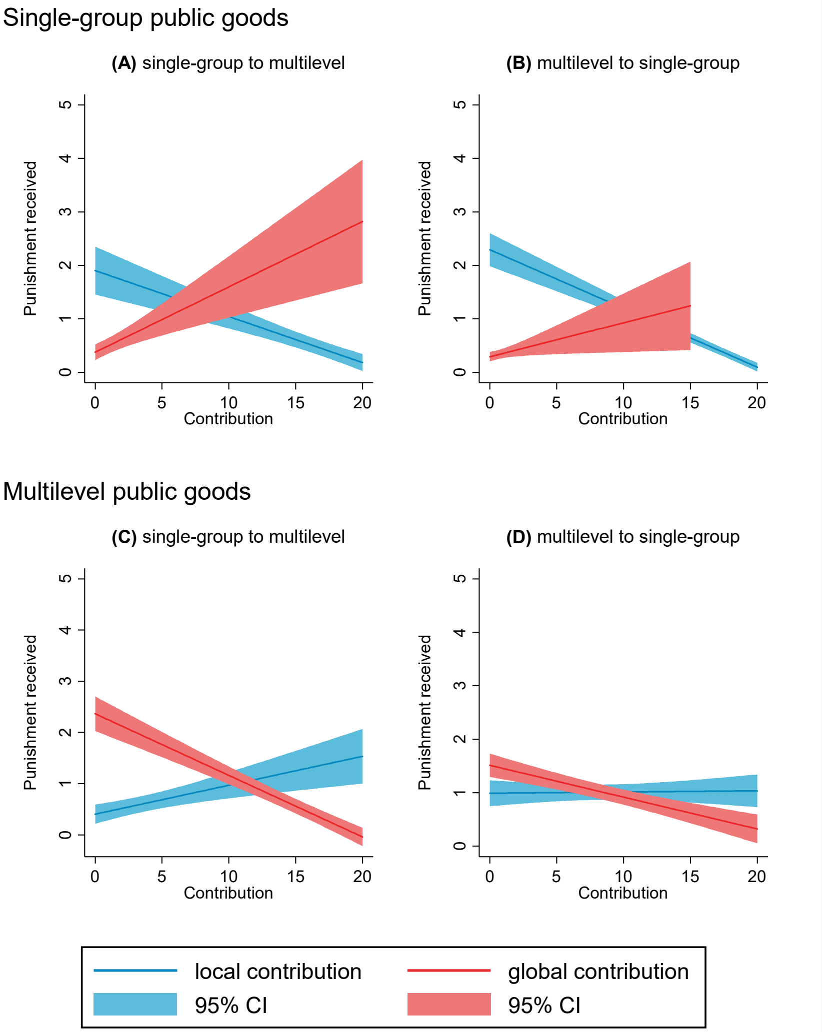


Figure 5.4. Punishment patterns in single-group and multilevel public goods games.
Note: We present the linear relationships between received punishment and contribution behavior, and 95% confidence intervals are included via shaded areas. Received punishment refers to the total number of punishment points received in a round by a player; each punishment point reduces the punished player's payoff by 3 MU. In single-group public goods games (panels A and B), we find that punishment increases with lower local contributions and higher global contributions. In multilevel public goods games (panels C and D), we find that punishment increases with lower global contributions and higher local contributions, in particular in the condition where a single-group public goods game comes before the multilevel public goods game (panel C)

5.2.4. Norms in single-group and multilevel public goods games

Results on the contribution and punishment behavior of the participants suggest a norm of local contributions in single-group PGGs and a norm of global contributions in multilevel PGGs when a single-group PGG comes before the multilevel PGG. We now examine whether the norms suggested by these contribution and punishment patterns are reflected in the direct measurements of norms. To measure norms, we asked participants to report their normative expectations, i.e., what they expect their group members to think are appropriate contributions to the local and global account in both PGGs (single-group and multilevel). To incentivize this measurement, participants earned money if their expectations were correct (see also Methods). The results are shown in Figure 5.5. We see that for single-group PGGs, participants hold normative expectations of high local contributions. These normative expectations are significantly higher with punishment (2.90 MU higher in condition *single-group to multilevel*, p = .003; 1.69 MU higher in condition *multilevel to single-group*, p = .06; Table D.10), suggesting that punishment may not only act as a means of norm enforcement but can also increase the perception of the norm.

For multilevel public goods, we see that normative expectations are higher regarding global contributions than local contributions in all four conditions. This norm of global contributions is strongest in the condition where a single-group PGG comes before the multilevel PGG and where punishment is possible. That is, the normative expectations of global contributions in this condition are significantly higher than in the other three conditions (2.97 MU higher than in condition *single-group to multilevel* without punishment, p = .02; 3.55 MU higher than condition *multilevel to single-group* with punishment, p = .003; 4.01 MU higher than in condition *multilevel to single-group* without punishment, p = .001; Table D.11). Recall that groups in condition *single-group to multilevel* with punishment also achieved the highest actual levels of global contributions (Figure 5.3) and most clearly punished deviations from global contributions (Figure 5.4). Groups in the other three conditions also started with relatively high global contributions, but showed a decline in global contributions over time. These lower levels of global contributions are consistent with these three conditions having lower norms of global contributions (Figure 5.5). Altogether, the normative expectations resemble the observed cooperation and punishment patterns rather closely for both the single-group and multilevel PGGs.

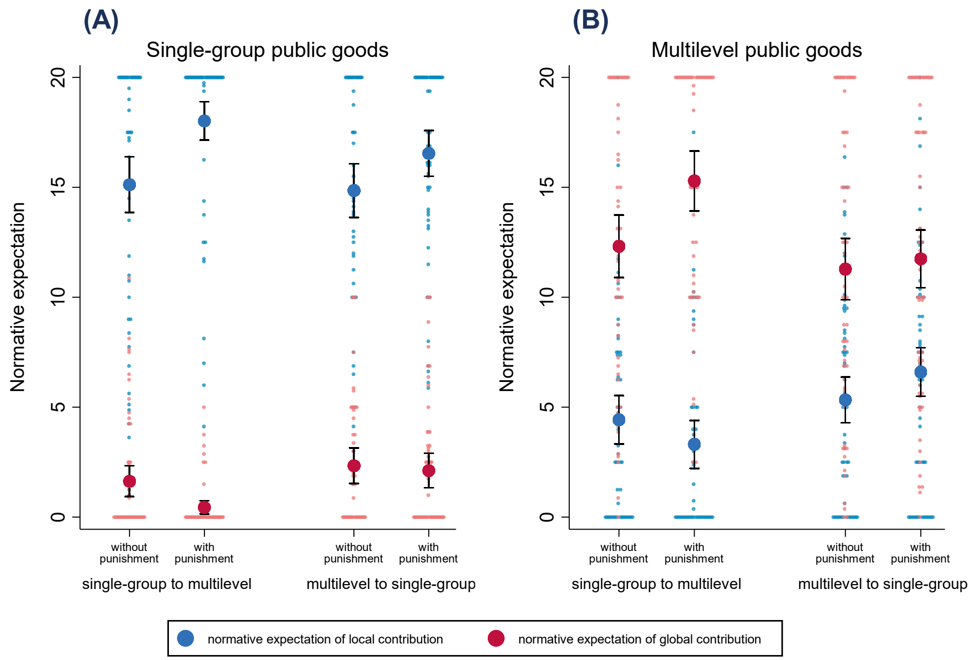


Figure 5.5. Norms in single-group and multilevel public goods games.
Note: We present the normative expectations averaged over the first and second sets of 10 rounds. The results separated by measurement moment are available in the Supplementary Material Figure D.2 and D.3. The large markers show mean values and the small markers show the values of individual participants. The 95% confidence intervals are included via capped spikes. In single-group public goods games (A), we find that normative expectations of local cooperation are higher if peer punishment is possible. In multilevel public goods games (B), we find that normative expectations of global cooperation are higher if peer punishment is possible, but only in the condition where participants interacted in a single-group public goods game before the multilevel public goods game.

To assess whether there is more normative disagreement in multilevel PGGs than in single-group PGGs, we examine the overlap in normative expectations between group members. Normative disagreement is lowest when all group members hold the same normative expectations and is highest when group members all hold different normative expectations. In Supplementary Material Figure D.4, we show the extent of overlap in normative expectations between group members for single-group and multilevel PGGs in all four conditions. We indeed find that normative disagreement is generally higher in multilevel PGGs than in single-group PGGs. The only exception is when a single-group PGG comes before the multilevel PGG and punishment is possible. Participants in this condition have a relatively high overlap in normative expectations between members even in multilevel PGGs. This again suggests that initial experience with solving single-group PGGs may help to prevent normative disagreement in subsequent multilevel PGGs.

5.3. Discussion

A large body of research shows that punishment promotes cooperation in single-group public goods problems (Balliet et al., 2011; Chaudhuri, 2011; Fehr & Gächter, 2002; Fehr & Schurtenberger, 2018; Henrich et al., 2006). Because several real-life public goods problems involve multiple group memberships, we studied whether punishment also promotes cooperation in multilevel public goods problems. We furthermore examined how the effect of punishment depends on whether groups are in a single-group public goods problem before the multilevel public goods problem or vice versa. Results show that while punishment promotes cooperation in single-group public goods problems, punishment does not promote cooperation in multilevel public goods problems if groups do not have prior experience in single-group public goods problems. In these groups, there is a roughly equal mix of local and global cooperation and there is a decay in global cooperation over time even with punishment. Hence, we observe some polarization into separate local goods at the expense of the more collectively beneficial global good. However, in groups that do have prior experience with single-group public goods problems, punishment does promote and sustain cooperation in the multilevel public goods problem. What is more, punishment leads these groups to develop and enforce collectively efficient norms of global cooperation and thus prevents polarized clusters of local cooperation.

The findings establish important boundary conditions for the effects of punishment on cooperation. Whereas we found punishment to promote cooperation in single-group public goods problems regardless of whether this problem appeared before or after the multilevel public goods problem, the effect of punishment in multilevel public goods problems does crucially depend on this order. That punishment does not promote cooperation in groups that start with multilevel public goods problems challenges the view that punishment unequivocally promotes cooperation. Moreover, the observed punishment patterns and norm measurements suggest that normative disagreement is generally higher in multilevel public goods problems than in single-group public goods problems. However, for groups that moved from a single-group public goods problem to a multilevel public goods problem, normative disagreement was not higher in the multilevel problem. What is more, experience with single-group public goods problems did not lead to 'sticky' norms of local cooperation in subsequent multilevel public goods problems as expected. On the contrary, initial experience with solving single-group public goods problems helped to develop and enforce norms of global cooperation in subsequent multilevel public goods problems.

Our findings suggest that 'starting small'– by moving from single-group public goods problems to multilevel public goods problems – is a promising strategy for achieving cooperation in multilevel public goods problems. One may wonder whether the initial experience with single-group public goods problems can be replaced by more experience with multilevel problems. That is, perhaps it just takes time to learn how to cooperate in multilevel public goods problems and even groups that start with a multilevel public goods problem will eventually learn to cooperate. However, our results do not support this view. In none of the conditions did we see groups achieving more global cooperation with more time spent in the multilevel public goods problem. If anything, global cooperation tended to decrease with more time spent in the multilevel public goods problem. This suggests that it is not experience with the multilevel problem in general that helps to achieve cooperation, but rather the 'starting small' with a single-group public goods problem before turning to the multilevel public goods problem.

This result complements previous research showing that 'starting small' helps to achieve cooperation in public good provision. For example, experiments show that achieving cooperation in large groups is facilitated by slowly increasing the group size over time instead of immediately starting with a large group size (Charness & Yang, 2014; Salmon & Weber, 2017; Weber, 2006), and that achieving cooperation in high-stake situations is facilitated by slowly increasing the stakes over time instead of starting immediately with high stakes (Ye et al., 2020). More generally, starting with simple versions of a problem before turning to more complex versions of a problem has been shown to help task performance (Yasarcan, 2009). Our findings are further in line with recent research suggesting that 'local-to-global' mechanisms help to achieve cooperation for global problems (Hauser et al., 2016; McGinnis & Ostrom, 2008). Finally, the results are in line with the idea of an 'expanding circle of cooperation'; people first cooperate within small units such as families and subgroups before being able to cooperate in larger social units involving other subgroups and strangers (Singer, 2011; Smith, 2017).

Multilevel structures in society often arise over time from membership changes between local groups. A well-known example is the increasingly multicultural structure of contemporary western societies, in which individuals from different ethnic groups increasingly need to cooperate. A common conjecture is that interaction in local groups leads to norms of cooperation specifically suited to the current group members (Bernhard et al., 2006; Choi et al., 2019, 2021b; Titlestad et al., 2019). When these groups end up in a multilevel problem through immigration from other groups, the preexisting norms of local cooperation may lead to polarized clusters of local cooperation and impede global cooperation across groups. Our findings suggest this need not always be the case. Initial interaction in local groups can also facilitate a norm of cooperation that maximizes efficiency, and groups realize that global cooperation maximizes efficiency when they end up in a multilevel problem through group changes. However, the attainment of global cooperation does crucially depend on the option to enforce norms via peer punishment. When punishment is not possible, we observe a downward trend in global cooperation, also in groups that have initial experience with single-group public goods problems.

Because we are the first to test the effect of punishment in multilevel public goods experiments, we stayed relatively close to the typical paradigm of public goods games with peer punishment. We offer a few suggestions for future research to test whether our findings generalize to other situations. First, we made use of arbitrary groupings (i.e., minimal group paradigm). Perhaps tensions between different groups are stronger in natural groups, which would make global cooperation even more difficult. Future research could conduct multilevel public goods experiments using natural groups, for example by letting participants from different nationalities or ethnicities play the game together (Buchan et al., 2011, 2009; Drouvelis et al., 2021). Additionally, the local groups in our multilevel public goods problem were of equal size and had equal punishing power to enforce norms. These conditions are not always realized in real-life groups or societies, where the incumbent group is usually the majority and has more power to impose its norms on the incoming members from different groups (Otten et al., 2021). In future experiments, these conditions can be reproduced by making one of the local groups larger in size than the other (Otten et al., 2021) and/or by alternating who can punish whom (Ozono et al., 2020). New insights also emerge from studying individuals in multiple non-nested groups (Bakker, 2019).

We found that punishment only promotes cooperation in multilevel public goods problems when groups have prior experience with solving single-group public goods problems. In an increasingly interconnected world, multiple group memberships may exacerbate the challenges of achieving collective cooperation. Different groups may pursue different goals, potentially leading to segregated clusters of local cooperation that prevent the joint benefits from global cooperation. Our results suggest that 'starting small' in combination with opportunities for norm enforcement may help to overcome such challenges.

5.4. Methods

5.4.1. Data collection

We conducted the experiment at the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University in October 2021. The experiment was programmed with z-Tree software (Fischbacher, 2007) and we recruited participants using the internet recruitment system ORSEE (Greiner, 2015). We ran 14 sessions with about 16 participants per session on average, obtaining a total of 220 participants. Sessions lasted about 1 hour and 15 minutes. Payment depended on choices during the experiment and chance. Participants earned about 17 euros on average (min = 8.5, max = 22.5). Almost all participants were taking courses at Utrecht University. Out of the 220 participants, 114 were Dutch and 106 were from various other countries. Participants were on average 24 years old, 153 participants were female, 63 were male, and 4 identified as another gender. The sample size was selected to be able to detect medium effect sizes (Cohen's d = 0.5). We regard this effect size as a reasonable benchmark, given that meta-analyses on related studies using PGGs have found this effect size on average (Balliet et al., 2011; Spadaro et al., 2022).

Upon arrival in the lab, participants were randomly allocated to individual cubicles and told that they should not communicate with other participants. They were informed about the experiment through written instructions, which are provided in the Supplementary Material section D.6. The written instructions inform the participants that the experiment is divided into two parts. The first part is explained in the written instructions, and instructions on the second part are provided on the computer screen after the first part ends. After reading the instructions on the first part, participants had to complete a quiz to test their understanding of the instructions. 181 out of 220 participants answered all questions correctly on the first try, 27 participants answered 80-83% correctly on the first try, 8 participants answered 60-67% correctly on the first try, and 4 participants answered less than 60% correctly on the first try. Participants were shown which questions they had answered incorrectly and had to correct these before they could continue with the experiment. The quiz is provided in the Supplementary Material section D.6. At the end of the experiment, participants also rated their own understanding of the experimental instructions. 192 participants reported a good understanding, 26 reported not bad, not good, and 2 reported bad. We did not exclude any participants or data points from the analyses to preserve random assignment.

5.4.2. Design

The main part of the experiment consists of a repeated PGG. Before the start of the game, we arbitrarily categorize participants into subgroups via three colors: blue, red, or green[2]. The color of each participant is fixed throughout the experiment, and participants can see the color of their group members. Per round of the game, each participant receives an endowment of 20 monetary units and needs to allocate this endowment between a private account that only benefits oneself, a color account that gives a return of 0.7 to members of one's own color, and a collective account that gives a return of 0.5 to all members regardless of their color. There are four members per group. Contributions to the color account indicate local cooperation, contributions to the collective account indicate global cooperation.

In the first part of the experiment, we let participants play 10 rounds of the PGG in their group. In half of the conditions, all group members have the same color, which means they are in a single-group PGG (Figures 5.1A and 5.1B). In the other half, groups consist of two members of one color and two members of another color, which means they are in a multilevel PGG (Figures 5.1C and 5.1D). To reduce endgame effects, we tell participants that the precise number of rounds remains unknown to them and falls between 8 and 12. After the first 10 rounds of the PGG, we replace two members per group with two new members from another group and start the second part of the experiment. In the conditions where group members all had the same color, the replacement leads the group to become mixed in terms of color (e.g., in a group with four blue members, two of the blue members are replaced by two red members). This means that the membership replacement changes the situation from a single-group PGG to a multilevel PGG (Figures 5.1A and 5.1B). In the conditions where groups were mixed in color before membership replacement, the replacement leads all group members to have the same color (e.g., in a group with two blue and two red members, the two red members are replaced by two blue members). This means that the membership replacement changes the situation from a multilevel PGG to a single-group PGG (Figures 5.1C and 5.1D). We inform the participants of the membership replacement and then let the reshaped groups in both conditions play another 10 rounds of the PGG. We additionally vary whether peer punishment is possible, giving a total of 2 x 2 = 4 conditions that are varied between participants (Figure 5.1, 52 participants in A, 52 participants in B, 56 participants in C, and 60 participants in D).

After each round, participants can see how much each of their group members contributed to the color and collective account. In the conditions with punishment, participants can additionally choose to assign deduction points to their group members. They can assign up to 10 deduction points to each group member per round. Each assigned deduction point costs the punisher 1 MU and the punished member 3 MU. As is common in related studies (Fehr & Gächter, 2000; Reuben & Riedl, 2013), participants see how many deduction points they received, but not by whom. This curbs punishment driven by revenge motives instead of by contribution behavior and thereby helps to interpret punishment as a way to enforce norms.

[2] Three colors are needed when there is an uneven number of groups in the laboratory and we want to switch members between groups such that each reshaped group consists of two members of one color and two members of another color. Half of the sessions had an uneven number of groups. With an even number of groups, we only use red and blue colors.

5.4.3. Norm measurements

Before playing the game and assigning the colors, we measured participants' personal normative views (which are different from normative expectations). To do so, we showed participants two hypothetical groups of four members, one in a single-group PGG (all four members the same color) and one in a multilevel PGG (two members with one color and two members with another color). We ask participants to report on their personal normative views for each of the two hypothetical groups with the question "In your view, what is the appropriate amount that each member should contribute to the color account and the collective account?". Participants could indicate a contribution to the color and collective account for each of the four members between 0 and 20. Participants can try out different combinations of contributions, and see how it affects the earnings of each group member in the two hypothetical groups (screenshots are provided in the Supplementary Material Figures D.5 and D.6).

Personal normative views are measured again before the 10th round of the first part of the experiment and also before the 10th round of the second part. Directly after these two measurements, we also asked participants to guess the personal normative views submitted by their group members. This measures participants' normative expectations rather than their own personal normative views (screenshots are provided in the Supplementary Material Figures D.7 and D.8). The measure of normative expectations is incentivized; if participants correctly guess the most frequent personal normative views among their group members, they receive a payment of 100 MU (~€1.40). The payment is separate for participants' guesses about their group members' views concerning (1) single-group PGGs and (2) multilevel PGGs. Hence, they can earn up to 200 MU (~€2.80) per measurement of normative expectations (400 MU in total because there are two measurement moments of normative expectations). The measurement of normative expectations is similar to prior studies (Krupka & Weber, 2013; Otten et al., 2021; Przepiorka et al., 2022; Reuben et al., 2015). Measurements of normative expectations were followed by measurements of empirical expectations, where participants guess what each of their group members will contribute to the color and collective account in the upcoming round (a screenshot is provided in Figure D.9). These expectations are also incentivized. Per participant, we randomly pick a guess about one of their three group members' contributions and compare this with the actual group member's contribution. If the guess and contribution are the same, participants receive a payment of 100 MU. Because normative expectations are highly correlated with both personal normative views (0.83, $p < .001$) and empirical expectations (0.85, $p < .001$), we do not analyze them separately. We focus on normative expectations, as these are central in most accounts of norms (Bicchieri & Chavez, 2010; Kimbrough & Vostroknutov, 2016; Krupka & Weber, 2013; Otten et al., 2021; Przepiorka et al., 2022; Reuben et al., 2015).

Because norms are measured directly before round 10, they can have an influence on contribution behavior in round 10. Indeed, prior research suggests that measuring norms can make them temporarily more salient and thereby increase cooperation (D'Adda et al., 2016; Krupka & Weber, 2009). Consistent with this research, we see an increase in cooperative behavior in round 10 compared to round 9, especially after the first set of norm measurements (before round 10 of the first set of rounds). In single-group PGGs, it is most cooperative to contribute locally (in terms of maximizing collective payoffs), and we see an increase in local contributions after the norm measurements. In multilevel PGGs, it is most cooperative to contribute globally (in terms of maximizing collective payoffs), and we see an increase in global contributions after the norm measurements.

After the experiment, we presented participants with some hypothetical contribution scenarios and asked them to rate how appropriate these contributions were. We also asked them for some sociodemographic characteristics. These post-experiment measures are not analyzed in the current study. The experiment was preregistered before data collection at Open Science Framework: https://osf.io/5kawt. Informed consent was obtained from all participants and the experimental procedures were approved by the Faculty Ethical Review Board of the Faculty of Social and Behavioral Sciences of Utrecht University. All research was in line with relevant regulations. All data are openly available at the Open Science Framework: https://doi.org/10.17605/OSF.IO/2DKGP.

5.4.4. Analyses

We run linear regression models to estimate whether behavior and normative perceptions depend on experimental conditions and the PGG (single-group or multilevel). For analyses that have repeated observations within individuals and groups, we run two types of regression models. The first account for repeated observations by estimating individual-cluster robust standard errors. The second account for repeated observations by estimating individual-level and group-level random effects (i.e., multilevel regression). Because the results of both types of models are substantively similar, we only report on regression models with cluster-robust standard errors in the main text. The multilevel regression models can be found in Supplementary Material sections D.4 and D.5 (Tables D.12 through D.16).

Chapter 6. Double standards in facilitating norm violations: A field experiment on attempted free-riding in public transport[1]

Abstract

Norms regulate cooperation in society but are sometimes threatened by norm violators. To sustain cooperation, people need to be willing to enforce norms by sanctioning norm violators. However, people may also act against the norm by facilitating the norm violator's actions. Why people sometimes react to norm violators by enforcing the norm and other times by acting against the norm is an unresolved puzzle. We study whether responses to norm violations depend on the norm violator's ethnic group membership. In a field experiment, confederates violate the norm of paying for public transport by attempting to free-ride. Confederates approach travelers who are about to go through check-in gates at Dutch train stations and request to follow them without checking in themselves. We observe whether travelers enforce the norm by rejecting this request or facilitate violating the norm by helping the confederates to free-ride. In total, 801 travelers were individually approached at three train stations by ten different confederates, five with a visibly native-majority background and five with a visibly ethnic-minority background. We find that confederates with a native-majority background are more likely to receive help with free-riding than confederates with an ethnic-minority background across all locations, travelers of all age groups, and all experimental sessions. Content analysis of the travelers' responses to the free-riding request shows that these were more disapproving and less helpful toward confederates with an ethnic-minority background. Our findings reveal double standards in real-life behaviors toward norm violators with different ethnic backgrounds.

[1] This chapter is based on a paper written by Kasper Otten, Vincent Buskens, Wojtek Przepiorka, and Naomi Ellemers. The paper has been submitted to an international peer-reviewed journal. Otten wrote the manuscript, developed and executed the experiment, and did the analysis. All authors contributed to experiment development and manuscript writing. We are very grateful to Rumi Ünal, Kevin Christiaans, Ayad Murad, Rens Heilema, Alexander van Reeken, Bas Nillesen, Tarik Rahali, Oussama Dib, Silvan Zomerdijk, and Feyyaz Bademoglu for their participation in the experiment as confederates. We also thank Ana Macanovic for comments and suggestions on the content analysis. The study is preregistered at doi.org/10.17605/OSF.IO/9EVBM. All data and code are openly available at doi.org/10.17605/OSF.IO/VQ9HZ.

6.1. Introduction

Cooperation in society is regulated by social norms prescribing how people should behave (Bicchieri, 2006; Fehr & Schurtenberger, 2018; House et al., 2020; Young, 2015). Norms pervade almost every aspect of social life; they tell us to wait in line, throw our trash in the bin, respect others' personal space, and so on. Such norms help to make social life predictable and keep social order. However, compliance with norms is not always guaranteed, and when compliance breaks down, it can lead to disorder, conflict, and harm (Álvarez-Benjumea & Winter, 2020; Pettigrew, 1991). Indeed, sustained cooperation is often found to depend crucially on people's willingness to enforce cooperation norms by verbally disapproving of norm violators and other sanctioning behaviors (Chaudhuri, 2011; Fehr & Gächter, 2002; Henrich et al., 2006). However, observing norm violations by others can also erode the norm, e.g., when people condone, approve, or facilitate the norm violator's actions (Diekmann et al., 2015; Dimant, 2019; Keizer et al., 2008). We still have a limited understanding of what factors cause people to react to norm violators by enforcing the norm or by facilitating norm violations. Here, we investigate this issue with a field experiment in which we confront members of the general public with a confederate requesting help with a norm violation.

Two strands of field research have examined how people react to norm violators. The first strand focused on norm enforcement, i.e., whether people are willing to sanction norm violators. In these field experiments, researchers let confederates publicly violate a norm and then examine whether bystanders sanction the confederate by communicating verbal disapproval (Balafoutas et al., 2014, 2016; Balafoutas & Nikiforakis, 2012; Berger & Hevenstone, 2016; Brauer & Chekroun, 2005; Milgram et al., 1986; Mujcic & Frijters, 2020; Przepiorka & Berger, 2016; Schmitt et al., 1992; Winter & Zhang, 2018; Wolbring et al., 2013). These studies show that some people are indeed willing to enforce the norm, although the majority do not enforce the norm. For example, studies showed that some people tell litterers to pick up their trash (Balafoutas et al., 2014; Balafoutas & Nikiforakis, 2012; Winter & Zhang, 2018) or to turn off loud music in quiet zones of trains (Przepiorka & Berger, 2016). These studies are designed to examine norm enforcement and do not include conditions that invite the active facilitation of norm violations.

The second strand of field research does examine whether people who observe norm violations by another person may be more likely to engage in norm-violating behaviors themselves, either by directly copying the norm violation or by facilitating the other person's norm violation (Cialdini et al., 1990; Dannick, 1973; Lefkowitz et al., 1955; Mullen et al., 1990). For example, researchers examined if people are more likely to jaywalk if they see a confederate jaywalking (Dannick, 1973; Mullen et al., 1990), lie to cover up another person's norm violation (DePaulo et al., 1996), or help fellow students who are trying to cheat on exams or copy homework (Scrimpshire et al., 2017). These studies suggest that observing norm violations indeed increases the chance to engage in norm-violating behaviors oneself by copying the norm violation or by helping the other person to violate the norm.

Both strands of prior field research make clear that responding to norm violators is not a trivial matter because it involves multiple, potentially conflicting, normative considerations. On the one hand, confronting norm violators helps to maintain and regulate societal norms of cooperation. For example, confronting fare dodgers helps to keep public transport affordable, and confronting litterers helps to keep public spaces clean. On the other hand, confronting norm violators can mean restricting another person's agency and stopping them from reaching their goals. A recent study of 57 societies suggests that people hold mixed views about the appropriateness of directly confronting norm violators, with a significant share of people disapproving of informal norm enforcers (Eriksson et al., 2021). Accordingly, many people report that they are hesitant to confront norm violators because they fear a negative response from the norm violator (Balafoutas & Nikiforakis, 2012). In situations where a person requests help to violate the norm, enforcing the norm even means deviating from the norm of helping individuals in need. This is the case, for example, when a fellow student asks to copy your homework or when fare dodgers ask paying travelers if they can follow them through check-in gates. Hence, it is not self-evident that people will choose to enforce societal norms, because this may be seen to go against other norms, such as helping a person in need or respecting their autonomy.

In sum, previous field experiments show that observing norm violations may trigger two opposing behaviors: behaviors that uphold the norm (e.g., sanctioning the norm violator) and behaviors that erode the norm (e.g., facilitating the norm violator's actions). We have little understanding of the factors that affect the likelihood of these behaviors. As far as we are aware, there are no field experiments that study both types of reactions to norm violations and the factors that drive people to one of these behavioral reactions over the other. We have seen that people may go against the norm instead of upholding it out of concern for the norm violators, namely to help them or to respect their autonomy. A potentially relevant factor, then, is who the norm violators are. Perhaps people are more attuned to the interests or perspectives of norm violators who belong to their group. We study whether the norm violators' group membership plays a role in whether people respond to norm violators by enforcing the norm or by facilitating the norm violator's actions.

Group membership plays a role in how people interpret the intentions underlying others' behaviors, including others' norm violations (DeRidder & Tripathi, 1992). Intergroup attribution theory suggests that people are more likely to assume good reasons for ingroup members' norm violations and ill intentions for outgroup members' norm violations (Hewstone, 1990; Pettigrew, 2020). For example, when observing someone jaywalking, people may be more likely to think the jaywalker is in a hurry if the jaywalker belongs to their ingroup, whereas they may be more likely to think the jaywalker has a disregard for the rules when the jaywalker belongs to an outgroup. These attributions can shape how people will react to the norm violator. They may be more likely to regard violations as permissible if they assume good intentions, and they may be more likely to disapprove of the norm violator if they assume bad intentions. Intergroup biases may thus drive reactions to norm violators, with more leniency toward ingroup norm violators than outgroup norm violators (DeRidder & Tripathi, 1992). Altogether, we test the following two, pre-registered hypotheses:

1. Outgroup norm violators are less likely to receive help with the norm violation than ingroup norm violators.

2. Outgroup norm violators are more likely to receive disapproval than ingroup norm violators.

6.1.1. Setting and research design

We conduct a field experiment with confederates as alleged norm violators in front of Dutch train stations. The norm that confederates allegedly violate is to pay for public transport. For many train stations in the Netherlands, people can only enter the station if they open one of several gates by swiping a chip card past a sensor at the gate. Checking in this way ensures that the use of public transport is registered and the associated costs are deducted from the chip card. We place confederates near one of these gates and let them approach travelers who are about to check in and hence open the gate. The confederates say they want to catch a train and ask if they can follow the traveler when the traveler opens the gate. This means that the confederates will not check in themselves and hence leave the impression that they do not pay for the use of public transport, i.e., that they will free-ride. We thus regard this as an attempt to violate the norm of paying for public transport. If the traveler agrees to open the gate, we see this as facilitating the norm violation. Travelers who agree to open the gate also engage in norm-violating behavior themselves. Opening the gate for other people than yourself is not allowed and is considered assisted fare-dodging. This can lead to a fine of €90, similar to the fine for free-riding yourself. If the traveler rejects the request to open the gate, we regard this as norm enforcement. Additionally, we can assess the strength of norm enforcement through content analysis of whether and how travelers express verbal disapproval of the request.

We operationalize the group membership of the norm violator in terms of visible ethnic background. In the Netherlands, there is a perception that free-riding in public transport occurs more frequently among (young) persons with a Turkish and Moroccan background (Heyblom, 2020; Kruissink & Essers, 2004). Moreover, both Dutch and Turkish persons believe that persons with a native Dutch background are more likely to sanction such free-riders than persons with a Turkish background (DeRidder & Tripathi, 1992). Hence, regarding public transport use, there are perceived differences in both norm conformity and norm enforcement between members of the native-majority group and the ethnic-minority group, suggesting this is a relevant group membership in this context. The perceptions of ethnic differences in free-riding frequencies of course need not reflect reality, but as long as people believe them, they can act upon them (Quillian, 2006). A further indication of perceived ethnic differences regarding free-riding in public transport comes from the Dutch term that is used for free-riding in public transport, namely ‘zwartrijden’. This term translates to ‘riding black’ and is increasingly subject to controversy because of potential discriminatory connotations (Heyblom, 2020). We manipulate whether the confederate has a visibly native-majority (Dutch) background or a visibly ethnic-minority background (Turkish or Moroccan). A recent field experiment suggests that people can estimate other people’s ethnic background with relatively high accuracy (Winter & Zhang, 2018), and this is also the case in our experiment (see Methods).

We hired 10 male actors between the age of 19 and 27 as confederates, with 5 having a native-majority background and 5 having an ethnic-minority background. The field experiment took place in front of 3 train stations in the Netherlands: Amersfoort, Hilversum, and Woerden (with respectively 25 thousand, 15 thousand, and 8 thousand travelers per day; *Number of Daily Travelers at Dutch Train Stations*, 2022). These train stations were selected because (i) they are located in areas predominantly inhabited by citizens of the native-majority group, (ii) can only be entered by passing through the check-in gates, and (iii) were centrally located in the

Netherlands which helped to find enough confederates willing to participate. Sessions took place from 16:00 to 19:00 on weekdays in the period from April 2022 to June 2022. Over a total of 14 sessions, confederates approached 801 travelers with the request to follow them through the check-in gates (277 in Amersfoort, 281 in Woerden, and 243 in Hilversum). The request was always made in a neutral tone using the same sentence "I want to catch a train, can I walk behind you when you check in?".

In a typical session, two confederates were present, one with a native-majority background and one with an ethnic-minority background (in 4 of the 14 sessions this was not the case because one of the two confederates canceled). The confederates alternated in approaching persons. By having both confederates present within a session and alternating between confederates, we ensured that the treatment (the norm violator's group membership) was random and not confounded by session-specific characteristics (e.g., day of the week, time of the day, or location). A confederate only approached persons traveling alone, so we included only one-to-one interactions. This way we prevented potential diffusion-of-responsibility complications in norm enforcement that could occur when people were traveling in groups (Berger & Hevenstone, 2016; Przepiorka & Diekmann, 2018). Confederates furthermore only approached persons who were clearly over 18 years old and male. Having only male participants and confederates ensures that group membership in terms of gender is fixed, allowing for clearer inferences of the effects of group membership in terms of ethnic background. At the beginning of a session, confederates approached the first traveler that met these criteria. After having noted down the response of this traveler, they approached the next traveler that met the criteria, and so on until the end of the session. Confederates only approached the next traveler once the previous traveler and previous bystanders had left.

After a traveler responded to the confederates' request, the confederates revealed that their request was part of a scientific study and informed the traveler that they did not actually need to follow them through the check-in gates. Confederates subsequently filled in a form on a mobile phone in which they indicated whether the traveler allowed them through the check-in gates, how the traveler phrased their response, and their guess of the traveler's age, national background, and understanding of the confederate's request. A researcher was always present during the experimental sessions and monitored the confederates and how they recorded the participants' reactions. In only 14 out of 801 cases, did the researcher disagree with how the confederate coded a participant's reaction. In these cases, the researcher's perception was followed. Confederates were blind to the study's predictions so as not to influence their behavior. They were only informed that the study was about how people respond to norm-violating behaviors in public transport. We obtained ethical approval from the Faculty Ethics Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University and we preregistered the study at the Open Science Framework: doi.org/10.17605/OSF.IO/9EVBM.

To analyze Hypothesis 1 – that outgroup norm violators are less likely to receive help with the norm violation than ingroup norm violators – we analyze the behavioral responses of the travelers with two-sided *t*-tests. That is, we examine whether participants are more likely to allow native-majority confederates through the gates than ethnic-minority confederates. To analyze Hypothesis 2 – that outgroup norm violators receive more disapproval than ingroup norm violators – we used content analysis of the way the travelers phrased their response to the confederates. To not impose our own views on what constitutes disapproval, we recruited 57 independent coders to rate the answers that the confederates obtained. These coders were informed that the travelers' answers were obtained in a study about how people respond to norm-violating behaviors in public transport, but did not know whether the travelers' answers were directed to confederates with a native-majority or ethnic-minority background. The coders rate the answers in terms of verbal disapproval, helping intentions, and whether the norm to pay your fare was explicitly mentioned, and they classified the verbal answers into pre-defined categories. We use two-sided *t*-tests to determine whether the ratings of verbal expressions toward native-majority confederates differ from those toward ethnic-minority confederates.

After travelers answered the confederate's request, they were also asked whether they are willing to take part in a short questionnaire about (i) the motivation underlying their response, (ii) their personal view on the appropriateness of free-riding in public transport, (iii) their expectation on other people's views on the appropriateness of free-riding in public transport, and (iv) generalized trust. Out of all 801 travelers, 265 travelers agreed to participate in the post-experimental questionnaire, which amounts to a response rate of 33%. The answers to the post-experimental questionnaire are also presented below.

Our field experiment combines the advantages of the experimental approach with the advantages of studying behavior in a natural environment (Baldassarri & Abascal, 2017). The experimental approach ensures that treatment assignment is random and allows for causal inferences about the impact of the group membership of confederates. The field environment allows us to observe realistic behaviors toward norm violations and free narratives explaining these behaviors, without having to impose our own assumptions about the perceived costs, benefits, or motives underlying these behaviors. By focusing on the norm violation of free-riding in public transport, we study a norm violation that is relatively prevalent in everyday life, widespread in all countries, costs billions of dollars worldwide, and is widely considered inappropriate (Dai et al., 2018). The norm against free-riding in public transport is considered a prime example of a civic norm across countries and has received considerable scientific attention (Herrmann et al., 2008; Knack & Keefer, 1997). Yet, how citizens behave when they observe violations of this norm is largely unstudied. Another advantage of focusing on norm violations in public transport is the accessibility of the target population. A very large share of the population in the Netherlands uses public transport, with a wide variety in terms of age, education level, income level, and occupation. Because travelers are observed in the 'wild' and are unaware of participating in an experimental study, we avoid both demand and selection effects.

6.2. Results

6.2.1. Behaviors toward norm violators

Out of the 801 travelers, 400 were approached by native-majority norm violators and 401 by ethnic-minority norm violators. Native-majority norm violators were helped in 67% of the cases compared to 49% for the ethnic-minority norm violators (the difference is 18 percentage points, $t(799) = 5.20$, $p < .001$). The effect size represents a Cohen's d of 0.37, which is typically regarded as a medium-sized effect in behavioral research (Cohen, 1988; Hemphill, 2003). The finding supports Hypothesis 1 that outgroup members are less likely to receive help with the norm violation than ingroup norm violators. After controlling for estimates of the travelers' demographic characteristics, their understanding of the request, and session effects, the magnitude of this difference remains (17 percentage points, $t(799) = 4.65$, $p < .001$, Supplementary Table E.1). When we select only sessions in which both a native-majority and an ethnic-minority confederate were present, we also find a similar difference (15 percentage points, $t(638) = 3.78$, $p < .001$). In Figure 6.1, we show the difference in helping behaviors toward native-majority and ethnic-minority norm violators per (i) location, (ii) travelers' age, and (iii) session. We consistently observe differential treatment depending on confederate group membership; it is visible in all three locations, all age groups, and across all experimental sessions.

To assess to what extent this differential treatment is driven by travelers' own group membership, we can examine to what extent it is present for travelers with a native-majority Dutch background (601 travelers), travelers with a Turkish/Moroccan ethnic-minority background (68 travelers), and travelers with another ethnic background such as German or Polish (130 travelers) (travelers' ethnic backgrounds as estimated by the confederates, see Methods). Travelers with a native-majority background help native-majority norm violators in 65% of the cases and ethnic-minority norm violators in 45% of the cases (the difference is 20 percentage points, $t(601) = 4.97$, $p < .001$). Travelers with a Turkish/Moroccan ethnic-minority background help native-majority norm violators in 66% of the cases and ethnic-minority norm violators in 72% of the cases (the difference is 6.5 percentage points, $t(66) = 0.58$, $p = .72$). Finally, travelers with another ethnic background (e.g., German or Polish) help native-majority norm violators in 75% of the cases and ethnic-minority norm violators in 53% of the cases (the difference is 22 percentage points, $t(128) = 2.64$, $p = .009$). Hence, it seems that the only group of travelers that does not have a bias against norm violators with a Turkish/Moroccan background consists of travelers whom themselves have a Turkish/Moroccan background.

In sum, native-majority citizens are more likely to help native-majority members to violate norms while ethnic-minority norm violators are more likely to face norm enforcement by having their free-riding requests rejected. However, in itself the decision to help or reject norm violators does not inform us about the strength of norm enforcement. We next turn to content analyses of the travelers' verbal expressions to examine the strength of norm enforcement in the form of communicating verbal disapproval. The content analyses of travelers' expressions also help to reveal potential mechanisms that are stated to motivate their norm enforcement decisions.

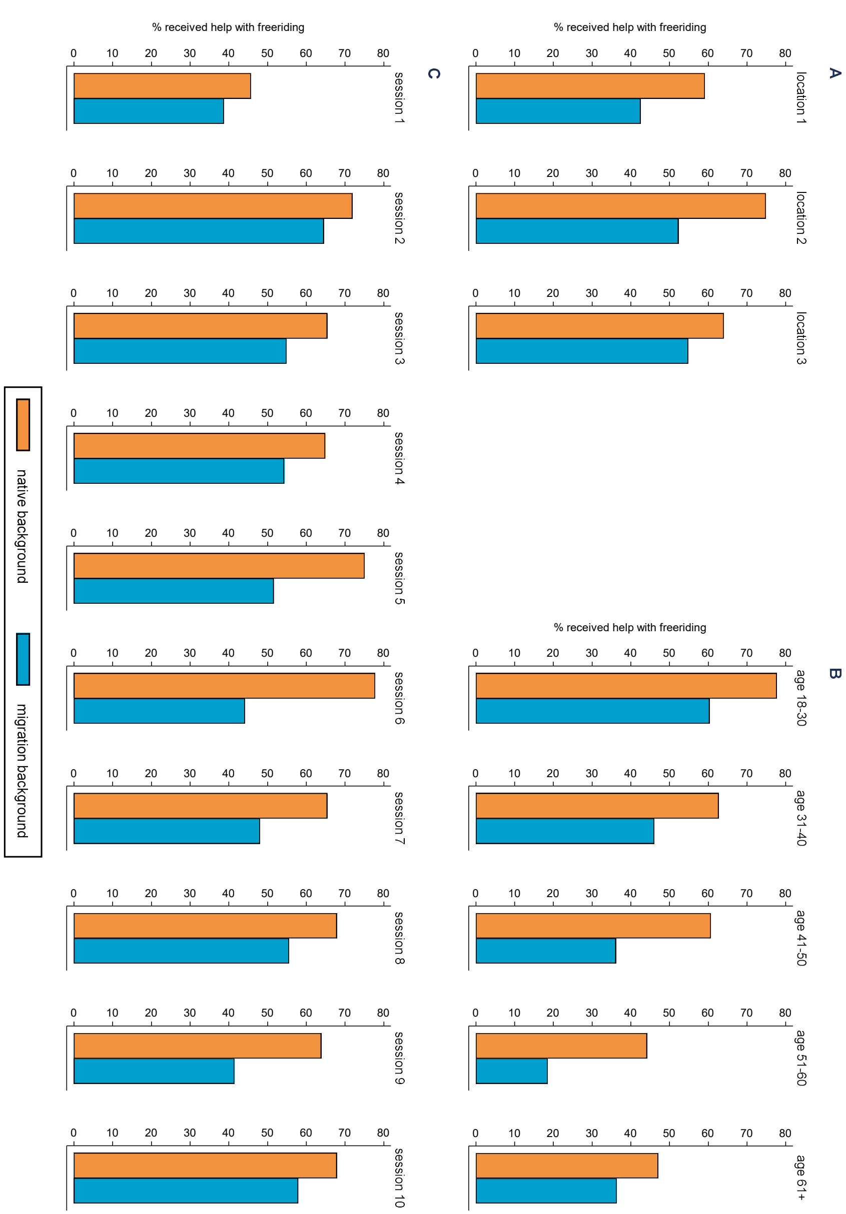


Figure 6.1. Behaviors toward native-majority and ethnic-minority norm violators.
Note: **(A)** Behavior by location; location 1 is Amersfoort, location 2 is Woerden, and location 3 is Hilversum. **(B)** Behavior by travelers' age. Confederates estimated the travelers' age in categories of 18-30, 31-40, 41-50, 51-60, 61-70, and 71+. Because the 71+ age category only had 5 observations, we could not reliably estimate the rate of differential treatment for this category and merged it with the 61-70 age category, creating an age category 61+. **(C)** Behavior by experimental session for all sessions in which both a native-majority and ethnic-minority confederate were present.

6.2.2. Verbal reactions toward norm violators

The 801 travelers gave a total of 513 unique verbal answers (some answers such as a simple 'yes' or 'no' occur multiple times). We recruited 57 independent coders to each rate 57 randomly selected verbal answers, resulting in a total of 3249 ratings. This means that each verbal answer is rated by 6 to 7 coders. We mainly use Krippendorff's alpha to assess intercoder reliability. This is a standard intercoder reliability measure with a rule of thumb that scores should be above 0.67 to be considered sufficiently reliable (Hayes & Krippendorff, 2007; Macanovic & Przepiorka, 2022). The main aspect that was rated is verbal (dis)approval. Coders answered the question "How disapproving or approving is the traveler's answer according to you?" on a 5-point Likert scale ranging from 1 (very disapproving) to 5 (very approving). Krippendorff's alpha for ratings of verbal disapproval is 0.74, suggesting decent intercoder reliability. We take the average across coders' ratings to assign each verbal answer a score on verbal (dis)approval. To give an indication of what type of answers coders deemed to be disapproving and approving, we display the answers with the lowest and highest approval ratings in Table 6.1.

Table 6.1. Examples of verbally disapproving and approving answers.

Disapproving
No of course not.
No, absolutely not.
My train is already here and I find this highly inappropriate.
No no surely not.
No you have to buy a ticket yourself.
No piss off.
No no no, I won't be a part of that.
Absolutely not.
No get lost.
Of course not idiot.
Approving
That's completely fine man.
Yes I would do the same.
Yes of course! Let's go!
Hahahahaha yes let's go sure.
Yes sure absolutely.
Yes of course that's fine.
Yes of course you may, just walk behind me!
Yes, that's fine man!
Yes sure, let's go!
Yes of course.

Note: The question asked by the confederates was "I want to catch a train, can I walk behind you when you check in?" We show verbal answers on which all coders who saw the answers agreed that they were very disapproving (first ten rows) or very approving (last ten rows). The original answers, which were given in Dutch, are provided in Supplementary Table E.5.

We find a clear link between behavior toward norm violators and verbal expressions of (dis)approval. Travelers who helped the norm violator express relatively high approval (M = 3.88 on a scale ranging from 1 = "very disapproving" to 5 = "very approving", SD = 0.73) whereas travelers who rejected the norm violator expressed relatively high disapproval (M = 2.11, SD = 0.59). This difference is statistically significant ($t(799) = 36.75$, $p < .001$) and large; helping behavior toward norm violators correlates with verbal expressions of approval at 0.79 ($p < .001$). We find a small but significant difference in verbal expressions of (dis)approval depending on confederate group membership; the scores are 3.22 for native-majority and 3.06 for ethnic-minority norm violators (the difference is 0.16 points, $t(799) = 2.04$, $p = .04$). This difference remains significant when we select only travelers with a native-majority background (the average approval score is 3.19 toward native-majority norm violators vs. 2.98 toward ethnic-minority norm violators, $t(601) = 2.30$, $p = .02$) or all travelers without a Turkish/Moroccan background (3.21 vs. 3.02, $t(731) = 2.36$, $p = .02$). This suggests that verbal answers directed toward ethnic-minority norm violators are less approving across the board.

In Figure 6.2, we show the distribution of ratings of verbal (dis)approval toward native-majority and ethnic-minority norm violators. We see that the modal verbal answer is neutral (26% and 27% of answers received by native-majority and ethnic-minority norm violators respectively). We furthermore see that answers that are very disapproving are relatively rare (14% both for native-majority and ethnic-minority norm violators). Answers that are moderately disapproving are more likely to be received by ethnic-minority norm violators than native-majority norm violators (21.0% vs. 15.7%, $t(799) = 3.30$, $p = .001$). Native-majority norm violators are more likely to receive very approving answers (23.3% vs. 17.6%, $t(799) = 3.08$, $p = .002$). These findings support Hypothesis 2 that outgroup norm violators are more likely to receive disapproval than ingroup norm violators. This result is driven by ethnic-minority confederates receiving more rejections, not harsher rejections. When we select only travelers who reject the free-riding request, ethnic-minority confederates do not receive more disapproval than native-majority confederates (see Supplementary Section E.7). Note also that differential treatment in terms of verbal expressions is less than differential treatment in actual behavior. We come back to this finding in the Discussion.

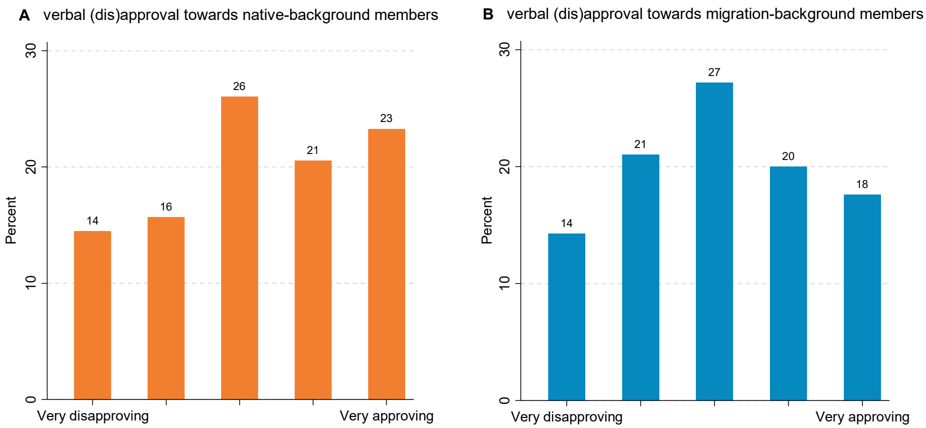


Figure 6.2. Verbal (dis)approval toward native-majority and ethnic-minority norm violators.
Note: **(A)** Distribution of approval ratings for verbal answers toward native-majority norm violators. **(B)** Distribution of approval ratings for verbal answers toward ethnic-minority norm violators.

The coders also rated to what extent travelers verbally expressed helping intentions, whether the norm to pay your fare was explicitly mentioned, and they also classified the verbal expressions into pre-defined categories. The results for these additional ratings are summarized here and given in full in Supplementary Section E.7. Verbal expressions toward native-majority confederates as opposed to ethnic-minority confederates were rated as more helpful (3.20 vs. 2.84 on a Likert-scale ranging from 1 “Very hesitant to help” to 5 “Very willing to help”, $t(799) = 3.64$, $p < .001$). Whether the norm to pay your fare was explicitly mentioned in the verbal answer did not differ between native-majority and ethnic-minority confederates (6.9% vs 5.2%, $t(799) = 1.16$, $p = .25$). The norm to pay your fare was rarely mentioned in both experimental conditions. This suggests that travelers being more inclined to reject free-riding requests from ethnic-minority confederates is unlikely to be driven by a motive to teach them the norm.

Finally, three categorical ratings are of particular relevance for the underlying theory of intergroup attribution bias. Recall that this theory suggests that people assume better intentions underlying norm violations of ingroup members than outgroup members. The three ratings relevant to this theory concern whether the travelers’ verbal expressions (i) assume good intentions underlying the violator’s request, (ii) assume bad intentions underlying the violator’s request, or (iii) inquire about the intentions underlying the violator’s request. We indeed find that verbal responses more often contain expressions that assume good intentions toward native-majority norm violators than toward ethnic-minority norm violators (7.0% vs. 4.9%, $t(799) = 2.42$, $p = .02$), but there is no difference in assumptions of negative intentions (7.4% vs. 7.3%, $t(799) = 0.16$, $p = .87$). Verbal responses that inquire about the intentions underlying the violator’s request are more likely toward native-majority than ethnic-minority norm violators (3.3% vs. 0.6%, $t(799) = 3.41$, $p < .001$). In sum, ethnic-minority norm violators are somewhat less likely to receive the benefit of the doubt and travelers are also less likely to ask them about their intentions. However, these differences are small and intercoder agreement was mixed on these ratings (see Supplementary Section E.7). Hence, we again find that differences in verbal expressions are more subtle than the actual behavioral differences observed.

6.2.3. Post-experimental questionnaire

As mentioned, 265 travelers agreed to participate in the post-experimental questionnaire (33% of all travelers). Travelers who helped the norm violator were more likely to participate in the questionnaire than travelers who rejected the norm violator (46% vs 16%, $t(799) = 9.24$, $p < .001$). However, participation in the questionnaire did not depend on the norm violator's background; 35% of the travelers approached by native-majority confederates participated and 31% of the travelers approached by ethnic-minority confederates participated ($t(799) = 1.00$, $p = .32$). The first (open-ended) question that travelers answered was why they responded to norm violators with helping or rejecting the free-riding request. Most of the travelers who helped the norm violator to free-ride answered that they did so out of kindness (e.g., "I like to help people") or because they do not care about other people free-riding in public transport (e.g., "I don't really care, it is not my problem"). Most of the travelers who rejected the norm violator's request to free-ride answered that they did so because free-riding is against their principles and they believe everybody should pay their share (e.g., "It is not honest and people work hard for their money"). All original open-ended answers and automatic translations to English are available at doi.org/10.17605/OSF.IO/VQ9HZ.

Travelers next answered to what extent they find it socially appropriate or inappropriate to pass through the check-in gates without checking in on a 5-point Likert scale ranging from 1 (very inappropriate) to 5 (very appropriate). They were also asked to what extent they think others find this to be socially appropriate or inappropriate. The average score for travelers' own opinion is 2.24 and for their expectation of others' opinion is 2.16 (the difference is .08 points, $t(264) = 1.08$, $p = .28$). This suggests that travelers find, and expect others to find, the norm violators' behavior to be moderately inappropriate on average. Note, however, that travelers who helped the norm violator were overrepresented among those who participated in the questionnaire. When we select travelers who rejected the norm violator, we find the norm violation to be judged very to moderately inappropriate ($M = 1.59$, $SD = 0.71$). The correlation between travelers' own opinions and their expectations of others' opinions is 0.29 ($p < .001$). In general, travelers expect others to find the norm violation moderately inappropriate regardless of whether they themselves find the norm violation to be inappropriate (see Supplementary Figure E.10). Whether the travelers help or reject the norm violation is associated with their own stated view of whether passing through the gates without checking in is inappropriate (correlation = .32, $p < .001$), but not with their expectation of others' views (correlation = -0.04, $p = .50$). Hence, people seem to base their decision on how to behave toward norm violators more so on their own opinion of what is right and wrong.

6.3. Discussion

Field experiments show that observing norm violations may trigger two opposing behaviors: behaviors that uphold the norm (e.g., sanctioning the norm violator) and behaviors that erode the norm (e.g., facilitating the norm violator's actions). Little is known about the factors that affect the likelihood of these behaviors, or about the motives and concerns driving them. Using a field experiment at Dutch train stations, we find that group membership is an important factor in how people behave when facing the norm violation of attempted free-riding. Native-majority travelers are more likely to facilitate the attempt to free-ride if the norm violator belongs to the native-majority ingroup than if the norm violator belongs to the ethnic-minority outgroup. Conversely, native-majority travelers are more likely to enforce the norm toward ethnic-minority norm violators by refusing the request to free-ride and by expressing verbal disapproval of the request. The greater willingness of native-majority travelers to help native-majority free-riders is visible across all studied locations, travelers of all age groups, and all experimental sessions. Hence, we find robust evidence of double standards toward norm violators.

In most field experiments on norm enforcement and norm violations, there is little to no verbal interaction with the participants (Balafoutas et al., 2014, 2016; Balafoutas & Nikiforakis, 2012; Berger & Hevenstone, 2016; Brauer & Chekroun, 2005; Mujcic & Frijters, 2020; Przepiorka & Berger, 2016; Schmitt et al., 1992; Winter & Zhang, 2018; Wolbring et al., 2013). Because our field experiment involves asking participants directly for help with free-riding, we could almost always record a verbal reaction from the participant. This allowed us to complement our behavioral analyses with content analyses of the verbal expressions of the participants, revealing their stated concerns and motives. Independent human coders rated the travelers' verbal reactions for expressions of (dis)approval, among others. Although these content analyses revealed that native-majority norm violators are less likely to receive verbal disapproval than ethnic-minority norm violators, the difference in verbal disapproval is much smaller than the difference in actual helping behavior. Hence, the ingroup bias is subtle and difficult to see by analyzing the way people talk to out-group members, but it becomes more apparent when analyzing their actual behavior. This is in line with other studies showing the implicit and covert nature of everyday discrimination (Aidenberger & Doehne, 2021; Baldassarri, 2020; Choi et al., 2021b, 2021a; Koopmans & Veit, 2014; Quillian, 2006; Zhang et al., 2019, 2022), and is also in line with people's self-views that they do not (intend to) discriminate even when their behavior shows that they do (Crandall & Eshleman, 2003; Czopp et al., 2006). Furthermore, our finding that most travelers do not express strong verbal disapproval is in line with previous field experiments showing that people prefer weak forms of norm enforcement (e.g., rejecting the request to free-ride without expressing verbal disapproval) over strong forms of norm enforcement (e.g., verbally reproaching the norm violator) (Balafoutas et al., 2014; Berger & Hevenstone, 2016).

Intergroup attribution theory suggests that people are more likely to assume good intentions for ingroup members' norm violations and ill intentions for outgroup members' norm violations (DeRidder & Tripathi, 1992; Hewstone, 1990; Pettigrew, 2020). The finding that native-majority travelers are more likely to help native-majority confederates to free-ride than ethnic-minority confederates is in line with this theory. Even though we found that differential treatment is more difficult to detect in verbal expressions than in actual behavior, the content analyses did provide some evidence for intergroup attributions. The content analyses suggest that travelers were more likely to assume good intentions from native-majority free-riders while travelers were less interested in the intentions of ethnic-minority free-riders. The latter observation resonates with prior research suggesting that people are generally less inclined to 'mentalize' or consider the concerns and perspectives of others to account for their behavior, when these others belong to a different group (Molenberghs & Louis, 2018). Interestingly, we did not find that travelers are more likely to assume negative intentions from ethnic-minority members. Hence, the difference in assumed intentions seems to occur mostly in the positive domain instead of the negative domain. Of course, travelers may also assume good or bad intentions without revealing them in their verbal reactions, which is precisely why analyzing behavior remains important.

While persistent discrimination in institutional settings such as the labor and housing market has long been documented (Auspurg et al., 2019; Pager & Shepherd, 2008; Zschirnt & Ruedin, 2016), field research showing everyday discrimination in more informal, unstructured encounters has only recently been gaining traction (Aidenberger & Doehne, 2021; Baldassarri, 2020; Choi et al., 2021b, 2021a; Koopmans & Veit, 2014; Zhang et al., 2019, 2022). There is already evidence that law enforcement targets free-riding by outgroup members more so than by ingroup members, due to so-called 'ethnic profiling' (Carter & Johnson, 2021; Mujcic & Frijters, 2020). As a high-profile example, it recently came to light that the Dutch government targeted citizens with an ethnic-minority background more so in controls for free-riding than citizens with a native-majority background (tax evasion in particular; European Parliament, 2022). Ethnic profiling appears to be persistent even in the face of bias-training programs (Lai & Lisnek, 2023). Our findings highlight that members of the native-majority background are not only more likely to get away with free-riding in institutional contexts, but also in informal contexts. This is especially important because many social situations are governed by social norms rather than legal norms. Whereas legal norms are regulated by law enforcement, social norms are regulated by informal norm enforcement from ordinary citizens. We find that ordinary citizens discriminate by applying double standards in norm enforcement toward free-riding. This brings native-majority members at a double advantage; their heightened capacity to free-ride allows them to 'cut corners' both in institutional and informal settings. This unfair advantage in free-riding can increase intergroup inequality (e.g., in terms of paid taxes and fares) and undermine societal cooperation.

6.4. Materials and Methods

6.4.1. Field experiment

Before the experimental sessions, confederates were given written instructions (provided in Supplementary Section E.1). Before each session, the researcher would go through these instructions together with the confederates on location. The researcher also practiced three scenarios with confederates that participated for the first time (confederates that already participated in a previous session did not have to do this practice anymore). The scenarios are (i) a traveler agrees to the confederate's request to pass through the check-in gates but does not want to participate in the questionnaire, (ii) a traveler rejects the confederate's request to pass through the check-in gates but does want to participate in the questionnaire, and (iii) the traveler asks "Why?" in response to the confederate's request. The researcher also walked the confederates through the form that they had to fill in on a mobile phone to conduct the post-experimental questionnaire and to note down the response of the traveler after each interaction (provided in Supplementary Section E.2).

Confederates were requested to wear casual clothes (as far as possible wearing white sneakers, jeans, and a dark coat). We recruited only male confederates in order not to complicate the experiment further by adding the factor of gender. We chose male rather than female confederates because there is the perception that males are more likely to free-ride in public transport (Kruissink & Essers, 2004), making the experiment potentially more realistic than if we had chosen females. Out of all 10 confederates, 5 had a native Dutch background (native-majority) and 5 had a Turkish or Moroccan ethnic background (3 Turkish and 2 Moroccan). Photos of the confederates are provided in Supplementary Figure E.2. Prior research suggests that people can guess the ethnic background of others with relatively high accuracy (Winter & Zhang, 2018). Our finding that travelers discriminate in their helping behavior based on the ethnic background of the confederates also suggests that they can guess the confederates' ethnic background with sufficient accuracy. As reported in the main text, the received help differed significantly between the native-majority and ethnic-minority confederates. Among the ethnic-minority confederates, we do not find a difference in received help between confederates with a Turkish and Moroccan background (50.2% received help vs. 46.1%, $t(399) = 0.76$, $p = .45$).

Confederates were positioned next to the entry of the station, which is a few meters from the check-in gates at all three locations. They then approached the first traveler who met the three criteria: (i) over 18 years old, (ii) traveling alone, and (iii) male. Confederates only approached people who travel alone to avoid diffusion-of-responsibility complications in norm enforcement that can occur when people are traveling in groups (Berger & Hevenstone, 2016; Przepiorka & Diekmann, 2018). That confederates only approach males was requested by our Faculty Ethics Review Board because female participants may feel more uncomfortable when asked by a male to share the small space of a check-in gate. Additionally, having only male confederates and participants ensures that group membership in terms of gender is fixed, allowing for clearer inferences of the effects of group membership in terms of ethnic background.

Out of all travelers, 90% gave a clear affirmative or rejecting answer to the confederate's first request to follow the travelers through the check-in gates. The remaining 10% did not give a clear answer immediately, for example, because they wanted to know the reasons underlying the confederate's request or because they did not hear the first request well. In those cases, the confederates repeated the request one time. If the answer was then still not clear, we regarded it as a rejecting answer. The percentage of travelers whose answer was not clear after the confederates' first request does not differ between native-majority and ethnic-minority confederates (11% vs 10%, $t(799) = 0.47$, $p = .64$). Regardless of whether travelers agreed or disagreed to let the confederate through the check-in gates, confederates would not actually go through the check-in gates. Instead, the confederates revealed that the request was part of a scientific study after having obtained a response to the request. The confederates then delivered the debriefing to the traveler and were available to answer questions. The debriefing text explains the purpose of the study, what will be done with the data, the possibility of being removed from the data, and whom to contact in case of questions (see Supplementary section E.3). To make the debriefing less cumbersome for the travelers, we printed a QR code and weblink on business cards and had the confederates give these to travelers. Scanning the QR code or using the link led to the debriefing text. If the traveler had questions that the confederate could not answer, he directed the traveler to the researcher who was also present on location. Because travelers do not actually pass the check-in gates, the experiment took place in the public space.

To get an indication of whether the confederates' guesses of the travelers' (ethnic) migration background were accurate, we can compare the actual migration composition of the three study locations with the estimated migration composition based on the confederates' guesses. Information on the composition of our study locations in terms of migration background is available from Statistics Netherlands (2022). In Amersfoort, 25.9% of the population had a migration background on January 1, 2022 compared to an estimated 24.9% by our confederates. In Hilversum, 30.6% of the population had a migration background on January 1, 2022 compared to an estimated 30.5% by our confederates. In Woerden, 16.5% of the population had a migration background on January 1, 2022 compared to an estimated 19.6% by our confederates. The differences between the actual and estimated migration backgrounds are small in each location, and the actual rank order between locations is the same as the estimated rank order (from low to high percentage of migration background: Hilversum → Amersfoort → Woerden).

Related field experiments often recruit 1-2 confederates (Berger & Hevenstone, 2016; Brauer & Chekroun, 2005; Dannick, 1973; Lefkowitz et al., 1955), 3-5 confederates (Aidenberger & Doehne, 2021; Balafoutas et al., 2016; Balafoutas & Nikiforakis, 2012; Przepiorka & Berger, 2016), or 6-8 confederates (Balafoutas et al., 2014; Winter & Zhang, 2018). We recruited 10 confederates to reduce the chance that the differences in travelers' behaviors toward native-majority and ethnic-minority confederates are driven by idiosyncratic confederate differences other than ethnic background. While we find some differences between native-majority confederates (among the 5 confederates, the average received help varied from 57.4% to 74.7%) and between ethnic-minority confederates (among the 5 confederates, the average received help varied from 43.1% to 56.8%), the average received help for each native-majority confederate is higher than the average received help for each ethnic-minority confederate (see Supplementary Figure E.3). This suggests that it is indeed ethnic background that is mainly driving the results and not confounding by other confederate differences. After the confederates participated for the first time, they could themselves decide whether they wanted to participate in another session. All confederates decided to participate again (7 confederates participated in 2 sessions, 1 confederate participated in 3 sessions, and 2 confederates participated in 4 sessions).

We aimed to collect about 600 observations, which would give 80-85% power to detect effect sizes of Cohen's $d \sim 0.25$, which is on the lower end of effect sizes reported in related field experiments (Aidenberger & Doehne, 2021; Winter & Zhang, 2018; Zhang et al., 2019). Because the confederates were somewhat faster than expected in approaching travelers and noting down their responses, we ended up with 801 observations. We ended the experiment when all confederates had the chance to participate in at least two sessions. In all analyses, we use the standard $p < .05$ for two-sided tests criterium to determine if the results are significantly different from those expected if the null hypothesis were correct.

6.4.2. Post-experimental questionnaire

We kept the post-experimental questionnaire relatively short, because asking for more than a few minutes of a person's time may drastically reduce participation rates, especially because persons may not have much time when they intend to travel. All participants in the post-experimental questionnaire gave informed consent (see also Supplementary Section E.2). As reported in the main text, 265 out of 801 travelers participated in the post-experimental questionnaire. We find that travelers participating in the questionnaire do not differ from travelers not participating in the questionnaire in their ethnic background, the ethnic background of the confederate, or the experimental session, but they are somewhat younger than travelers not participating (Table E.2). Because older participants find it somewhat more inappropriate to pass the gates without checking in (Table E.3), the travelers not participating in the questionnaire may have deemed the confederates' request even more inappropriate.

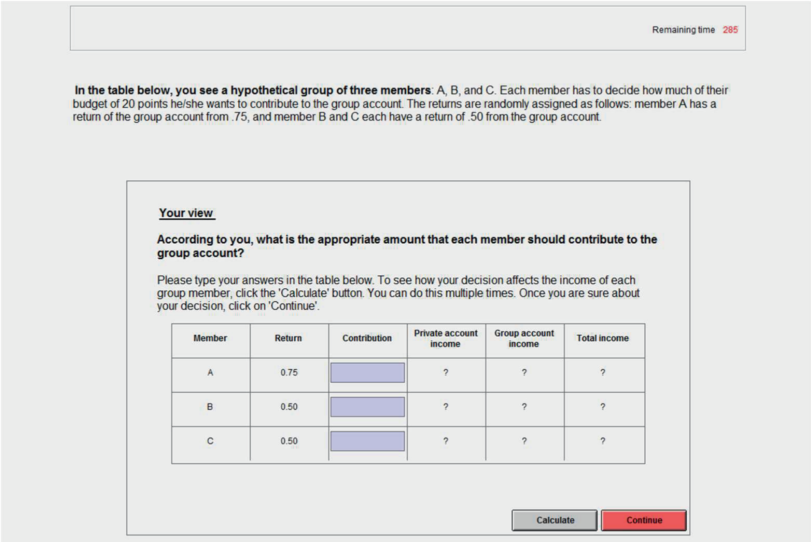
6.4.3. Coding of the verbal answers

We recruited 57 independent human coders to rate the verbal answers received by the confederates in our field experiment. The coders were recruited via our Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University. The average age among coders was 23 years (*min* = 18, *max* = 65), 75% were female, and 89% were students at Utrecht University. We find that ratings of approval do not depend on the coder's age, sex, political orientation, and whether the coder is a student (Supplementary Table E.4). Coders were invited to our lab in 4 sessions from September 2022 to October 2022. Sessions lasted about 45 minutes and coders received 15 euros for completing the task. They were assigned to an individual cubicle and were informed about the task through written instructions (provided in Supplementary Section E.4). These instructions explained the field experiment and the context in which the travelers' answers were obtained. The 57 coders each rated 57 answers in terms of (i) verbal disapproval, (ii) helping intentions, (iii) mentions of the norm to pay your fare, and (iv) some other details. The full set of items on which the coders rated the travelers' answers is provided in Supplementary Section E.5. All data are openly available at doi.org/10.17605/OSF.IO/VQ9HZ.

Appendix A. Supplementary material for Chapter 2

A.1. Figures and tables

(a) before calculating the payoff consequences of one's normative view



(b) after calculating the payoff consequences of one's normative view

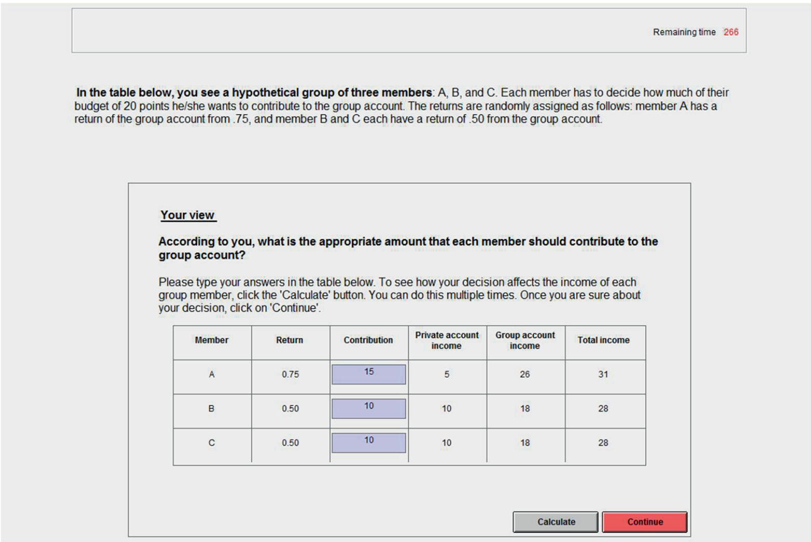


Figure A.1. Screenshots of experimental normative view measurement.

Table A.1. Changes in normative views for minority and majority participants.

	Position of participant		Mann-Whitney test of difference by return-type	
	minority	majority	*z*-statistic	*p*-value
Absolute change between rounds 1 - 10				
normative view on high-return members	2.38	2.70	0.94	.35
	(4.35)	(3.56)		
normative view on low-return members	2.50	2.44	-0.23	.82
	(3.57)	(3.42)		
Relative change between rounds 1 - 10				
normative view on high-return members	2.38	1.39	-1.12	.26
	(4.35)	(4.26)		
normative view on low-return members	0.94	1.65	0.78	.44
	(4.27)	(3.87)		

Note: estimates are from condition disagreement because only when there is disagreement can we say one participant holds a minority view and the other two hold a majority view. For the absolute change, both positive and negative changes are regarded as positive levels of change. For the relative change, the negative changes are subtracted from the positive ones. Positive changes indicate that the participants' normative views prescribe higher contributions in round 10 than in round 1. Standard deviations are in parentheses.

Table A.2. Mann-Whitney test for contributions and punishments by condition.

Contributions - individual level	
z-statistic	-2.235
Contributions - group level	
z-statistic	-0.994
Punishment - individual level	
z-statistic	2.049
Punishment - group level	
z-statistic	1.372

Note: We test whether the contribution and punishment decisions over rounds 1-10 differ between conditions with non-parametric Mann-Whitney ranksum tests, both when using individual decisions and group-mean decisions. We find no significant differences according to conventional standards: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (Bonferroni-adjusted $p/4$, two-tailed tests)

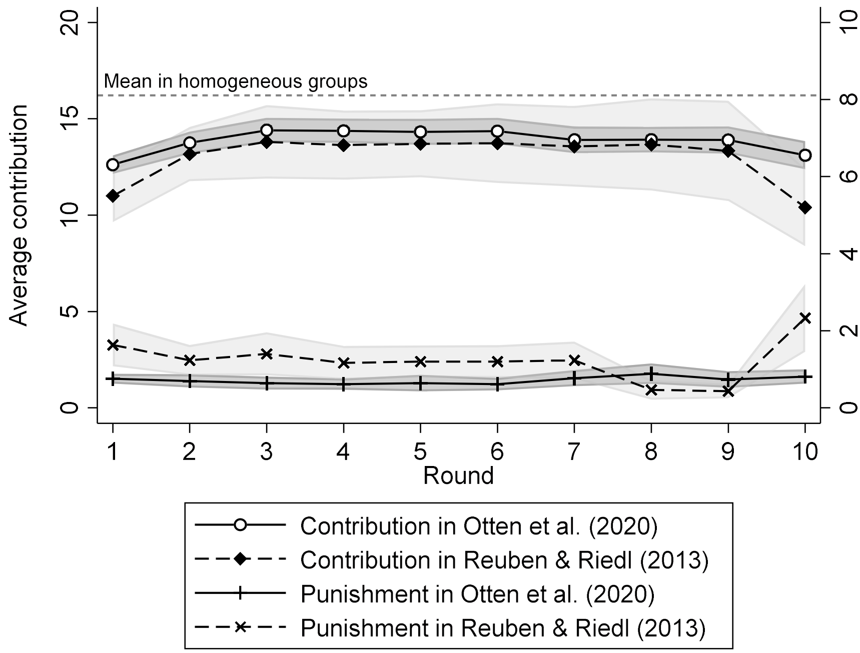


Figure A.2. Behavioral trends in Otten et al. (2020) and Reuben and Riedl (2013).
Note: N = 1920 (192 participants x 10 rounds) for Otten et al. (2020) and N = 300 (30 participants x 10 rounds) for Reuben and Riedl (2013). The main experimental parameter values were exactly the same in both studies, but Otten et al. (2020) measured norms before play whereas Reuben and Riedl (2013) did not. That the behavioral patterns are largely similar across the two studies suggests that the norm measurement did not affect behavior. 95% confidence intervals are added for all outcomes on the group-level with grey area shading. A dashed horizontal line is added displaying the average contribution level found in the homogeneous groups with peer punishment in Reuben and Riedl (2013). The contribution levels in the heterogeneous conditions are below that of the homogeneous condition. Otten et al. (2020) refers to the current manuscript, Reuben and Riedl (2013) refers to: Reuben, E. & Riedl, A. Enforcement of contribution norms in public good games with heterogeneous populations. *Games Econ. Behav.* **77**, 122–137 (2013).

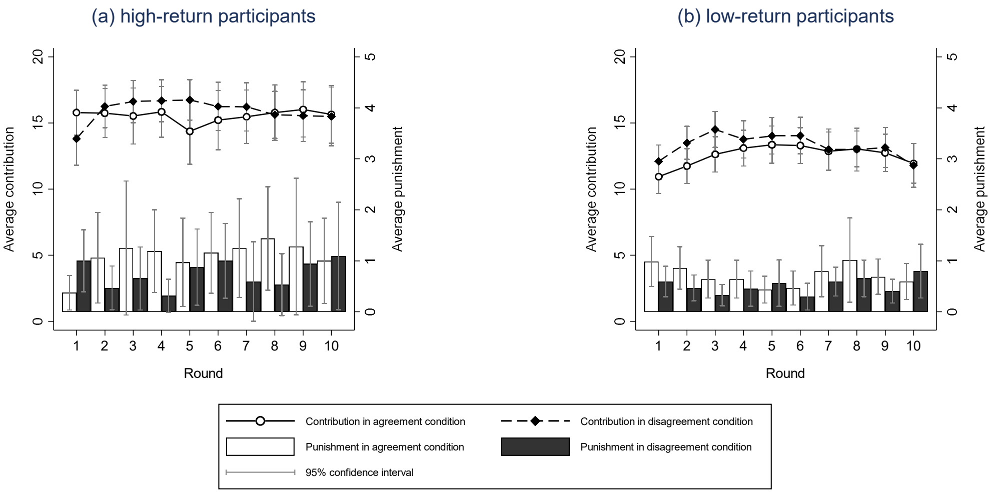


Figure A.3. Average contribution and punishment per round and condition and return-type.

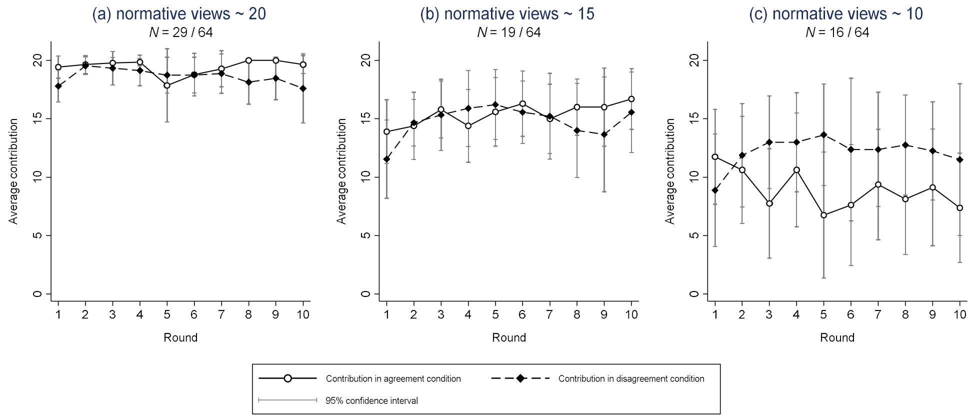


Figure A.4. Contributions by normative views of high-return participants.

Note: We examine contribution levels for high-return participants according to their first-measured normative views of what high-return participants should contribute, i.e., what they think would be appropriate contributions for themselves. We distinguish three levels: (1) normative views that prescribe contributions of ~ 20 (≥ 17.5), (2) normative views that prescribe contributions of ~ 15 (< 17.5 & ≥ 12.5), and (3) normative views that prescribe contributions of ~ 10 (< 12.5 & ≥ 7.5). 1 high-return participant reported that a contribution of 5 is appropriate for high-return participants. We did not think it useful to create a separate category for 1 participant, and therefore included the participant in the category with normative views ~ 10.

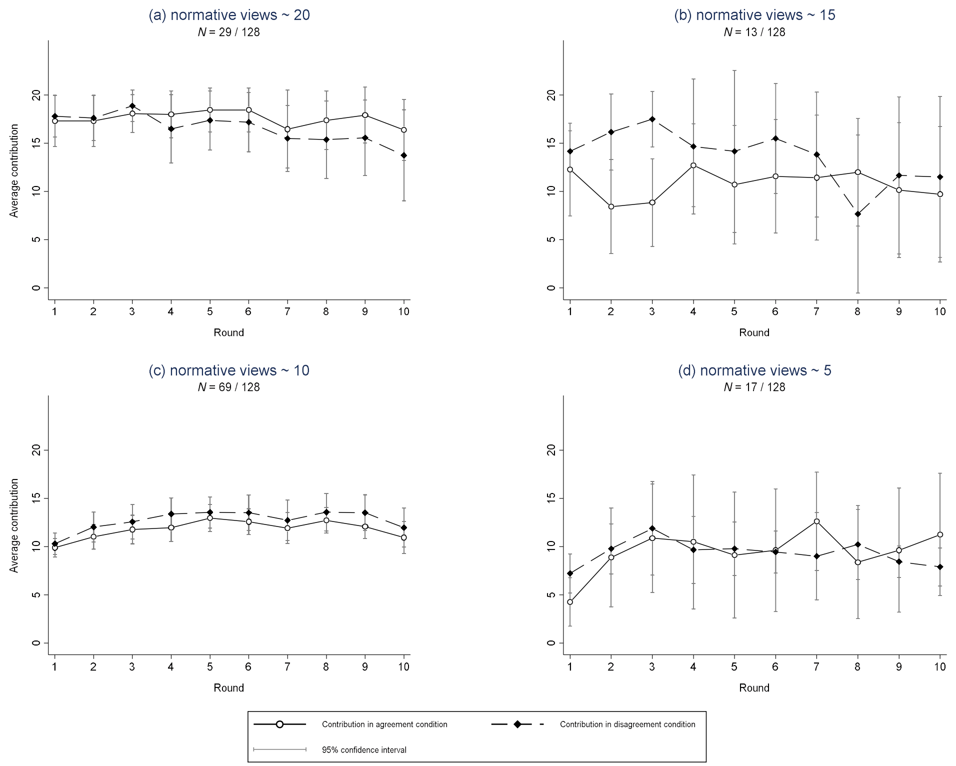


Figure A.5. Contributions by normative views of low-return participants.

Note: We examine contribution levels for low-return participants according to their first-measured normative views of what low-return participants should contribute, i.e., what they think what would be appropriate contributions for themselves. We distinguish four levels: (1) normative views that prescribe contributions of ~ 20 (≥ 17.5), (2) normative views that prescribe contributions of ~ 15 (< 17.5 & ≥ 12.5), (3) normative views that prescribe contributions of ~ 10 (< 12.5 & ≥ 7.5), and (4) normative views that prescribe contributions of ~ 5 (< 7.5).

A.2. Exploratory analyses

Because virtually all variation in contribution levels is within conditions, we turn to exploratory analyses to examine what might explain this variation. In the explanatory analyses, we conceptualized normative disagreement based on normative views about equal-contributions and equal-earnings as is common in related studies. We now take three main alternative approaches to assess how normative disagreement can affect cooperation in the PGG. First, some perspectives suggest that rather than to personal normative views (i.e., what participants themselves deem appropriate), participants respond to normative expectations (i.e., what participants think their group members deem appropriate). The corresponding prediction would be that disagreement in normative expectations, rather than personal views, will negatively affect the level of public good provision. Second, next to disagreement on how much high-return members should contribute relative to low-return members (equal-earnings vs equal-contributions), we examine disagreement on absolute contribution norms. That is, there are multiple absolute levels of contributions that may satisfy the rules of equal-earnings or equal-contributions, and people may disagree about the appropriate absolute level as well. Third, for normative disagreement to have the potential to reduce public good provision, norm conformity must be conditional on the conformity of others. An alternative perspective is that people conform to their normative views unconditionally, in which case the group-mean normative view, rather than disagreement on these views, predicts public good provision.

Altogether, we thus examine the dimensions of (1) normative views vs normative expectations, (2) absolute versus relative contribution rules, and (3) group-mean norms versus group-disagreement on norms. The combination of all three dimensions gives us 8 (2x2x2) potential ways in which norms may influence public good provision. We conduct a random-effects Tobit regression with each of the 8 norm-conceptualizations as independent variables and the group-mean contribution per round as the dependent variable. All variables are standardized, to allow for comparisons of their effect sizes. Figure A.6 presents the results of the analysis (see also Table A.3).

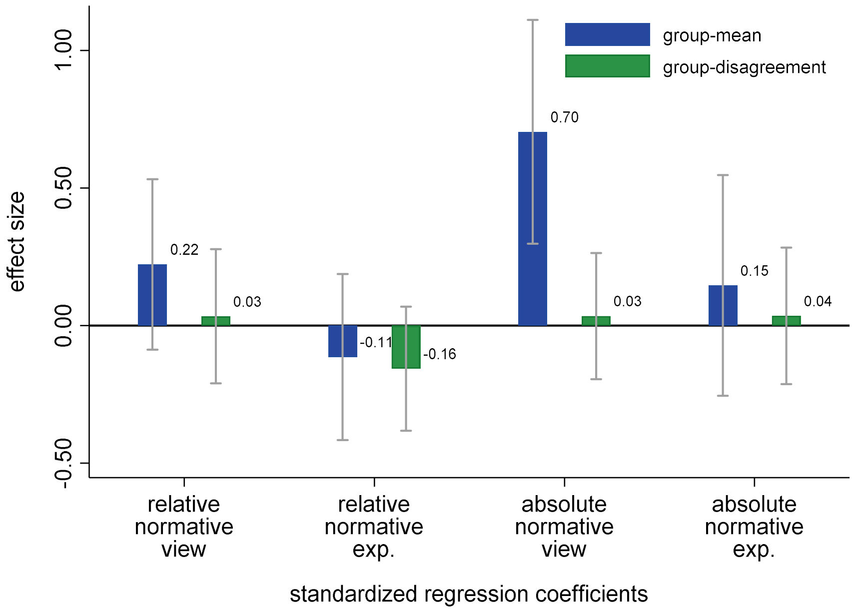


Figure A.6. Effects of normative views and expectations on public good provision.
Note: Per predictor, the mean group-score is the average of all 3 participants, and the disagreement score is the group's maximum score subtracted by the group's minimum. Relative normative view (and expectation) refers to the difference in the appropriate contribution between the high- and low-return members. Absolute normative view (and exp.) refers to the appropriate absolute level of contributions averaged over the low- and high-return members. The participants' normative views are what they themselves deem appropriate, the participants' normative expectations are what they expect their group members to deem appropriate. The 95% confidence intervals are included and Bonferroni adjusted for multiple testing. Full model specification is shown in Table A.3, column 5.

Figure A.6 corroborates the conclusion that normative disagreement between the relative contribution rules of equal-contributions and equal-earnings does not affect public good provision: group-disagreement on relative normative views does not significantly affect the group-mean contribution. Because there are multiple absolute levels of contributions that may satisfy the rules of equal-earnings or equal-contributions, people may disagree about the appropriate absolute level as well. In Figure 2.2a of the main-text, we indeed see that among supporters of each rule, there is variation in the absolute contribution levels considered appropriate. However, in Figure A.6 we see that such group-disagreement on the absolute normative views does not affect the contribution level. Similarly, whether disagreement is in terms of normative views or normative expectations also does not seem to matter: both group-disagreements on views and expectations are unrelated to public good provision. Thus, regardless of how we conceptualize normative disagreement, it does not seem to significantly affect public good provision.

Table A.3. Random-effects Tobit regression of mean public good provision.

	(1) relative normative view	(2) relative normative exp.	(3) absolute normative view	(4) absolute normative exp.	(5) all norms w/o controls	(6) all norms with controls
[mean] relative normative view	-.081				.222	.183
	(.138)				(.113)	(.109)
[dis] relative normative view	-.145				.034	.044
	(.138)				(.089)	(.087)
[mean] relative normative exp.		-.190			-.114	-.115
		(.127)			(.110)	(.105)
[dis] relative normative exp.		-.194			-.157	-.231*
		(.127)			(.083)	(.080)
[mean] absolute normative view			.809***		.704***	.666***
			(.085)		(.149)	(.148)
[dis] absolute normative view			-.014		.034	-.025
			(.084)		(.084)	(.083)
[mean] absolute normative exp.				.732***	.146	.130
				(.096)	(.147)	(.139)
[dis] absolute normative exp.				-.111	.035	-.024
				(.094)	(.091)	(.089)
[mean] social value orientation						-.137
						(.085)
[dis] social value orientation						-.132
						(.088)
[mean] age						.094
						(.076)
[mean] political orientation						-.131
						(.078)
proportion of males						.170
						(.080)
round	.013	.013	.013	.013	.013	.013
	(.008)	(.008)	(.008)	(.008)	(.008)	(.008)
constant	.033	.034	.035	.037	.035	.034
	(.137)	(.133)	(.094)	(.103)	(.089)	(.083)
sigma_u	1.017***	.987***	.638***	.722***	.586***	.528***
	(.095)	(.092)	(.062)	(.069)	(.058)	(.053)
sigma_e	.567***	.567***	.567***	.567***	.567***	.567***
	(.019)	(.019)	(.019)	(.019)	(.019)	(.019)
N observations	640	640	640	640	640	640
N groups	64	64	64	64	64	64
rho	.763	.752	.559	.619	.517	.464

Note: We conduct random-effects Tobit regressions with each of the norm-conceptualizations as independent variables (see main text) and the group-mean contribution per round as the dependent variable. For each variable, [mean] refers to the group-mean score and [dis] refers to the group-disagreement score. In all models, 1 observation is left-censored and 123 observations are right-censored. In columns 1-4, the group-mean and group-disagreement on the different norm-conceptualizations are tested in separate models. In column 5, all norm-conceptualizations are included within a single model, and in column 6 control variables are added. We control for the group composition on several variables measured at the end of the experiment: social value orientation, sex, age, and political orientation (self-reported from 1 = very left to 10 = very right). All variables are standardized (except round), to allow for comparisons of their effect sizes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Bonferroni-adjusted $p/8$, two-tailed tests). Standard errors in parentheses. As can be seen, the only variable strongly related to the group-mean contribution is the group-mean absolute normative view. The only exception is the group-mean absolute normative expectation in model 4, but that is because normative expectations are highly correlated with normative views as reported in the main text. The effect of group-mean normative expectations disappears when controlling for group-mean normative views.

Table A.4. OLS regression of mean public good provision.

	(1) relative normative view	(2) relative normative exp.	(3) absolute normative view	(4) absolute normative exp.	(5) all norms w/o controls	(6) all norms with controls
[mean] relative normative view	-.018				.230	.205
	(.117)				(.090)	(.098)
[dis] relative normative view	-.106				.048	.064
	(.102)				(.062)	(.059)
[mean] relative normative exp.		-.116			-.073	-.080
		(.105)			(.105)	(.096)
[dis] relative normative exp.		-.181			-.151	-.220
		(.122)			(.084)	(.089)
[mean] absolute normative view			.660***		.589***	.556***
			(.067)		(.125)	(.134)
[dis] absolute normative view			-.007		.042	-.010
			(.072)		(.068)	(.063)
[mean] absolute normative exp.				.594***	.123	.110
				(.076)	(.125)	(.119)
[dis] absolute normative exp.				-.092	.038	-.017
				(.076)	(.081)	(.080)
[mean] social value orientation						-.127
						(.085)
[dis] social value orientation						-.129
						(.078)
[mean] age						.078
						(.052)
[mean] political orientation						-.097
						(.074)
proportion of males						.154
						(.077)
round	.002	.002	.002	.002	.002	.002
	(.012)	(.012)	(.012)	(.012)	(.013)	(.013)
constant	-.012	-.012	-.012	-.012	-.012	-.012
	(.112)	(.114)	(.067)	(.072)	(.068)	(.073)
N observations	640	640	640	640	640	640
N groups	64	64	64	64	64	64
R2	.010	.051	.437	.362	.495	.543

Note: We conduct ordinary least squares (OLS) regression with each of the norm-conceptualizations as independent variables (see main text) and the group-mean contribution per round as the dependent variable. For each variable, [mean] refers to the group-mean score and [dis] refers to the group-disagreement score. We account for repeated measures within groups by estimating cluster-robust standard errors. In columns 1-4, the group-mean and group-disagreement on the different norm-conceptualizations are tested in separate models. In column 5, all norm-conceptualizations are included within a single model, and in column 6 control variables are added. We control for the group composition on several variables measured at the end of the experiment: social value orientation, sex, age, and political orientation (self-reported from 1 = very left to 10 = very right). All variables are standardized (except round), to allow for comparisons of their effect sizes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Bonferroni-adjusted $p/8$, two-tailed tests). Standard errors in parentheses.

Table A.5. Population-averaged regression of mean public good provision.

	(1) relative normative view	(2) relative normative exp.	(3) absolute normative view	(4) absolute normative exp.	(5) all norms w/o controls	(6) all norms with controls
[mean] relative normative view	-.047				.191	.161
	(.109)				(.089)	(.086)
[dis] relative normative view	-.110				.034	.044
	(.109)				(.070)	(.069)
[mean] relative normative exp.		-.112			-.049	-.052
		(.100)			(.086)	(.083)
[dis] relative normative exp.		-.157			-.137	$-.202^{*}$
		(.100)			(.065)	(.063)
[mean] absolute normative view			$.640^{***}$		$.568^{***}$	$.537^{***}$
			(.067)		(.116)	(.117)
[dis] absolute normative view			.007		.047	-.007
			(.067)		(.066)	(.065)
[mean] absolute normative exp.				$.573^{***}$	.124	.114
				(.073)	(.114)	(.109)
[dis] absolute normative exp.				-.077	.045	-.011
				(.073)	(.070)	(.069)
[mean] social value orientation						-.119
						(.067)
[dis] social value orientation						-.107
						(.070)
[mean] age						.096
						(.060)
[mean] political orientation						-.109
						(.061)
proportion of males						.137
						(.063)
round	.011	.011	.010	.010	.009	.009
	(.016)	(.016)	(.014)	(.014)	(.013)	(.013)
constant	-.198	-.194	-.150	-.161	-.138	-.129
	(.134)	(.131)	(.101)	(.108)	(.095)	(.090)
N observations	640	640	640	640	640	640
N groups	64	64	64	64	64	64

Note: We conduct population-averaged regressions with each of the norm-conceptualizations as independent variables (see main text) and the group-mean contribution per round as the dependent variable. For each variable, [mean] refers to the group-mean score and [dis] refers to the group-disagreement score. In columns 1-4, the group-mean and group-disagreement on the different norm-conceptualizations are tested in separate models. In column 5, all norm-conceptualizations are included within a single model, and in column 6 control variables are added. We control for the group composition on several variables measured at the end of the experiment: social value orientation, sex, age, and political orientation (self-reported from 1 = very left to 10 = very right). All variables are standardized (except round), to allow for comparisons of their effect sizes. The within-subject working correlation matrix is autoregressive of order 1. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Bonferroni-adjusted $p/8$, two-tailed tests). Standard errors in parentheses.

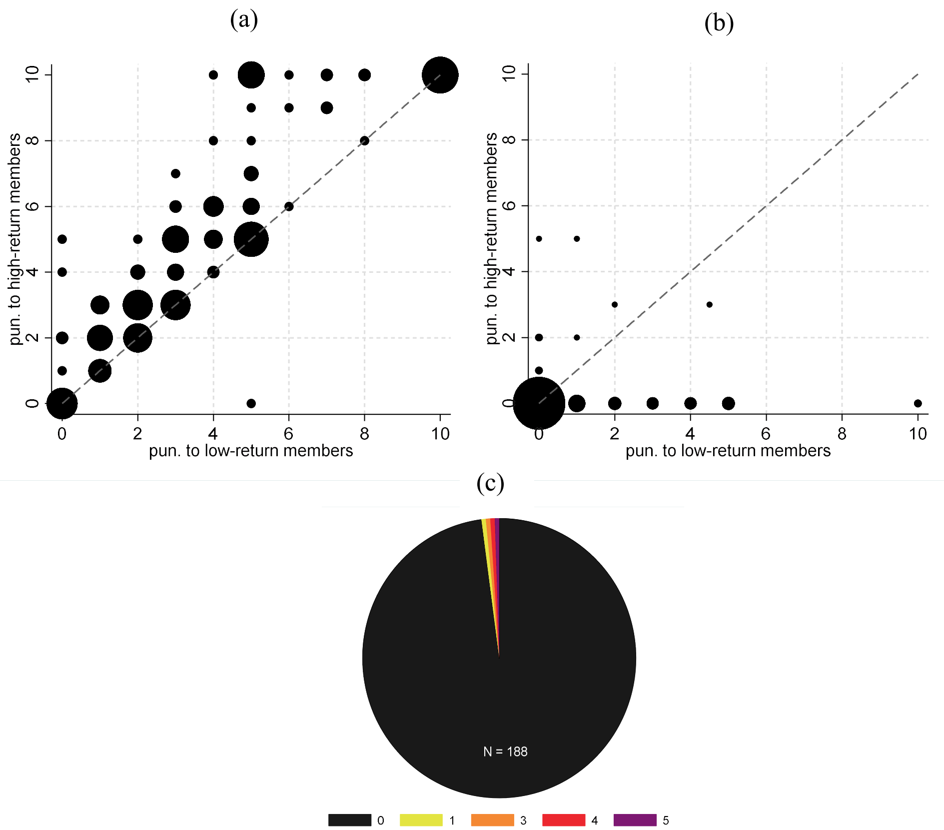


Figure A.7. Normative views on punishment of free-riders (a), adherents to equal-earnings norm (b), and unconditional prosocials (c).
Note: Participants were asked to report on their normative views and expectations of punishment in two hypothetical scenarios after completion of the experiment. In both scenarios, one unit of punishment costs the punisher one MU and reduces the punished member's income by 3 MUs, as in the actual experiment. In the first scenario, 1 low-return member contributed all 20 points and the two other members (one low- and one high-return member) contributed 0. These last two members are the free-riders, and the normative views on how much they should be punished are plotted in panel (a). Appropriate number of punishment points are plotted on the x-axis for the low-return members and on the y-axis for the high-return member. The member who contributed 20 in this scenario is the unconditional prosocial. The distribution of normative views on the appropriate punishment for this member is presented via a pie chart in panel (c). In the second hypothetical scenario, the high-return member contributed 20 points and the two low-return members each contributed 10 points, which leads to equal earnings for all. The normative views on the appropriate punishment for these low- and high-return members who adhere to the equal-earnings norm are presented in panel (b). Appropriate number of punishment points are plotted on the x-axis for the low-return members and on the y-axis for the high-return member. Because normative views and expectations on punishment are highly correlated, we only present the normative views in all three panels. We see that, while most participants agree that some level of punishment is appropriate for free-riders (although there is a large variation in how much is considered appropriate), most participants agree that no punishment is appropriate for adherents to the equal-earnings norm. Yet, when asking about contribution norms, only a minority of participants reports that equal-earnings is the appropriate norm (the majority balances equal-contributions and equal-earnings, see Figure 2a in main text). This suggests that participants are tolerant of behaviors that conform to other normative views, as long as these normative views are to some extent prosocial. A norm of free-riding is largely condemned, but the equal-earnings norm is tolerated even by those who do not subscribe to the equal-earnings norm themselves. Punishment of members who contribute fully while others are free-riding is virtually never supported.

A.3. Experimental instructions

Experimental Laboratory for Sociology and Economics

Welcome

Welcome to this experiment and thank you for coming. Please read the following instructions carefully. These instructions are the same for all participants. The instructions state everything you need to know in order to participate in the experiment. If you have any questions, please raise your hand. One of the experimenters will approach you and answer your question.

The experiment is about group decision making. You can earn money by means of earning points during the experiment. The number of points that you earn depends on your own choices, the choices of other participants in your group, and chance. At the end of the experiment, the total number of points that you earned will be exchanged at a rate of:

70 points = 1 Euro

The money you earn will be rounded up to whole euros and paid out in cash at the end of the experiment. There is a minimum payment of 5 euros, and a maximum payment of 23 euros. Other participants will not see how much you have earned. During the experiment you are not allowed to communicate with other participants. Please turn off your mobile phone. You may only use functions on the computer screen that are necessary to carry out the experiment.

First, we introduce the decision situation in which you will interact. You will learn about the procedure of the experiment later. **A decision situation consists of 2 stages: a contribution stage followed by a review stage.** In the contribution stage, you decide how many points you contribute to a group account. In the review stage, you learn how much the other members of your group contributed to the group account. We will first explain the contribution stage.

Contribution stage

You are a member of a group of **3 participants**. You and the two other members of your group are **each given 20 points**. Each of you can **choose how many points to keep for yourself in a private account and how many points to contribute to a group account**.

Your points from the private account
You will earn 1 point for each point you keep in your private account.
For example, if you keep all 20 points into your private account (and therefore do not contribute to the group account), your income will amount to exactly 20 points out of your private account. If you keep 6 points into your private account, your income from this account will be 6 points. **No one except you earns something from your private account.**

Your income from the group account
Each group member will profit from points you contribute to the group account. You will also profit from the other group members' contributions. Just like in real life, some persons profit more from contributions to the group account than others.

For each point contributed to the group account (by you and the other members):
1 member earns 0.75 points and 2 members earn 0.50 points each.

Whether you are a member with a return of 0.50 or 0.75 from the group account will be randomly determined at the start of the experiment, **and will stay the same for the entire duration of the experiment**.

For example, if the 3 members combined contribute in **total 40 points to the group account**,

1 member receives: **0.75 times 40 = 30 points** from the group account,
2 members each receive: **0.50 times 40 = 20 points** from the group account.

<u>Your total income from the private account and group account</u>

Each member can choose any number of points to contribute to the group account, from 0 to 20 points. Every point a member does not contribute to the group account will automatically remain in his/her private account. **Each member's total income from the contribution stage is the combined income from his/her private account and the group account.**

Table 1 gives an arbitrary example of how each member's income from the private account, group account, and the total income are calculated when the total contributions to the group account are 40 (15+15+10).

Table 1 – example

Member	Return	Contribution	Private account income	Group account income	Total income
A	0.75	15	5	30	35
B	0.50	15	5	20	25
C	0.50	10	10	20	30

Review stage

Each contribution stage is followed by a review stage. In the review stage, everyone in the group will see how much each of the other group members contributed to the group account as well as their income from the contribution stage. Then, all group members have a chance to **decrease** the income of each other group member. You can decide if you want to spend points to decrease the income of the other two group members, for example because you disagree with how much they contributed or earned.

If you want to decrease another member's income you do that by assigning deduction points. **Every deduction point assigned to another group member reduces his/her income by 3 points, and your own income by 1 point**. Similarly, every deduction point that one of your group members assigns to you decreases your income by 3 points and costs the group member 1 point. Note that this might imply that you or other participants lose income in a particular round. If you do not want to decrease the income of a group member, you must assign him/her

0 deduction points. Every participant can assign up to a maximum of 10 deduction points to each group member, regardless of the income from the contribution stage.
For example, if you assign 2 deduction points to a group member this costs you 2 points and reduces the group member's income by 6 points (2 times 3). Another example: if one of your group members assigns 3 deduction points to you, this reduces the group member's income by 3 points and your income by 9 points (3 times 3).

After everyone has made a decision, you will see how many deduction points were assigned to you by the other group members and also what your total income for the round is. You will not see which individual participant assigned deduction points to you, you can only see the total number of deduction points assigned to you and how that affected your income. Similarly, if you assigned deduction points to one or more of your group members, they will not see that you are the one who assigned the points.

Overview of the Session

The experiment consists of **2 parts**, and in total lasts about **1 hour and 45 minutes**.

In the 1st part you will play 10 rounds of the decision situation (10 contribution and 10 review stages).

Before you play these 10 rounds, we will first ask you to answer some questions about the decision situation. These questions concern:

- your understanding of the decision situation,
- your view on the appropriate amount that each group member should contribute to the group account,
- your guess of what the other participants think are appropriate contributions.

Some questions appear multiple times throughout the experiment. You do not have to be consistent with your answers to these questions. Your answers may or may not have changed during the experiment. Similarly, what you view as appropriate contributions may or may not be the same as what the other participants think are appropriate contributions.

After this 1st part in which you answer questions about the decision situation and play 10 rounds of it, **you will receive new instructions on your computer screen for the 2nd part of the experiment.** The 2nd part of the experiment is of similar length to the 1st part.

Because you play together with other persons, you will sometimes have to wait until the other persons have made their decision. These waiting times are incorporated in the total expected duration of 1 hour and 45 minutes for the experiment.

Appendix B. Supplementary material for Chapter 3

B.1. Distribution and disagreement of normative views

In one condition, groups are in agreement before membership change and in disagreement after membership change. In the other condition, groups are in disagreement before membership change and in agreement after membership change. When we present data on the whole experiment (before and after membership change), we will refer to the former condition as agreement-disagreement and the latter condition as disagreement-agreement. When we only present data on the experiment after membership change, we will simply use the latter part of the labels (so agreement-disagreement becomes disagreement, and disagreement-agreement becomes agreement).

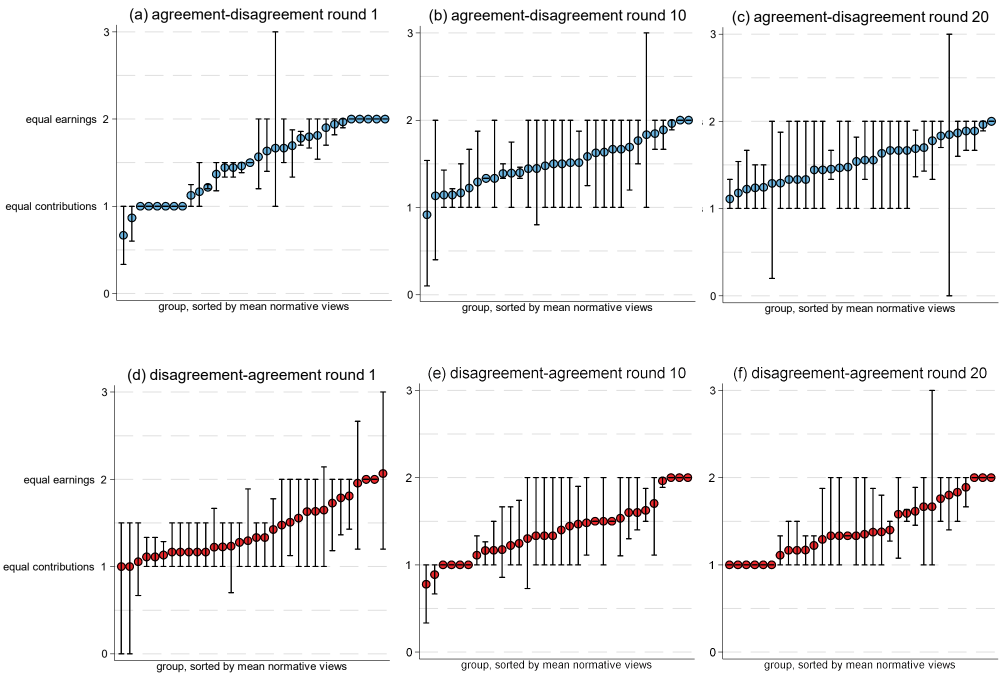


Figure B.1. Distribution of normative views.

Note: Each participant reported on the appropriate contribution that a high-return member should make and the appropriate contribution that each of the low-return members should make. We divide the participant's appropriate contribution for the high-return member by the participant's average appropriate contribution for the low-return members to achieve a ratio that indicates the participant's position on the equal-contributions to equal-earnings spectrum on the y-axis. With the return distribution in our experiment, high-return members should contribute twice as much as low-return members to equalize earnings. Participants supporting equal-earnings therefore have a ratio of 2, while participants supporting equal-contributions have a ratio of 1. Each circle represents a group, and the groups are sorted on the x-axis based on their mean ratio. The circle provides the group-mean ratio, and the capped spikes provide the range between the group-min and group-max ratio (i.e., the extent of group-disagreement). There are 32 groups in each condition, and we show the ratios for both conditions at the three measurement moments, i.e., before round 1, before round 11, and after round 20. We see that almost all group-average normative views fall within the spectrum of equal-contributions (y = 1) to equal-earnings (y = 2). Both before and after membership change (first measurement vs second and third measurement), there is considerable variation between groups in the average normative views; several groups support equal-contributions, several groups support equal-earnings, and several groups support a balance between these two rules. Comparison of the capped spikes in Figure B.1a and Figure B.1d corroborates that within-group normative disagreement is larger in condition disagreement-agreement before membership change. Comparison of the capped spikes in Figure B.1b-c and Figure B.1e-f corroborates that within-group normative disagreement is larger in condition agreement-disagreement after membership change.

B.2. Experimental outcomes before and after the membership change

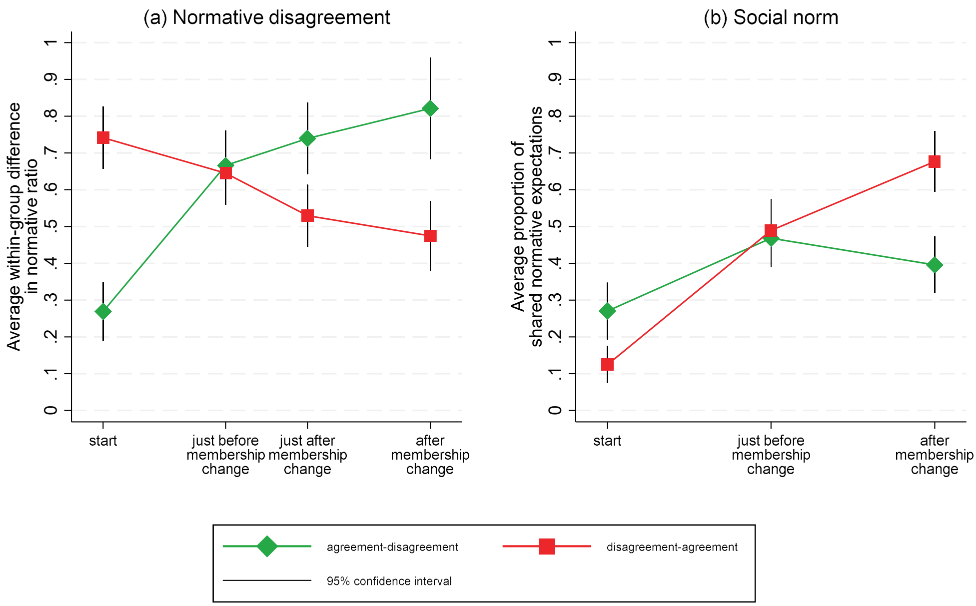


Figure B.2. Normative disagreement and social norms.

Note: normative views and expectations are measured before the start of round 1, after round 10 (before membership change), and after round 20. The measurement of normative views does not ask about one's current group members, but the measurement of normative expectations does (i.e., one has to predict the views of one's current group members). This means that normative views measured after round 10 can be used for assessing disagreement in groups both before the membership change and after the membership change, as we do in panel (a), while normative expectations measured after round 10 can only be used for assessing similarity among the current group members, i.e., before membership change, as we do in panel (b).

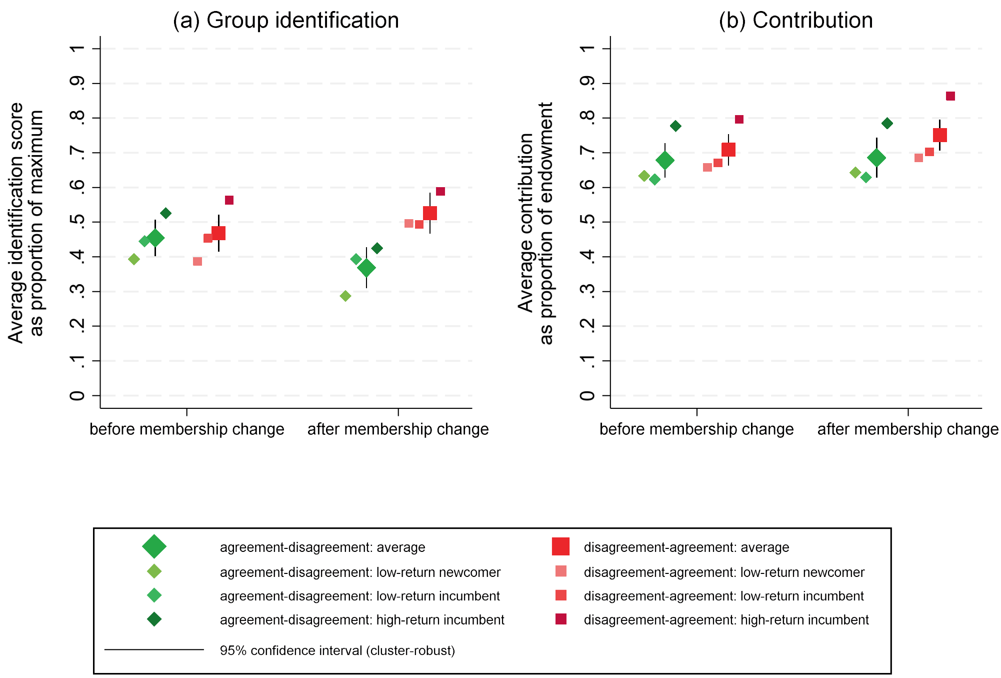


Figure B.3. Group identification and contribution.

B.3. The effect of disagreement on contributions

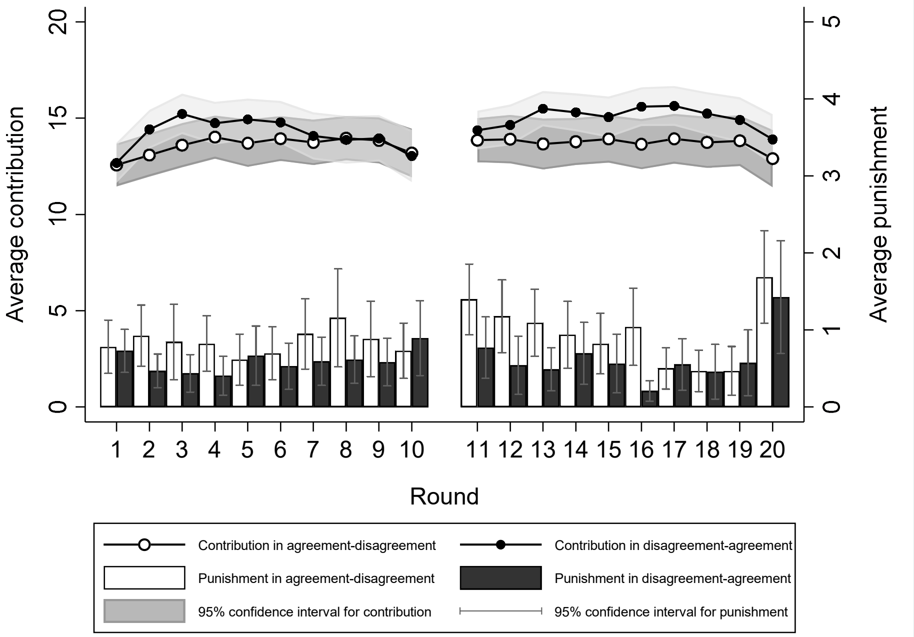


Figure B.4. Average contribution and punishment per round and condition.
Note: We see that both before (rounds 1-10) and after membership change (rounds 11-20), there is hardly a difference in contribution levels between conditions. Although in rounds 11-20, the contribution level seems to be slightly lower in condition agreement-disagreement (which is consistent with our hypothesis that contribution levels are lower when newcomers and incumbents are in normative disagreement), the confidence intervals of both conditions overlap in all rounds, and the difference is already discernible before membership change. Average punishment points received per round and condition are also included (for each participant the punishment is the combined punishment points received from both other group members).

Table B.1. Population-averaged model tests of hypothesis.

	(1) ind. all rounds	(2) ind. first rounds	(3) ind. end rounds	(4) group. all rounds	(5) group. first rounds	(6) group. end rounds
Condition & time period						
a. agreement before membership change	13.652***	12.562***	13.188***	13.561***	12.563***	13.187***
	(.428)	(.540)	(.679)	(.689)	(.582)	(.905)
b. disagreement after membership change	14.181***	13.854***	12.896***	13.715***	13.854***	12.896***
	(.453)	(.540)	(.679)	(.689)	(.582)	(.905)
c. disagreement before membership change	14.168***	12.677***	13.031***	14.170***	12.677***	13.031***
	(.428)	(.540)	(.679)	(.689)	(.582)	(.905)
d. agreement after membership change	15.041***	14.375***	13.896***	15.014***	14.375***	13.896***
	(.453)	(.540)	(.679)	(.689)	(.582)	(.905)
Hypothesis						
(c – d) – (a – b) < 0	-.344	-.406	-1.156	-.691	-.406	-1.156
chi2(1)	.881	.332	1.218	5.364	.332	.687

Note: We take the contribution decision as the dependent variable and as independent variable a factor indicating whether the decision was made in the condition that moves from agreement to disagreement after membership change (from **a** to **b**) or the condition that moves from disagreement to agreement after membership change (from **c** to **d**). This allows us to estimate whether the change in contribution levels before and after membership change differs significantly by experimental condition. Across six models, we vary whether the contribution level is on the individual-level (models 1-3) or group-level (models 4-6) and whether we include as observations all rounds (models 1 & 4), only the first rounds (round 1 before membership change and round 11 after membership change, models 2 & 5), or only the last rounds (round 10 before membership change and round 20 after membership change, models 3 & 6). Regardless of which model is used, we find no significant difference in the change in contribution levels before and after membership change between conditions according to conventional standards: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (Bonferroni-adjusted $p/6$, two-tailed tests). Coefficients are predictive margins. Standard errors in parentheses. The within-subject working correlation matrix is unstructured in all models except for model 4, where results do not converge with the unstructured matrix and the exchangeable matrix is used instead.

Table B.2. Conditional test of the effect of normative disagreement on change in contributions after membership change.

	Model 1	**Model 2**
Contribution level prior to membership change	-0.58^{***}	-0.28^{**}
	(0.14)	(0.10)
Disagreement: experimental condition	-8.90^{**}	
	(2.77)	
Disagreement: experimental condition ×	0.58^{**}	
Contribution level prior to membership change	(0.19)	
Disagreement: group-level change		-5.86
		(4.09)
Disagreement: group-level change ×		0.40
Contribution level prior to membership change		(0.27)
Intercept	8.99^{***}	4.30^{**}
	(2.03)	(1.47)
R^2	.23	.13

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. N = 64 groups. **Model 1:** The results suggest that normative agreement has a positive effect in low-contributing groups. For groups with an average contribution of 0 before membership change, being in the condition with agreement among newcomers and incumbents is estimated to lead to an 8.90 higher average contribution after membership change (8.99 for condition agreement vs 8.99 – 8.90 = 0.09 for condition disagreement). The opposite does not hold: normative disagreement among newcomers and incumbents does not harm contributions in groups that had high contribution levels before membership change. For example, for groups with an average contribution of 15 before membership change, disagreement is estimated to lead to an -8.90 + (15 × 0.58) = 0.2 lower average contribution, which is insignificant (p = .80). Also for groups with an average contribution of 20 before membership change, disagreement is not significantly related to a change in contribution (-8.90 + (20 × 0.58) = 2.70, p = .06). **Model 2:** here change in disagreement is measured not by experimental condition but by taking the difference between the group-level disagreement score (group max normative ratio – group min normative ratio; see main-text) just before membership change and just after membership change. We use the norm measurement after round 10, as this is the most up-to-date normative view before membership change. To measure group-disagreement before membership change, we compare this normative view among members in the 'old' group compositions. To measure group-disagreement after membership, we compare this normative view among members in the new group composition. Because we use the same norm measurement moment for comparing the group composition before and after membership change, the change in norm composition cannot be related to participants changing their normative views, i.e., it is purely a compositional change induced by the newcomer replacing an incumbent. The results using this variable suggest no conditional effect of a change in disagreement on a change in contribution levels; the change in group-level disagreement and its interaction with the group's prior contribution level are both insignificant.

Table B.3. The change in contributions after newcomer entry by the change in group-level disagreement and normative views.

	Model 1	Model 2
Change in group-average of:		
Participants' average normative view	1.06^{***}	0.92^{**}
	(0.32)	(0.33)
Participants' view of equal-contributions vs equal-earnings	-0.02	-0.24
	(2.15)	(2.20)
Change in group-disagreement on:		
Participants' average normative view	0.04	0.07
	(0.13)	(0.13)
Participants' view of equal-contributions vs equal-earnings	-0.60	-0.10
	(0.94)	(0.98)
Change in group-average punishment		-0.35
		(0.50)
Change in group-average identification		0.31
		(0.35)
Intercept	0.46	0.51
	(0.39)	(0.39)
R^2	.29	.32

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. $N = 64$ groups. The outcome variable measures how much the group-average contribution changes after newcomer entry (i.e., the group-average contribution in the second 10 rounds minus the group-average contribution in the first 10 rounds). **Model 1.** The predictors measure how much the group-average and group-disagreement (group-max – group-min) of the participants' average normative view and the participants' view on equal-contributions vs equal earnings change after newcomer entry. A participant's average normative view measures what a participant deems to be the appropriate average contribution in the group. For example, if a participant thinks the high-return member should contribute 20 and the two low-return members should contribute 10, the participant's average normative view would be (20+10+10)/3=13.33. A participant's view on equal-contributions vs equal-earnings measures the ratio of how much more the participant thinks high-return members should contribute than low-return members. For example, if a participant thinks the high-return member should contribute 20 and the two low-return members should contribute 10, the participant's ratio would be 20/10=2. To compare the group-composition (group-average and group-disagreement) on normative views before and after membership change, we use the norm measurement after round 10, as this is the most up-to-date normative view before membership change. If the newcomer's average normative view measured after round 10 is higher than the replaced incumbent's average normative view measured after round 10, the group-average normative view is increased because of the membership replacement. Because we use the same norm measurement moment for comparing the group composition before and after membership change, the change in norm composition cannot be related to participants changing their normative views, i.e., it is purely a compositional change induced by the newcomer replacing an incumbent. We find that the change in the group-average normative view that is brought about by the membership replacement is an important factor in predicting whether the membership replacement increases or decreases the group-average contribution. For every point increase in the group-average normative view, the group-average contribution is estimated to increase by 1.06. We find no relationship between the change in group-level disagreement and change in group-average contributions, neither when examining average normative views nor when examining views on equal-contributions vs equal-earnings. **Model 2.** We additionally add the change in group-average punishment (i.e., the group-average punishment in the second 10 rounds minus the group-average punishment in the first 10 rounds) and the change in group-average identification (i.e., the group-average identification as measured after the second 10 rounds minus the group-average identification as measured after the first 10 rounds). Neither of these variables is significantly related to the change in the group-average contribution while the change in group-average normative views remains a strong and significant predictor.

B.4. Group identification

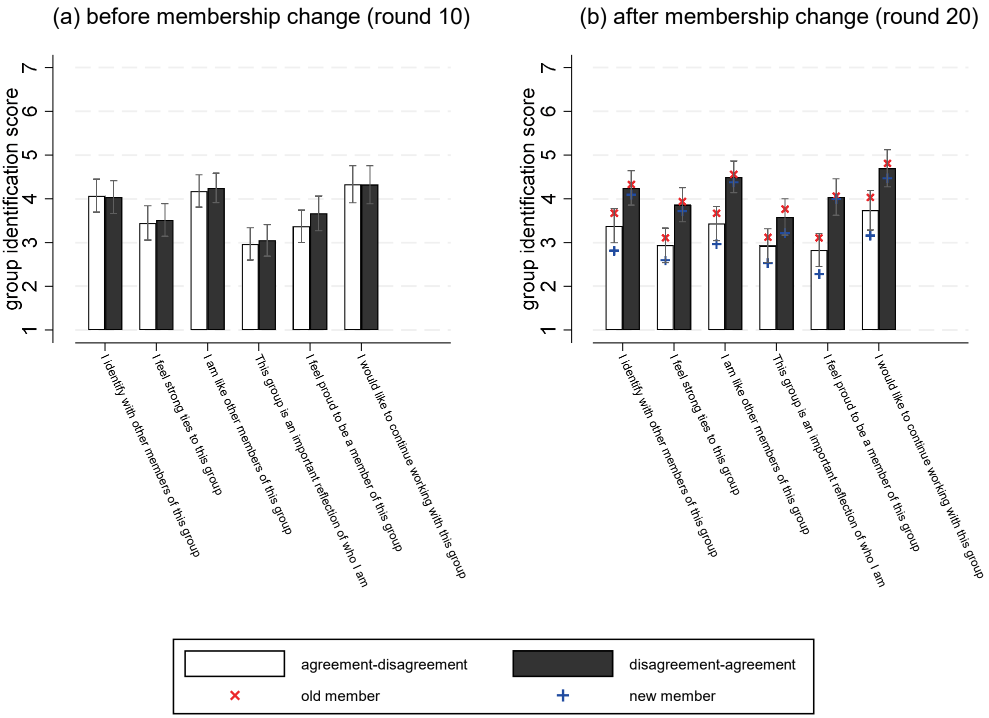


Figure B.5. Group identification by condition before and after membership change for each item.

Note: We see that there are no differences by condition in the scores on the group identification items before membership change (panel a). We see that there is a consistent difference by condition in the scores on the group identification items after membership change (panel b), with lower scores for condition agreement-disagreement. We furthermore see that new members score lower on the group identification items than old members. The 95% confidence intervals are included via capped spikes.

Table B.4. Group identification by group-level disagreement.

	Individual-level			Group-level		
	1	**2**	**3**	**4**	**5**	**6**
Group disagreement	-0.70***	-0.59**	-0.47*	-0.70*	-0.56*	-0.37
	(0.21)	(0.21)	(0.20)	(0.27)	(0.26)	(0.23)
Average contribution		0.07**	0.05*		0.09*	0.08*
		(0.02)	(0.02)		(0.04)	(0.03)
Average punishment			-0.45***			-0.64***
			(0.10)			(0.14)
Intercept	4.13***	3.05***	3.56***	4.13***	2.77***	3.26***
	(0.18)	(0.42)	(0.42)	(0.24)	(0.60)	(0.53)
R^2	0.06	0.09	0.18	.10	0.18	0.39

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. $N = 192$ participants for Models 1-3 and $N = 64$ groups for Models 4-6. We relate group identification (min=1, max=7, see main-text) to group-level disagreement (see main-text) in two types of models and assess whether the relationship remains when adding average contribution and average punishment as predictors. In the individual-level models, group identification is the participant's individual score on group identification measured after the second set of 10 rounds and the average contribution and punishment are the participant's average contribution and average received punishment over the second set of 10 rounds. In the group-level models, group identification is the group's average score on group identification measured after the second set of 10 rounds and the average contribution and punishment are the group's average contribution and average received punishment over the second set of 10 rounds. In both types of models, group-level disagreement is the group's disagreement (group max normative ratio – group min normative ratio; see main-text) measured after the second set of 10 rounds. We see that group-level disagreement has a significant negative relationship with group identification in five out of six models. The relationship largely remains when controlling for contribution. The relationship does become weaker (and insignificant in Model 6) when controlling for punishment, which suggests that the relationship between disagreement and group identification runs partly via punishment. Punishment is negatively related to group identification, the contribution is positively related to group identification.

B.5. Punishment

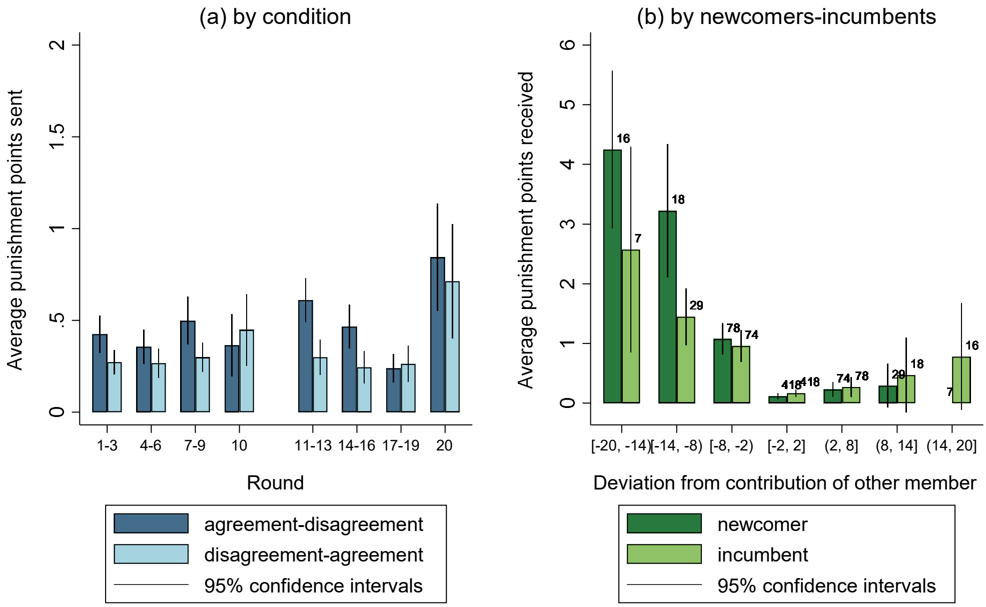


Figure B.6. Punishment by condition and newcomer-incumbent division.

Note: **(a)** The number of observations per round is 192, leading to a total of 3840 observations (20 rounds x 192 participants). We see that before membership change (rounds 1-10), there are no discernible differences between conditions in punishment levels. After membership change, punishment is considerably higher in condition agreement-disagreement in the early rounds (11-16). However, the difference disappears for the final rounds (17-20). **(b)** We examine punishment as a function of how much the newcomer deviates from the low-return incumbent's contribution and vice versa. We only compare the contributions of the low-return newcomer with those of the low-return incumbent, i.e., we leave out the contributions of high-return incumbents. This is to prevent confounding of newcomer-incumbent differences with return-rate differences, as we know that many normative views prescribe higher contributions for high-return members. The numbers above the bars indicate the number of observations underlying the bars. For example, newcomers deviated between -20 and -14 points from the contribution of the low-return incumbent in 16 cases. We see that newcomers are more strongly punished for negatively deviating from the incumbents' contribution than the other way around. This holds for large deviations (more than 8 contribution points), but not for smaller deviations (between 8 and 2 points deviation).

Table B.5. Mann-Whitney tests for punishment by condition and incumbent/newcomer status.

Test	Number of observations	z	p
(1) Punishment in rounds 11-13 by condition	1152 (192 participants x 3 rounds x two co-members to punish	5.53	<.001
(2) Punishment of low-contributors (≤25%) by newcomer/incumbent	326 (newcomers contributed ≤25% 186 times in the last 10 rounds, incumbents contributed ≤25% 140 times in the last 10 rounds)	3.86	<.001
(3) Punishment of low-contributors (≤25%) by newcomer/incumbent in disagreement condition	222 (newcomers contributed ≤25% 120 times in the last 10 rounds, incumbents contributed ≤25% 102 times in the last 10 rounds)	3.76	<.001
(4) Punishment of low-contributors (≤25%) by newcomer/incumbent in agreement condition	104 (newcomers contributed ≤25% 66 times in the last 10 rounds, incumbents contributed ≤25% 38 times in the last 10 rounds)	1.37	.17

Note: **(1)** we find that punishment levels are significantly higher in rounds 11-13 in condition disagreement than in condition agreement. **(2)** We find that low-contributors are punishment significantly more if they are (low-return) newcomers instead of (low-return) incumbents. **(3)** We find that low-contributors are punished significantly more if they are (low-return) newcomers instead of (low-return) incumbents in condition disagreement. **(4)** We find that low-contributors are not punished significantly more if they are (low-return) newcomers instead of (low-return) incumbents in condition agreement.

Table B.6. Punishment by group-level disagreement.

	Individual-level			Group-level		
	1	**2**	**3**	**4**	**5**	**6**
Group disagreement	-0.02	-0.19	-0.30	-0.02	-0.08	-0.29
	(0.44)	(0.38)	(0.39)	(0.49)	(0.49)	(0.48)
Round	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)	(0.05)
Group disagreement	0.02	0.02	0.02	0.02	0.02	0.02
× Round	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Contribution		-0.10^{***}	-0.09^{***}		-0.04	-0.01
		(0.02)	(0.02)		(0.02)	(0.02)
Group identification			-0.18^{***}			-0.36^{***}
			(0.05)			(0.09)
Intercept	0.87	2.43^{***}	3.02^{***}	0.87	1.46	2.52^{**}
	(0.47)	(0.56)	(0.56)	(0.62)	(0.74)	(0.72)
R^2	0.01	0.08	0.10	0.02	0.03	0.12

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. Results are for the second set of 10 rounds. $N = 1920$ (192 participants × 10 rounds) for Models 1-3 and $N = 640$ (64 groups × 10 rounds) for Models 4-6. We account for repeated measures within participants in Models 1-3 by estimating participant cluster-robust standard errors. We account for repeated measures within groups in Models 4-6 by estimating group cluster-robust standard errors. We relate punishment to group-level disagreement and add controls for contribution and group identification. Because results using the experimental condition suggest that punishment is higher with more disagreement only in the initial rounds (see main-text), we interact the group-level disagreement with the round. In the individual-level models, punishment is the participant's total punishment received per round, contribution is the participant's contribution per round, and group identification is the participant's group identification score as measured after the second set of 10 rounds. In the group-level models, punishment is the group's average punishment received per round, contribution is the group's average contribution per round, and group identification is the group's average score on group identification as measured after the second set of 10 rounds. The analyses using group-level disagreement suggest no significant relationship between disagreement and punishment, also not in the early rounds after membership change (given the insignificant interaction between group-level disagreement and round).

B.6. Predicting contribution levels

To examine what alternatively predicts contribution levels after membership change, and how that differs between newcomers and incumbents, we use ordinary least square (OLS) regression models. We account for repeated measures within participants by estimating cluster-robust standard errors. In Model 1 of Table B.7, we run a regression with the contribution decision as dependent variable and the participants' own normative view (as measured just before membership change) and the lagged average contribution of their group members as predictors. The normative view was shown to be an important predictor of behavior in the study on the first part of the experiment (Otten et al., 2020) and the group members' average contribution is a main predictor of behavior in research on public goods games (Chaudhuri, 2011). Recall that there is one low-return newcomer, one low-return incumbent, and one high-return incumbent in every group. We only compare the contributions of the low-return newcomer with those of the low-return incumbent, i.e., we leave out the high-return incumbents. This is to prevent confounding of newcomer-incumbent differences with return-rate differences (but we will show that results also hold when comparing the newcomer to the high-return incumbent).

We see that the two predictors, the participants' own normative view and the lagged contribution of their group members, together explain more than half of the variation in contribution levels, for both newcomers and incumbents. We see that newcomers are more influenced by the contribution of others than incumbents. A one-point increase in the contribution of others increases one's own subsequent contribution by .77 points for newcomers and by .53 for incumbents (Wald test for difference between incumbents and newcomers, $p = .001$). We furthermore find that incumbents contribute more in line with their own normative view than newcomers. A one-point increase in what the participants themselves think they should contribute (i.e., their own normative view) increases their contribution by .15 points for newcomers and by .43 points for incumbents (Wald test for difference between incumbents and newcomers, $p = .015$). Thus, newcomers seem to be predominantly influenced by the contribution of others while incumbents are influenced both by the contribution of others and their own normative view.

In Model 2 of Table B.7, we add as predictors the participant's own punishment received in the prior round, the own contribution decision in the prior round, and the interaction between these two variables. The variables are uncentered, meaning that the effect of the punishment received is the effect of punishment when the participant did not contribute to the public good (and vice versa, the effect of one's prior contribution is the effect when no punishment was received in the prior round). We see that punishment received in the prior round has a significant positive effect on the subsequent contribution when the participant did not contribute to the public good. In this case, every punishment point received increases the newcomer's subsequent contribution by .79 points and the incumbent's subsequent contribution by .60 points (the difference between newcomers and incumbents does not reach significance: Wald test, $p = .59$). There is a significant negative interaction between received punishment and own contribution. This means that the positive effect of punishment decreases as the own contribution increases. Already when the participant contributed 10 points, the received punishment no longer has a significant effect on subsequent contributions (incumbents: coefficient = .26, p = .08; newcomers: coefficient = .12, p = .54), nor does the effect reach conventional significance values ($p < .05$) for higher contribution levels.

The negative interaction between the punishment received and the own contribution also implies that the effect of one's own contribution decreases as the received punishment increases. The participant's own prior contribution is a strong predictor of the current contribution when received punishment is zero (coefficient = .67 and .66 for newcomers and incumbents respectively). The participant's own prior contribution only weakly predicts the current contribution when received punishment is 10 (coefficient = .07 and .35 for newcomers and incumbents respectively). Whether participants deviate from their own contribution thus strongly depends on the punishment they receive. The three newly added predictors in Model 2 add about 20 percent explained variance to Model 1, making the total explained variance about 70 percent for both newcomers and incumbents.

Table B.7. Predictors of contribution levels for newcomers and low-return incumbents.

	Model 1		Model 2	
	Newcomer	Incumbent	Newcomer	Incumbent
own normative view	.15*	.43***	.01	.13**
	(.06)	(.09)	(.04)	(.04)
contribution of others [a]	.77***	.53***	.32***	.22***
	(.06)	(.07)	(.07)	(.05)
received punishment [a]			.79**	.60*
			(.25)	(.24)
own contribution [a]			.67***	.66***
			(.08)	(.07)
own contribution [a] ×			-.07**	-.03*
received punishment [a]			(.02)	(.02)
Number of observations	576	576	576	576
R^2	.52	.59	.67	.73

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, [a] lagged. We account for repeated measures within participants by estimating cluster-robust standard errors. Standard errors are in parentheses. The newcomer has a low-return rate from the public good, one incumbent has the same low-return rate, and another incumbent has a high-return rate. We only compare the newcomer with the low-return incumbent, i.e., we leave out the high-return incumbents. This is to prevent confounding of newcomer-incumbent differences with return rate differences. The variable 'contribution of others' refers to the contribution of the newcomer for the low-return incumbent and refers to the contribution of the low-return incumbent for the newcomer. The variable 'own normative view' indicates what the participants think members with a return like themselves should contribute, i.e., how much they think low-return members should contribute if they have a low-return themselves; how much high-return members should contribute if they have a high-return themselves.

The results are robust to alternative model specifications. In Table B.8, we show that results remain when controlling for differences between members that may have developed in the first 10 rounds of the experiment (before membership change). We do this by estimating the difference-in-differences for the coefficients of own normative view and the contribution of others. That is, we estimate the difference between newcomers and incumbents in the difference in the coefficients between the first and last 10 rounds of the game (before and after membership change). In Table B.9, we show that the results also remain when comparing newcomers with high-return incumbents. Results do not substantively differ when subdividing by experimental condition, see Table B.10.

Table B.8. Predictors of contribution levels for newcomers and incumbents when controlling for potential differences developed before membership change.

	newcomer			incumbent			Wald test
	R1-10	R11-20	dif	R1-10	R11-20	dif	dif-in-dif
own normative view	.21**	.15*	-.06	.27***	.43***	.16	-.22#
	(.07)	(.06)		(.07)	(.09)		-
contribution of others [a]	.58***	.77***	.19*	.58***	.53***	-.05	.24*
	(.07)	(.06)		(.06)	(.07)		-
Number of observations	576	576		576	576		-
R^2	.39	.52		.50	.59		-

Note: # $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$, [a] lagged. R1-10 = rounds 1-10. R11-20 = rounds 11-20. We account for repeated measures within participants by estimating cluster-robust standard errors. Standard errors are in parentheses. The newcomer has a low-return rate from the public good, one incumbent has the same low-return rate, and another incumbent has a high-return rate. We only compare the newcomer with the low-return incumbent, i.e., we leave out the high-return incumbents. This is to prevent confounding of newcomer-incumbent differences with return rate differences. The variable 'contribution of others' refers to the contribution of the newcomer for the low-return incumbent and refers to the contribution of the low-return incumbent for the newcomer.

Table B.9. Predictors of contribution levels for newcomers and high-return incumbents.

	Model 1		Model 2	
	Newcomer	Incumbent	Newcomer	Incumbent
own normative view	.23*	.66***	.00	.14**
	(.11)	(.10)	(.05)	(.05)
contribution of others [a]	.57***	.25**	.15**	.09**
	(.09)	(.08)	(.04)	(.03)
received punishment [a]			1.13***	.58*
			(.24)	(.27)
own contribution [a]			.85***	.77***
			(.05)	(.05)
own contribution [a] × received punishment [a]			-.08***	-.03**
			(.02)	(.01)
N observations	576	576	576	576
R2	.29	.51	.65	.72

Note: * $p < .05$, ** $p < .01$, *** $p < .001$, [a] lagged. We account for repeated measures within participants by estimating cluster-robust standard errors. Standard errors are in parentheses. The newcomer has a low-return rate from the public good, one incumbent has the same low-return rate, and another incumbent has a high-return rate. We here compare the newcomer with the high-return incumbent, i.e., we leave out the low-return incumbents. The variable 'contribution of others' refers to the contribution of the newcomer for the high-return incumbent and refers to the contribution of the high-return incumbent for the newcomer.

Table B.10. Predictors of contribution levels for newcomers and incumbents by condition.

	Newcomer				Incumbent			
	Model 1		Model 2		Model 1		Model 2	
	con. 0	con. 1	con. 0	con. 1	con. 0	con. 1	con. 0	con. 1
own normative view	.18*	.15	.08	-.05	.48**	.40***	.11$^{\#}$	.17*
	(.08)	(.12)	(.05)	(.05)	(.15)	(.11)	(.06)	(.06)
contribution of others [a]	.83***	.71***	.38***	.25*	.57***	.46***	.21***	.23**
	(.07)	(.11)	(.10)	(.10)	(.08)	(.11)	(.06)	(.08)
received punishment [a]			.74*	.83*			.84***	.35
			(.31)	(.32)			(.20)	(.34)
own contribution [a]			.60***	.75***			.73***	.56***
			(.12)	(.11)			(.07)	(.11)
own contribution [a] ×			-.03	-.10***			-.05***	-.01
received punishment [a]			(.02)	(.02)			(.01)	(.03)
N observations	288	288	288	288	288	288	288	288
R2	.56	.48	.67	.72	.63	.54	.79	.65

Note: $^{\#}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$, [a] lagged. con. 0 = condition disagreement. con. 1 = condition agreement. We account for repeated measures within participants by estimating cluster-robust standard errors. Standard errors are in parentheses. The newcomer has a low-return rate from the public good, one incumbent has the same low-return rate, and another incumbent has a high-return rate. We only compare the newcomer with the low-return incumbent, i.e., we leave out the high-return incumbents. This is to prevent confounding of newcomer-incumbent differences with return rate differences. The variable 'contribution of others' refers to the contribution of the newcomer for the low-return incumbent and refers to the contribution of the low-return incumbent for the newcomer.

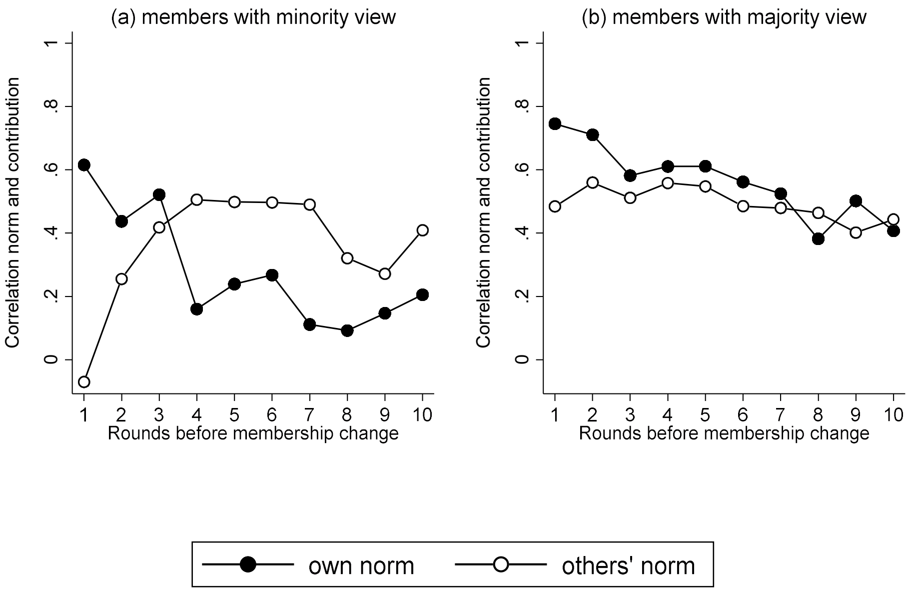


Figure B.7. Correlation between contribution and own or others' normative views before membership change.

Note: All participants provided their normative view on the appropriate contribution that a high-return member should make and the appropriate contribution that each of the low-return members should make. We examine if participants contribute in line with their normative view on how much they themselves should contribute (i.e., if they have a low-return, how much they think low-return members should contribute; if they have a high-return, how much high-return members should contribute) or in line with how much their group members think they should contribute. We look at condition disagreement-agreement before membership change, i.e., at newly formed groups with normative disagreement. This means there are no incumbent-newcomer divisions yet, and per group one member disagrees with two other members. We show how the participants' contribution correlates with their own normative views and with their group members' view. We separate participants holding a minority view in their group (panel a) and participants holding a majority view in their group (panel b). We use the normative views as measured just before the start of the game (round 1). We show that we find no clear differences between minority and majority participants in how contribution decisions correlate with their own normative view and their group members' normative views. Thus, newcomers conceding to the normative views of incumbents seems not just to be the result of the newcomers being a minority.

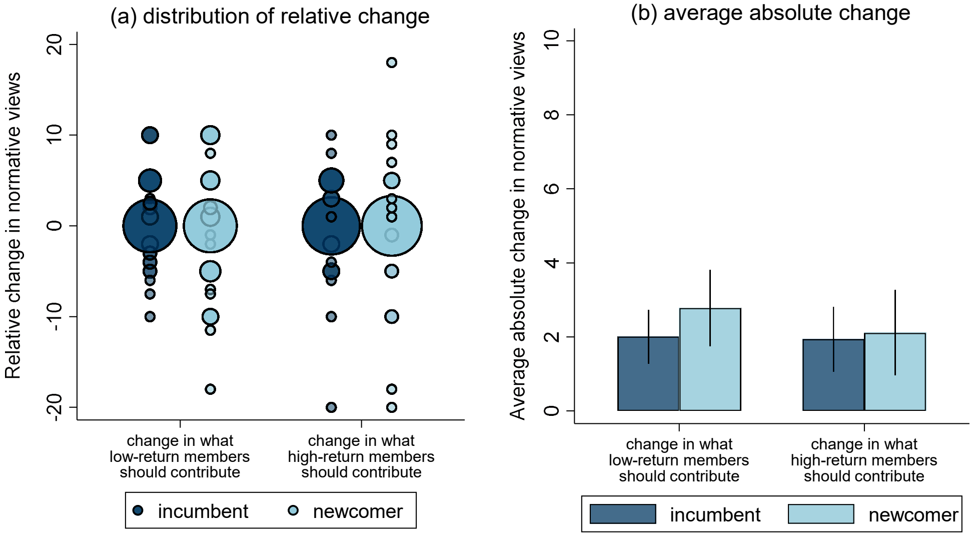
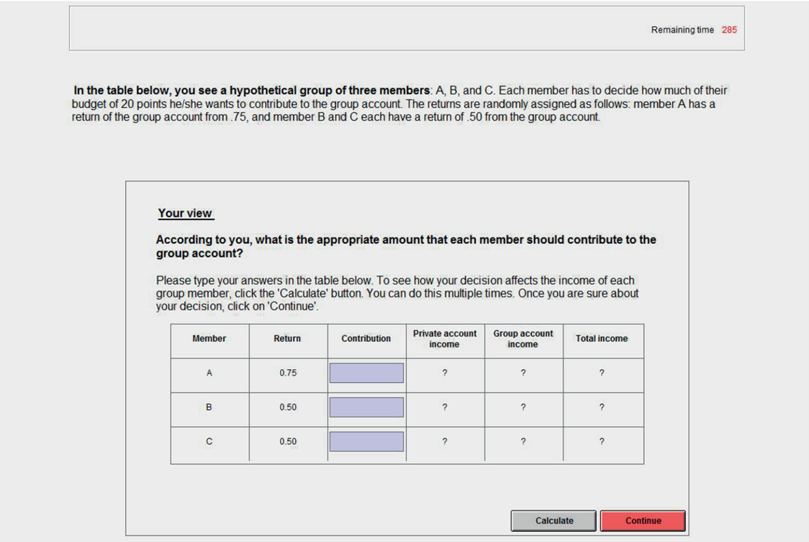


Figure B.8. Change in normative views for incumbents and newcomers.

Note: We compare low-return incumbents with (low-return) newcomers. For each of the four categories (change for views on low- or high-return members x relative or absolute change), the number of observations is 128 (64 low-return incumbents and 64 newcomers). Marker size is weighted by the number of observations in panel (a). We show that both newcomers and incumbents hold relatively stable normative views. The average temporal change in what they view to be appropriate contributions (when comparing views before and after the 10 rounds they interact together) is about 2 to 3 contribution points out of 20 for both incumbents and newcomers. Thus, while newcomers may change their behavior to adapt to incumbents, their normative views do not change much.

B.7. Screens and experimental instructions

(a) before calculating the payoff consequences of one's normative view



(b) after calculating the payoff consequences of one's normative view

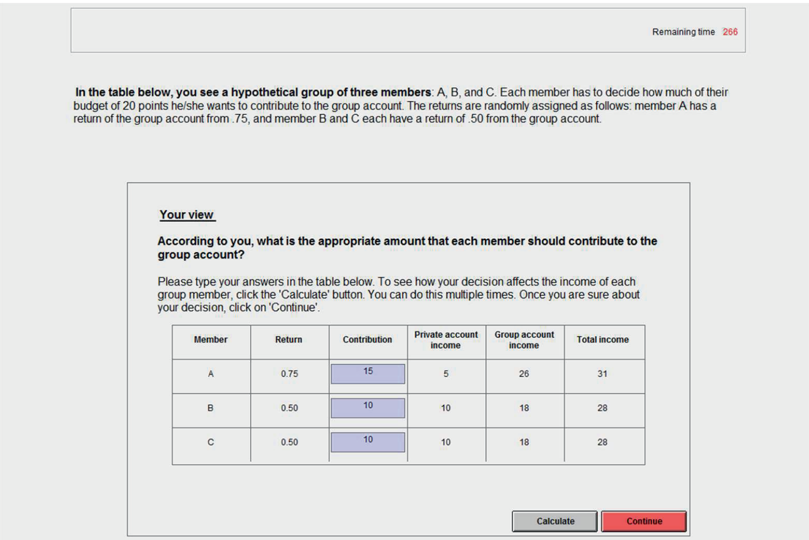


Figure B.9. Screenshots of experimental normative view measurement.

The instructions can be found in the next page. Sentences in square brackets are comments from the authors about the instructions and were not presented to the participants.

Experimental Laboratory for Sociology and Economics Universiteit Utrecht

- Instructions -

Welcome

Welcome to this experiment and thank you for coming. Please read the following instructions carefully. These instructions are the same for all participants. The instructions state everything you need to know in order to participate in the experiment. If you have any questions, please raise your hand. One of the experimenters will approach you and answer your question.

The experiment is about group decision making. You can earn money by means of earning points during the experiment. The number of points that you earn depends on your own choices, the choices of other participants in your group, and chance. At the end of the experiment, the total number of points that you earned will be exchanged at a rate of:

70 points = 1 Euro

The money you earn will be rounded up to whole euros and paid out in cash at the end of the experiment. There is a minimum payment of 5 euros, and a maximum payment of 23 euros. Other participants will not see how much you have earned. During the experiment you are not allowed to communicate with other participants. Please turn off your mobile phone. You may only use functions on the computer screen that are necessary to carry out the experiment.

First, we introduce the decision situation in which you will interact. You will learn about the procedure of the experiment later. **A decision situation consists of 2 stages: a contribution stage followed by a review stage.** In the contribution stage, you decide how many points you contribute to a group account. In the review stage, you learn how much the other members of your group contributed to the group account. We will first explain the contribution stage.

Contribution stage

You are a member of a group of **3 participants**. You and the two other members of your group are **each given 20 points**. Each of you can **choose how many points to keep for yourself in a private account and how many points to contribute to a group account**.

Your points from the private account
You will earn 1 point for each point you keep in your private account.
For example, if you keep all 20 points into your private account (and therefore do not contribute to the group account), your income will amount to exactly 20 points out of your private account. If you keep 6 points into your private account, your income from this account will be 6 points. **No one except you earns something from your private account.**

Your income from the group account
Each group member will profit from points you contribute to the group account. You will also profit from the other group members' contributions. Just like in real life, some persons profit more from contributions to the group account than others.

For each point contributed to the group account (by you and the other members):
1 member earns 0.75 points and 2 members earn 0.50 points each.

Whether you are a member with a return of 0.50 or 0.75 from the group account will be randomly determined at the start of the experiment, and will stay the same for the entire duration of the experiment.

For example, if the 3 members combined contribute in **total 40 points to the group account**,

1 member receives: **0.75 times 40 = 30 points** from the group account,
2 members each receive: **0.50 times 40 = 20 points** from the group account.

Your total income from the private account and group account

Each member can choose any number of points to contribute to the group account, from 0 to 20 points. Every point a member does not contribute to the group account will automatically remain in his/her private account. **Each member's total income from the contribution stage is the combined income from his/her private account and the group account.**

Table 1 gives an arbitrary example of how each member's income from the private account, group account, and the total income are calculated when the total contributions to the group account are 40 (15+15+10).

Table 1 – example

Member	Return	Contribution	Private account income	Group account income	Total income
A	0.75	15	5	30	35
B	0.50	15	5	20	25
C	0.50	10	10	20	30

Review stage

Each contribution stage is followed by a review stage. In the review stage, everyone in the group will see how much each of the other group members contributed to the group account as well as their income from the contribution stage. Then, all group members have a chance to **decrease** the income of each other group member. You can decide if you want to spend points to decrease the income of the other two group members, for example because you disagree with how much they contributed or earned.

If you want to decrease another member's income you do that by assigning deduction points. **Every deduction point assigned to another group member reduces his/her income by 3 points, and your own income by 1 point**. Similarly, every deduction point that one of your group members assigns to you decreases your income by 3 points and costs the group member 1 point. Note that this might imply that you or other participants lose income in a particular round. If you do not want to decrease the income of a group member, you must assign him/her

0 deduction points. Every participant can assign up to a maximum of 10 deduction points to each group member, regardless of the income from the contribution stage.
For example, if you assign 2 deduction points to a group member this costs you 2 points and reduces the group member's income by 6 points (2 times 3). Another example: if one of your group members assigns 3 deduction points to you, this reduces the group member's income by 3 points and your income by 9 points (3 times 3).

After everyone has made a decision, you will see how many deduction points were assigned to you by the other group members and also what your total income for the round is. You will not see which individual participant assigned deduction points to you, you can only see the total number of deduction points assigned to you and how that affected your income. Similarly, if you assigned deduction points to one or more of your group members, they will not see that you are the one who assigned the points.

Overview of the Session

The experiment consists of **2 parts**, and in total lasts about **1 hour and 45 minutes**.

In the 1st part you will play 10 rounds of the decision situation (10 contribution and 10 review stages).

Before you play these 10 rounds, we will first ask you to answer some questions about the decision situation. These questions concern:

- your understanding of the decision situation,
- your view on the appropriate amount that each group member should contribute to the group account,
- your guess of what the other participants think are appropriate contributions.

Some questions appear multiple times throughout the experiment. You do not have to be consistent with your answers to these questions. Your answers may or may not have changed during the experiment. Similarly, what you view as appropriate contributions may or may not be the same as what the other participants think are appropriate contributions.

After this 1st part in which you answer questions about the decision situation and play 10 rounds of it, **you will receive new instructions on your computer screen for the 2nd part of the experiment.** The 2nd part of the experiment is of similar length to the 1st part.

Because you play together with other persons, you will sometimes have to wait until the other persons have made their decision. These waiting times are incorporated in the total expected duration of 1 hour and 45 minutes for the experiment.

[After reading these instructions, participants were presented with a quiz about the instructions and answered questions about their normative views and expectations (see Figure B.9). Before the start of the first 10 decision rounds, we provided participants with the following information on their computer screen]:

Next part of the experiment [on screen only]

You will now play **10** rounds of the decision situation yourself.

Each round consists of a contribution stage followed by a review stage. In the contribution stage you decide how much you contribute to the group account. In the review stage you learn the contributions of your group members and can assign them deduction points.

Every group receives a color.
4 groups received color **blue** and 4 groups received color **orange**. You and your two group members are in a **blue group**.

In your group, the returns from the group account are randomly assigned as follows:

You: 0.50
Member 1: 0.50
Member 2: 0.75

Your return will remain the same for the entire experiment.
You will play with the same two group members all 10 rounds.

Once you are ready, please click 'Continue'.

[Note that this is an example of a participant in a blue group. The blue text was replaced by orange text and vice versa if the participant's group was orange. In the example, the participant was assigned a return of 0.50. If the participant was assigned a return of 0.75, the returns assigned for Member 1 and Member 2 would be 0.50].

[After completing the ten rounds, participants again answered questions about their normative views and expectations, and the group identification items described in the methods. We then presented them with the information for the second part of the experiment. We will present the instructions both for incumbents and newcomers].

[Incumbent instructions (on screen only)]:

<u>2nd part of the experiment</u>

The 2nd part of the experiment starts now. You will play another set of **10** rounds of the decision situation.

Remember that there are **4** groups in the room with color **blue** and **4** groups in the room with color **orange**.

So far you have been playing in a group with color **blue**.

One group member will now leave your group, and be replaced by a new member from a group with color orange.

Group member 1 remains the same. **The new group member receives number 2.**
The returns from the group account are as follows:

You: 0.50
Member 1: 0.75
Member 2: 0.50

You will play with this group all next 10 rounds.

Before we continue to the 10 rounds of the decision situation, we will ask you:

One question about the <u>new member's view</u> on the appropriate amount that each member in a group should contribute to the group account.

After the 10 rounds of the decision situation, we ask you to complete 4 final tasks and to fill in a questionnaire. Once you are ready, please click 'Continue'.

[Note that this is an example of a participant coming from a blue group. The blue text was replaced by orange text and vice versa if the participant's group was orange. In the example, the participant was assigned a return of 0.50. If the participant was assigned a return of 0.75, the returns assigned for Member 1 and Member 2 would be 0.50. The 4 final tasks refer to post-experiment measures (e.g., social value orientation), and can be found in the experiment's pre-registration and in the openly available dataset].

[Newcomer instructions (on screen only)]:

2nd part of the experiment

The 2nd part of the experiment starts now. You will play another set of **10** rounds of the decision situation.

Remember that there are **4** groups in the room with color **blue** and **4** groups in the room with color **orange**.

So far you have been playing in a group with color **blue**.

You will now enter a group with color orange, and therefore play with two new group members.

These two group members have played the past 10 rounds together, but you have not interacted with them yet.

The returns from the group account are as follows:

You: 0.50
Member 1: 0.50
Member 2: 0.75

You will play with this group all next 10 rounds.

Before we continue to the 10 rounds of the decision situation, we will ask you:

One question about the view of your two new members on the appropriate amount that each member in a group should contribute to the group account.

After the 10 rounds of the decision situation, we ask you to complete 4 final tasks and to fill in a questionnaire. Once you are ready, please click ‘Continue’.

[Note that this is an example of a participant coming from a blue group. The blue text was replaced by orange text and vice versa if the participant’s group was orange. In the example, the participant was assigned a return of 0.50. If the participant was assigned a return of 0.75, the returns assigned for Member 1 and Member 2 would be 0.50.]

Appendix C. Supplementary material for Chapter 4

C.1. More details on public good provision in Ikariam

In Table C.1, we show an example of what group members see about the contributions to the sawmill. In this example, the current public good level is 24, a total contribution of 1855942 units of wood is necessary to move from public good level 24 to 25, of which 700909 have been contributed so far (column 1 in Table C.1). The total prior contributions that have been made to get the sawmill from level 1 to 24 are also displayed (column 6 in Table C.1). We can see there that 4 of the 6 group members have contributed to the public good so far. Members are known to each other via self-chosen nicknames and names of their towns (columns 2-3 in Table C.1). If persons look at the contributions of other group members, they can also see each other's (town hall) level, indicating how far they have progressed in the game (column 4 in Table C.1).

The production rate of the public good is the same for each group member, but it is up to each member whether to collect what has been produced for him/her. Each group member has simulated citizens at his/her disposal that can be assigned as workers to gather the resources produced by the public goods or to do other tasks. For example, if a public good is at level 1 and therefore produces 30 units of wood per hour for each member, some members may still choose not to obtain their 30 units of wood per hour by not assigning their simulated citizens as workers to gather the wood. A person may choose to do so, for example, because the person wants to use his/her simulated citizens for other tasks, e.g., as soldiers to fight other persons. If persons look at the contributions of other group members, they can see how many workers each member has assigned to the public good (column 5 in Table C.1). Note that the number of workers a member has assigned to the public good does not affect how much other members can collect from the public good, i.e., the good is non-rival.

In total, persons thus see their group members' nicknames, town nicknames, contributions to the public good, individual level, and usage of the public good. This works exactly the same for the sawmill and the island-specific public good (wine, marble, crystal glass, or sulfur). More information on the group structure can be found in Figure C.1.

Table C.1. Example table of public good provision in Ikariam.

Saw mill	**Other players on this island**				
Level:	**Player**	**Town**	**Level**	**Workers**	**Donated**
24	Atlas-2812	Colony 42	Level 28	313 workers	0
Required for next level:	Brochis	Delta	Level 8	100 workers	50,000
1,855,942	Esprit	Ophelia	Level 30	655 workers	100,909
Available:	jessZeta	Zeta 1	Level 26	683 workers	551,484
700,909		Zeta 2	Level 26	696 workers	
Donations:		Zeta 3	Level 20	340 workers	
	Maspero	Judah	Level 16	0 workers	450,000
Donate for expansion		Sumeria	Level 29	696 workers	
	pseudonym	Polis	Level 1	0 workers	0

Note: Column 1 shows the level of one of the public goods (the sawmill, level 24), the threshold to reach the next level of the public good (1855942 units of wood), and the resources already contributed to the public good (700909 units of wood). Columns 2-5 shows player nicknames, towns that are on this island, level of towns, and workers assigned to make use of the public good. Column 6 shows the total of contributions to the public good of all players during their time on the island (e.g. player ‘Esprit’ contributed 100909, whereas ‘pseudonym’ contributed 0).

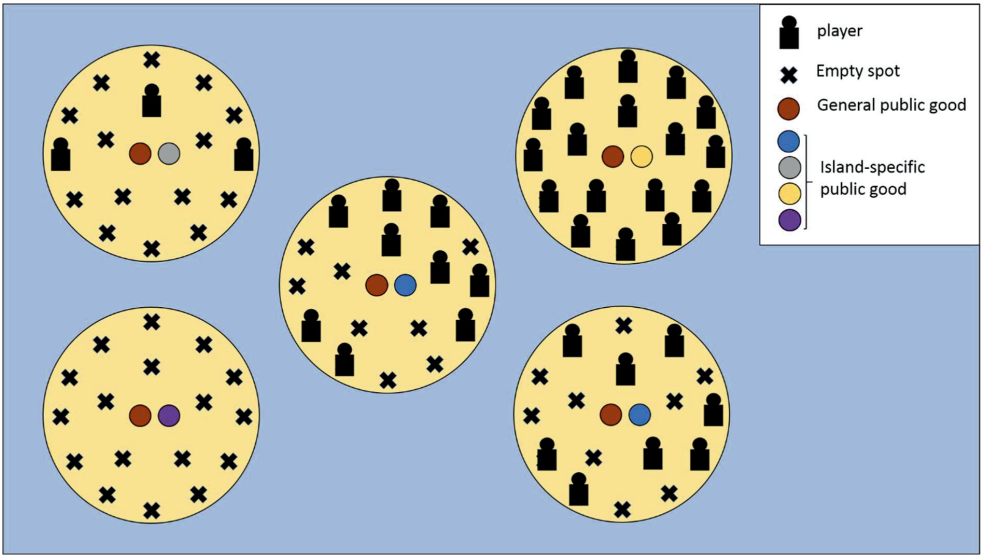


Figure C.1. Example groups in Ikariam and their public goods.

Note: Public good provision occurs on islands in Ikariam. Each island contains two public goods. The first is the same on every island, namely a sawmill that provides wood (indicated with brown circles). The second public good provides an island-specific resource, namely crystal glass, marble, sulfur, or wine (indicated with blue, grey, yellow, and purple circles respectively). Up to 17 players can inhabit an island. Players have certain information available about other islands which they can use to decide which islands to inhabit next when building a new town. They know which players and how many players inhabit an island, which island-specific public good is produced on the island, the level of the public goods on the island, and how far the island is from the island(s) they currently inhabit. The players cannot see how much players are currently contributing to the public good on islands that they do not inhabit themselves, although the public good level may provide some indication of this. Players can also choose to build a new town on an island that they already inhabit.

C.2. Thresholds and step-returns

There are 50 thresholds (public good levels) in Ikariam that can be surpassed in succession. Each succeeding threshold requires more contributions than the previous threshold. For example, while increasing the public good from level 1 (return of 30 units of wood per hour) to level 2 (return of 38 units of wood per hour) requires only 394 contributed units of wood, further increasing to level 3 (return of 50 units of wood per hour) requires 992 contributed units of wood. Although increasing increments in thresholds are not common in public good games, some studies do incorporate them (Asher et al., 2009; Ye et al., 2020).

One important determinant of contributions to public goods is the value of the public good relative to the costs of producing it. For continuous public good games, this tradeoff between value and costs has been formalized in the multiplication factor and the marginal per capita return (MPCR). The parallel concept in threshold public good games is the step return, which is the total group payoff from the public good divided by the total contribution threshold (Croson & Marks, 2000).

In Ikariam, the total group payoff is mostly a function of time (and group size) because increasing the public good level increases the hourly production of resources. Since the multiplication factor and MPCR are about the tradeoff between value and costs, we denote the step return in Ikariam as the inverse of the number of days until the increase in total group payoffs from levelling up the public good matches the total contribution threshold. In Table C.2, we provide the step returns for a medium-sized group (8 members) associated with each public good level. A step return of 0.2 would mean that it takes a medium-sized group 5 days until the increase in the total group payoffs from levelling up the public good breaks even with the contributions that were required to surpass the threshold.

Table C.2. Thresholds and step-returns of public good provision.

Public good level	Threshold	Step Return
1	-	-
2	394	4.678172589
3	992	2.787096774
4	1732	1.862355658
5	2788	1.322238164
6	3783	0.974464711
7	5632	0.736363636
8	8139	0.566162919
9	10452	0.44087256
10	13298	0.346518273
11	18478	0.274315402
12	23213	0.218360401
13	29038	0.174557476
14	39494	0.140011141
15	49107	0.112603091
16	66010	0.090749886
17	81766	0.073262725
18	101146	0.059225278
19	134598	0.04792939
20	154304	0.038822066
21	205012	0.031467426
22	270839	0.025520697
23	311541	0.020707387
24	411229	0.016808153
25	506475	0.013647268
26	665201	0.011083567
27	767723	0.009003247
28	1007959	0.007314583
29	1240496	0.005943429
30	1526516	0.004829822
31	1995717	0.003925206
32	2311042	0.003190249
33	3020994	0.002593054
34	3935195	0.002107748
35	4572136	0.001713335
36	5624478	0.001392769
37	7325850	0.00113221
38	9011590	0.000920415
39	11085051	0.000748251
40	13635408	0.000608299
41	17704143	0.000494528
42	20630781	0.00040204
43	26786470	0.000326852
44	32948197	0.000265726
45	40527121	0.000216033
46	52472840	0.000175634
47	61315353	0.00014279
48	79388129	0.000116088
49	97648282	0.000094380
50	120108270	0.000076731

C.3. Relationship between total number of newcomers and group-average contributions

Table C.3 shows the relationship between the group-average contribution percentage over all 28 time periods and the total number of newcomers that entered during this time. We see that there is a negative relationship between the number of newcomers and the average contribution percentage.

Table C.3. Group-average contribution across all time periods by total number of newcomers, average group size, and average period.

Total number of newcomers	-0.45***
	(0.03)
Average group size	-0.03
	(0.13)
Average period	0.94***
	(0.11)
Intercept	29.29***
	(2.11)
N observations	11348
R^2	0.06

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests.

C.4. Robustness analyses

We conduct several robustness analyses to examine whether the negative relationship between the number of newcomers and the contribution percentage also holds under different operationalizations of the number of newcomers and contribution behavior. First, we replace the dichotomous distinction between newcomers and incumbents with a continuous variable indicating what we call tenure. This variable counts how many prior time periods an individual has been present in the group. A longer average tenure in a group, when controlling for group size and period, indicates more stability in group composition. We show that a group's average tenure is positively related to its contribution percentage, suggesting that fewer group changes are related to higher contributions (Table C.4). Second, we repeat the analyses for separate country-specific game servers (Table C.5-C.6) and the two public goods separately (Table C.7-C.9). Third, we repeat the analyses when taking the first-lag of the number of newcomers (Table C.10). Fourth, we repeat the analyses when leaving out outliers with very high contribution percentages (Table C.11). Fifth, we conduct analyses with crossed fixed effects that account simultaneously for between-group and between-individual confounders (Table C.12-C.13). Finally, we conduct analyses in which we control for the public good level (Table C.14). The negative relationship between the number of newcomers and contributions to the public goods is robust to all these different specifications.

Table C.4. OLS regression model of the average contribution percentage by group tenure.

Average tenure	3.84^{***}
	(0.18)
Average group size	-3.07^{***}
	(0.11)
Average snapshot	1.12^{***}
	(0.10)
Intercept	15.72^{***}
	(2.15)
N observations	11348
R^2	0.08

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. We see that the average group tenure is positively related to the contribution percentage (B = 3.84, SE = 0.18, $p < .001$), indicating that more stability in group composition is associated with higher contribution percentages.

Table C.5. Contribution percentage by the number of newcomers per period, separated by country.

	(1) Germany	(2) England	(3) France	(4) Greece	(5) Turkey
Number of newcomers	-2.73***	-3.87***	-2.65***	-2.70***	-2.22***
	(0.09)	(0.12)	(0.09)	(0.10)	(0.06)
Group size	-1.04***	0.02	-0.75***	0.24***	-0.08*
	(0.06)	(0.09)	(0.06)	(0.06)	(0.04)
Period	-0.37***	-0.46***	-0.23***	-0.18***	-0.17***
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)
Intercept	32.97***	41.30***	35.30***	37.61***	27.04***
	(0.16)	(0.23)	(0.16)	(0.18)	(0.11)
N observations	35728	29801	37256	37176	59569
R^2 (overall)	0.06	0.03	0.06	0.02	0.03
R^2 (within)	0.04	0.04	0.03	0.02	0.03

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are groups per period. We have data for five country-specific servers, namely: Germany, England, France, Greece, and Turkey. The negative relationship between the number of newcomers and the contribution percentage holds in all of the five servers (Germany, B = -2.73, SE = 0.09, p < .001; England, B = -3.87, SE = 0.12, p < .001; France, B = -2.65, SE = 0.09, p < .001; Greece, B = -2.70, SE = 0.10, p < .001; Turkey, B = -2.22, SE = 0.06, p < .001).

Table C.6. Contribution percentage by the total number of newcomers across all periods, separated by country.

	(1) Germany	(2) England	(3) France	(4) Greece	(5) Turkey
Total number of newcomers	-0.39***	-0.40***	-0.49***	-0.62***	-0.32***
	(0.08)	(0.08)	(0.10)	(0.11)	(0.04)
Average group size	-0.35	0.33	-0.59	0.70*	0.15
	(0.32)	(0.33)	(0.35)	(0.34)	(0.21)
Average period	1.09***	0.92***	0.41	0.53*	1.93***
	(0.24)	(0.26)	(0.26)	(0.26)	(0.19)
Intercept	36.10***	41.64***	38.61***	38.45***	29.07
	(0.73)	(1.00)	(0.87)	(0.89)	(0.54)
N observations	2119	1798	2184	2063	3184
R^2	0.07	0.03	0.06	0.03	0.09

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are groups. We have data for five country-specific servers, namely: Germany, England, France, Greece, and Turkey. The negative relationship between the number of newcomers and the contribution percentage holds in all of the five servers (Germany, B = -0.39, SE = 0.08, p < .001; England, B = -0.40, SE = 0.08, p < .001; France, B = -0.49, SE = 0.10, p < .001; Greece, B = -0.62, SE = 0.11, p < .001; Turkey, B = -0.32, SE = 0.04, p < .001).

Table C.7. Contribution percentage by the number of newcomers per period, separated by public good.

	(1) sawmill	(2) island-specific good
Number of newcomers	-1.48***	-1.39***
	(0.02)	(0.02)
Group size	-0.15***	-0.05***
	(0.02)	(0.02)
Period	-0.39***	0.12***
	(0.01)	(0.01)
Intercept	18.20***	16.49***
	(0.05)	(0.04)
N observations	199582	199618
R^2 (overall)	0.02	0.04

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients of independent variables and intercept are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are groups per period. Public good 1 is the sawmill, public good 2 is the island-specific public good (i.e., marble, crystal glass, sulfur, wine). The negative relationship between the contribution percentage and the number of newcomers holds for both public good types (sawmill, B = -1.48, SE = 0.02, $p < .001$; island-specific good, B = -1.39, SE = 0.02, $p < .001$).

Table C.8. Contribution percentage by the total number of newcomers across all periods, separated by public good.

	(1) sawmill	(2) island-specific good
Total number of newcomers	-0.24***	-0.22***
	(0.02)	(0.02)
Average group size	0.15	-0.02
	(0.08)	(0.07)
Average period	0.12	0.82***
	(0.07)	(0.06)
Intercept	21.04***	8.26***
	(1.33)	(1.22)
N observations	11349	11348
R^2	0.03	0.06

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients of independent variables and intercept are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are groups. Public good 1 is the sawmill, public good 2 is the island-specific public good (i.e., marble, crystal glass, sulfur, wine). The negative relationship between the contribution percentage and the number of newcomers holds for both public good types (sawmill, B = -0.24, SE = 0.02, $p < .001$; island-specific good, B = -0.22, SE = 0.02, $p < .001$).

Table C.9. Individual contribution percentage by newcomer status per public good.

	(1) sawmill	(2) island-specific good
Newcomer	-10.55^{***}	-8.45^{***}
	(0.09)	(0.08)
Period	0.22^{***}	0.55^{***}
	(0.01)	(0.01)
Group size	-0.19^{***}	-0.38^{***}
	(0.01)	(0.01)
Intercept	16.62^{***}	10.97^{***}
	(0.17)	(0.16)
N observations	1576025	1581248
R^2 (overall)	0.01	0.01

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period.

Table C.10. Average contribution by lagged number of newcomers, group size, and period.

	Model 1	Model 2
Lagged number of newcomers	-1.69***	-2.66***
	(0.04)	(0.06)
Group size	-0.79***	-0.62***
	(0.03)	(0.03)
Period	-0.36***	-0.37***
	(0.01)	(0.01)
Lagged number of newcomers × group size		0.20***
		(0.01)
Lagged number of newcomers × period		-0.04***
		(0.01)
Intercept	34.21***	34.21***
	(0.07)	(0.07)
N observations	186992	186992
R^2 (overall)	0.05	0.05
Rho	0.63	0.63

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients of independent variables and intercept are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. Results include 10942 groups (fewer than in Table 1 because here we can only include groups that were present in at least two consecutive time periods due to taking the lag of the number of newcomers), with these groups existing on average for ~17 periods, giving a total number of observations of 186992 group-period combinations. We find a negative relationship between the lagged number of newcomers and the group-average contribution percentage (Model 1, B = -1.69, SE = 0.04, $p < .001$; Model 2, B = -2.66, SE = 0.06, $p < .001$).

Because individuals can move resources between their groups, it can happen that they contribute more to the public good of a group than the total resources they had available in that group, i.e., individuals can end up with contribution percentages above 100 percent. Likewise, because we only have snapshots of an individual's available resources instead of a continuous-time overview of an individual's resources, it is possible that an individual had more (or fewer) resources available than we see at the snapshot, which can also lead to contribution percentages above 100 percent. This happens in 6.5% of our analyzed cases. In Table C.11, we show that the negative relationship between the number of newcomers and the group-average contribution percentage is robust to adjusting for these outliers or excluding them.

Table C.11. Contribution percentage by the number of newcomers, adjusting for outliers.

	Model 1	Model 2	Model 3	Model 4
Number of newcomers	-2.32***	-1.89***	-1.43***	-1.97***
	(0.02)	(0.02)	(0.01)	(0.02)
Group size	-0.03*	-0.50***	-0.40***	0.23***
	(0.02)	(0.01)	(0.01)	(0.01)
Period	-0.16***	-0.21***	-0.13***	-0.06***
	(0.01)	(0.01)	(0.00)	(0.01)
Intercept	29.93***	25.11***	18.51***	25.82***
	(0.04)	(0.03)	(0.03)	(0.04)
N observations	199530	199530	193660	188463
R^2 (overall)	0.05	0.11	0.10	0.04
R^2 (within)	0.05	0.07	0.07	0.05

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. In Model 1, we set group-average contribution percentages above 100% to 100%. In Model 2, we set individual contribution percentages above 100% to 100% before calculating the group-average contribution. In Model 3, we exclude individuals with contribution percentages above 100% when calculating the group-average contribution percentage. In Model 4, we exclude groups with average contribution percentages above 100%. We see that the negative relationship between the number of newcomers and the group-average contribution is strongly robust to adjusting for, or excluding, outliers (Model 1, B = -2.32, SE = 0.02, $p < .001$; Model 2, B = -1.89, SE = 0.02, $p < .001$; Model 3, B = -1.43, SE = 0.01, $p < .001$; Model 4, B = -1.97, SE = 0.02, $p < .001$).

Table C.12. Crossed fixed effects for contribution percentage by number of newcomers.

Number of newcomers	-1.51***
	(0.02)
Group size	0.66***
	(0.02)
Period	1.42***
	(0.01)
Intercept	2.99***
	(0.28)
N observations	1492659
R^2	0.04

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with crossed fixed effects to account for repeated measures within groups and individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. There are 80075 singleton observations, which are excluded. We find a negative relationship between the number of newcomers and the contribution percentage (B = -1.51, SE = 0.02, $p < .001$).

Table C.13. Crossed fixed effects for contribution percentage by newcomer status.

Newcomer	-13.42***
	(0.12)
Group size	0.35***
	(0.02)
Period	1.31***
	(0.01)
Intercept	7.50***
	(0.29)
N observations	1492659
R^2	0.05

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with crossed fixed effects to account for repeated measures within groups and individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. There are 80075 singleton observations, which are excluded. We find that newcomers have lower contribution percentages than incumbents (B = -13.42, SE = 0.12, $p < .001$).

Table C.14. Regression model of average contribution percentages with group fixed effects.

	Model 1	Model 2
Number of newcomers	-2.72***	-1.35***
	(0.04)	(0.06)
Group size	-0.32***	-1.91***
	(0.03)	(0.03)
Period	-0.27***	-3.44***
	(0.01)	(0.03)
Public good level		5.67***
		(0.05)
Number of newcomers × Public good level		-0.14***
		(0.01)
Intercept	33.75***	33.75***
	(0.07)	(0.07)
R^2 (overall)	0.04	0.03
Rho	0.58	0.66

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients of independent variables and intercept are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. Results include 11348 groups, with groups existing on average for 17-18 periods, giving a total number of observations of 199530 group-period combinations.

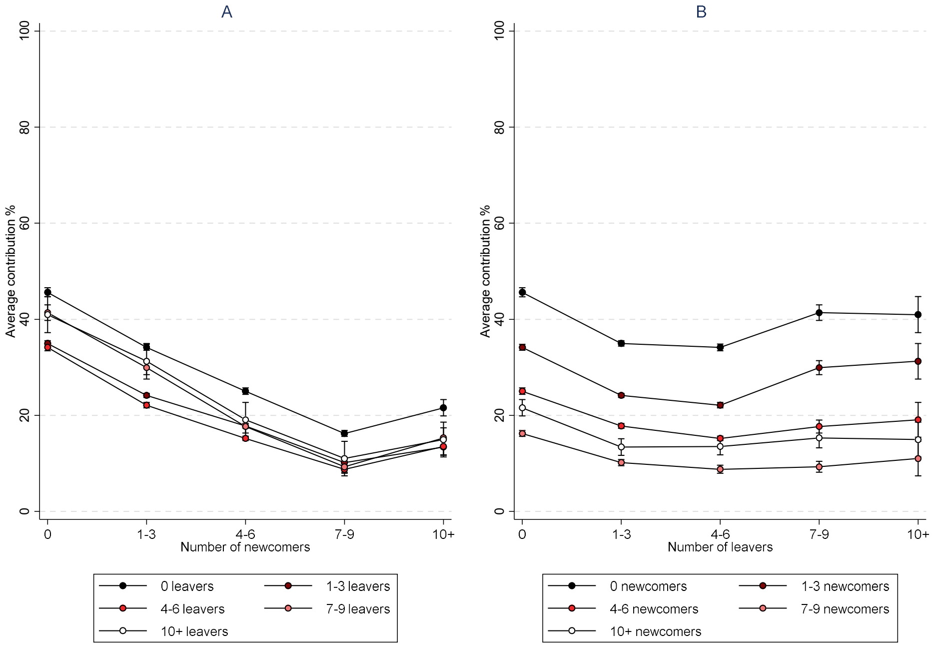


Figure C.2. Contribution percentage by the number of newcomers and the number of leavers.

Note: We show the estimated margins of an OLS regression with the group's average contribution percentage as the dependent variable and the number of newcomers and the number of leavers (persons that left the group in the prior snapshot), group size, and snapshot as predictive factors (not including the interactions between the factors). The data are discretized in this figure and its underlying model (Table C.15) for visualization purposes, the non-discretized analyses can be found in Table C.16. Data are presented as mean values and we account for repeated observations within groups by estimating cluster-robust 95% confidence intervals. **(A)** We see that, controlling for the number of leavers, the number of newcomers has a negative relationship with the contribution percentage. **(B)** We see that, controlling for the number of newcomers, there is not a clear relationship between the number of leavers and the contribution percentage. The contribution decreases somewhat when moving from 0 to 1-3 leavers, but does not change much between 1-3 leavers and 10+ leavers. Results include 10942 groups, with groups existing on average for ~17 periods, giving a total number of observations of 187875 group-period combinations. The number of observations differs from analyses without the number of leavers, because we can only see if a group has leavers if the group exists in more than two periods, meaning that we have to leave out the few groups that exist for only one period (see also Table C.15 and Table C.16).

Table C.15. Regression model underlying Figure C.2.

1-3 newcomers	-8.73***
	(0.26)
4-6 newcomers	-15.50***
	(0.41)
7-9 newcomers	-23.76***
	(0.51)
10+ newcomers	-20.47***
	(0.96)
1-3 leavers	-7.05***
	(0.29)
4-6 leavers	-8.95***
	(0.35)
7-9 leavers	-7.30***
	(0.72)
10+ leavers	-7.92***
	(1.87)
Group sizes 6-10	-12.52***
	(0.59)
Group sizes 11-17	-11.26***
	(0.73)
Periods 6-10	-2.27***
	(0.44)
Periods 11-15	-4.23***
	(0.51)
Periods 16-20	-4.30***
	(0.55)
Periods 21-28	-6.07***
	(0.65)
Intercept	55.10***
	(0.84)
N observations	187875
R^2	0.07

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients are marginal effects with group cluster-robust standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. Reference category for number of newcomers are groups with 0 newcomers, reference category for number of leavers are groups with 0 leavers, reference category for group size are groups with 1-5 members, reference category for time periods are groups at periods 1-5. Data were discretized for visualization of Figure C.2, the non-discretized analyses are presented in Table C.16.

Table C.16. Regression model of average contribution percentages with group fixed effects.

Number of newcomers	-2.76***
	(0.05)
Number of leavers	-1.10***
	(0.05)
Group size	-0.56***
	(0.03)
Period	-0.42***
	(0.01)
Intercept	34.16***
	(0.07)
R^2 (overall)	0.06
Rho	0.63

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with group fixed effects to account for repeated measures within groups. Coefficients of independent variables and intercept are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. Results include 10942 groups, with groups existing on average for ~17 periods, giving a total number of observations of 187875 group-period combinations. The number of observations differs from analyses without the number of leavers, because we can only see if a group has leavers if the group exists in more than two periods, meaning that we have to leave out the few groups that exist for only one period. Results show that the coefficient for the number of newcomers is 2-3 times the size of the coefficient for the number of leavers (B = -2.76, SE = 0.05, $p < .001$ and B = -1.10, SE = 0.05, $p < .001$ respectively).

C.5. Analyses for the incumbent-newcomer difference in contributions

In Table C.17, we examine the relationship between the contribution percentages of individual players and their group members and how this depends on whether players are newcomers and time spent in the group. In Table C.18, we examine if newcomers contribute higher percentages if they already know incumbents from their other groups. In Table C.19, we examine the newcomer-incumbent contribution difference when selecting players who belong to only one group. In Table C.20, we examine whether newcomers' contribution is related to the group-average contribution in their prior groups. In Table C.21, we examine to what extent the difference in contribution percentage between newcomers and incumbents remains when adjusting for private good levels. A player's private good level on an island is captured by the player's town hall level on the island. Every increase in the town hall level increases the maximum number of citizens allowed in one's town. Each citizen works to produce resources and pays taxes to give gold (the latter of which can be used to buy yet more resources). Hence, a higher-level town hall means more resources. In this sense, the town hall level is akin to the endowment in regular public good games. Comparing newcomers and incumbents of similar private good levels (town hall levels) is not trivial because incumbents typically have higher private good levels, as is shown in Figure C.3. Especially at the high end of the private good levels, there are almost no newcomers (the converse is less problematic: there are enough incumbents with low private good levels). We see that town hall levels above 10 are achieved by only 1% of the newcomers, making it difficult to find enough newcomers above this level to compare with incumbents of the same town hall level. Therefore, when adding private good level as a control variable in Table C.21, we restrict the analyses to town hall levels between 1 and 10, which captures 99% of the newcomers and 50% of the incumbents. In Table C.22, we examine to what extent the difference in contribution percentage between newcomers and incumbents remains when adjusting for public good levels. Because the public good level does not differ between newcomers and incumbents, we do not have to restrict this analysis to a certain range of public good levels.

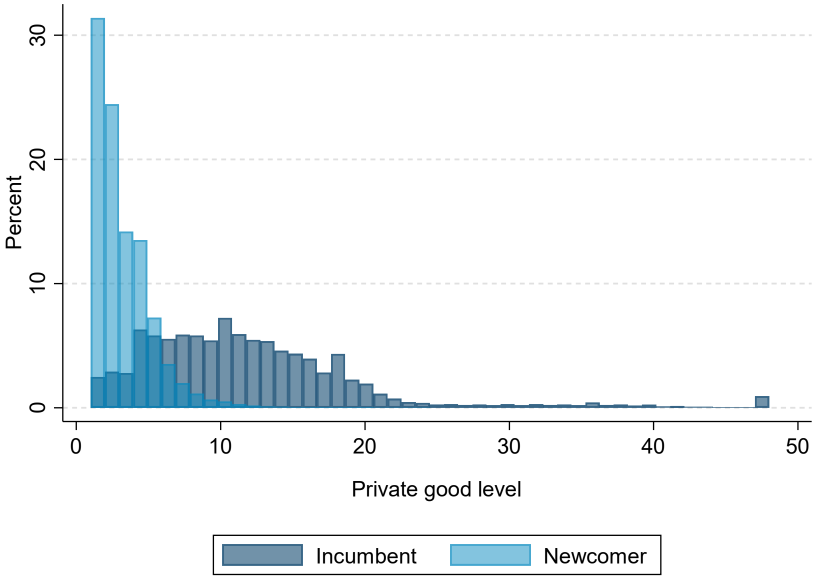


Figure C.3. Histogram of private good levels for incumbents and newcomers.

Table C.17. Individual contribution percentage by group's average contribution.

	Model 1	Model 2
Group-average contribution	0.20***	0.21***
	(0.00)	(0.00)
Newcomer	-18.19***	
	(0.14)	
Newcomer ×	-0.25***	
Group-average contribution	(0.00)	
Individual tenure		1.99***
		(0.01)
Individual tenure ×		0.01***
Group-average contribution		(0.00)
Period	0.73***	-0.53***
	(0.01)	(0.01)
Group size	-0.36***	-0.18***
	(0.02)	(0.02)
Intercept	28.44***	28.44***
	(0.04)	(0.04)
N observations	1536892	1536892
R^2 (overall)	0.04	0.06
R^2 (within)	0.04	0.05

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period. Model 1 shows that the relationship between individuals' contribution percentage and their groups' average contribution percentage is lower for newcomers (B = -0.25, SE < 0.01, $p < .001$). Model 2 shows that the relationship between individuals' contribution percentage and their groups' average contribution percentage increases with tenure (number of periods spent in the group; B = 0.01, SE < 0.01, $p < .001$).

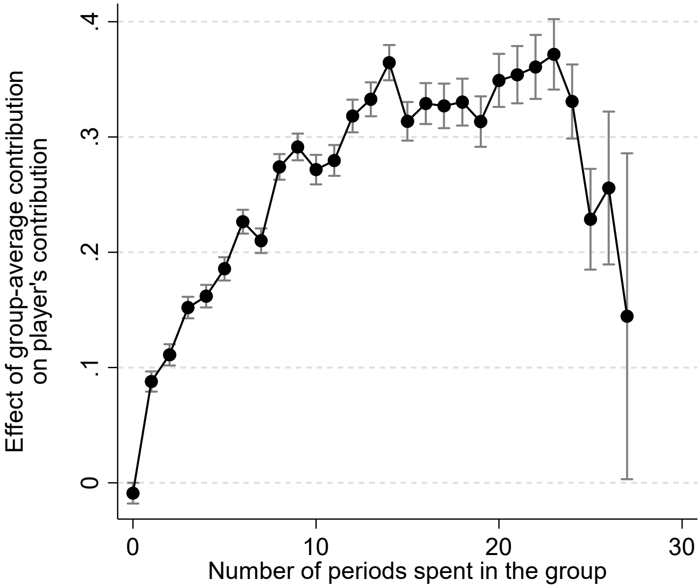


Figure C.4. The relationship between players' contribution percentage and their group members' contribution percentage by the players' time spent in the group.
Note: Results are derived from a regression model on the individual contribution percentage per period. We account for repeated measures within individuals by estimating individual fixed effects. 95% confidence intervals are provided with vertical capped spikes. Tenure (number of periods spent in the group) is added as a factor variable and interacted with the group's average contribution percentage. We find a clear increase in the relationship between the group's and player's contribution percentage up until a tenure of about 25. After tenure reaches 25, we see a decrease in this relationship, but there are few observations for tenure > 25, leading to large confidence intervals for these final observations. Results include 1536892 contribution decisions made by 134471 players.

Table C.18. Newcomers' contribution percentage by the proportion of incumbents they already know from prior islands.

Proportion of incumbents already known	0.74*
	(0.37)
Group size	-0.80***
	(0.03)
Period	-0.68***
	(0.02)
Intercept	23.25***
	(0.48)
N observations	226263
R^2 (overall)	0.01
R^2 (within)	0.01

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period. We find a very small effect of newcomers already knowing incumbents from the other islands that they inhabit on their contribution percentage. The difference in the contribution percentage between newcomers who know no incumbents and newcomers who know all incumbents on the new island is only 0.74 percentage points (B = 0.74, SE = 0.37, p = .043).

Table C.19. Individual contribution percentage for players who belong to only one group.

Newcomer	-34.77***
	(0.76)
Group size	3.41***
	(0.18)
Period	-0.19**
	(0.06)
Intercept	7.16**
	(2.19)
N observations	209471
R^2	0.07

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression. Coefficients are marginal effects with individual cluster-robust standard errors in parentheses. We have 209471 observations in which players inhabit only one island, i.e., are part of only one group. We find that newcomers have considerably lower contribution percentages than incumbents when selecting players who are part of only one group (B = -34.77, SE = 0.76, $p < .001$).

Table C.20. Newcomers' contribution percentage by the group-average contributions of their prior group(s).

	Model 1	Model 2
Group-average contribution in most recent prior group	0.02***	
	(0.00)	
Group-average contribution in all prior groups		-0.01*
		(0.01)
Group size	-1.03***	-0.81***
	(0.03)	(0.04)
Period	-0.74***	-0.70***
	(0.02)	(0.02)
Intercept	28.51***	26.09***
	(0.49)	(0.55)
N observations	108256	105351
R^2 (overall)	0.04	0.02
R^2 (within)	0.03	0.02

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period. We find very small effects of the group-average contribution in prior groups on the newcomers' contribution in their current group. A one-percentage-point increase in the prior groups' contribution percentage is associated with a 0.02 increase (B = 0.02, SE = < 0.01, $p < .001$) or 0.01 decrease in the newcomers' contribution percentage (B = -0.01, SE = 0.01, $p = .016$), depending on whether we take the average contribution percentage of the newcomers' most recent prior group or the average over all prior groups. In both cases, effects are so small as to consider them negligible.

Table C.21. Individual contribution percentage by newcomer status, with and without controlling for the private good level.

	Model 1	Model 2	Model 3
Newcomer	-18.22***	-16.88***	-8.63***
	(0.13)	(0.16)	(0.19)
Period	0.70***	0.93***	0.70***
	(0.01)	(0.02)	(0.02)
Group size	-0.61***	-0.67***	-0.92***
	(0.01)	(0.02)	(0.02)
Private good level			2.93***
			(0.03)
Intercept	27.79***	27.05***	14.94***
	(0.26)	(0.39)	(0.41)
N observations	1572734	898030	898030
R^2 (overall)	0.01	0.01	0.01
R^2 (within)	0.03	0.03	0.04

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients of independent variables are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period. Model 1 does not control for private good level; Model 2 also does not control for private good level but includes only observations with private good levels up to 10; Model 3 controls for private good level and includes only observations with private good levels up to 10. Comparing Model 1 and 2, we see that limiting the analysis to observations up to private good levels of 10 does not appreciably change the newcomer-incumbent difference in the contribution percentage. Comparing Model 2 and 3, we see that controlling for private good level roughly halves the newcomer-incumbent difference in contribution percentage. When controlling for private good level, we limit the analysis to private good levels of up to 10 because there are not enough newcomers with private good levels above 10 to reliably compare them with incumbents per private good level (~1%, see also Figure C.3).

Table C.22. Individual contribution percentage by newcomer status and public good level.

	Model 1	Model 2
Newcomer	-18.22***	-9.44***
	(0.13)	(0.23)
Period	0.70***	-0.68***
	(0.01)	(0.01)
Group size	-0.61***	-1.46***
	(0.01)	(0.02)
Public good level		2.35***
		(0.02)
Newcomer × Public good level		-0.89***
		(0.03)
Intercept	27.79***	32.50***
	(0.26)	(0.27)
N observations	1572734	1572734
R^2 (overall)	0.01	0.03
R^2 (within)	0.03	0.04

Note:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Linear regression with individual fixed effects to account for repeated measures within individuals. Coefficients in Model 1 are marginal effects with standard errors in parentheses. Statistical significance is calculated using two-sided t-tests. The units of analysis are individuals per period. The main effect of the newcomer variable in Model 2 is the effect of being a newcomer (instead of incumbent) at the lowest public good level (1). The difference between newcomers and incumbents increases with higher public good levels, as indicated by the negative interaction between the newcomer variable and the public good level (B = -0.89, SE = 0.03, $p < .001$).

Appendix D. Supplementary material for Chapter 5

D.1. Linear models of contribution behavior

Table D.1. Linear regression of total cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Punishment	3.40***	3.75***	4.85***	1.43*
	(0.89)	(0.86)	(0.74)	(0.72)
Intercept	14.10***	14.67***	13.24***	15.45***
	(0.68)	(0.76)	(0.59)	(0.61)
N	4400			
R^2	.09			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Number of observations is 4400 (220 participants in two sets of 10 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.2. Linear regression of local cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Punishment	3.81***	3.97***	-1.56*	1.24
	(0.96)	(0.94)	(0.76)	(0.89)
Intercept	13.10***	13.97***	4.94***	7.20***
	(0.73)	(0.79)	(0.54)	(0.66)
N	4400			
R^2	.41			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Number of observations is 4400 (220 participants in two sets of 10 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.3. Linear regression of global cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Punishment	-0.40	-0.22	6.41***	0.20
	(0.33)	(0.35)	(0.97)	(1.01)
Intercept	1.00***	0.70*	8.31***	8.25***
	(0.26)	(0.32)	(0.60)	(0.70)
N	4400			
R^2	.48			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Number of observations is 4400 (220 participants in two sets of 10 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.4. Linear regression of total cooperation by round.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Without punishment				
Round	-0.51***	-0.56***	-0.54***	-0.42***
	(0.14)	(0.10)	(0.12)	(0.10)
Intercept	16.43***	17.72***	16.07***	17.52***
	(0.75)	(0.75)	(0.68)	(0.60)
With punishment				
Round	-0.15	0.04	-0.06	-0.03
	(0.10)	(0.05)	(0.07)	(0.08)
Intercept	18.15***	18.24***	18.54***	17.09***
	(0.54)	(0.36)	(0.46)	(0.44)
N	3960			
R^2	.12			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. We exclude round 10 because norm measurements occurred just before this round and these norm measurements distort the time trend (see also main text). Number of observations is 3960 (220 participants in two sets of 9 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.5. Linear regression of local cooperation by round.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Without punishment				
Round	-0.43**	-0.46***	0.14	0.31**
	(0.15)	(0.11)	(0.11)	(0.11)
Intercept	15.04***	16.47***	4.31***	5.84***
	(0.89)	(0.88)	(0.55)	(0.76)
With punishment				
Round	-0.03	0.12	0.11	0.24**
	(0.13)	(0.06)	(0.11)	(0.09)
Intercept	16.88***	17.35***	2.83***	7.28***
	(0.80)	(0.52)	(0.63)	(0.70)
N	3960			
R^2	.42			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. We exclude round 10 because norm measurements occurred just before this round and these norm measurements distort the time trend (see also main text). Number of observations is 3960 (220 participants in two sets of 9 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.6. Linear regression of global cooperation by round.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Without punishment				
Round	-0.07	-0.10*	-0.68***	-0.73***
	(0.05)	(0.04)	(0.14)	(0.11)
Intercept	1.39**	1.25*	11.76***	11.68***
	(0.42)	(0.50)	(0.87)	(0.91)
With punishment				
Round	-0.12	-0.08**	-0.18	-0.28**
	(0.07)	(0.03)	(0.12)	(0.10)
Intercept	1.28*	0.89***	15.72***	9.81***
	(0.57)	(0.24)	(0.86)	(0.78)
N	3960			
R^2	.50			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. We exclude round 10 because norm measurements occurred just before this round and these norm measurements distort the time trend (see also main text). Number of observations is 3960 (220 participants in two sets of 9 rounds). We use individual-cluster robust standard errors to take repeated observations into account.

Table D.7. Linear regression of local and global cooperation in multilevel public goods.

	Single-group to multilevel	Multilevel to single-group
Without punishment		
Local cooperation	4.94***	7.20***
	(0.54)	(0.66)
Global cooperation	8.31***	8.25***
	(0.60)	(0.70)
With punishment		
Local cooperation	3.38***	8.43***
	(0.54)	(0.59)
Global cooperation	14.71***	8.45***
	(0.77)	(0.74)
N	4400	
R^2	.18	

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Number of observations is 4400 (220 participants making 10 local contribution decisions and 10 global contribution decisions in the multilevel public goods game). We use individual-cluster robust standard errors to take repeated observations into account.

D.2. Linear models of punishment behavior

Table D.8. Linear regression of punishment frequency by public goods game (single-group or multilevel) and order (single-group to multilevel or multilevel to single-group).

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Punishment percentage	15.58^{***} (2.95)	16.33^{***} (2.87)	20.96^{***} (3.65)	36.83^{***} (4.39)
N	2240			
R^2	.04			

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Standard errors in parentheses. $N = 2240$ because there are 112 participants in the punishment conditions and each participant makes punishment decisions in two sets of 10 rounds. We use individual-cluster robust standard errors to take repeated observations into account.

Table D.9. Linear regression models of received punishment points by contributions.

	Single-group public goods		Multilevel public goods	
	Local	Global	Local	Global
Contribution	-0.09^{***}	0.12^{***}	0.06^{***}	-0.12^{***}
	(0.01)	(0.02)	(0.02)	(0.01)
Order[a]	0.39	-0.08	0.59^{***}	-0.85^{***}
	(0.29)	(0.09)	(0.16)	(0.21)
Contribution ×	-0.02	-0.06	-0.05^{**}	0.06^{***}
Order[a]	(0.02)	(0.05)	(0.02)	(0.01)
Intercept	1.90^{***}	0.37^{***}	0.40^{***}	2.36^{***}
	(0.18)	(0.06)	(0.09)	(0.18)
N	1120	1120	1120	1120
R^2	.12	.03	.02	.14

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. [a] Order indicates whether the single-group PGG came before the multilevel PGG (0) or vice versa (1). Estimates underlying Figure 4 in the main text. Standard errors in parentheses. N = 1120 for the single-group and the multilevel PGG because the punishment conditions involve 112 participants who interacted in 10 rounds in the single-group and multilevel PGG.

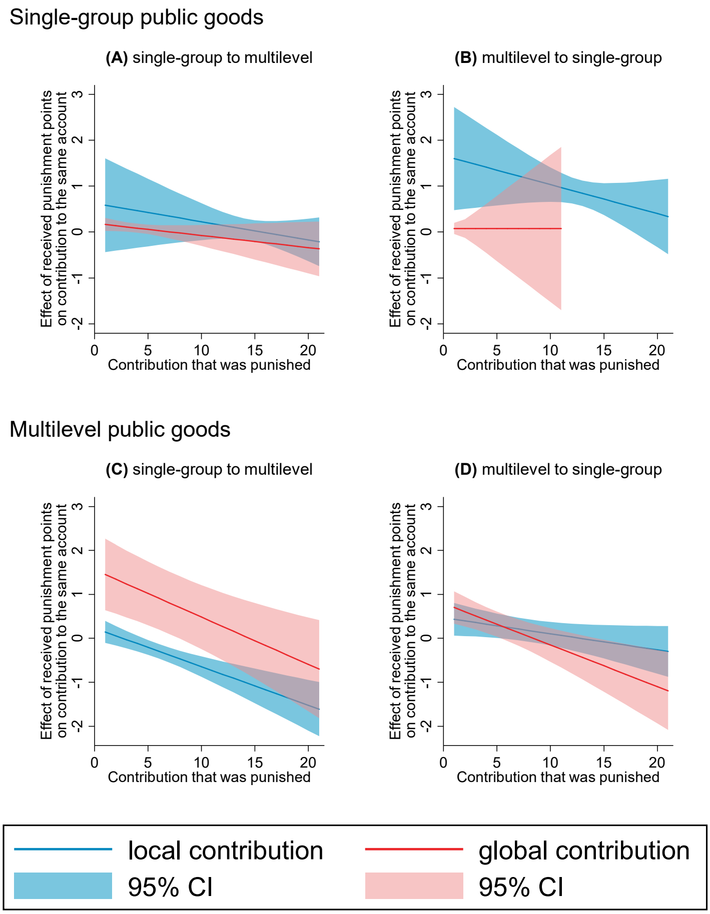


Figure D.1. Reactions to received punishment by own contribution behavior.
Note: We present the linear relationships between the effect of received punishment and the contribution to the local (blue) and global (red) account that was punished. The effect of received punishment is conceptualized as the change in contribution to the local (blue) or global (red) account for each received punishment point. For example, in panel B, the effect of ~2 on the blue line at 'Contribution that was punished' = 0 means that for participants who contributed 0 to the local account, each received punishment point is associated with a 2 MU higher contribution to the local account in the next round. As can be expected, we find a negative relationship between the effect of received punishment and contribution behavior in most cases. That is, punishment targeted toward low contributors is more effective in increasing subsequent contributions than punishment targeted toward high contributors. For single-group public goods in condition 'single-group to multilevel' (A), we find that received punishment does not have a notable effect on cooperation. In condition 'multilevel to single-group' (B), we find that punishment has a positive effect on local contributions if the punished participants made low local contributions. For multilevel public goods in condition 'single-group to multilevel' (C), we find that punishment has a positive effect on global contributions if the punished participants made low global contributions. We also find that punishment of high local contributors is associated with these punished participants subsequently making lower local contributions. In condition 'multilevel to single-group' (D), punishment only has a positive effect on global contributions if the punished participants made very low global contributions (<5).

D.3. Norm measurements

Table D.10. Normative expectations of local cooperation in single-group public goods games.

	Estimate
Single-group to multilevel	
Without punishment	15.12***
	(0.68)
With punishment	18.01***
	(0.68)
Multilevel to single-group	
Without punishment	14.85***
	(0.65)
With punishment	16.54***
	(0.63)
N	220
R^2	.06

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. We average the normative expectations measured at round 10 and round 20, which means there is 1 observation per participant, giving a total $N = 220$.

Table D.11. Normative expectations of global cooperation in multilevel public goods games.

	Estimate
Single-group to multilevel	
Without punishment	12.32***
	(0.88)
With punishment	15.29***
	(0.88)
Multilevel to single-group	
Without punishment	11.28***
	(0.84)
With punishment	11.75***
	(0.82)
N	220
R^2	.06

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. We average the normative expectations measured at round 10 and round 20, which means there is 1 observation per participant, giving a total $N = 220$.

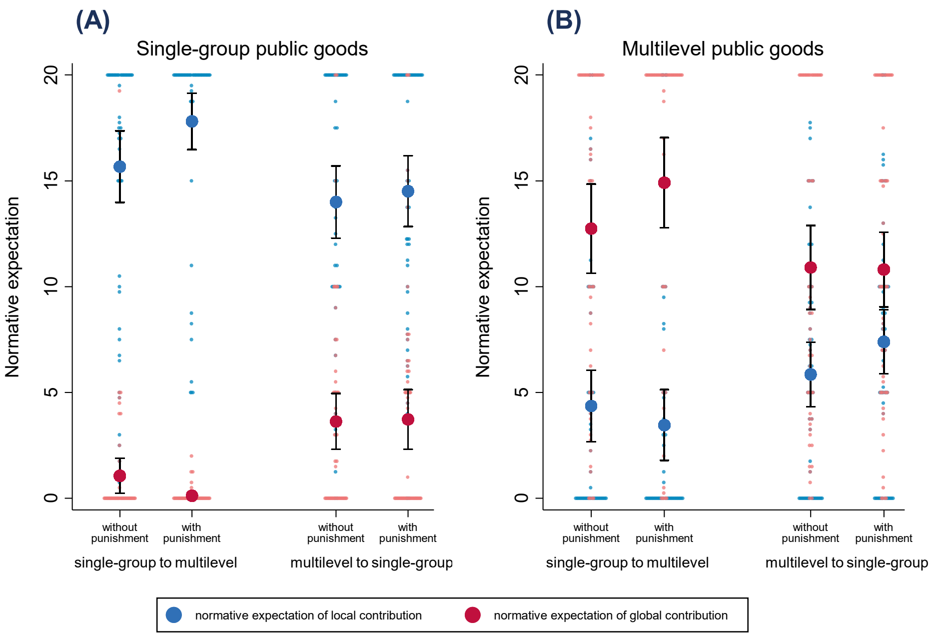


Figure D.2. Norms in single-group and multilevel public goods games measured in round 10.

Note: We present the normative expectations measured in round 10 of the first set of rounds. The large markers show mean values, and the small markers show values of individual participants. 95% confidence intervals are included via capped spikes. In single-group public goods games (A), we find that normative expectations of local cooperation are higher if peer punishment is possible in condition ‘single-group to multilevel’. In multilevel public goods games (B), we find that normative expectations of global cooperation are somewhat higher if peer punishment is possible in condition ‘single-group to multilevel’.

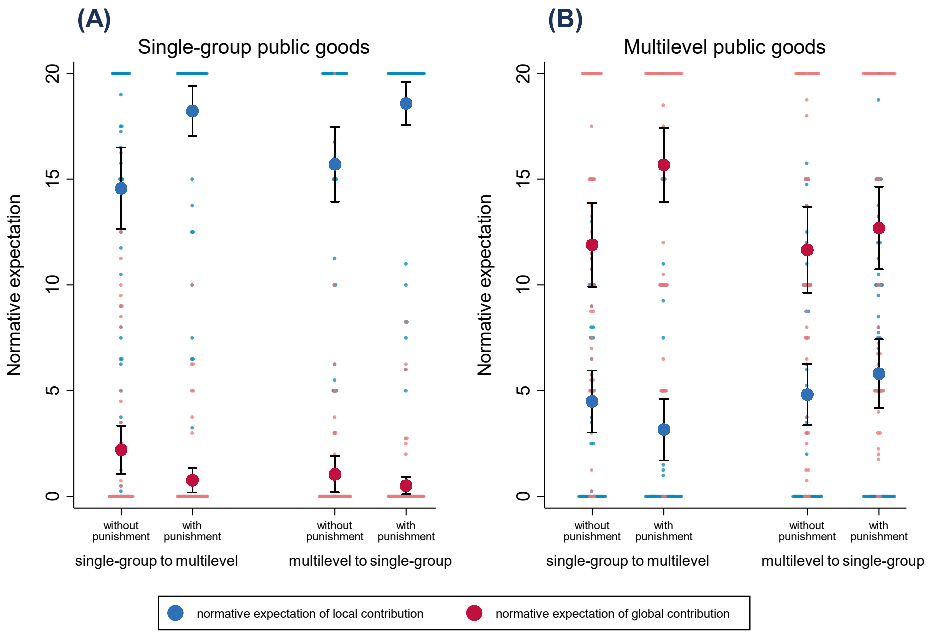


Figure D.3. Norms in single-group and multilevel public goods games measured in round 20.

Note: We present the normative expectations measured in round 10 of the second set of rounds. The large markers show mean values, and the small markers show values of individual participants. 95% confidence intervals are included via capped spikes. In single-group public goods games (A), we find that normative expectations of local cooperation are higher if peer punishment is possible in condition ‘single-group to multilevel’ and condition ‘multilevel to single-group’. In multilevel public goods games (B), we find that normative expectations of global cooperation are higher if peer punishment is possible in condition ‘single-group to multilevel’.

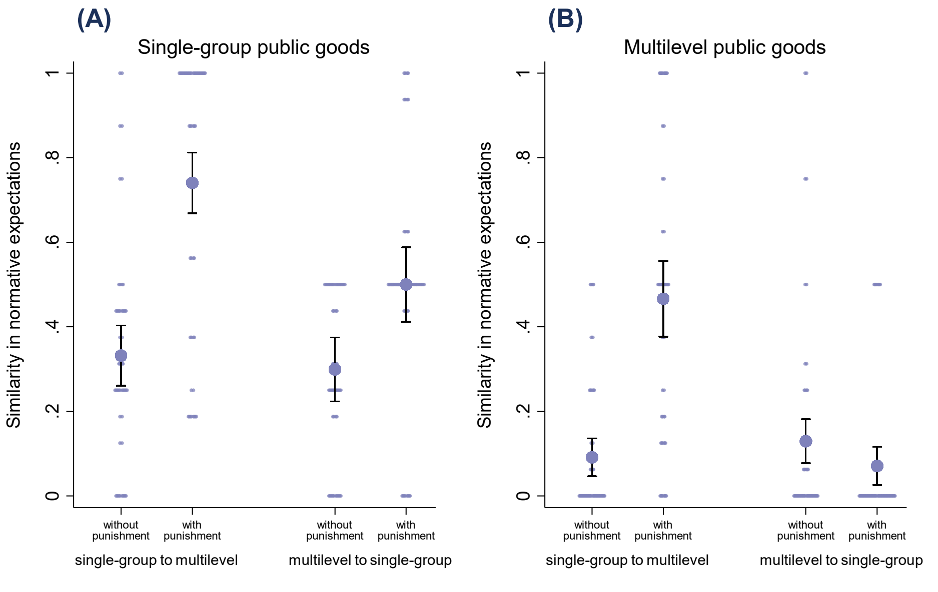


Figure D.4. Shared normative expectations in single-group and multilevel public goods games.

Note: Similarity in normative expectations of groups averaged over the first and second set of 10 rounds. As mentioned in the main text, all participants were asked to guess what their group members deemed the appropriate contribution to the local good and global good for a hypothetical group of four members in a single-group public goods game and a multilevel public goods game. Per hypothetical group, a participant thus makes four guesses about the appropriate contribution to the local good and the global good, and each of these guesses may be the same as, or different from, the guesses made by the other three group members. We examine the proportion of these guesses that were exactly the same between all three members of a group. The higher this proportion, the more the members share their normative expectations and hence share a social norm. For single-group public goods games (A), we find that similarity in normative expectations is higher when peer punishment is possible for conditions ‘single-group to multilevel’ and ‘multilevel to single-group’. For multilevel public goods games (B), we find that similarity in normative expectations is higher when peer punishment is possible only in condition ‘single-group to multilevel’.

D.4. Multilevel models of contribution behavior

Table D.12. Multilevel regression of total cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Predictors				
Punishment	3.40*	3.75*	4.85***	1.43
	(1.62)	(1.54)	(0.96)	(1.15)
Intercept	14.10***	14.67***	13.24***	15.45***
	(1.22)	(1.36)	(0.74)	(0.94)
Random effects				
Group	10.45			
	(1.72)			
Individual	5.44			
	(0.98)			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Number of observations is 4400 (220 participants in two sets of 10 rounds). Robust standard errors in parentheses. Repeated observations within individuals and groups are taken into account by estimating random effects for individuals and groups.

Table D.13. Multilevel regression of local cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Predictors				
Punishment	3.81*	3.97*	-1.56	1.24
	(1.68)	(1.64)	(1.35)	(1.49)
Intercept	13.10***	13.97***	4.94***	7.20***
	(1.22)	(1.36)	(0.90)	(1.09)
Random effects				
Group	14.38			
	(2.02)			
Individual	5.57			
	(1.02)			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Number of observations is 4400 (220 participants in two sets of 10 rounds). Robust standard errors in parentheses. Repeated observations within individuals and groups are taken into account by estimating random effects for individuals and groups.

Table D.14. Multilevel regression of global cooperation by punishment.

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Predictors				
Punishment	-0.40	-0.22	6.41***	0.20
	(0.35)	(0.40)	(1.68)	(1.73)
Intercept	1.00***	0.70*	8.31***	8.25***
	(0.28)	(0.34)	(0.91)	(1.11)
Random effects				
Group		8.94		
		(1.82)		
Individual		4.96		
		(0.88)		

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Number of observations is 4400 (220 participants in two sets of 10 rounds). Robust standard errors in parentheses. Repeated observations within individuals and groups are taken into account by estimating random effects for individuals and groups.

Table D.15. Multilevel regression of local and global cooperation in multilevel public goods.

	Single-group to multilevel	Multilevel to single-group
Without punishment		
Local cooperation	4.94***	7.20***
	(0.91)	(1.10)
Global cooperation	8.31***	8.25***
	(0.91)	(1.11)
With punishment		
Local cooperation	3.38**	8.43***
	(1.01)	(1.01)
Global cooperation	14.71***	8.45***
	(1.42)	(1.33)
Random effects		
Group	1.39	
	(0.31)	
Individual	0.21	
	(0.43)	

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Robust standard errors in parentheses. Number of observations is 4400 (220 participants making 10 local contribution decisions and 10 global contribution decisions each in the multilevel public goods game). Repeated observations within individuals and groups are taken into account by estimating random effects for individuals and groups.

D.5. Multilevel models of punishment behavior

Table D.16. Multilevel regression of punishment frequency by public goods game (single-group or multilevel) and order (single-group to multilevel or vice versa).

	Single-group public goods		Multilevel public goods	
	Single-group to multilevel	Multilevel to single-group	Single-group to multilevel	Multilevel to single-group
Punishment percentage	15.58*** (3.91)	16.33*** (4.46)	20.96*** (5.59)	36.83*** (6.88)
Random effects				
Group	3.04 (0.72)			
Individual	2.88 (0.64)			

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. $N = 2240$ because there are 112 participants in the punishment conditions and each participant makes punishment decisions in two sets of 10 rounds (112 × 10 × 2). Robust standard errors in parentheses. Repeated observations within individuals and groups are taken into account by estimating random effects for individuals and groups.

D.6. Experimental instructions and screens

The experimental instructions are presented on the next page. The text in square brackets was only presented to participants in conditions with punishment. This text was not in square brackets when the instructions were presented to these participants.

Experimental Laboratory for Sociology and Economics Universiteit Utrecht

Welcome

Welcome to this experiment and thank you for coming. Please read the following instructions carefully. These instructions are the same for all participants. The instructions state everything you need to know in order to participate in the experiment. If you have any questions, please raise your hand. One of the experimenters will approach you and answer your question.

The experiment is about group decision making. You can earn money by means of earning points during the experiment. The number of points that you earn depends on your own choices and the choices of other participants. At the end of the experiment, the total number of points that you earned will be exchanged at a rate of: **70 points = 1 Euro**.

The money you earn will be paid out in cash at the end of the experiment. There is a minimum payment of 5 euros and a maximum payment of 25 euros. Other participants will not see how much you have earned. During the experiment, you are not allowed to communicate with other participants. Please turn off your mobile phone. You may only use functions on the computer screen that are necessary to carry out the experiment.

Overview of the Session

The experiment consists of **2 parts** and lasts about **1 hour and 15 minutes in total**. Before each part of the experiment, you will receive instructions explaining that particular part.

In the **1st part**, you will participate in 8 to 12 rounds of a decision situation that will be explained below. The exact number of rounds remains unknown to you. Before and after you participate in these rounds, we will ask you to answer some questions. Because you participate together with other persons, you will sometimes have to wait until the other persons have made their decisions. These waiting times are incorporated in the total duration of the experiment.

Instructions for part 1

Each participant will be randomly assigned a color: **Blue**, **Red**, or **Green**. You will be randomly assigned to a group of 4 members (including you) that either consists of all members of the same color or a mix of 2 members of your color and 2 members of another color. We will tell you your color and the color of your group members before part 1 starts. Each participant's color will stay the same for the entire experiment. We will now describe the decision situation that you will be facing in each round. A decision situation consists of 2 stages: a contribution stage followed by a review stage. We start with describing the contribution stage.

Contribution stage

You and the three other members of your group are **each given 20 points**. Everyone has to decide how to allocate their points. You will have three alternative ways to use your points:

1. Keep them for yourself in a **private account**.
2. Contribute them to the **color account**, which provides points to you and other **members of the same color as you**.
3. Contribute them to the **collective account**, which provides points to you and all other **members, regardless of their color**.

There is no right or wrong way to allocate your points. Furthermore, you do not have to allocate points to all three alternatives; you can also allocate any amount to two alternatives or only one alternative. However, the way you allocate your points will affect your income and your group members' income. We explain this in detail next.

Income from the private account

Each point you keep in your private account provides income for yourself. For example, if you keep 6 points in your private account, your income from this account will be 6 points. **No one except you earns anything from your private account.**

Income from the color account

Each member **of the same color as you** will earn something from points you contribute to the color account. You will also earn something from the contributions to the color account made by you and other members of the same color as you.

For each point contributed to the color account (by you and other members of the same color): **you and the other members of the same color will each obtain 0.7 points**.

For example, suppose a group consists of two Red members and two Blue group members and that each member contributes 5 points to the color account. Contributions to the color account by Red members only benefit Red members, so each Red member obtains (5+5)*0.7 = 7 points. Similarly, contributions to the color account by Blue members only benefit Blue members, so each Blue member obtains (5+5)*0.7 = 7 points from the color account.

Income from the collective account

Each member, **regardless of their color**, will earn something from points you contribute to the collective account. You will also earn something from the contributions to the collective account made by you and all other group members.

For each point contributed to the collective account (by you or by any other member): **you and all other members will each obtain 0.5 points**

For example, suppose a group consists of two Red members and two Blue group members and that each member contributes 5 points to the collective account. Contributions to the collective account benefit all members, regardless of color. So each Blue and each Red member obtains (5+5+5+5)*0.5 = 10 points from the collective account.

Total income from the private account, the color account, and the collective account

Each member's total income from the contribution stage is the combined income from his/her private account and the color account and the collective account, that is:
your income from your private account + your income from the color account + your income from the collective account.

Review stage

Each contribution stage is followed by a review stage. In the review stage, everyone in the group will see for the current round how much each of the other group members contributed to the color account and the collective account and how much they kept for themselves in their private account.

[Then, all group members have the possibility to **decrease** each other's incomes. That is, you can decide if you want to spend points to decrease the income of other members in your group and so can the other three members in your group.

If you want to decrease another member's income, you do that by assigning deduction points. **Every deduction point assigned to another group member reduces his/her income by 3 points, and your own income by 1 point**. Similarly, every deduction point that one of your group members assigns to you decreases your income by 3 points and costs the group member 1 point. Note that this might imply that you or other participants lose income in a particular round. If you do not want to decrease the income of a group member, you must assign him/her 0 deduction points. Every participant can assign up to 10 deduction points to each group member, regardless of the income from the contribution stage.

For example, if you assign 2 deduction points to a group member, this costs you 2 points and reduces the group member's income by 6 points (2 times 3). Another example: if one of your group members assigns 3 deduction points to you, this reduces the group member's income by 3 points and your income by 9 points (3 times 3).

After everyone made a decision, you will see on a new screen how many deduction points were assigned to you by the other group members and also what your total income for the round is. You will not see which other participant assigned deduction points to you; you can only see the total number of deduction points assigned to you and how that affected your income. Similarly, if you assigned deduction points to one or more of the other group members, they will not see that you are the one who assigned the points.]

Instructions for part 2 of the experiment will be provided on your computer screen after the 8 to 12 rounds of the decision situation.

Instructions for the second part of the experiment

You will now participate in another set of **8 to 12** rounds of the same decision situation.

Before you start, we will change the group composition.
You and another member of your group will be matched with two new members.
These two other members have interacted in the past 8 to 12 rounds together (in a group of 4), but you and your remaining group member have not interacted with them yet.

These two new members have a different color from you and your remaining group member. This means that your group will consist of a mix of 2 Red members and 2 Blue members. In this group, you will interact for the 8 to 12 remaining rounds of the decision situation.*

The rest of the decision situation will remain the same. You again have to choose how many points to keep for yourself in a private account and how many points you contribute to the color account and the collective account.

If you have read and understood these instructions enter 17 in the box below and click continue.

Remember:
You will earn 1 point for each point you keep in your private account.
For each point contributed to the color account (by you and other members of the same color): you and the other members of the same color will each obtain 0.7 points.
For each point contributed to the collective account (by you and all other members): you and all other members will each obtain 0.5 points.

* [note about instructions: The combination of colors may differ (e.g., 2 red and 2 green are also possible) depending on the colors assigned to the participants. In the conditions where groups move from a heterogeneous to homogeneous composition due to membership, the paragraph is replaced by the following paragraph (colors may again differ, e.g., 4 red members are also possible):

These two new members have the same color as you and your remaining group member. This means that your group will consist of 4 Blue members. In this group, you will interact for the 8 to 12 remaining rounds of the decision situation.]

Quiz

1. Suppose you are **Red** and one of your **Red** group members contributes 10 points to the **color account**. How much does this contribution increase your income (if any)?

a) 10 * 0.7 = 7 points
b) 10 * 0.5 = 5 points
c) not increased

2. Suppose you are **Blue** and one of your **Blue** group members contributes 10 points to the **color account**. How much does this contribution increase your income (if any)?

a) 10 * 0.7 = 7 points
b) 10 * 0.5 = 5 points
c) not increased

3. Suppose you are **Blue** and one of your **Red** group members contributes 10 points to the **color account**. How much does this contribution increase your income (if any)?

a) 10 * 0.7 = 7 points
b) 10 * 0.5 = 5 points
c) not increased

4. Suppose you are **Blue** and one of your **Blue** group members contributes 10 points to the **collective account**. How much does this contribution increase your income (if any)?

a) 10 * 0.7 = 7 points
b) 10 * 0.5 = 5 points
c) not increased

5. Suppose you are **Blue** and one of your **Red** group members contributes 10 points to the **collective account**. How much does this contribution increase your income (if any)?

a) 10 * 0.7 = 7 points
b) 10 * 0.5 = 5 points
c) not increased

6. How much is your income reduced if your group members together assigned you a total of 2 deduction points?

a) 2 points
b) 6 points
c) 9 points

[Question 6 was only asked in conditions with punishment. The correct answers are: 1a, 2a, 3c, 4b, 5b, 6b].

Measurement of personal normative views

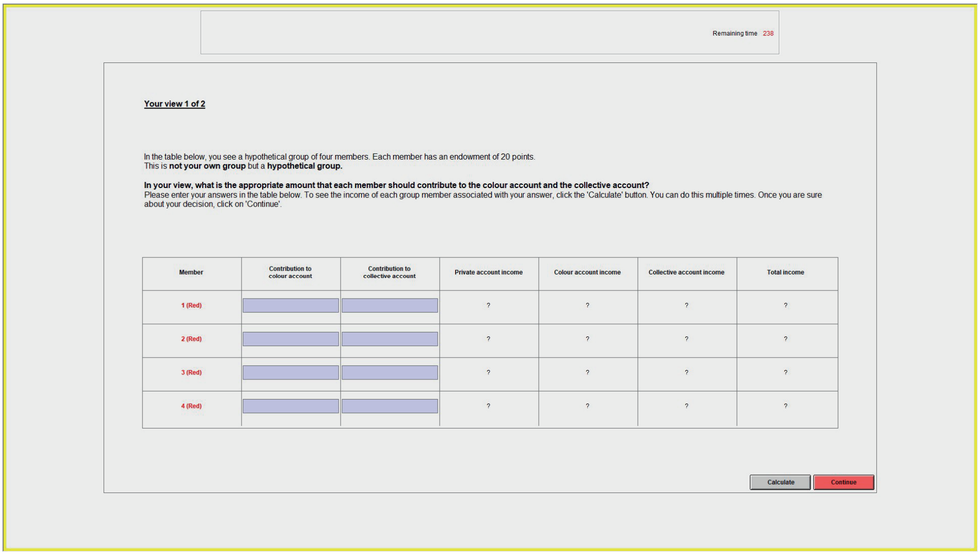


Figure D.5. Screen for the measurement of personal normative views about homogeneous groups.

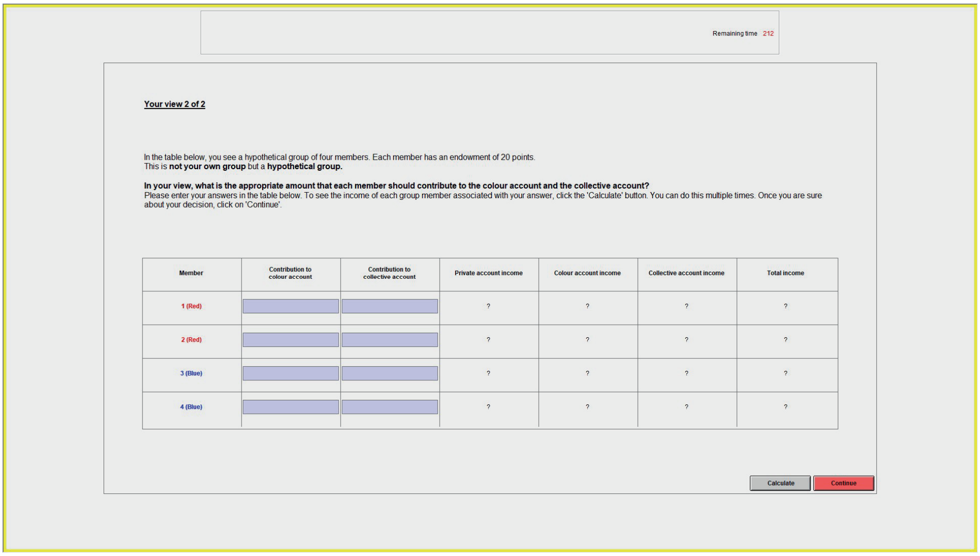


Figure D.6. Screen for the measurement of personal normative views about heterogeneous groups.

Measurement of normative expectations

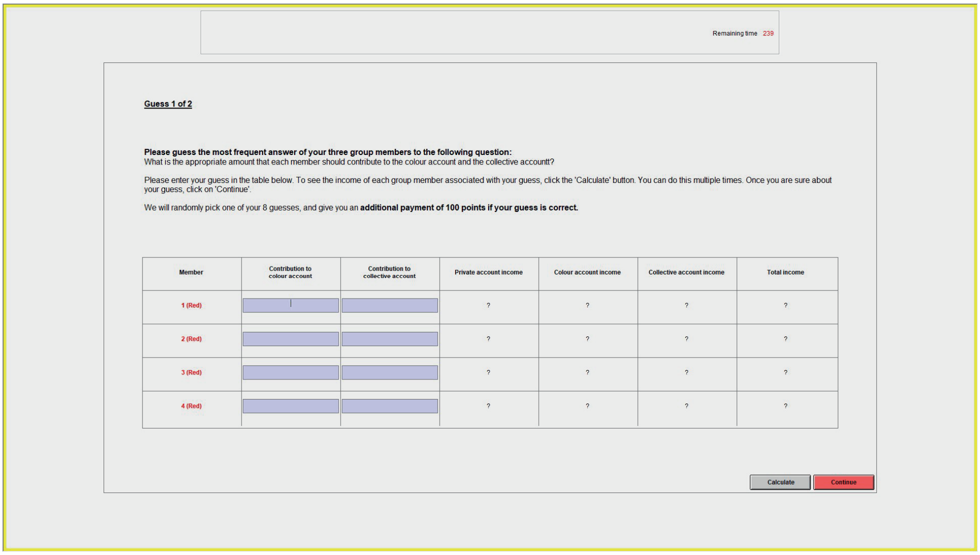


Figure D.7. Screen for the measurement of normative expectations about homogeneous groups.

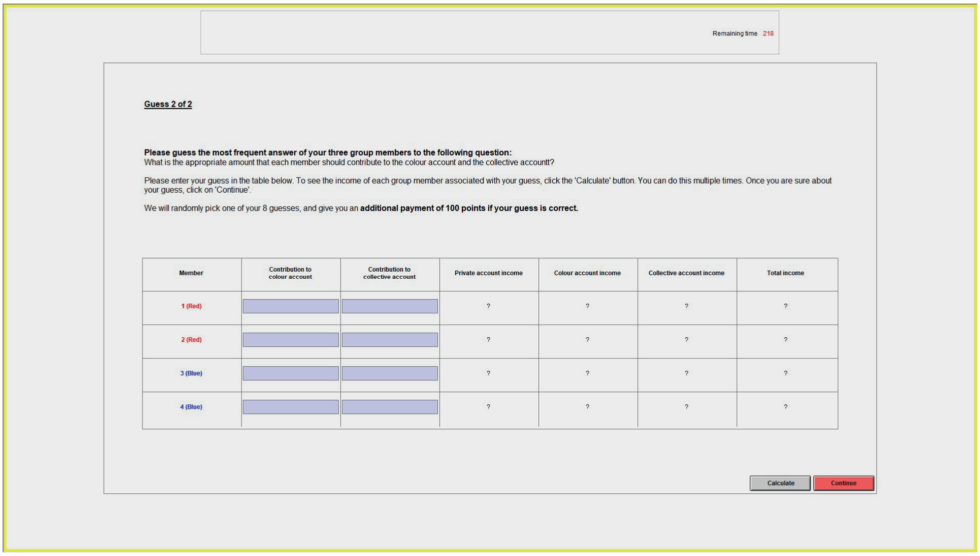


Figure D.8. Screen for the measurement of normative expectations about heterogeneous groups.

Measurement of empirical expectations

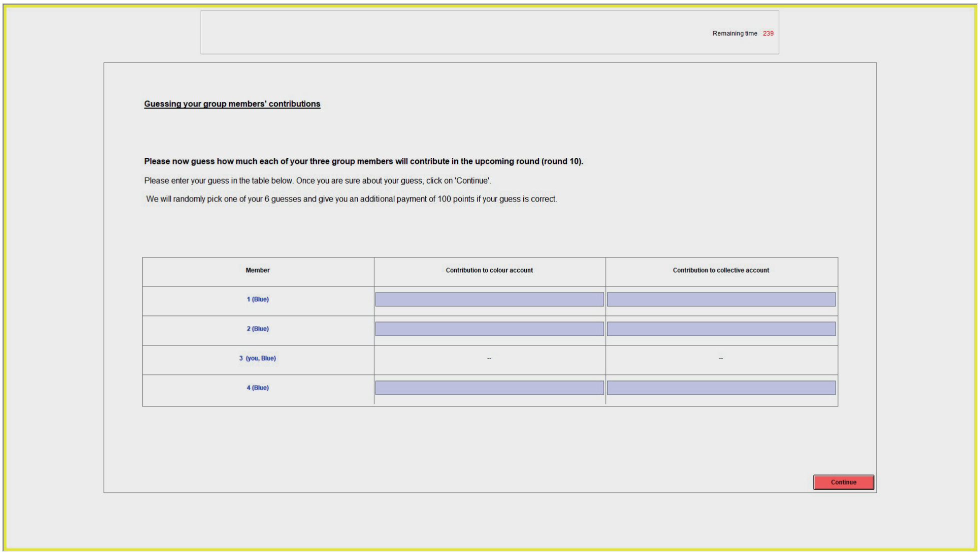


Figure D.9. Screen for the measurement of empirical expectations.
Note: This is an example for a participant that interacted in a homogeneous group. Participants that interacted in a heterogenous group would see different colors in the column 'Member'.

Appendix E. Supplementary material for Chapter 6

E.1. Instructions for confederates

The instructions as presented to the confederates are provided in this section. We provide both the original Dutch version and a version translated into English.

Instructies acteurs [Original Dutch version]
Je gaat personen benaderen in de buurt van treinstations, met het verzoek om achter hen aan te lopen tijdens het inchecken. Ook geef je een informatiekaart en stel je indien mogelijk een paar vragen. Bij elke sessie zijn twee acteurs aanwezig, die om de beurt personen aanspreken.

We benaderen alleen personen die aan de volgende drie eigenschappen voldoen:

1. Mannen
2. Boven de 18 jaar oud
3. Die alleen reizen (dus niet koppels of groepen personen).

Zodra we het experiment beginnen, benader je de eerste persoon die richting het station loopt en aan deze eisen voldoet. Je spreekt de persoon aan met de volgende zin:

"Ik wil een trein halen, mag ik achter u aanlopen wanneer u incheckt?".

Indien de reactie van de persoon onduidelijk is of de persoon een vraag stelt, herhaal je het verzoek één keer (als het daarna nog steeds niet duidelijk is, rekenen we dat als een reactie dat je **niet** achter der persoon aan mag lopen).

Afhankelijk van de reactie van de participant, doe je het volgende:

1. Als de persoon direct wegloopt na het horen of beantwoorden van je vraag:
Hiermee eindigt de interactie. Je vult een formulier in op een mobiele telefoon die wij je geven.

2. Als de persoon niet direct wegloopt na zijn reactie:
Je geeft de informatiekaart en spreekt de onderstaande zin uit.

"Ik moest het vragen voor een onderzoek van de Universiteit Utrecht. Mag ik u nog 3 korte vragen stellen?".

Let op: je gaat dus nooit echt door de OV-poortjes. Je geeft namelijk altijd direct na de reactie van de participant aan dat het om een onderzoek gaat.

2.1 Indien de persoon mee wil doen met de vragenlijst: je neemt de korte vragenlijst af en vult de antwoorden in op een mobiele telefoon. Daarna bedank je de persoon voor zijn tijd.

2.2 Indien de persoon niet mee wil doen: bedank de persoon voor zijn tijd. Hiermee eindigt de interactie. Je vult een kort formulier in op een mobiele telefoon.

Er zullen geen video opnames gemaakt worden. Je noteert wel de reactie van de persoon op een mobiele telefoon.

Na de interactie
De andere acteur is nu aan de beurt om een persoon aan te spreken.

Zodra de interactie tussen de andere acteur en persoon afgerond is, kan je de volgende persoon benaderen.

Als een persoon op locatie vragen heeft over het onderzoek, kan je hem naar de onderzoeker sturen die ook op locatie is.

Maak je geen zorgen als niet alles meteen duidelijk is, we nemen het samen op locatie nog even door en zullen ook even oefenen.

De informatiekaart
Voor het experiment krijg je van ons informatiekaarten die je aan de participanten geeft nadat ze hebben gereageerd op je vraag om achter hen aan te lopen. De informatiekaart is hieronder te zien:


Universiteit Utrecht

Voor meer informatie over het onderzoek kunt u de QR-code scannen of gaan naar: www.uu.nl/organisatie/sociologie/debriefing.

Indien u vragen of opmerkingen heeft, kunt u mailen naar Kasper Otten: k.d.otten@uu.nl.

Instructions confederates [English translation]
You will approach people near train stations, with the request to follow them while they check in. You also give an information card and ask a few questions if possible. Two actors are present at each session, who take turns approaching people.

We only approach people who meet the following three characteristics:

4. Men
5. Above 18 years old
6. Who travel alone (not couples or groups of people).

Once we start the experiment, you approach the first person who walks toward the station and meets these requirements. You address the person with the following sentence:

"I want to catch a train, can I follow you when you check in?".

If the person's response is unclear or the person asks a question, repeat the request once (if it is still not clear after that, we will count that as a response that you are not allowed to follow the person).

Depending on the participant's response, you do the following:

1. If the person walks away immediately after hearing or answering your question:
This ends the interaction. You fill in a form on a mobile phone that we give you.

2. If the person does not walk away immediately after his reaction:
You give the information card and deliver the sentence below.

"I had to ask it for a research project conducted by Utrecht University. May I ask you 3 short questions?".

Please note: you never really go through the check-in gates. You always indicate immediately after the response of the participant that it concerns a research project.

2.1 If the person wants to participate in the questionnaire: you take the short questionnaire and fill in the answers on a mobile phone. Afterward, you thank the person for their time.

2.2 If the person does not want to participate: thank the person for his time. This ends the interaction. You fill out a short form on a mobile phone.

No video recordings will be made. You do note down the person's response on a mobile phone.

After the interaction
It is now the turn of the other actor to approach a person.

Once the interaction between the other actor and person is complete, you can approach the next person.

If a person on location has questions about the research, you can send them to the researcher who is also on location.

Do not worry if not everything is immediately clear, we will go through it together on location and will also practice.

The information card
Before the experiment you will receive information cards from us that you give to the participants after they have responded to your request to follow them. The information card can be seen below:


Universiteit Utrecht

Voor meer informatie over het onderzoek kunt u de QR-code scannen of gaan naar: www.uu.nl/organisatie/sociologie/debriefing.

Indien u vragen of opmerkingen heeft, kunt u mailen naar Kasper Otten: k.d.otten@uu.nl.

E.2. Form and questionnaire

The form that confederates had to fill in is provided in this section. The form has conditional elements such that it could contain both the optional post-experimental questionnaire and the elements that confederates themselves had to note down after each interaction with a traveler. We provide the original form in Dutch and a version translated into English.

Original Dutch version

Mocht je achter de persoon aanlopen?

- Ja, na mijn eerste verzoek
- Ja, na het herhalen van mijn verzoek
- Nee, na mijn eerste verzoek
- Nee, na het herhalen van mijn verzoek

Wil de persoon meedoen aan de vragenlijst?

- Ja
- Nee
- Persoon ging weg voor ik het kon vragen

[Als de acteur 'Ja' selecteert, gaat hij automatisch naar **Block 1** en daarna pas naar **Block 2**. Als de acteur 'Nee' of 'Persoon ging weg voor ik het kon vragen' selecteert, gaat hij automatisch naar **Block 2.**]

Block 1

Lees de volgende tekst voor:

We zijn geïnteresseerd in de beweegredenen achter uw reactie op mijn verzoek om achter u aan te lopen door de OV-poortjes. Ik zal u hier enkele vragen over stellen, wat ongeveer 2 minuten kost.

Er zijn geen risico's verbonden aan uw deelname, we vragen niet naar persoonsgegevens, en we zullen uw antwoorden anoniem verwerken. De resultaten van de vragenlijst zullen meegenomen worden in een openbare wetenschappelijke studie. Deelname is vrijwillig en u kunt ten alle tijden stoppen.

Heeft u vragen over uw deelname?

Gaat u akkoord met deelname aan de vragenlijst?

Als de persoon **geen vragen heeft en akkoord gaat**, lees je de eerste vraag op de volgende pagina.

[Indien de acteur 'Ja, na mijn eerste verzoek' of 'Ja, na het herhalen van mijn verzoek' heeft geselecteerd]:

1. Waarom heeft u het verzoek om achter u aan te lopen door de OV-poortjes toegestaan?

[Indien de acteur 'Nee, na mijn eerste verzoek' of 'Nee, na het herhalen van mijn verzoek' heeft geselecteerd]:

1. Waarom heeft u het verzoek om achter u aan te lopen door de OV-poortjes geweigerd?

2A. In hoeverre vindt u het sociaal gewenst of ongewenst om door de OV-poortjes te gaan zonder in te checken? Kiest u alstublieft uit de volgende antwoord categorieën:

- Heel erg ongewenst
- Een beetje ongewenst
- Niet ongewenst en niet gewenst
- Een beetje gewenst
- Heel erg gewenst

2B. In hoeverre **verwacht u dat anderen** het sociaal gewenst of ongewenst vinden om door de OV-poortjes te gaan zonder in te checken? Kiest u alstublieft uit de volgende antwoord categorieën:

- Heel erg ongewenst
- Een beetje ongewenst
- Niet ongewenst en niet gewenst
- Een beetje gewenst
- Heel erg gewenst

3. Denkt u, in het algemeen, dat de meeste mensen te vertrouwen zijn, of dat je niet voorzichtig genoeg kunt zijn in de omgang met mensen? Wilt u een antwoord geven van 0 tot 10, waarbij 0 betekent dat je niet voorzichtig genoeg kunt zijn en 10 dat de meeste mensen te vertrouwen zijn.

Beëindig de vragenlijst met de volgende zinnen:

Dit is het einde van de vragenlijst. Heel erg bedankt voor uw deelname.
Daarmee eindigt de interactie. Ga alsjeblieft terug naar je plek en vul de rest van het formulier in.

Block 2

Voornaam acteur:

Probeer hieronder te herhalen wat de persoon precies zei als reactie op je verzoek om hem door de OV-poortjes te volgen:

Heb je nog opmerkingen over de interactie met de persoon? Zo ja, dan kan je die hieronder geven. Bijvoorbeeld als je onzeker was over de intentie van de persoon om je wel of niet door de OV-poortjes te laten.

Hoe goed denk je dat de persoon je (laatste) verzoek om achter hem aan te lopen begreep?

- o Goed
- o Niet goed, niet slecht
- o Slecht

Wat denk je dat de leeftijd van de persoon is?

- o 18-30 jaar
- o 31-40 jaar
- o 41-50 jaar
- o 51-60 jaar
- o 61-70 jaar
- o 71+

Wat denk je dat de (migratie)achtergrond van de persoon is?

- ☐ Nederland
- ☐ Turkije
- ☐ Marokko
- ☐ Suriname
- ☐ Indonesië
- ☐ Duitsland
- ☐ Polen
- ☐ Anders, vul hieronder in:

English translation

Were you allowed to follow the person?

- Yes, after my first request
- Yes, after repeating my request
- No, after my first request
- No, after repeating my request

Does the person want to participate in the questionnaire?

- Yes
- No
- Person left before I could ask

[If the confederate selects 'Yes', he is automatically directed to **Block 1** and afterward to **Block 2**. If the confederate selects 'No' or 'Person left before I could ask', he is automatically directed to **Block 2**.]

Block 1

Read the following text out loud:

We are interested in the motivations behind your reaction to my request to follow you through the check-in gates. I will ask you a few questions about this, which will take about 2 minutes.

There are no risks associated with your participation, we do not ask about personal information, and we will handle your answers anonymously. The results of the questionnaire will be incorporated in an openly available scientific article. Participation is voluntary and you can stop at all times.

Do you have any questions about your participation?

Do you agree with participating in the questionnaire?

If the person **does not have questions and agrees**, you read the first question on the following page.

[If the confederate selected 'Yes, after my first request' or 'Yes, after repeating my request']:

1. Why did you agree to the request to follow you through the check-in gates?

[If the confederate selected 'No, after my first request' or 'No, after repeating my request']:

1. Why did you deny the request to follow you through the check-in gates?

2A. To what extent do you find it socially appropriate or inappropriate to pass through the check-in gates without checking in? Please choose from the following answer categories:

- Very inappropriate
- Moderately inappropriate
- Not inappropriate and not appropriate
- Moderately appropriate
- Very appropriate

2B. To what extent **do you expect others** to find it socially appropriate or inappropriate to pass through the check-in gates without checking in? Please choose from the following answer categories:

- Very inappropriate
- Moderately inappropriate
- Not inappropriate and not appropriate
- Moderately appropriate
- Very appropriate

3. Do you think, in general, that most people can be trusted, or that you cannot be careful enough in dealing with people? Please give an answer ranging from 0 to 10, where 0 means that you cannot be careful enough and 10 means that most people can be trusted.

End the questionnaire with the following sentences:

This is the end of the questionnaire. Many thanks for your participation.

With this ends the interaction. Please return to your sport and fill in the rest of the form.

Block 2

First name confederate:

Try to repeat below what the person said exactly in response to your request to follow them through the check-in gates.

Do you have any comments about the interaction with the person? If yes, you can add them below. For example, if you were unsure about whether the person intended to let you pass through the check-in gates or not.

How well do you think the person understood your (last) request to follow them?

- Good
- Not good, not bad
- Bad

What do you think the age of the person is?

- 18-30 years
- 31-40 years
- 41-50 years
- 51-60 years
- 61-70 years
- 71+

What do you think is the (migration)background of the person?

- ☐ The Netherlands
- ☐ Turkey
- ☐ Morocco
- ☐ Surinam
- ☐ Indonesia
- ☐ Germany
- ☐ Poland
- ☐ Other, please specify below:

E.3. Debriefing text

We provide both the original debriefing text in Dutch and a translation into English.

Original Dutch version

U bent zojuist benaderd met het verzoek om een persoon door de OV-poortjes te laten zonder dat deze persoon incheckt. Dit verzoek is onderdeel van een wetenschappelijke studie van de Universiteit Utrecht, uitgevoerd door Kasper Otten (k.d.otten@uu.nl).

In dit onderzoek zijn we geïnteresseerd in hoe mensen reageren op het verzoek van een ander persoon om door de OV-poortjes te gaan zonder in te checken, waarbij dit verzoek wordt gezien als een vorm van normafwijkend gedrag.

In het onderzoek laten wij onderzoeksassistenten het verzoek doen om door de OV-poortjes te gaan zonder in te checken. Om de integriteit van het onderzoek te waarborgen, konden we u pas informeren over het onderzoek nadat u gereageerd heeft op de onderzoeksassistent.

Er worden tijdens het onderzoek geen persoonsgegevens genoteerd. Mocht u toch willen dat uw reactie op de onderzoeksassistent niet wordt meegenomen in het onderzoek, dan kunt u dat laten weten aan de onderzoeksassistent. Hij zal dan de notities van uw reactie verwijderen.

We willen u hartelijk bedanken voor uw medewerking in het onderzoek. Omdat het onderzoek nog loopt, zouden we het erg waarderen als u de achtergrondinformatie in deze brief niet deelt met anderen. **Wij vragen u dus uw kennis van dit onderzoek vertrouwelijk te houden.**

Als u vragen of opmerkingen heeft, kunt u die delen met de onderzoeksassistent of contact opnemen met Kasper Otten op emailadres: k.d.otten@uu.nl. Nogmaals hartelijk dank voor uw medewerking.

English translation

You were just approached with the request to let a person through the check-in gates without that person checking in. This request is part of a scientific study of Utrecht University, conducted by Kasper Otten (k.d.otten@uu.nl).

In this study, we are interested in how people react to the request of another person to go through the check-in gates without checking in, with this request being seen as a form of norm-violating behavior.

In this study, we let research assistants make the request to go through the check-in gates without checking in. To secure the integrity of the research, we could only inform you about the study after you responded to the research assistant.

No personal information is recorded during the study. Should you nevertheless want your reaction to the research assistant to be removed from the study, you can let this know to the research assistant. He will then remove the notes of your reaction.

We want to thank you very much for your participation in the research. Because the research is still ongoing, we would very much appreciate it if you do not share the background information in this letter with others. **We therefore ask you to keep your knowledge of this research confidential.**

If you have questions or comments, you can share them with the research assistant or contact Kasper Otten at the email address: k.d.otten@uu.nl. Many thanks again for your participation.

E.4. Instructions for coders

The instructions as presented to the coders are provided in this section. We provide both the original Dutch version and a version translated into English.

Experimental Laboratory for Sociology and Economics

Universiteit Utrecht

- Instructies [Original Dutch] –

Welkom

Welkom bij deze sessie en bedankt voor het komen! Leest u alstublieft de volgende instructies goed door. In de instructies wordt alles uitgelegd wat u moet weten om deel te nemen aan de sessie. Als u iets wilt vragen, steek dan uw hand op. Een van de onderzoekers zal naar u toe komen om uw vraag te beantwoorden.

In deze sessie laten we u interacties zien die plaatsvonden tijdens een sociaal experiment op treinstations in Nederland. In het sociaal experiment hebben wij acteurs ingehuurd om reizigers te benaderen die gingen inchecken bij de OV-poortjes op een treinstation. De acteurs stelden daarbij de volgende vraag aan de reizigers:

“Ik wil een trein halen, mag ik achter u aanlopen wanneer u incheckt?”

Wij hebben uw hulp nodig om de antwoorden die de reizigers gaven op deze vraag te beoordelen. Wij laten u daarom achtereenvolgens 57 antwoorden zien die de reizigers hebben gegeven. U verdient geld door uw mening te geven over deze antwoorden.

We vragen uw mening over (1) de mate waarin de antwoorden goedkeuring of afkeuring laten zien, (2) ze bereidheid om te helpen laten zien, (3) of ze duidelijk maken dat je hoort te betalen voor het openbaar vervoer, en (4) enkele andere details over de antwoorden.

U ontvangt 15 euro als u uw mening heeft gegeven over deze 57 antwoorden. De sessie duurt naar verwachting ongeveer 1 uur. Het is erg belangrijk dat u zorgvuldig uw mening geeft.

Het geld dat u verdient zal aan het einde van de sessie contant worden uitbetaald. Het is tijdens de sessie niet toegestaan met andere personen te communiceren. Zet alstublieft ook uw mobiele telefoon uit. Het is niet toegestaan andere functies van uw computer te gebruiken dan degene die noodzakelijk zijn voor de sessie. Hartelijk dank.

Meer informatie over het sociaal experiment

Het sociaal experiment is uitgevoerd tussen april 2022 en juni 2022. De acteurs waren mannen tussen de 19 en 27 jaar. De acteurs benaderden alleen reizigers die aan de volgende drie eigenschappen voldoen:

1. Mannen
2. Boven de 18 jaar oud
3. Die alleen reizen (dus niet koppels of groepen).

De reizigers die werden benaderd door de acteurs wisten aanvankelijk niet dat ze meededen aan een sociaal experiment. Pas nadat ze een antwoord hadden gegeven op de vraag van de acteurs, werd aan hen verteld dat dit voor een sociaal experiment was.

In de meeste gevallen laten we u het letterlijke antwoord van de reiziger zien. In sommige gevallen is niet het letterlijke antwoord van de reiziger gegeven, maar is beschreven hoe de reiziger reageerde (bijvoorbeeld ‘hij vond het prima’ of ‘hij knikte ja’). Op de achterzijde van deze pagina zijn twee voorbeeldinteracties te zien.

Voorbeeld Interactie 1

Vraag van acteur:

Ik wil een trein halen, mag ik achter u aanlopen wanneer u incheckt?

Antwoord van reiziger:

Nee je kan ook een kaartje kopen

Voorbeeld Interactie 2

Vraag van acteur:

Ik wil een trein halen, mag ik achter u aanlopen wanneer u incheckt?

Antwoord van reiziger:

Ja hoor loop maar achter me aan

Definities

Hieronder worden definities gegeven voor afkeuring, goedkeuring, en helpen.

<u>Afkeuring</u>
Personen laten afkeuring zien door aan te geven dat ze het ergens niet mee eens zijn, iets niet goed vinden, of een negatief oordeel hebben.

<u>Goedkeuring</u>
Personen laten goedkeuring zien door aan te geven dat ze het ergens mee eens zijn, iets goed vinden, of een positief oordeel hebben.

<u>Helpen</u>
Personen laten zien bereid te zijn om een ander te helpen door de ander bij te staan, te ondersteunen, of te assisteren. Dit staat los van of personen afkeuring/goedkeuring laten zien.

Experimental Laboratory for Sociology and Economics

Universiteit Utrecht

- Instructions [English translation] –

Welcome

Welcome to this session and thank you for coming! Please read the following instructions carefully. In the instructions, everything you need to know to participate in the session will be explained. If you want to ask something, raise your hand. One of the researchers will come to you to answer your question.

In this session, we will show you interactions that took place during a social experiment at train stations in the Netherlands. In the social experiment, we hired actors to approach travelers who were about to pass through the check-in gates at a train station. The actors asked the following question to the travelers:

"I want to catch a train, can I walk behind you when you check in?"

We need your help to rate the answers that travelers gave to this question. We will therefore show you consecutively 57 answers that the travelers gave. You earn money by giving your opinion about these answers.

We ask your opinion about (1) the extent to which the answers show approval or disapproval, (2) if they show intentions to help, (3) if they make clear that you should pay to use public transport, and (4) some other details about the answers.

You earn 15 euros once you have given your opinion about these 57 answers. The sessions will take about 1 hour on expectation. It is very important that you give your opinion diligently.

The money that you earn will be paid in cash at the end of the session. It is not allowed during the session to communicate with other persons. Please also turn off your mobile phone. It is not allowed to use other functions of the computer other than the ones necessary for the session. Many thanks.

More information about the social experiment

The social experiment was conducted between April 2022 and June 2022. The actors were men between 19 and 27 years old. The actors only approached travelers who met the following three criteria:

1. Men
2. Above 18 years old
3. Who were traveling alone (so no couples or groups).

The travelers who were approached by the actors did initially not know that they were part of a social experiment. Only after they gave an answer to the question of the actors, were they told that this was for a social experiment.

In most cases, we show you the literal answer of the traveler. In some cases, not the literal answer of the traveler is shown, but instead it is described how the traveler reacted (for example 'he thought it was fine' or 'he nodded yes'). On the backside of this page, we show you two example interactions.

Example interaction 1

Actor's request:

I want to catch a train, can I walk behind you when you check in?

Traveler's answer:

No you can also buy a ticket

Example interaction 2

Actor's request:

I want to catch a train, can I walk behind you when you check in?

Traveler's answer:

Yes sure just walk behind me

Definitions

Below, definitions are given for disapproval, approval, and helping.

<u>Disapproval</u>
Persons show disapproval by indicating that they disagree with something, that they think something is not good, or by having a negative judgment.

<u>Approval</u>
Persons show approval by indicating that they agree with something, that they think something is good, or by having a positive judgment.

<u>Helping</u>
Persons show a willingness to help another person by having the person's back, supporting the person, or assisting the person. This is independent of whether persons show approval/disapproval.

E.5. Questions for coders

The questions that coders had to answer to rate the travelers' verbal answers are provided in this section. We provide the original questions in Dutch and a version translated into English. We also provide a screenshot with an example screen of how the questions were presented to the coders on the computer screen, using z-tree software (Fischbacher, 2007).

Original Dutch version

1. Hoe afkeurend of goedkeurend is volgens u het antwoord van de reiziger? Geef uw antwoord op een schaal van 1 (heel afkeurend) tot 5 (heel goedkeurend).

Heel afkeurend	1	2	3	4	5	Heel goedkeurend

2. In welke mate geeft het antwoord aan dat de reiziger bereid is de acteur te helpen? Geef uw antwoord op een schaal van 1 (heel terughoudend) tot 5 (heel bereid).

Heel terughoudend	1	2	3	4	5	Heel bereid

3. Verwijst de reiziger naar de regel dat je een kaartje moet hebben?
   - o Nee
   - o Een beetje
   - o Ja

4. Welke van de volgende aspecten zijn te vinden in de reactie van de reiziger? U kunt meerdere categorieën selecteren.

   a) Hoe het gebruik van het openbaar vervoer werkt
   b) Dat het ongepast is om zonder kaartje te rijden
   c) Dat hij denkt dat de acteur wil profiteren zonder te betalen
   d) Dat hij denkt dat de acteur een goede reden heeft voor zijn verzoek
   e) Waarom hij wel of niet wil helpen
   f) Dat de acteur zelf zijn probleem moet oplossen
   g) Dat hij wil weten waarom de acteur dit vraagt
   h) De reiziger geeft een reactie zonder toelichting (bijvoorbeeld 'ja hoor', 'nee')
   Anders, namelijk: [ ]

5. Heeft u opmerkingen over de reactie van de reiziger of uw antwoorden? Als u geen opmerkingen heeft, kunt u het onderstaande vak leeg laten.

[ ]

English translation

1. How disapproving or approving is the traveler's answer according to you? Give your answer on a scale from 1 (very disapproving) to 5 (very approving).

Very disapproving 1 2 3 4 5 Very approving

2. To what extent does the answer indicate that the traveler is willing to help the actor? Give your answer on a scale from 1 (very hesitant) to 5 (very willing).

Very hesitant 1 2 3 4 5 Very willing

3. Does the traveler refer to the rule that you need to have a ticket?

- No
- A little
- Yes

4. Which of the following aspects can be found in the traveler's reaction? You can select multiple categories.

a) How the use of public transport works
b) That it is inappropriate to travel without a ticket
c) That he thinks the actor wants to profit without paying
d) That he thinks that the actor has a good reason for his request
e) Why he does or does not want to help
f) That the actor himself has to solve his problem
g) That he wants to know why the actor asks this
h) The traveler gives a reaction without clarification (for example 'yes sure', 'no')

Other, namely:

5. Do you have comments about the traveler's reaction or your answers? If you have no comments, you can leave the box below empty.

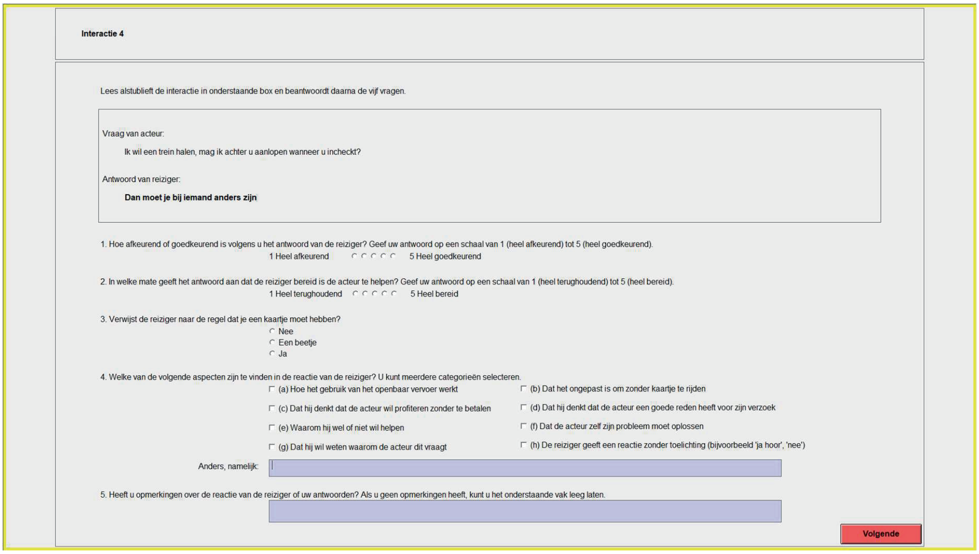


Figure E.1. Example screen of coders' items.

E.6. Further analyses of behavior

Figure E.2 provides photos of the confederates and their assigned numbers. Figure E.3 provides the percentage of help each confederate received on average. Tables E.1-E.3 provide regression analyses on (1) whether travelers helped the confederate, (2) whether travelers participated in the questionnaire, and (3) to what extent travelers find free-riding (dis)appropriate.

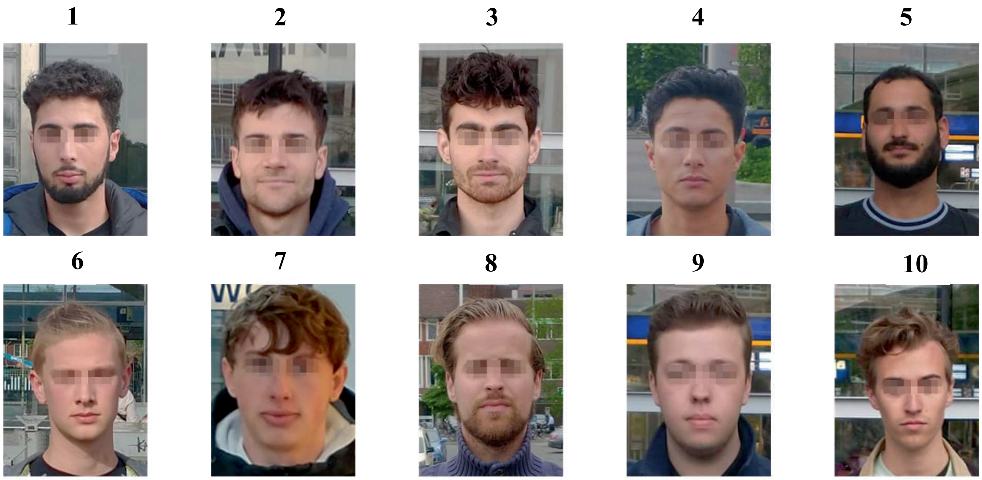


Figure E.2. Confederates in the field experiment.
Note: Confederates are numbered based on their percentage of received help (see also Figure E.3). The first five confederates have an ethnic-minority background, and the second five confederates have a native-majority background. The eyes are pixelated for anonymization.

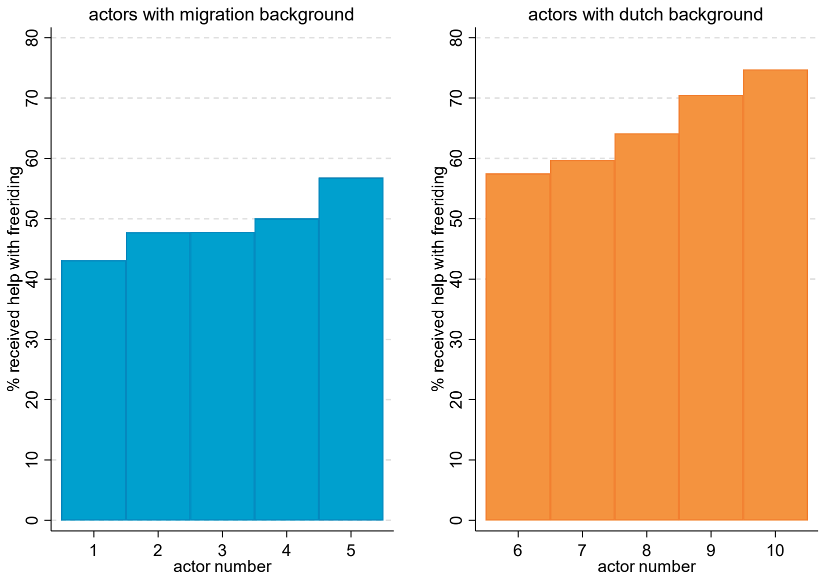


Figure E.3. Helping rates by confederate.

Note: Confederates are numbered based on their received help (see also Figure E.2). The first five confederates have an ethnic-minority background, and the second five confederates have a native-majority background. We see that all five confederates with an ethnic-minority background received lower helping rates than all five confederates with a native-majority background.

Table E.1. Percentage of received help with free-riding with control variables.

	estimate
Confederate's migration background	-17.05***
	(3.66)
Age (ref. is 18-30)	
Age 31-40	-14.36***
	(4.10)
Age 41-50	-20.03***
	(4.70)
Age 51-60	-32.18***
	(6.12)
Age 61+	-27.48**
	(9.15)
Origin (ref. is Dutch origin)	
Moroccan/Turkish origin	18.21**
	(5.97)
Other origin	9.46*
	(4.60)
Session (ref. is Session 1)	
Session 2	-0.42
	(8.09)
Session 3	19.07*
	(8.17)
Session 4	11.16
	(8.39)
Session 5	10.18
	(7.93)
Session 6	-15.78
	(11.73)
Session 7	7.61
	(10.47)
Session 8	14.68
	(8.46)
Session 9	7.80
	(7.95)
Session 10	15.07
	(8.76)
Session 11	9.90
	(8.19)
Session 12	6.44
	(7.84)
Session 13	28.19**
	(9.50)
Session 14	10.81
	(8.13)
Understanding (ref. is Good understanding)	
Moderate understanding	-23.98***
	(6.14)
Bad understanding	-44.38***
	(9.61)
Intercept	67.92***
	(6.39)
N observations	801
R^2	0.17

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Reference category for age is 'Age 18-30', for ethnic background is 'Dutch origin', for session is 'Session 1', and for understanding is 'Good understanding'.

Table E.2. Probability of participating in questionnaire with control variables.

	estimate
Confederate's migration background	-0.03
	(0.04)
Age (ref. is 18-30)	
Age 31-40	-0.09*
	(0.04)
Age 41-50	-0.20***
	(0.05)
Age 51-60	-0.22***
	(0.06)
Age 61+	-0.17
	(0.09)
Origin (ref. is Dutch origin)	
Moroccan/Turkish origin	0.00
	(0.06)
Other origin	0.03
	(0.05)
Session (ref. is Session 1)	
Session 2	-0.13
	(0.08)
Session 3	0.06
	(0.08)
Session 4	-0.12
	(0.08)
Session 5	-0.03
	(0.08)
Session 6	-0.10
	(0.12)
Session 7	-0.03
	(0.10)
Session 8	0.12
	(0.08)
Session 9	-0.02
	(0.08)
Session 10	-0.07
	(0.09)
Session 11	0.04
	(0.08)
Session 12	0.01
	(0.08)
Session 13	0.01
	(0.09)
Session 14	0.16*
	(0.08)
Understanding (ref. is Good understanding)	
Moderate understanding	-0.15*
	(0.06)
Bad understanding	-0.21*
	(0.10)
Intercept	0.44***
	(0.06)
N observations	801
R^2	0.09

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Reference category for age is 'Age 18-30', for ethnic background is 'Dutch origin', for session is 'Session 1', and for understanding is 'Good understanding'.

Table E.3. Personal opinion of the appropriateness of free-riding with control variables.

	estimate
Confederate's migration background	0.18
	(0.14)
Age (ref. is 18-30)	
Age 31-40	-0.07
	(0.15)
Age 41-50	-0.46*
	(0.21)
Age 51-60	-0.61*
	(0.29)
Age 61+	-0.82*
	(0.39)
Origin (ref. is Dutch origin)	
Moroccan/Turkish origin	0.23
	(0.23)
Other origin	0.34*
	(0.17)
Session (ref. is Session 1)	
Session 2	0.00
	(0.37)
Session 3	-0.01
	(0.30)
Session 4	0.24
	(0.36)
Session 5	0.57
	(0.31)
Session 6	0.75
	(0.46)
Session 7	0.49
	(0.40)
Session 8	0.71*
	(0.30)
Session 9	0.87**
	(0.30)
Session 10	0.09
	(0.38)
Session 11	0.18
	(0.30)
Session 12	0.51
	(0.30)
Session 13	0.38
	(0.34)
Session 14	0.51
	(0.28)
Understanding (ref. is Good understanding)	
Moderate understanding	0.30
	(0.31)
Bad understanding	1.40*
	(0.58)
Intercept	1.77***
	(0.23)
N observations	265
R^2	0.18

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Scores range from 1 (very disapproving) to 5 (very approving). Reference for age is 'Age 18-30', for ethnic background is 'Dutch origin', for session is 'Session 1', and for understanding is 'Good understanding'.

E.7. Further analyses of verbal expressions

The analyses on ratings of verbal (dis)approval are reported in the main text. The second aspect on which coders rated the verbal answers is helping intentions, by answering the question "To what extent does the answer indicate that the traveler is willing to help the actor?" on a 5-point Likert scale ranging from 1 (very hesitant) to 5 (very willing). The intercoder reliability is good for this rating (Krippendorff's alpha = 0.78). We take the average across the coders' ratings to assign each verbal answer a score on helping intentions. We find a significant difference in verbal expressions of helping intentions toward native-majority and ethnic-minority norm violators; the scores are 3.20 and 2.84 respectively (difference is 0.36 points, $t(799) = 3.64$, $p < .001$). The distribution of ratings for helping intentions toward native-majority and ethnic-minority norm violators is provided in Figure E.4. The figure shows that travelers are more likely to indicate that they are willing or very willing to help native-majority norm violators than ethnic-minority norm violators (53% vs 40%).

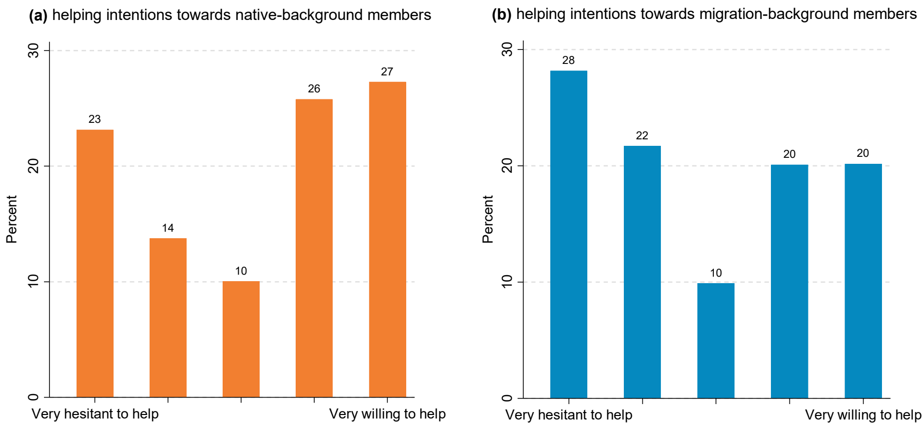


Figure E.4. Verbal expressions of helping intentions toward native-majority and ethnic-minority norm violators.

Note: **(A)** Distribution of ratings for helping intentions toward native-majority norm violators. **(B)** Distribution of ratings for helping intentions toward ethnic-minority norm violators.

The third aspect on which coders rated the travelers' verbal answers is mentions of the norm to pay for public transport. The raters answered the question "Does the traveler refer to the rule that you need to have a ticket?" with the answer categories being "Yes", "A little", and "No". The intercoder reliability is slightly lower than the threshold of 0.67 for sufficient reliability (Krippendorff's alpha = 0.62). The distribution of ratings for mentions of the rule to pay your fare toward native-majority and ethnic-minority norm violators is provided in Figure E.5. We see that the large majority of travelers do not explicitly mention the rule in their answers toward the norm violators. We do not find a significant difference between the likelihood of the three possible answer categories ("Yes", "A little", and "No") between native-majority and ethnic-minority norm violators ($F(2, 800) = 2.72$, $p = .07$).

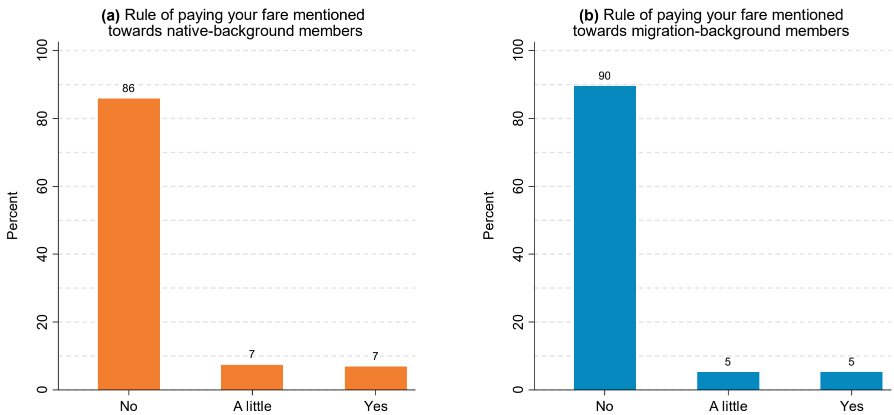


Figure E.5. Verbal expressions of the rule to pay your fare toward native-majority and ethnic-minority norm violators.

Note: **(A)** Distribution of ratings for mentions of the rule to pay your fare toward native-majority norm violators. **(B)** Distribution of ratings for mentions of the rule to pay your fare toward ethnic-minority norm violators.

The fourth question the raters answered was "Which of the following aspects can be found in the traveler's reaction?". The raters had to select all relevant categories from the list below. For each category on the list, we add between parentheses what the category was intended to measure, Krippendorff's alpha (KA), and the prevalence-and bias-adjusted kappa (PABAK) (this content between parentheses was not shown to the raters). PABAK is added as a measure of intercoder reliability because several answer categories on the list are selected very infrequently by the coders, and PABAK adjusts for this. PABAK values above 0.6 are generally considered to indicate sufficient reliability (Landis & Koch, 1977). In general, the intercoder reliability is mixed for this categorical item; Krippendorff's alpha suggests low reliability for most categories while PABAK suggests sufficient reliability for most categories (except the category 'no clarification').

Answer categories for item "Which of the following aspects can be found in the traveler's reaction?"

- Why he does or does not want to help (self-focused reaction, KA = 0.52, PABAK = 0.63)
- That the actor himself has to solve his problem (actor-focused reaction, KA = 0.24, PABAK = 0.63)
- That it is inappropriate to travel without a ticket (norm-focused reaction, KA = 0.37, PABAK = 0.69)
- How the use of public transport works (information-focused reaction, KA = 0.33, PABAK = 0.70)
- That he wants to know why the actor asks this (intention-seeking reaction, KA = 0.60, PABAK = 0.95)
- That he thinks the actor wants to profit without paying (assume negative intention, KA = 0.19, PABAK = 0.76)
- That he thinks that the actor has a good reason for his request (assume positive intention, KA = 0.16, PABAK = 0.76)
- The traveler gives a reaction without clarification (no clarification, KA = .57, PABAK = 0.57)
- Other, please specify (open-ended)

Figure E.6 presents the results for this categorical item. The figure shows that native-majority norm violators compared to ethnic-minority norm violators are more likely to receive verbal expressions that seek intentions behind the violator's request (3.3% vs. 0.6%, $t(799) = 3.41$, $p < .001$) and assume good intentions behind the violator's request (7.0% vs. 4.9%, $t(799) = 2.42$, $p = .02$). There are no significant differences between native-majority and ethnic-minority norm violators in terms of whether the travelers' answer was self-focused (17.2% vs. 21.0%, $t(799) = -1.81$, $p = .07$), actor-focused (11.0% vs. 9.0%, $t(799) = 1.51$, $p = .13$), norm-focused (13.5% vs. 12.8%, $t(799) = 0.46$, $p = .64$), information-focused, (10.1% vs. 10.7%, $t(799) = -0.43$, $p = .67$), assumed negative intentions (7.4% vs. 7.3%, $t(799) = 0.16$, $p = .87$), or provided no clarification (60.3% vs. 62.5%, $t(799) = -0.81$, $p = .42$).

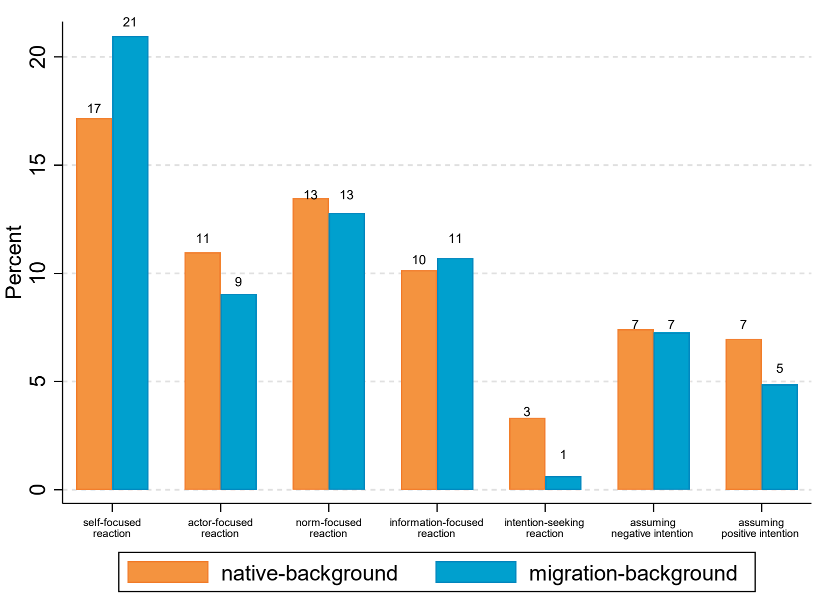


Figure E.6. Verbal expressions classified into several categories by the raters.
Note: Results are separated by verbal expressions targeted toward native-majority norm violators (orange bars, left bar for each category) and ethnic-minority norm violators (blue bars, right bar for each category).

Results separated by whether travelers helped the norm violators

Figures E.7-E.9 present the ratings for verbal expressions of (dis)approval (E.7), helping intentions (E.8), and mentions of the norm (E.9), separated not only by whether the confederate had a native-majority or ethnic-minority background, but also by whether the traveler did or did not allow the confederate through the gates. As can be expected, verbal expressions are much more approving and helpful from travelers who allow the confederate through the gates, and these travelers are also slightly less likely to mention the norm to pay the fare. We do not find evidence that rejections toward ethnic-minority norm-violators are more disapproving. If anything, the reverse seems to be the case (rejections seem to be a bit less disapproving toward ethnic-minority norm violators). That we still find that ethic-minority norm violators receive more verbal disapproval than native-majority norm violators is because ethic-minority norm violators' requests are more likely to be rejected, and rejections are more disapproving than non-rejections.

However, note that comparisons between native-majority and ethnic-minority confederates within the two subsets of travelers (travelers who did and did not let confederates through the gates) do not allow for causal inferences. The reason is that treatment assignment (the confederate's background) is only random across the entire group of travelers, and not the subset of travelers who rejected the norm violation (or the subset of travelers who helped the norm violation). Travelers who reject the request of native-majority members may differ from travelers who reject the request of ethnic-minority members, and this can confound effects when assessing discrimination within this group. For example, travelers who rejected the request of an ethnic-minority norm violator may want to prevent the appearance of discrimination by signaling helping intentions instead of disapproval.

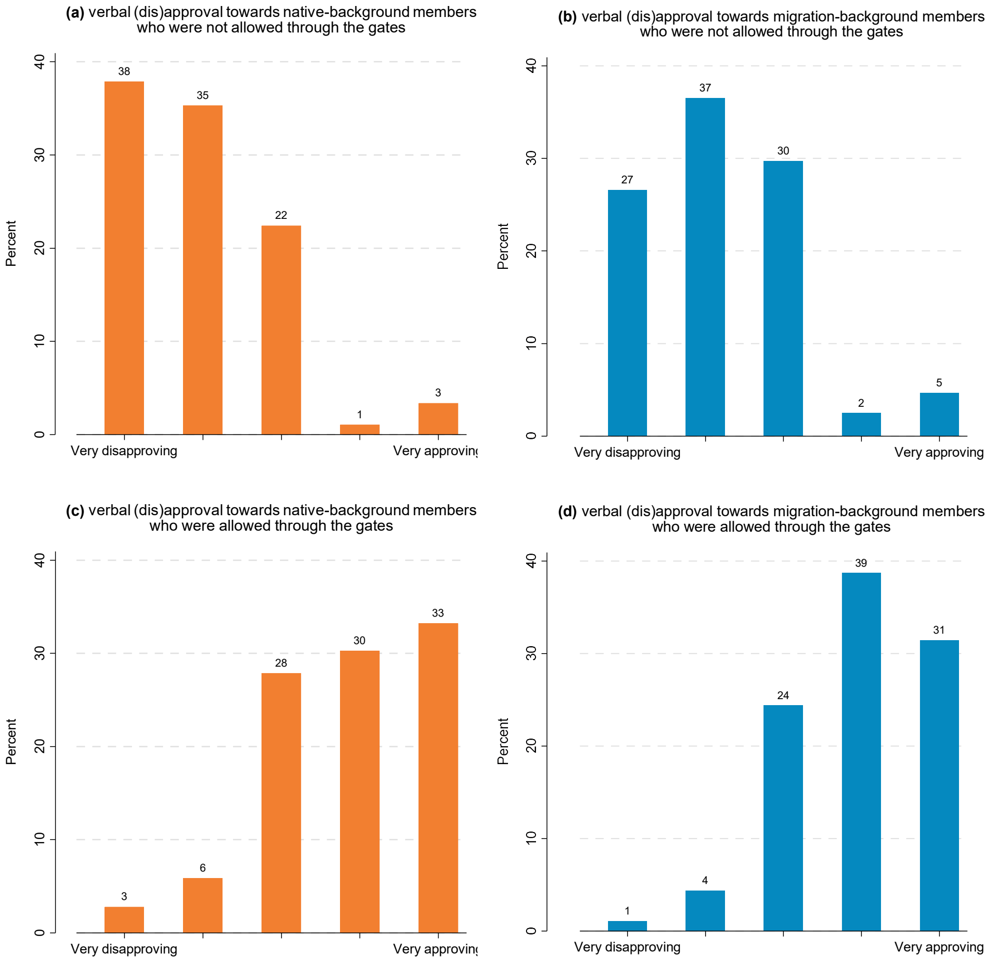


Figure E.7. Verbal (dis)approval toward native-majority and ethnic-minority norm violators by helping behavior of travelers.

Note: **(A)** Distribution of approval ratings for verbal answers toward native-majority norm violators who were not allowed through the gates. **(B)** Distribution of approval ratings for verbal answers toward ethnic-minority norm violators who were not allowed through the gates. **(C)** Distribution of approval ratings for verbal answers toward native-majority norm violators who were allowed through the gates. **(D)** Distribution of approval ratings for verbal answers toward ethnic-minority norm violators who were allowed through the gates.

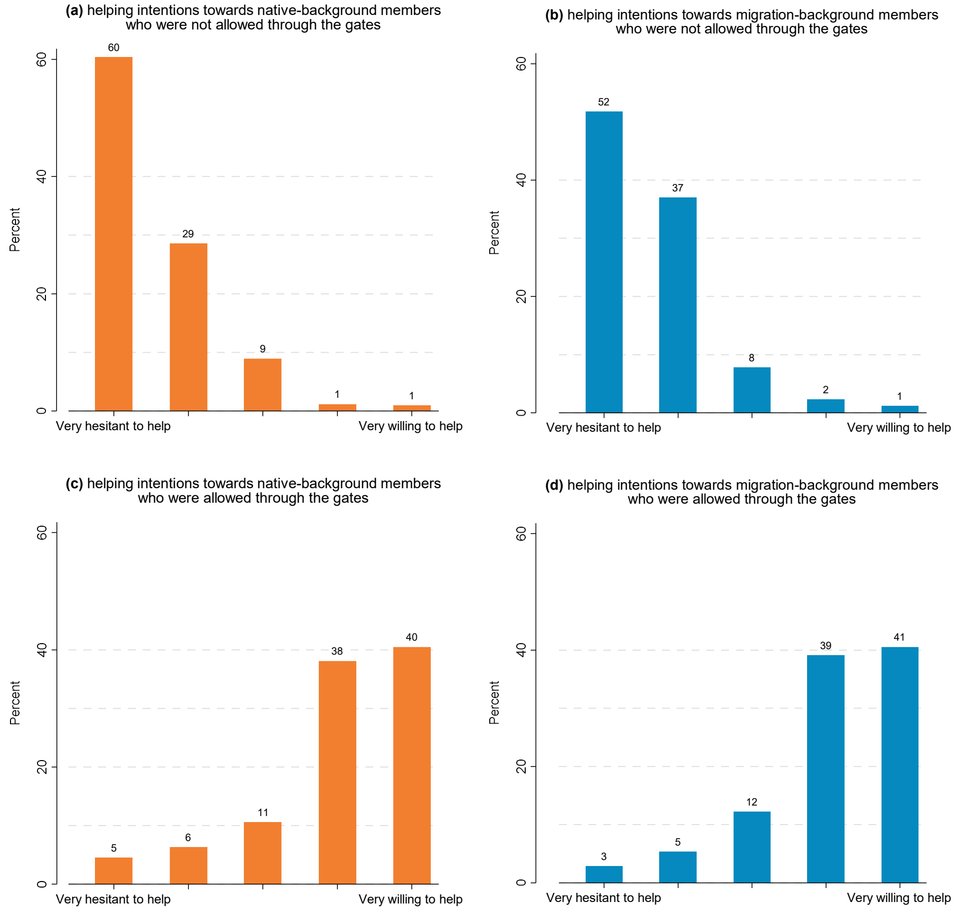


Figure E.8. Verbal expressions of helping intentions toward native-majority and ethnic-minority norm violators by helping behavior of travelers.

Note: **(A)** Distribution of ratings for helping intentions toward native-majority norm violators who were not allowed through the gates. **(B)** Distribution of ratings for helping intentions toward ethnic-minority norm violators who were not allowed through the gates. **(C)** Distribution of ratings for helping intentions toward native-majority norm violators who were allowed through the gates. **(D)** Distribution of ratings for helping intentions toward ethnic-minority norm violators who were allowed through the gates.

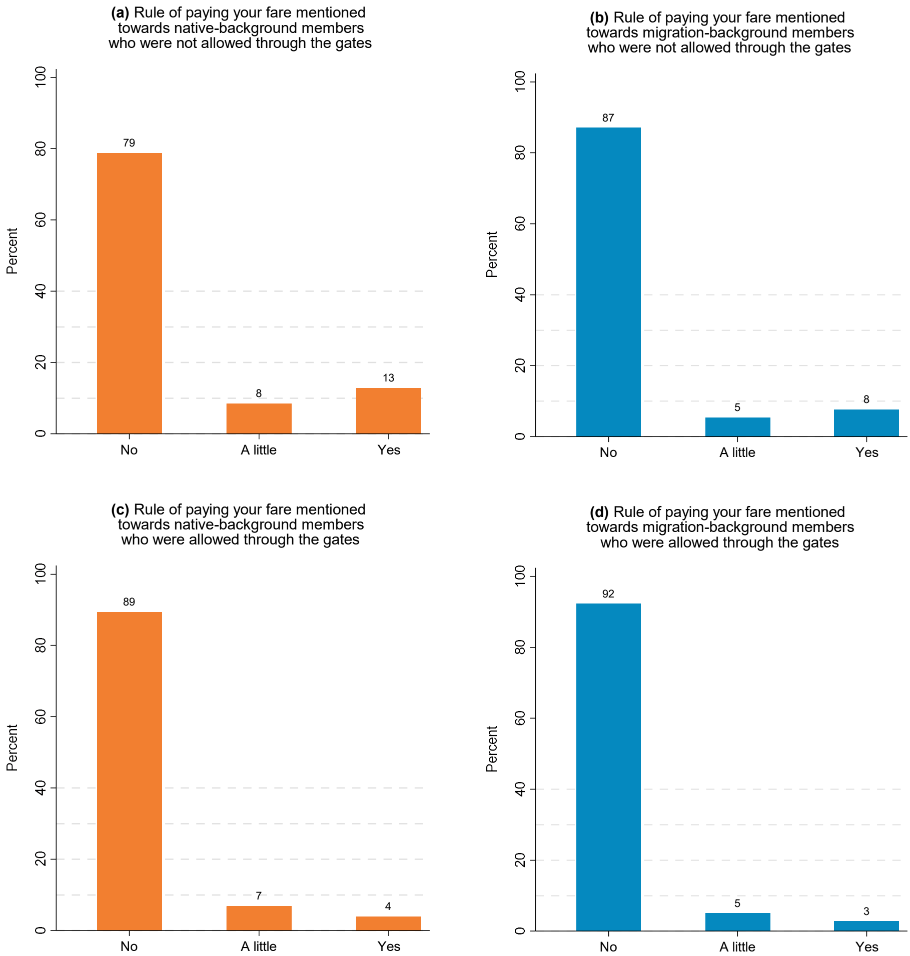


Figure E.9. Verbal expressions of the rule to pay your fare toward native-majority and ethnic-minority norm violators by helping behavior of travelers.

Note: **(A)** Distribution of ratings for mentions of the rule to pay your fare toward native-majority norm violators who were not allowed through the gates. **(B)** Distribution of ratings for mentions of the rule to pay your fare toward ethnic-minority norm violators who were not allowed through the gates. **(C)** Distribution of ratings for mentions of the rule to pay your fare toward native-majority norm violators who were allowed through the gates. **(D)** Distribution of ratings for mentions of the rule to pay your fare toward ethnic-minority norm violators who were allowed through the gates.

Table E.4. Approval ratings by coder demographics.

	estimate
Age	0.01
	(0.00)
Female	-0.03
	(0.06)
Student	0.10
	(0.10)
Political orientation	-0.00
	(0.01)
Intercept	2.90***
	(0.16)
N observations	3249
R^2	0.00

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses. Coders answered the question "How disapproving or approving is the traveler's answer according to you?" on a 5-point Likert scale ranging from 1 (very disapproving) to 5 (very approving). Age is continuous, female and student are binary variables. Political orientation is measured on a 10-point scale ranging from 1 (Left) to 10 (Right).

Table E.5. Examples of verbally disapproving and approving verbal answers in Dutch.

Disapproving
Nee tuurlijk niet
Nee, absoluut niet
Mijn trein staat er al en ik vind dit zeer ongepast
Nee nee zeker niet
Nee je moet zelf een kaartje kopen
Nee op tiefen
Nee nee nee, daar doe ik niet aan mee
Absoluut niet
Nee rot op
Nee natuurlijk niet idioot
Approving
Helemaal prima jongen
Ja ik zou het zelf ook doen
Ja natuurlijk! Kom maar!
Hahahahaha ja kom maar joh
Ja hoor absoluut
Ja tuurlijk mag dat
Ja tuurlijk mag jij dat, loop maar achter me aan!
Ja, is goed man!
Ja hoor, kom maar!
Ja natuurlijk

Note: The question asked by the confederates was "I want to catch a train, can I walk behind you when you check in?" We show verbal answers on which all coders who saw the answers agreed that they were very disapproving (first ten rows) or very approving (last ten rows).

E.8. Further analyses of post-experimental questionnaire

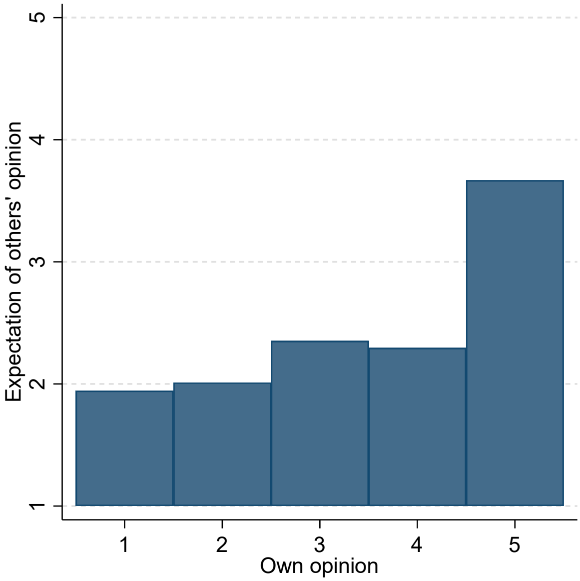


Figure E.10. Travelers' own opinion and their expectations about others' opinion regarding free-riding in public transport.

Note: Travelers answered to what extent they find it socially appropriate or inappropriate to pass through the check-in gates without checking in on a 5-point Likert scale ranging from 1 (very inappropriate) to 5 (very appropriate) (horizontal axis). They were also asked to what extent they expect others to find this to be socially appropriate or inappropriate, using the same Likert scale (vertical axis). 265 travelers reported their opinions and expectations ($N = 71$ for own opinion = 1, $N = 94$ for own opinion = 2, $N = 74$ for own opinion = 3, $N = 17$ for own opinion = 4, and $N = 9$ for own opinion = 5). We find that travelers expect others to judge free-riding to be moderately inappropriate on average (a score of 2 on the vertical axis) regardless of their own opinion. The only exception is formed by the 9 travelers who judge free-riding to be very appropriate (own opinion = 5 on the horizontal axis); they expect others to be approximately neutral toward free-riding in public transport (a score of 3-4 on the vertical axis).

Generalized trust

After reporting their own opinion and expectations about others' opinion regarding free-riding in public transport, travelers answered an item on general trust ("In general, do you think that most people can be trusted or that you cannot be too careful when dealing with people?") on a scale from 0 (no trust) to 10 (complete trust). Travelers report moderate to high trust on average and this does not depend on whether they helped the norm violator or not (6.54 vs. 6.65, $t(263) = 0.36$, $p = .72$) or whether the norm violator had a native-majority or ethnic-minority background (6.37 vs. 6.77, $t(263) = 1.64$, $p = .10$). Generalized trust does not significantly correlate with travelers' own opinions (correlation = -0.04, $p = .56$) or travelers' expectations about others' opinions (correlation = 0.01, $p = .90$).

Nederlandse samenvatting

Achtergrond

Publieke goederen

Het bijdragen aan publieke goederen is door de tijd heen een cruciale activiteit geweest. Jager-verzamelaars droegen bij aan publieke goederen wanneer ze op grote schaal samen jaagden, hun voedsel deelden en werkten aan collectieve voorzieningen zoals jachtnetten en irrigatie (Boyd & Richerson, 2022; Hawkes et al., 1993; Hill, 2002). Mensen dragen bij aan cruciale publieke goederen wanneer ze hun leven riskeren in een oorlog om hun land te verdedigen (Mathew & Boyd, 2011), wanneer ze hun baan riskeren om te staken voor betere collectieve arbeidsomstandigheden (Naylor, 1989) en wanneer ze meedoen aan protesten en revoluties tegen dictatoriale regimes (Muller & Opp, 1986). In het dagelijks leven dragen mensen bij aan publieke goederen wanneer ze vrijwilligerswerk doen voor lokale evenementen of clubs, de afwas doen in gedeelde woonruimtes, hun deel van teamtaken doen of maatregelen nemen om hun ecologische voetafdruk te verkleinen (Van Vugt et al., 2000). En in moderne samenlevingen dragen mensen via belastingen bij aan nationale publieke goederen zoals openbaar vervoer, openbaar onderwijs, handhaving van de openbare orde en sociale zekerheid (Kallhoff, 2014).

Hoewel al deze activiteiten op talloze manieren verschillen, hebben ze een gemeenschappelijk kenmerk – ze vereisen individueel kostbare bijdrages die ten goede komen aan de groep als geheel, inclusief leden die zelf niet hebben bijgedragen aan het publiek goed. De twee kenmerken die vaak gebruikt worden om publieke goederen te definiëren zijn dat (1) leden niet uitgesloten kunnen worden van het voordeel van het goed (niet-uitsluitbaarheid) en (2) de voordelen die een bepaald lid geniet niet ten koste gaan van andere leden (niet-rivaliserend). Deze twee kenmerken zijn gradueel van aard. Een goed is zelden perfect niet-uitsluitbaar en niet-rivaliserend (Malkin & Wildavsky, 1991). Zolang mensen van het goed kunnen profiteren zonder er zelf aan bij te dragen, beschouwen wetenschappers het over het algemeen als een publiek goed (Cornes & Sandler, 1994, 1996). Dit kenmerk van publieke goederen betekent dat de groep als geheel beter af is als groepsleden aan het publiek goed bijdragen, maar dat individuele leden in de verleiding kunnen komen om mee te liften op de bijdrages van anderen. Er is dus een sociaal dilemma dat centraal staat bij publieke goederen: kiezen tussen het eigen belang (meeliften) en het algemeen belang (bijdragen). In hoeverre mensen kunnen samenwerken voor publieke goederen ondanks de verleiding om mee te liften is een kwestie die nog steeds veel aandacht ontvangt, zowel vanuit de maatschappij als de wetenschap.

Eerder onderzoek heeft zich voornamelijk gericht op de omstandigheden die coöperatie voor publieke goederen bevorderen, maar onderzoek naar hoe deze coöperatie in stand kan worden gehouden wanneer de omstandigheden veranderen is nog schaars (Lazega et al., 2022; Olsson et al., 2015). We weten vooral nog weinig over hoe coöperatie in stand kan worden gehouden wanneer groepen veranderen door de komst van nieuwkomers wiens opvattingen en behoeften verschillen van die van de bestaande leden (leden die al deel uitmaakten van de groep). In dit proefschrift bestuderen we of en hoe coöperatie voor publieke goederen in stand wordt gehouden in veranderende groepen bestaande uit nieuwkomers en bestaande leden. Voordat we ingaan op de nieuwkomers, lichten we eerst kort toe hoe coöperatie tot stand kan komen voor publieke goederen.

Coöperatiemechanismes voor publieke goederen

Eerder onderzoek heeft meerdere manieren laten zien om coöperatie voor publieke goederen tot stand te brengen. Een bekende manier is een centrale autoriteit die de bijdrages aan publieke goederen verplicht, bijvoorbeeld de overheid. Landelijke publieke goederen zoals openbaar onderwijs, defensie en veiligheid, en sociale zekerheid zijn vaak centraal gereguleerd via wetshandhaving van de overheid. Centrale regulatie en wetshandhaving op zichzelf genomen zijn echter ook publiek goederen, die alleen kunnen functioneren als burgers het eens zijn over de regels die gehandhaafd moeten worden (Cowen, 1992; Tullock, 1971) en daaraan vrijwillig bijdragen via politieke participatie (bijvoorbeeld door te stemmen). Aan publieke goederen georganiseerd door een centrale autoriteit gaat dus een publiek goed zonder centrale autoriteit vooraf. Bovendien zijn er veel situaties waarbij publieke goederen niet centraal georganiseerd kunnen worden, variërend van kleinschalige samenwerkingsproblemen zoals teamwerk tot internationale problemen zoals het bestrijden van de klimaatcrisis. Het is daarom belangrijk om te kijken naar manieren waarop coöperatie voor publieke goederen tot stand kan komen zonder een centrale autoriteit.

Ten eerste kunnen individuen samenwerken zonder centrale autoriteit als ze familie zijn. Vanuit een evolutionair perspectief is samenwerking met familieleden goed voor de eigen genen, omdat mensen een deel van hun genen delen met hun familieleden (Hamilton, 1964). Veel van de eerdergenoemde voorbeelden van publieke goederen vinden echter niet plaats tussen familieleden. Een mechanisme dat samenwerking tussen niet-familieleden bevordert, is directe wederkerigheid. Bij directe wederkerigheid kunnen individuen zich coöperatief gedragen tegenover anderen omdat dit de kans vergroot dat deze anderen daar iets voor terugdoen (Axelrod, 1984; Trivers, 1971). Jager-verzamelaars die meer prooi vangen dan ze zelf kunnen eten bijvoorbeeld, kunnen hun overschot delen met anderen in de hoop dat deze anderen hetzelfde zullen doen in de toekomst. Directe wederkerigheid berust dus op herhaaldelijke interacties tussen *dezelfde* individuen. Wanneer herhaaldelijke interacties plaatsvinden tussen *verschillende* individuen, kan coöperatie plaatsvinden via indirecte wederkerigheid (Nowak & Sigmund, 1998). Bij indirecte wederkerigheid kunnen individuen zich coöperatief opstellen zelfs tegen personen die ze niet nog een keer tegenkomen omdat dit hun reputatie verbetert. Een goede reputatie kan anderen namelijk stimuleren om zich coöperatief op te stellen ten opzichte van degene die deze reputatie heeft (Buskens & Weesie, 2000; Granovetter, 1985; Raub & Weesie, 1990).

Sociale normen als centraal coöperatiemechanisme voor publieke goederen

Familiebanden, directe wederkerigheid en indirecte wederkerigheid zijn belangrijke mechanismen voor coöperatie, niet alleen bij mensen maar ook bij andere dieren (Bshary & Grutter, 2006; Carter, 2014). Het mechanisme dat echter vaak als het belangrijkst wordt gezien om de unieke grootte en diversiteit van coöperatie tussen mensen te verklaren, bestaat uit sociale normen (Bicchieri, 2006; Elster, 1989; Fehr & Schurtenberger, 2018; House et al., 2019). Sociale normen zijn informele regels die voorschrijven hoe mensen zich wel en niet moeten gedragen (Elster, 1989; Fehr & Fischbacher, 2004a; Tomasello & Vaish, 2013; Young, 2015). Sociale normen vertellen ons bijvoorbeeld dat we in de rij moeten wachten bij winkelkassa's, dat we ons afval in de vuilnisbak moeten gooien in plaats van op de grond en dat we de persoonlijke ruimte van anderen moeten respecteren. Het leren van en zich conformeren aan sociale normen gaat bij veel mensen natuurlijk en vaak onbewust (Boyd et al., 2011; Henrich, 2015; House et al., 2019). Wanneer normen sociaal kunnen worden afgedwongen door normovertreders te straffen en normconformisten te belonen, kan coöperatie worden bereikt zelfs in situaties waarin reputaties niet een rol spelen (Fehr & Gächter, 2002). Het vermogen om sociale normen te ontwikkelen en handhaven wordt zelfs gezien als een van de onderscheidende kenmerken van de mens (Fehr & Fischbacher, 2004b; Quervain et al., 2004). Natuurlijk zijn de verschillende mechanismen die coöperatie bevorderen niet volledig onafhankelijk van elkaar en kunnen ze op elkaar inwerken. Sociale normen kunnen bijvoorbeeld directe wederkerigheid voorschrijven (Kimbrough & Vostroknutov, 2016) en aangeven welk gedrag leidt tot goede en slechte reputaties (Ohtsuki & Iwasa, 2006).

De meningen verschillen over hoe sociale normen ontstaan (Brennan et al., 2013), maar veel wetenschappers zijn het erover eens dat herhaalde interactie tussen dezelfde groepsleden het ontstaan van sociale normen bevordert (Baldassarri & Abascal, 2020; Bendor & Swistak, 2001; Coleman, 1988; Ostrom, 2000; Przepiorka et al., 2022; Titlestad et al., 2019; Voss, 2001). Dat wil zeggen, herhaalde interactie met dezelfde groepsleden faciliteert een leerproces dat leden in staat stelt om tot overeenstemming te komen van wat gepast of "goed" gedrag is (Duffy & Ochs, 2009). Dit helpt groepsleden weer om te weten wat ze van anderen kunnen verwachten en om elkaar verantwoordelijk te houden voor oncoöperatief gedrag (Brennan et al., 2013). Stabiliteit in de groepssamenstelling is dus een belangrijke factor in het bevorderen van sociale normen en daarmee coöperatie voor publieke goederen. Maar wat gebeurt er als groepen veranderen door het vertrek van leden of de aankomst van nieuwe leden? Dit is wat we onderzoeken in dit proefschrift.

Groepsveranderingen en de gevolgen voor sociale normen en coöperatie

Hoewel stabiliteit in de groepssamenstelling belangrijk is voor sociale normen en het leveren van publieke goederen, krijgen vrijwel alle groepen op een bepaald moment te maken met vertrekkende leden en nieuwe leden. Landen, steden, en buurten veranderen van samenstelling door migratie, arbeidsorganisaties nemen nieuwe werknemers aan en laten bestaande werknemers gaan door pensioen, ontslag, of een nieuwe baan, vrijwilligersorganisaties en lokale clubs trekken nieuwe leden aan en zien andere leden vertrekken, enzovoort. Bovendien komen groepsveranderingen steeds vaker voor door recente migratiepatronen, globalisering, en technologische veranderingen (Tannenbaum et al., 2012). Immigratie naar West-Europa en Noord-Amerika is de afgelopen decennia toegenomen, wat heeft geleid tot discussie over de vraag of en hoe immigrantengroepen en de ontvangende bevolking kunnen samenwerken voor publieke goederen (Baldassarri & Abascal, 2020). De teloorgang van de “baan voor het leven” heeft geleid tot meer personeelsverloop in werkorganisaties, wat vragen oproept over hoe werkteams een succesvolle samenwerking tussen nieuwe en oude werknemers in stand kunnen houden (Berton & Garibaldi, 2012; Cahuc et al., 2016). Groepsveranderingen worden het "nieuwe normaal" genoemd (Huckman et al., 2009) en zijn tegenwoordig eerder regel dan uitzondering (Grund et al., 2018).

Het belang van een stabiele groepssamenstelling voor sociale normen en publieke goederen suggereert dat veranderingen in de groepssamenstelling publieke goederen juist bedreigen. Het idee dat publieke goederen en de sociale normen die deze ondersteunen bedreigd worden door veranderingen in de groepssamenstelling komt inderdaad vaak voor.

Elinor Ostrom schrijft bijvoorbeeld: "Emigratie kan de economische levensvatbaarheid van een regime veranderen door het verlies van degenen die de benodigde middelen bijdragen. Immigratie kan nieuwe deelnemers met zich meebrengen die anderen niet vertrouwen en zich niet snel sociale normen eigen maken die over een lange periode zijn vastgesteld." (Vertaald vanuit het Engels; Ostrom, 2000, p. 153). Normatieve verschillen tussen nieuwkomers en bestaande leden worden dan ook vaak aangehaald als een grote uitdaging voor coöperatie (Collier, 2013; Habyarimana et al., 2009; Hainmueller & Hopkins, 2014).

Omdat nieuwkomers vaak niet volledig op de hoogte zijn van de sociale normen in de ontvangende groep of gewend zijn aan andere normen, zullen ze zich aanvankelijk niet altijd conformeren aan de coöperatienorm in de groep en daarmee mogelijk worden gezien als meelifters. Dit bedreigt de coöperatie, omdat mensen vaak alleen bereid zijn om bij te dragen aan publieke goederen als ze erop vertrouwen dat anderen ook zullen bijdragen. Ze zullen daarom hun eigen bijdrage verlagen als ze denken dat anderen niet genoeg bijdragen (Gächter, 2007; Thöni & Volk, 2018). Deze conditionele bijdrages aan publieke goederen impliceert dat coöperatie fragiel is. Verondersteld meeliften van sommige leden kan meeliften onder andere leden aanwakkeren en zo een neerwaartse spiraal in gang zetten die leidt tot het verlies van sociale normen en de neergang van publieke goederen. Al met al is coöperatie voor publieke goederen kwetsbaar wanneer nieuwkomers niet bijdragen volgens de sociale norm vanwege onwetendheid of het ondersteunen van andere normen.

Een gerelateerde dreiging is dat de heersende norm en vorm van publieke goederenvoorziening afgestemd zijn op de behoeften en wensen van de bestaande leden, en niet op die van de nieuwkomers. Normen bevoordelen vaak de eigen groep boven anderen (Bernhard et al., 2006; Romano et al., 2021). Zulk bevoordeling van de eigen groep kan leiden tot een achtergestelde positie voor nieuwkomers. Feestdagen in een groep kunnen bijvoorbeeld gericht zijn op de religie van de bestaande leden, terwijl de religie van de nieuwkomers grotendeels genegeerd wordt (bijvoorbeeld betaald verlof tijdens Kerstmis en niet tijdens het Suikerfeest). Onderzoek suggereert dat nieuwkomers in eerste instantie niet als volwaardige groepsleden worden gezien door de bestaande leden (Rink et al., 2013). Pas na een socialisatieproces, waarbij de bestaande leden proberen om de nieuwkomers zich te laten conformeren aan hun normen, zien bestaande leden en nieuwkomers elkaar als volwaardige leden van de groep (Levine & Moreland, 1994; Pratsinakis, 2018). Er wordt dus gezegd dat nieuwkomers pas na verloop van tijd een overgang maken van “outsiders naar insiders” (Bauer et al., 2007). Onderzoek naar de percepties van de ontvangende bevolking ten opzichte van immigrantengroepen suggereert dat dit proces tientallen jaren kan duren, en zelfs dan worden bepaalde immigrantengroepen nog steeds niet als volledig onderdeel van de groep beschouwd (Coenders & Scheepers, 2008; De Coninck et al., 2021; Lee & Fiske, 2006; Storm et al., 2017). Zolang nieuwkomers niet tot de bestaande groep behoren, kan bevoordeling van bestaande leden een barrière zijn voor een inclusieve vorm van publieke goederenvoorziening die alle leden ten goede komt. Kortom, veranderende groepssamenstellingen kunnen niet alleen coöperatie voor publieke goederen bedreigen, maar ook leiden tot een vorm van publieke goederenvoorziening die nieuwkomers benadeelt.

Onderzoeksvragen

Er is nog maar weinig empirisch onderzoek gedaan naar de relatie tussen groepsveranderingen en coöperatie voor publieke goederen. Het meeste onderzoek naar publieke goederen richt zich ofwel op stabiele groepen of op kortstondige groepen waarbij mensen elkaar maar één keer ontmoeten en daarna weer elk hun eigen weg gaan. Hoewel dergelijk onderzoek onze kennis van publieke goederen sterk heeft verbeterd, laat het een type groep buiten beschouwing dat in het dagelijkse leven vaak voorkomt – groepen met een geleidelijk veranderende samenstelling door de komst van nieuwkomers en het vertrek van bestaande leden. Er zijn maar een paar studies die veranderingen in groepssamenstelling in verband brengen met coöperatie voor publieke goederen, en die rapporteren tot nu toe gemengde resultaten. Hoewel die studies informatief zijn, bevatten ze slechts een kleine deel van de mogelijke manieren waarop een groep kan veranderen.

We hebben bovendien weinig bewijs over het onderliggende mechanisme van verschillen in normatieve opvattingen tussen nieuwkomers en bestaande leden die de coöperatie zou bedreigen. Het meeste bewijs is afkomstig van observationele en casestudies, wat het trekken van causale conclusies moeilijk maakt (Collier, 2013; Habyarimana et al., 2009). Nieuwkomers verschillen namelijk vaak niet alleen van bestaande leden in hun normatieve opvattingen, maar ook in hun hulpbronnen, sociale status, sociale kansen en verschillende andere factoren. Wanneer deze verschillen samen voorkomen, is het moeilijk om vast te stellen welke van deze verschillen verantwoordelijk zijn voor coöperatieproblemen tussen nieuwkomers en bestaande leden. Een recente literatuurstudie suggereert inderdaad dat de veronderstelde negatieve effecten van normatieve verschillen tussen immigrantengroepen en de ontvangende bevolking op publieke goederen vertekend zijn door andere aspecten van interetnische groepen zoals armoede en politieke instabiliteit (Baldassarri & Abascal, 2020). Longitudinaal en experimenteel onderzoek is van vitaal belang voor een causaal begrip van coöperatie voor publieke goederen in veranderende groepen.

In dit proefschrift willen we bijdragen aan de kennis van coöperatie voor publieke goederen in veranderende groepen. Het proefschrift bevat vijf empirische hoofdstukken waarin de volgende drie hoofdvragen aan de orde komen:

1) Wat is de relatie tussen groepsveranderingen en coöperatie voor publieke goederen?
2) Wat is het effect van normatieve verschillen tussen nieuwkomers en bestaande leden op coöperatie voor publieke goederen?
3) In hoeverre is er bij publieke goederen in veranderende groepen sprake van bevoordeling van de eigen groep?

Het proefschrift integreert inzichten uit verschillende disciplines, waaronder de sociologie, sociale psychologie en gedragseconomie. Met betrekking tot sociologie raakt het proefschrift aan de drie belangrijkste sociologische thema's cultuur, sociale relaties en ongelijkheid (Van Tubergen, 2020). Het heeft betrekking op (1) cultuur in de vorm van normen, (2) sociale relaties in de vorm van veranderende sociale netwerken en hun implicaties voor samenwerking en (3) ongelijkheid in de vorm van ongelijke voordelen van publieke goederen tussen nieuwkomers en bestaande leden. Met betrekking tot sociale psychologie bestudeert het proefschrift wat de komst van nieuwkomers doet met (1) gevoelens van groepsidentiteit tussen bestaande leden en nieuwkomers, (2) de ontwikkeling van gedeelde groepsdoelen en normen en toewijding daaraan en (3) sociale invloed en conformiteit tussen nieuwkomers en bestaande leden. Tot slot bevat het proefschrift aspecten van gedragseconomie via (1) de focus op het leveren van publieke goederen als centrale uitkomstmaat, (2) het gebruik van spel-theoretische experimenten en (3) het meenemen van psychologische inzichten in de economische besluitvorming.

De grenzen tussen deze disciplines zijn niet eenduidig en worden steeds vager (Klein, 2017). De factoren die in de ene discipline relevant worden geacht, werken niet onafhankelijk van de relevante factoren in andere disciplines. Inzichten uit de psychologie over hoe nieuwkomers en bestaande leden zich identificeren met hun groep zijn bijvoorbeeld belangrijk om sociologisch te verklaren of leden zich conformeren aan normen en afzien van hun economische prikkel om mee te liften op publieke goederen. In plaats van psychologische, sociologische en economische factoren te bestuderen als afzonderlijke invloeden in aparte disciplinaire hoofdstukken, bestuderen we ze daarom als onderling gerelateerde factoren in elk hoofdstuk. Op deze manier proberen we de discipline-overschrijdende wisselwerking tussen individueel gedrag, groepsnormen en collectieve actie voor publieke goederen mee te nemen.

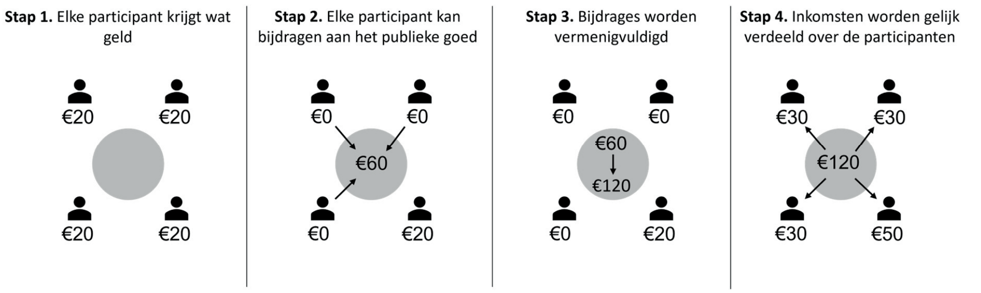
Methoden

Dit proefschrift gebruikt meerdere methodologische benaderingen. We combineren met name drie methoden: (1) lab-experimenten, (2) longitudinale analyse van een grootschalig en gecontextualiseerd online spel, en (3) een veldexperiment over een publiek goed in het dagelijkse leven. Elk van deze methoden heeft zijn voor- en nadelen en ze hebben allemaal een verschillend niveau van abstractie. Eenvoudige lab-experimenten bevinden zich aan het ene uiteinde van het abstractiespectrum en echte publieke goederen in het dagelijkse leven aan het andere uiteinde. Hoewel elke methode belangrijke inzichten kan opleveren, ontstaat een beter begrip door verschillende methoden van verschillende abstractieniveaus te combineren. Als de resultaten overeenkomen tussen de verschillende benaderingen, kunnen we er meer op vertrouwen dat ze niet enkel worden gedreven door de beperkingen van een specifieke benadering (Buskens & Raub, 2013; Jackson & Cox, 2013). Hieronder bespreken we kort elk van de drie methoden en hun voor- en nadelen.

Lab-experimenten

Lab-experimenten zijn een veelgebruikte methode in de natuurwetenschappen. Met uitzondering van de psychologie zijn lab-experimenten nog relatief ongebruikelijk in de sociale wetenschappen, hoewel ze wel tijd steeds populairder worden (Falk & Heckman, 2009; Jackson & Cox, 2013). Lab-experimenten zijn bij uitstek geschikt om causale verbanden te testen, omdat ze een mate van controle over variabelen bieden die niet mogelijk is in observationele studies. Wanneer in observationele studies wordt gevonden dat variabele A variabele B voorspelt, is het moeilijk om te bepalen of B wordt veroorzaakt door A, of A wordt veroorzaakt door B (omgekeerde causaliteit), of dat een andere weggelaten variabele C zowel A als B veroorzaakt (ook wel ‘confounding’ genoemd). Onderzoekers kunnen proberen causale conclusies te trekken uit observationele gegevens door statistisch te controleren voor verschillende ‘confounders’. Echter, tenzij alle relevante controlevariabelen worden vastgelegd en deze controlevariabelen foutloos worden gemeten, zal dit slechts gedeeltelijk corrigeren voor 'confounding' (Westfall & Yarkoni, 2016). En tenzij de observationele data perfect overeenkomen met de statistische modelspecificaties, zal het toevoegen van controlevariabelen 'confounding' ook niet voorkomen (Miller & Chapman, 2001; Westfall & Yarkoni, 2016). Longitudinale observationele data helpen om de kans op omgekeerde causaliteit en 'confounding' te verkleinen, maar sluiten ze helaas niet uit (Leszczensky & Wolbring, 2022; VanderWeele & An, 2013). Lab-experimenten kunnen de problemen van omgekeerde causaliteit en 'confounding' voorkomen door deelnemers willekeurig aan verschillende condities toe te wijzen. Zoals we zullen zien, hebben lab-experimenten ook beperkingen, in het bijzonder met betrekking tot realisme en generaliseerbaarheid. Ze kunnen echter dienen als een nuttige aanvulling op observationele methoden, met name wanneer het doel is om causale verbanden en mechanismen te toetsen die moeilijk op grote schaal te observeren zijn.

Onze lab-experimenten gebruiken het publiekgoedspel. Het publiekgoedspel is een veelgebruikte manier om het dilemma te bestuderen tussen het eigenbelang (meeliften) en het algemeen belang (bijdragen aan het publiek goed). Een bekende versie van het publiekgoedspel wordt weergegeven in Figuur 1. Aan het begin van het spel ontvangt elke deelnemer een geldbedrag (stap 1). Elke deelnemer moet vervolgens beslissen hoeveel van het geldbedrag bij te dragen aan een groepspot (stap 2). De experimentleider vermenigvuldigt alle bijdrages met een getal boven de 1 en kleiner dan het aantal deelnemers (stap 3). Tot slot worden de vermenigvuldigde bijdrages gelijk verdeeld onder alle leden, inclusief degenen die niet hebben bijgedragen aan de groepspot (stap 4). Omdat bijdrages aan het publiek goed worden vermenigvuldigd met een factor groter dan 1, verhogen bijdrages de totale uitbetaling voor de groep als geheel, en verdient de groep het meeste geld als iedereen het volledige geldbedrag bijdraagt. Maar omdat de vermenigvuldigingsfactor kleiner is dan het aantal deelnemers, verdient elke individuele deelnemer het meeste geld door mee te liften op de bijdrages van anderen. In hoofdstukken 2, 3 en 5 maken we gebruik van verschillende vormen van het publiekgoedspel.



Figuur 1. Het publiekgoedspel.

Elke deelnemer ontvangt een geldbedrag, in dit voorbeeld 20 euro (stap 1). Elke deelnemer moet dan beslissen hoeveel van dit geldbedrag bij te dragen aan een publiek goed (stap 2). In dit voorbeeld dragen drie deelnemers al hun 20 euro bij, terwijl één deelnemer de 20 euro zelf houdt. De experimentleider vermenigvuldigt vervolgens alle bijdrages met een bepaald getal, in dit voorbeeld met 2 (stap 3). Tot slot worden de vermenigvuldigde bijdrages gelijk verdeeld onder alle leden (30 euro voor elke deelnemer in dit voorbeeld), inclusief degenen die niet hebben bijgedragen aan het publiek goed (stap 4).

Gecontextualiseerde online spellen

Het standaard publiekgoedspel is een abstract spel; weggelaten worden de complexiteit en context van het veld (dat wil zeggen, publieke goederen in het dagelijkse leven, buiten het lab). Dit voorkomt ongewenste invloeden van het veld die niet onder de controle zijn van de experimentleider, maar kan ook het realisme van het spel beperken. Een andere aanpak is het inbrengen van een rijkere context in het publiekgoedspel. Het publiekgoedspel kan bijvoorbeeld gecontextualiseerd worden als een groep werknemers die samen moet werken in een team (Chen et al., 2009; Van Gerwen et al., 2018) of een groep mensen die gestrand zijn op een eiland en samen een vlot moeten bouwen om het eiland te verlaten (Baptista et al., 2013). Dergelijke scenario's kunnen deelnemers helpen om zich in te leven in het spel en er meer aandacht aan te besteden, waardoor de ecologische validiteit toeneemt. Het risico is echter dat hun gedrag wordt gestuurd door specifieke kenmerken van het scenario in plaats van het algemene dilemma van publieke goederen tussen het eigen en algemeen belang. De combinatie van verschillende benaderingen kan het meest complete inzicht geven. Daarom gebruiken we als tweede methode in dit proefschrift gecontextualiseerde publiekgoedspellen.

Voor het gecontextualiseerde publiekgoedspel gebruiken we een bestaand ‘massive multiplayer online’ (MMO) spel. Dit zijn grootschalige online spellen waar miljoenen spelers wereldwijd aan meedoen vanuit hun computer thuis. Deze spelers weten niet dat hun gedrag in het spel wordt gebruik voor een onderzoek (maar weten wel dat hun data gedeeld kan worden). Dit voorkomt een bekend probleem van lab-experimenten – dat deelnemers zich mogelijk anders dan gebruikelijk gedragen omdat ze weten dat ze geobserveerd worden door onderzoekers. Bovendien kunnen lab-experimenten meestal niet langer dan 1 of 2 uur duren, terwijl MMO spellen meerdere maanden of zelfs jaren kunnen beslaan. Dit stelt ons in staat om coöperatie ook op de lange termijn te onderzoeken. Natuurlijk zijn er ook een aantal nadelen aan het gebruik van MMO spellen. Het is bijvoorbeeld mogelijk dat de groep spelers niet representatief is voor de bredere populatie, mensen kunnen zich anders gedragen in MMO spellen dan in het dagelijkse leven en de toegang tot de data is afhankelijk van het bedrijf dat het MMO beheert. In hoofdstuk 4 gebruiken we grootschalige en langetermijndata van het MMO spel *Ikariam*, waarin de spelontwerpers bewust publieke goederen hebben ingebouwd. Het spel speelt zich af in een oude Griekse archipel en elk eiland van de archipel heeft te maken met een publiek goed. Het publiek goed is gecontextualiseerd als een zagerij die hout produceert op een niveau dat afhangt van de bijdrages van alle spelers op het eiland.

Veldexperimenten

Tot slot gaan we in op veldstudies naar publieke goederen in het dagelijkse leven. Publieke goederen in het veld zijn complex, met vaak veel onderling gerelateerde factoren die samen beïnvloeden hoe mensen bijdragen aan het publiek goed. Dit maakt het moeilijk om het effect van een bepaalde factor te isoleren in observationele studies en is een reden dat lab-experimenten met (gecontextualiseerde) publiekgoedspellen nuttig zijn en gebruikt worden. Maar als we willen dat deze lab-experimenten informatie verschaffen over publieke goederen in het veld, is het verstandig om te testen of de resultaten uit het lab zich vertalen naar het dagelijkse leven. Vergeleken met de grote hoeveelheid lab-studies over publiekgoedspellen zijn veldstudies echter nog steeds zeldzaam. Bovendien zijn veldstudies vaak observationeel, wat de eerdergenoemde risico's van ‘confounding’ met zich meebrengt. Een klein maar groeiend onderzoeksgebied omvat het gebruik van experimenten met publieke goederen in het veld. Deze combinatie van de experimentele methode en praktijksituaties maakt het mogelijk om causale conclusies te trekken in een natuurlijke context.

De relatie tussen nieuwkomers en bestaande leden die het vaakst bestudeerd wordt in veldexperimenten is die tussen immigranten en de ontvangende bevolking. Veldexperimenten hebben laten zien dat mensen zonder migratieachtergrond zich vaker coöperatief opstellen tegenover elkaar dan tegenover immigranten (Aidenberger & Doehne, 2021; Choi et al., 2019, 2021a, 2021b; Zhang et al., 2019). Bovendien worden immigranten vaker bestraft voor het overtreden van normen dan mensen zonder migratieachtergrond (Aidenberger & Doehne, 2021; Mujcic & Frijters, 2020; Winter & Zhang, 2018). Deze studies suggereren dat immigranten een belangrijke groep nieuwkomers vormen in het dagelijkse leven en de coöperatievraagstukken die daar plaatsvinden. In hoofdstuk 6 gebruiken we een veldexperiment om te bestuderen hoe treinreizigers reageren op zwartrijden in het openbaar vervoer (een vorm van meeliften op het publiek goed van openbaar vervoer). Ingehuurde acteurs vragen reizigers die op het punt staan door de incheckpoortjes te gaan op treinstations of ze hen mogen volgen zonder zelf in te checken. We observeren of reizigers dit meeliftgedrag afwijzen, hun afkeuring erover uitspreken, of het juist toestaan. We kijken hierbij of de reactie afhangt van of de zwartrijder een migratieachtergrond heeft. In totaal worden 801 reizigers individueel benaderd op drie treinstations door tien verschillende acteurs, vijf met een migratieachtergrond (Turks/Marokkaans) en vijf zonder migratieachtergrond.

Bevindingen en conclusies

Onze eerste onderzoeksvraag – wat is de relatie tussen groepsveranderingen en coöperatie voor publieke goederen – werd beantwoord met grootschalige data van een ‘massive multiplayer online’ spel. De resultaten suggereren dat de relatie meestal negatief is op de korte termijn; mensen dragen minder bij als nieuwkomer dan als bestaand lid, wat betekent dat groepen met meer nieuwkomers minder bijdragen aan publieke goederen. Omdat individuen in het spel zowel de rol van nieuwkomer als die van bestaand lid vervullen, konden we binnen individuen vergelijken hoe de bijdrages aan publieke goederen veranderen afhankelijk van hun rol in de groep. Zo konden we individuele kenmerken grotendeels uitsluiten als verklaring voor verschillen tussen nieuwkomers en bestaande leden. Dat we een verschil in de bijdrages vinden tussen nieuwkomers en bestaande leden terwijl we individuele kenmerken uitsluiten, suggereert dat alleen al het hebben van een nieuwkomerrol de neiging van mensen om bij te dragen kan verminderen. Het langetermijnpotentieel van nieuwkomers is echter even hoog als dat van bestaande leden. Nieuwkomers verhogen hun bijdrages in de loop van de tijd en dragen uiteindelijk net zoveel bij als bestaande leden. Deze bevindingen komen overeen met modellen van groepssocialisatie, waarbij nieuwkomers een proces van leren en aanpassen ondergaan voordat ze zich gedragen als een bestaand lid (Levine & Moreland, 1994).

De resultaten suggereren dat groepsveranderingen inderdaad een relevante factor zijn voor publieke goederen, en een uitdaging kunnen vormen voor duurzame coöperatie. Dit betekent echter niet dat groepsveranderingen slecht zijn of vermeden moeten worden. Als een groep nooit nieuwkomers toe zou laten, zouden publieke goederen op den duur op houden te bestaan. Bestaande leden zullen op een gegeven moment de groep moeten verlaten, en als er niemand is om hen te vervangen, zullen de publieke goederen die ze leverden ophouden te bestaan. Pensioenstelsels en arbeidsorganisaties kunnen bijvoorbeeld alleen op de lange termijn in stand worden gehouden als vertrekkende leden worden vervangen door nieuwkomers. Nieuwkomers zijn dus nodig om publieke goederen in stand te houden over periodes die meer dan één generatie beslaan. Zoals we gezien hebben, dragen nieuwkomers aanvankelijk soms minder bij dan de bestaande leden. Maar we hebben ook gezien dat nieuwkomers een groot potentieel hebben om op de lange termijn bij te dragen. Groepen kunnen profiteren van dit potentieel op de lange termijn als ze geduldig zijn en de verleiding weerstaan om nieuwkomers die aanvankelijk geen bijdrage leveren uit te sluiten of te verdrijven. De kennis dat nieuwkomers na een aanpassingsperiode zullen bijdragen aan het publiek goed, kan bestaande leden mogelijk helpen om geduldig te zijn en spanningen tussen nieuwkomers en bestaande leden te verminderen.

Nieuwkomers de tijd geven om te integreren en ze in een positie plaatsen waarin ze effectief kunnen bijdragen, helpt in dit opzicht ook. Nieuwe buurtbewoners zijn bijvoorbeeld eerder geneigd om vrijwilligerswerk te doen bij buurtevenementen als ze gesetteld zijn en deel zijn gaan uitmaken van de gemeenschap (Ghimire & Skinner, 2019) en immigranten zijn eerder geneigd om bij te dragen aan goede doelen als ze een baan hebben en geïntegreerd zijn in het gastland (Bekkers & Wiepking, 2007; Hainmueller et al., 2016). Alleen door nieuwkomers op te nemen in de groep kan coöperatie voor publieke goederen op de lange termijn duurzaam zijn. Om de voorwaarden voor duurzame coöperatie in veranderende groepen verder te begrijpen, kunnen we ons wenden tot de resultaten over de onderliggende normdynamiek tussen nieuwkomers en bestaande leden. Dit brengt ons bij de tweede onderzoeksvraag.

De tweede onderzoeksvraag – wat is het effect van normatieve verschillen tussen nieuwkomers en bestaande leden op coöperatie voor publieke goederen – hebben we voornamelijk met lab-experimenten onderzocht. Eerder observationeel onderzoek suggereerde dat normatieve verschillen tussen bestaande leden en nieuwkomers een belangrijke belemmering zijn voor coöperatie (Collier, 2013; Habyarimana et al., 2009; Hainmueller & Hopkins, 2014). Maar omdat nieuwkomers en bestaande leden meestal verschillen op meerdere aspecten (bv. status, inkomen, religie), was het tot nu toe onduidelijk of normatieve verschillen verantwoordelijk zijn voor coöperatieproblemen of een van de andere verschillen. Onze experimenten suggereren dat normatieve verschillen de coöperatie voor publieke goederen niet verlagen. Dit impliceert dat normatieve verschillen niet het belangrijkste mechanisme vormen dat coöperatieproblemen in veranderende groepen veroorzaakt. In plaats daarvan lijkt het er vooral op dat nieuwkomers na verloop van tijd de norm van de groep leren kennen en de tijd die dit kost hangt af van de complexiteit van het publiek goed.

Het publiekgoedspel vormt een zeer eenvoudige omgeving, waardoor het relatief gemakkelijk is voor nieuwkomers om de coöperatienorm van een groep te leren. We zagen dan ook dat nieuwkomers erg snel de norm van de groep leerden in het publiekgoedspel en zich daaraan conformeerden (hoofdstuk 3). Het online spel Ikariam vormt een complexere omgeving, waarin de coöperatie voor publieke goederen gebeurt in een contextrijke omgeving, over een langere periode, en met een breder scala aan spelers met verschillende middelen en behoeften. Deze kenmerken maken het moeilijker om de coöperatienorm van de groep te leren, en we zagen inderdaad dat nieuwkomers in Ikariam aanzienlijk meer tijd nodig hadden om bij te dragen volgens de norm van de bestaande spelers (hoofdstuk 4). Dit suggereert dat nieuwkomers zich aanpassen aan de coöperatienorm van de bestaande leden met een snelheid die afhangt van de complexiteit van het publiek goed en de norm die het ondersteunt. Naast geduld is een mogelijke manier om coöperatie in veranderende groepen te bevorderen dus om nieuwkomers duidelijke en volledige informatie te geven over de heersende coöperatienorm.

We vonden echter ook dat de bestaande leden hun eigen coöperatienormen oplegden aan de nieuwkomers, dat nieuwkomers harder werden gestraft voor meeliften dan bestaande leden, en dat normatieve verschillen leidden tot lagere gevoelens van groepsidentiteit, vooral onder nieuwkomers. Deze bevindingen suggereren dat de manier waarop de publieke goederen tot stand komen in situaties van normatieve verschillen niet voor iedereen even waardevol is en ongewenste neveneffecten kan hebben op het gebied van groepsidentiteit. Deze laatste bevindingen geven enkele eerste aanwijzingen voor het antwoord op onze derde onderzoeksvraag – in hoeverre is er bij publieke goederen in veranderende groepen sprake van bevoordeling van de eigen groep? We gebruikten aanvullende lab- en veldexperimenten om deze vraag verder te beantwoorden. We richtten ons in het bijzonder op bevoordeling van de eigen groep in (1) sociale sancties voor het meeliften op publieke goederen en (2) de manier waarop publieke goederen worden vormgegeven. We vinden bewijs voor bevoordeling van de eigen groep in beide domeinen. In het eerste domein vinden we dat nieuwkomers vaker sociale sancties ontvangen voor meeliftgedrag dan bestaande leden, zowel bij abstracte lab-experimenten (hoofdstuk 3) als bij dagelijkse publieke goederen in het veld (hoofdstuk 6). In het tweede domein vinden we dat wanneer mensen moeten kiezen tussen bijdragen aan exclusieve publieke goederen van de eigen groep en inclusieve publieke goederen die ook ten goede komen aan (nieuwe) leden uit een andere groep, een significant deel kiest voor exclusieve publieke goederen van de eigen groep (hoofdstuk 5). Deze bevoordeling van de eigen groep kan echter voorkomen worden als – in plaats van mensen direct in situaties te plaatsen waarbij ze moeten samenwerken met leden van een andere groep – ze hun coöperatiecirkel geleidelijk kunnen uitbreiden.

Uit ons onderzoek blijkt dus ook dat het niet altijd wenselijk is dat nieuwkomers zich aanpassen aan de bestaande leden en niet andersom. Dit eenzijdige aanpassingsproces verwaarloost de behoeften van nieuwkomers en kan leiden tot een vorm van publieke goederenvoorziening die alleen van waarde is voor de bestaande leden. Dit nadeel is moeilijker te detecteren; ze wordt gemist als we ons alleen richten op de mate waarin wordt bijgedragen aan publieke goederen en niet op de vorm die publieke goederen aannemen. Een publiek goed kan veel bijdrages ontvangen, maar als het alleen voordelen biedt voor een deel van de leden, is het niet optimaal of duurzaam. Duurzame samenwerking vergt aanpassing aan veranderende omstandigheden; om publieke goederen te behouden ondanks de komst van nieuwkomers met nieuwe behoeften, kan het nodig zijn om de aard en inhoud van publieke goederen na verloop van tijd aan te passen (SCOOP, 2019). Voor sommigen kan dit voelen als een 'verlies' van publieke goederen dat spanningen oproept, maar op de lange termijn is het noodzakelijk voor veranderende samenlevingen om publieke goederen te leveren die nuttig zijn voor de gehele bevolking.

Door het gebruik van meerdere onderzoeksmethoden zijn we in staat om de robuustheid van de bevindingen over de verschillende methoden te bekijken. Als vergelijkbare resultaten worden verkregen met verschillende methoden, zijn we zekerder over de robuustheid. Het eerste robuuste resultaat is dat nieuwkomers aanvankelijk een andere bijdrage leveren aan publieke goederen dan bestaande leden. Nieuwkomers dragen aanvankelijk bij volgens andere normen dan bestaande leden in het standaard publiekgoedspel (hoofdstuk 2-3), ze dragen aanvankelijk minder bij dan bestaande leden in het gecontextualiseerde publiekgoedspel (hoofdstuk 4), en of personen meeliftgedrag tegengaan bij een dagelijks publiek goed (het openbaar vervoer) hangt ook af van of de persoon en de meelifter een migratieachtergrond hebben (hoofdstuk 6). De tweede robuuste bevinding is dat nieuwkomers zich na verloop van tijd conformeren aan de bestaande leden; dit resultaat zagen we zowel in het standaard als in het gecontextualiseerde publiekgoedspel (hoofdstuk 3 en 4). Als derde robuuste resultaat zagen we dat de publieke goederen bevoordeling van de eigen groep bevatten. Dit resultaat kwam terug in het standaard publiekgoedspel (hoofdstuk 3), in een publiekgoedspel waarbij deelnemers moeten kiezen tussen de eigen groep en een andere groep (hoofdstuk 5) en het veldexperiment over openbaar vervoer als publiek goed (hoofdstuk 6). Het vierde robuuste resultaat was dat nieuwkomers harder worden gestraft voor meeliftgedrag dan bestaande leden, wat te zien was in zowel het standaard publiekgoedspel als het veldexperiment in het openbaar vervoer (hoofdstuk 3 en 6).

Tot slot zagen we in alle studies dat of een persoon de rol van nieuwkomer of bestaand lid heeft belangrijk is voor hoe de persoon zich gedraagt of hoe de persoon door anderen wordt behandeld. Sommige theorieën suggereren dat verschillen tussen nieuwkomers en bestaande leden in hun bijdrages aan publieke goederen verklaard worden door onderliggende verschillen in persoonlijke eigenschappen (Collier, 2013; Gereke et al., 2021). Volgens deze opvatting zouden groepen zich vooral bezig moeten houden met het voorkomen van ‘slechte appels’ in de groep (Grund et al., 2018). Dit proefschrift laat zien dat verschillen tussen nieuwkomers en bestaande leden kunnen optreden zonder verschillen in hun onderliggende eigenschappen. In al onze studies kwamen nieuwkomers en bestaande leden uit dezelfde groep deelnemers en verschilden ze dus niet in hun onderliggende eigenschappen. In het online spel moesten spelers zowel de rol van nieuwkomer als die van bestaand lid innemen. We ontdekten dat dezelfde spelers een verschillende bijdrage leverden, afhankelijk van of ze de rol van nieuwkomer of bestaand lid hadden. In de lab-experimenten werden de rollen van nieuwkomers en bestaande leden willekeurig toegewezen, dus zijn er gemiddeld geen verschillen tussen hen in onderliggende eigenschappen. Toch zagen we dat nieuwkomers anders werden behandeld dan bestaande leden (bijvoorbeeld in hun invloed op publieke goederen en hoe hard ze werden gestraft voor meeliftgedrag). In het veldexperiment deden alle acteurs het verzoek om reizigers te volgen door de incheckpoortjes met dezelfde neutraal geformuleerde zin. Toch reageerden mensen anders op acteurs met een migratieachtergrond dan op acteur zonder migratieachtergrond. In geen van de hoofdstukken kunnen de geobserveerde verschillen tussen nieuwkomers en bestaande leden dus toegeschreven worden aan verschillen in hun onderliggende eigenschappen. In plaats daarvan lijkt alleen al de rol van de nieuwkomer een belangrijke invloed te hebben op zowel het eigen gedrag van nieuwkomers als op hoe ze worden behandeld door bestaande leden.

References

Aaldering, H., & Böhm, R. (2020). Parochial versus universal cooperation: Introducing a novel economic game of within- and between-group interaction. *Social Psychological and Personality Science*, *11*(1), 36–45. https://doi.org/10.1177/1948550619841627

Aaldering, H., Ten Velden, F. S., Van Kleef, G. A., & De Dreu, C. K. W. (2018). Parochial cooperation in nested intergroup dilemmas is reduced when it harms out-groups. *Journal of Personality and Social Psychology*, *114*(6), 909–923. https://doi.org/10.1037/pspi0000125

Agustín, Ó. G., & Jørgensen, M. B. (2019). *Solidarity and the "refugee crisis" in Europe*. Palgrave McMillan. https://doi.org/10.1007/978-3-319-91848-8_3

Ahn, T. ., Isaac, R. M., & Salmon, T. C. (2008). Endogenous group formation. *Journal of Public Economic Theory*, *10*(2), 171–194.

Aidenberger, A., & Doehne, M. (2021). Unveiling everyday discrimination. Two field experiments on discrimination against religious minorities in day-to-day interactions. *British Journal of Sociology*, *72*(2), 328–346. https://doi.org/10.1111/1468-4446.12811

Alexander, M., & Christia, F. (2011). Context modularity of human altruism. *Science*, *334*(6061), 1392–1394. https://doi.org/10.1126/science.1202599

Álvarez-Benjumea, A., & Winter, F. (2020). The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. *Proceedings of the National Academy of Sciences*, *117*(37), 22800–22804. https://doi.org/10.1073/pnas.2007977117

Andersen, S. C., & Moynihan, D. P. (2018). How do socially distinctive newcomers fare? Evidence from a field experiment. *Public Administration Review*, *78*(6), 874–882. https://doi.org/10.1111/puar.12957

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16). https://doi.org/10.1073/pnas.2014893118

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE*, *8*(6), 1–8. https://doi.org/10.1371/journal.pone.0064941

Arechar, A., Kouchaki, M., & Rand, D. G. (2018). Examining spillovers between long and short repeated prisoner's dilemma games played in the laboratory. *Games*, *9*(1), 5. https://doi.org/10.2139/ssrn.2979639

Ariely, D., & Norton, M. I. (2007). Psychology and experimental economics: A gap in abstraction. *Current Directions in Psychological Science*, *16*(6), 336–339. https://doi.org/10.1111/j.1467-8721.2007.00531.x

Arrow, H., & McGrath, J. E. (1993). Membership matters: How member change and continuity affect small group structure, process, and performance. *Small Group Research*, *24*(3), 334–361.

Asher, S., Casaburi, L., Nikolov, P., & Ye, M. (2009). *One step at a time: Do threshold patterns matter in public good provision?* (Working Paper). http://hdl.handle.net/10419/27486

Auspurg, K., Schneck, A., & Hinz, T. (2019). Closed doors everywhere? A meta-analysis of field experiments on ethnic discrimination in rental housing markets. *Journal of Ethnic and Migration Studies*, *45*(1), 95–114. https://doi.org/10.1080/1369183X.2018.1489223

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science*, *317*(5837), 472–476. https://doi.org/10.1126/science.1146930

Bakker, D. M. (2019). *Cooperation and social control: effects of preferences, institutions, and social structure*. [Thesis fully internal (DIV), University of Groningen]. Rijksuniversiteit Groningen. https://doi.org/10.33612/diss.98552819

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, *56*(8), 1773–1785. https://doi.org/10.1016/j.euroecorev.2012.09.008

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*, 1–6. https://doi.org/10.1038/ncomms13327

Baldassarri, D. (2020). Market integration accounts for local variation in generalized altruism in a nationwide lost-letter experiment. *Proceedings of the National Academy of Sciences*, *117*(6), 2858–2863. https://doi.org/10.1073/pnas.1819934117

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, *43*, 41–73. https://doi.org/10.1146/annurev-soc-073014-112445

Baldassarri, D., & Abascal, M. (2020). Diversity and prosocial behavior. *Science*, *369*(6508), 1183–1187. https://doi.org/10.1126/SCIENCE.ABB2432

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556–1581. https://doi.org/http://dx.doi.org/10.1037/a0037737

Baptista, M., Damas, H., Dimas, J., Raimundo, G., Prada, R., Martinho, C., Santos, P. A., & Peña, J. (2013). A serious game based on a public goods experiment. *2013 International Conference on Social Computing*, 774–781. https://doi.org/10.1109/SocialCom.2013.116

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753. https://doi.org/10.1098/rspb.2006.0209

Barr, A., Packard, T., & Serra, D. (2014). Participatory accountability and collective action: Experimental evidence from Albania. *European Economic Review*, *68*, 250–269. https://doi.org/10.1016/j.euroecorev.2014.01.010

Battaglini, M., Nunnari, S., & Palfrey, T. R. (2016). The dynamic free rider problem: A laboratory study. *American Economic Journal: Microeconomics*, *8*(4), 268–308. https://doi.org/10.1257/mic.20150126

Bauer, T. N., Bodner, T., Erdogan, B., Truxillo, D. M., & Tucker, J. S. (2007). Newcomer adjustment during organizational socialization : A meta-analytic review of antecedents, outcomes, and methods. *Journal of Applied Psychology*, *92*(3), 707–721. https://doi.org/10.1037/0021-9010.92.3.707

Bekkers, R., & Wiepking, P. (2007). Understanding philanthropy: A review of 50 years of theories and research. *35th Annual Conference of the Association for Research on*

Nonprofit and Voluntary Action, 1–32.

Belot, M., Duch, R., & Miller, L. (2015). A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior and Organization*, *113*, 26–33. https://doi.org/10.1016/j.jebo.2015.02.007

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, *106*(6), 1493–1545. https://doi.org/10.1016/j.jtbi.2005.11.028

Berger, J., & Hevenstone, D. (2016). Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows. *Rationality and Society*, *28*(3), 299–319. https://doi.org/10.1177/1043463116634035

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group affiliation and altruistic norm enforcement. *American Economic Review*, *96*(2), 217–221. https://doi.org/10.1257/000282806777212594

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. https://doi.org/10.1038/nature04981

Berton, F., & Garibaldi, P. (2012). Workers and firms sorting into temporary jobs. *Economic Journal*, *122*(562), 125–154. https://doi.org/10.1111/j.1468-0297.2012.2533.x

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*, 161–178. https://doi.org/10.1002/bdm.648

Bicchieri, C., Dimant, E., & Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior and Organization*, *188*, 209–235. https://doi.org/10.1016/j.jebo.2021.04.002

Bicchieri, C., Lindemans, J. W., & Jiang, T. (2014). A structured approach to a diagnostic of collective practices. *Frontiers in Psychology*, *5*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2014.01418

Binmore, K. (2006). *Natural Justice*. Oxford University Press.

Binmore, K. (2010). Social norms or social preferences? *Mind and Society*, *9*(2), 139–157. https://doi.org/10.1007/s11299-010-0073-2

Birch, S. (2018). Democratic norms, empirical realities and approaches to improving voter turnout. *Res Publica*, *24*(1), 9–30. https://doi.org/10.1007/s11158-017-9380-3

Blackwell, C., & McKee, M. (2003). Only for my own neighborhood? Preferences and voluntary provision of local and global public goods. *Journal of Economic Behavior and Organization*, *52*(1), 115–131. https://doi.org/10.1016/S0167-2681(02)00178-6

Bluffstone, R., Dannenberg, A., Martinsson, P., Jha, P., & Bista, R. (2020). Cooperative behavior and common pool resources: Experimental evidence from community forest user groups in Nepal. *World Development*, *129*, 104889. https://doi.org/10.1016/j.worlddev.2020.104889

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior.* Harvard University Press.

Böhm, R., Bornstein, G., & Koppel, H. (2014). Between-group conflict and other-regarding preferences in nested social dilemmas. *Jena Economic Research Papers*.

Böhm, R., Rusch, H., & Baron, J. (2020). The psychology of intergroup conflict : A review of theories and measures. *Journal of Economic Behavior and Organization*, *178*, 947–962. https://doi.org/10.1016/j.jebo.2018.01.020

Boosey, L. A. (2017). Conditional cooperation in network public goods experiments. *Journal of Behavioral and Experimental Economics* , *69*, 108–116. https://doi.org/10.1016/j.socec.2017.01.001

Botelho, A., Harrison, G. W., Costa Pinto, L. M., & Rutström, E. E. (2009). Testing static game theory with dynamic experiments: A case study of public goods. *Games and Economic Behavior*, *67*(1), 253–265. https://doi.org/10.1016/j.geb.2008.11.013

Bowles, S., & Gintis, H. (2013). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Boyd, R., & Richerson, P. J. (2022). Large-scale cooperation in small-scale foraging societies. *Evolutionary Anthropology: Issues, News, and Reviews*, *31*(4), 175–198.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(SUPPL. 2), 10918–10925. https://doi.org/10.1073/pnas.1100290108

Brauer, M., & Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, *35*(7), 1519–1539. https://doi.org/10.1111/j.1559-1816.2005.tb02182.x

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). Explaining norms. In *Oxford University Press*. Oxford University Press.

Bshary, R., & Grutter, A. S. (2006). Image scoring and cooperation in a cleaner fish mutualism. *Nature*, *441*(7096), 975–978. https://doi.org/10.1038/nature04755

Buchan, N. R., Brewer, M. B., Grimalda, G., Wilson, R. K., Fatas, E., & Foddy, M. (2011). Global social identity and global cooperation. *Psychological Science*, *22*(6), 821–828. https://doi.org/10.1177/0956797611409590

Buchan, N. R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., & Foddy, M. (2009). Globalization and human cooperation. *Proceedings of the National Academy of Sciences*, *106*(11), 4138–4142. https://doi.org/10.1073/pnas.0809522106

Burt, R. S. (2012). Network-related personality and the agency question: Multirole evidence from a virtual world. *American Journal of Sociology*, *118*(3), 543–591. https://doi.org/10.1086/667856

Buskens, V., & Raub, W. (2013). Rational choice social research on social dilemmas: Embeddedness effects on trust. In R. Wittek, T. A. B. Snijders, & V. Nee (Eds.), *Handbook of rational choice social research* (pp. 113–150). Stanford University Press.

Buskens, V., & Weesie, J. (2000). Cooperation via social networks. *Analyse & Kritik*, *22*(1), 44–74. https://doi.org/10.1515/auk-2000-0103

Cahuc, P., Charlot, O., & Malherbet, F. (2016). Explaining the spread of temporary jobs and its impact on labor turnover. *International Economic Review*, *57*(2), 533–572. https://doi.org/10.1111/iere.12167

Camerer, C. F. (1996). Rules for experimenting in psychology and economics, and why they differ. In W. Guth & E. Van Damme (Eds.), *Essays in honor of Reinhard Selten* (pp. 313–327). Springer-Verlag.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carpenter, J. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, *60*(1), 31–51. https://doi.org/10.1016/j.geb.2006.08.011

Carpenter, J., & Seki, E. (2011). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, *49*(2), 612–630. https://doi.org/10.1111/j.1465-7295.2009.00268.x

Carter, G. (2014). The reciprocity controversy. *Animal Behavior and Cognition*, *1*(3), 368. https://doi.org/10.12966/abc.08.11.2014

Carter, T. L. J., & Johnson, L. T. (2021). "Blacks can't jump": The racialization of transit police responses to fare evasion. *Race and Justice*. https://doi.org/10.1177/21533687211007548

Catola, M., Alessandro, S. D., Guarnieri, P., Pizziol, V., & Alessandro, S. D. (2021). *Personal and social norms in a multilevel public goods experiment* (Working Paper).

Cavaliere, M., Sedwards, S., Tarnita, C. E., Nowak, M. A., & Csikász-Nagy, A. (2012). Prosperity is associated with instability in dynamical networks. *Journal of Theoretical Biology*, *299*, 126–138. https://doi.org/10.1016/j.jtbi.2011.09.005

Charness, G., & Yang, C. (2014). Starting small toward voluntary formation of efficient large groups in public goods provision. *Journal of Economic Behavior and Organization*, *102*, 119–132. https://doi.org/10.1016/j.jebo.2014.03.005

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, *14*(1), 47–83. https://doi.org/10.1007/s10683-010-9257-1

Chen, X. P., Pillutla, M. M., & Yao, X. (2009). Unintended consequences of cooperation inducing and maintaining mechanisms in public goods Dilemmas: Sanctions and moral appeals. *Group Processes and Intergroup Relations*, *12*(2), 241–255. https://doi.org/10.1177/1368430208098783

Cherry, T. L., Kroll, S., & Shogren, J. F. (2005). The impact of endowment heterogeneity and origin on public good contributions: Evidence from the lab. *Journal of Economic Behavior and Organization*, *57*(3), 357–365. https://doi.org/10.1016/j.jebo.2003.11.010

Chesney, T., Chuah, S. H., & Hoffmann, R. (2009). Virtual world experimentation: An exploratory study. *Journal of Economic Behavior and Organization*, *72*(1), 618–635. https://doi.org/10.1016/j.jebo.2009.05.026

Choi, D. D., Poertner, M., & Sambanis, N. (2019). Parochialism, social norms, and discrimination against immigrants. *Proceedings of the National Academy of Sciences*, *116*(33), 16274–16279. https://doi.org/10.1073/pnas.1820146116

Choi, D. D., Poertner, M., & Sambanis, N. (2021a). Linguistic assimilation does not reduce discrimination against immigrants: Evidence from Germany. *Journal of Experimental Political Science*, *8*(3), 235–246. https://doi.org/10.1017/XPS.2020.20

Choi, D. D., Poertner, M., & Sambanis, N. (2021b). The hijab penalty: Feminist backlash to Muslim immigrants. *American Journal of Political Science*, *00*(0), 1–16. https://doi.org/10.1111/ajps.12627

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cimino, A. (2011). The evolution of hazing: Motivational mechanisms and the abuse of newcomers. *Journal of Cognition and Culture*, *11*(3–4), 241–267. https://doi.org/10.1163/156853711X591242

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*, 1421–1435. https://doi.org/10.1016/j.jpubeco.2004.05.011

Coenders, M., & Scheepers, P. (2008). Changes in resistance to the social integration of foreigners in Germany 1980-2000: Individual and contextual determinants. *Journal of Ethnic and Migration Studies*, *34*(1), 1–26. https://doi.org/10.1080/13691830701708809

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*, S95–S120. https://doi.org/10.1086/228943

Coleman, J. S. (1990). The demand for effective norms. In *Foundations of social theory* (pp. 241–265). Harvard University Press.

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. https://doi.org/10.1038/d41586-022-00150-2

Collier, P. (2013). The social consequences. In *Exodus: How migration is changing our world* (pp. 57–110). Oxford University Press.

Conell, C., & Cohn, S. (1995). Learning from other people's actions: environmental variation and diffusion in French coal mining strikes, 1890-1935. *American Journal of Sociology*, *101*(2), 366–403. https://doi.org/10.1086/230728

Cornes, R., & Sandler, T. (1994). Are public goods myths? *Journal of Theoretical Politics*, *6*(3), 369–385. https://doi.org/10.1177/0951692894006003006

Cornes, R., & Sandler, T. (1996). *The theory of externalities, public goods, and club goods*. Cambridge University Press.

Corten, R., Buskens, V., & Rosenkranz, S. (2020). Cooperation, reputation effects, and network dynamics: Experimental evidence. In V. Buskens, R. Corten, & C. Snijders (Eds.), *Advances in the sociology of trust and cooperation* (pp. 391–416). de Gruyter. https://doi.org/10.1515/9783110647495-017

Cotta, B. (2020). What goes around, comes around? Access and allocation problems in Global North–South waste trade. *International Environmental Agreements: Politics, Law and Economics*, *20*(2), 255–269. https://doi.org/10.1007/s10784-020-09479-3

Cowen, T. (1992). Law as a public good: The economics of anarchy. *Economics and Philosophy*, *8*(2), 249–267. https://doi.org/10.1017/S0266267100003060

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Croson, R. T. A., & Marks, M. B. (2000). Step returns in threshold public goods: A meta- and experimental analysis. *Experimental Economics*, *2*(3), 239–259. https://doi.org/10.1007/bf01669198

Cubitt, R. P., Drouvelis, M., Gächter, S., & Kabalin, R. (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics*, *95*(3–4), 253–264. https://doi.org/10.1016/j.jpubeco.2010.10.011

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate?: Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784–803. https://doi.org/10.1037/0022-3514.90.5.784

D'Adda, G., Drouvelis, M., & Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, *62*, 1–7. https://doi.org/10.1016/j.socec.2016.02.003

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, *64*(3), 1081–1100. https://doi.org/10.1287/mnsc.2016.2616

Dannick, L. I. (1973). Influence of an anonymous stranger on a routine decision to act or not to act: An experiment in conformity. *Sociological Quarterly*, *14*(1), 127–134. https://doi.org/10.1111/j.1533-8525.1973.tb02119.x

Dawes, R., McTavish, J., & Shaklee, H. (1977). Behavior, communication and assumptions about other people's behavior in a common dilemma situation. *Journal of Personality and Social Psychology*, *35*, 1–11. https://doi.org/10.5840/zfs193211/2122

De Coninck, D., Ogan, C., & d'Haenens, L. (2021). Can 'the Other' ever become 'One of Us'? Comparing Turkish and European attitudes toward refugees: A five-country study. *International Communication Gazette*, *83*(3), 217–237. https://doi.org/10.1177/1748048519895376

De Dreu, C. K. W., & Gelfand, M. J. (2008). Conflict in the workplace: Sources, functions, and dynamics across multiple levels of analysis. In C. K. W. De Dreu & M. J. Gelfand (Eds.), *The psychology of conflict and conflict management in organizations* (pp. 3–54). Erlbaum.

De Dreu, C. K. W., Gross, J., Fariña, A., & Ma, Y. (2020). Group cooperation, carrying-capacity stress, and intergroup conflict. *Trends in Cognitive Sciences*, *24*(9), 760–776. https://doi.org/10.1016/j.tics.2020.06.005

De Matos Fernandes, C. A., Flache, A., Bakker, D. M., & Dijkstra, J. (2022). A bad barrel spoils a good apple: How uncertainty and networks affect whether matching rules can foster cooperation. *Journal of Artificial Societies and Social Simulation*, *25*(1). https://doi.org/10.18564/jasss.4754

De Oliveira, A., Croson, R. T. A., & Eckel, C. C. (2009). *One bad apple: Uncertainty and heterogeneity in public good provision* (Working Paper). https://doi.org/10.2139/ssrn.1883766

De Sousa, Y. F., & Munro, A. (2012). Truck, barter and exchange versus the endowment effect: Virtual field experiments in an online game environment. *Journal of Economic Psychology*, *33*(3), 482–493. https://doi.org/10.1016/j.joep.2011.12.011

Degen, D., Kuhn, T., & Van der Brug, W. (2019). Granting immigrants access to social benefits? How self-interest influences support for welfare state restrictiveness. *Journal of European Social Policy*, *29*(2), 148–165. https://doi.org/10.1177/0958928718781293

DePaulo, B. M., Kirkendol, S. E., Kashy, D. A., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979–995. https://doi.org/10.1037/0022-3514.70.5.979

DeRidder, R. E., & Tripathi, R. C. E. (1992). *Norm violation and intergroup relations*. Clarendon Press/Oxford University Press.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, *2*(4), 265–279. https://doi.org/10.1177/002200275800200401

Diekmann, A., & Przepiorka, W. (2016). "Take one for the team!" Individual heterogeneity and the emergence of latent norms in a volunteer's dilemma. *Social Forces*, *94*(3), 1309–1333. https://doi.org/10.1093/sf/sov107

Diekmann, A., Przepiorka, W., & Rauhut, H. (2015). Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. *Rationality and Society*, *27*(3), 309–333. https://doi.org/10.1177/1043463115593109

Dimant, E. (2019). Contagion of pro- and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, *73*, 66–88. https://doi.org/10.1016/j.joep.2019.04.009

Dimant, E. (2022). Hate trumps love: The impact of political polarization on social preferences. *Management Science*. https://doi.org/10.2139/ssrn.3848335

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy*, *65*(2), 135–150. https://doi.org/10.1086/257897

Drouvelis, M., Malaeb, B., Vlassopoulos, M., & Wahba, J. (2021). Cooperation in a fragmented society: Experimental evidence on Syrian refugees and natives in Lebanon. *Journal of Economic Behavior and Organization*, *187*, 176–191. https://doi.org/10.1016/j.jebo.2021.04.032

Duffy, J., & Lafky, J. (2016). Birth, death and public good provision. *Experimental Economics*, *19*(2), 317–341. https://doi.org/10.1007/s10683-015-9439-y

Duffy, J., & Lafky, J. (2021). Social conformity under evolving private preferences. *Games and Economic Behavior*, *128*, 104–124. https://doi.org/10.1016/j.geb.2021.04.005

Duffy, J., & Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, *66*(2), 785–812. https://doi.org/10.1016/j.geb.2008.07.003

Duffy, J., Ochs, J., & Vesterlund, L. (2007). Giving little by little: Dynamic voluntary contribution games. *Journal of Public Economics*, *91*(9), 1708–1730. https://doi.org/10.1016/j.jpubeco.2007.04.001

Dugar, S. (2013). Non-monetary incentives and opportunistic behavior: Evidence from a laboratory public good game. *Economic Inquiry*, *51*(2), 1374–1388. https://doi.org/10.1111/j.1465-7295.2011.00422.x

Dur, R., & Sol, J. (2010). Social interaction, co-worker altruism, and incentives. *Games and*

Economic Behavior, *69*(2), 293–301. https://doi.org/10.1016/j.geb.2009.10.013

Efferson, C., & Vogt, S. (2018). Behavioural homogenization with spillovers in a normative domain. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879). https://doi.org/10.1098/rspb.2018.0492

Efferson, C., Vogt, S., & Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour*, *4*(1), 55–68. https://doi.org/10.1038/s41562-019-0768-2

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878. https://doi.org/10.1098/rspb.2007.1558

Ellemers, N. (2021). Science as collaborative knowledge generation. *British Journal of Social Psychology*, *60*(1), 1–28. https://doi.org/10.1111/bjso.12430

Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, *24*(1), 160–193. https://doi.org/10.1080/10463283.2013.841490

Ellemers, N., & Van der Toorn, J. (2015). Groups as moral anchors. *Current Opinion in Psychology*, *6*, 189–194. https://doi.org/10.1016/j.copsyc.2015.08.018

Elster, J. (1989). *The cement of society: A study of social order*. Cambridge University Press.

Engl, F., Riedl, A. M., & Weber, R. A. (2021). Spillover effects of institutions on cooperative behavior, preferences, and beliefs. *American Economic Journal: Microeconomics 2021, 13*(4), 261–299. https://doi.org/10.1257/mic.20180336

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., Anum, A., Arikan, G., Aycan, Z., Bagherian, F., Barrera, D., Basnight-Brown, D., Batkeyev, B., Belaus, A., Berezina, E., Björnstjerna, M., … Van Lange, P. A. M. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature Communications*, *12*(1), 1–11. https://doi.org/10.1038/s41467-021-21602-9

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *ELife*, *10*, 1–30. https://doi.org/10.7554/eLife.71601

Espinosa, M. P., Fatás, E., & Ubeda, P. (2019). Linguistic diversity and out-group discrimination in bilingual societies. *Journal of Behavioral and Experimental Economics*, *81*(July 2018), 102–127. https://doi.org/10.1016/j.socec.2019.06.002

European Parliament. (2022). *The Dutch childcare benefit scandal, institutional racism and algorithms*. https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html#def1

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

Fehl, K., Van der Post, D. J., & Semmann, D. (2011). Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters*, *14*(6), 546–551. https://doi.org/10.1111/j.1461-0248.2011.01615.x

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Fehr, E., & Gintis, H. (2007). Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology*, *33*, 43–64. https://doi.org/10.1146/annurev.soc.33.040406.131812

Fehr, E., & Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the Tragedy of the Commons. *Journal of Public Economics*, *95*(9–10), 1144–1155. https://doi.org/10.1016/j.jpubeco.2011.05.013

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468. https://doi.org/10.1038/s41562-018-0385-5

Fehr, E., & Williams, T. (2018). *Social norms, endogenous sorting and the culture of cooperation* (No. 11457; IZA Discussion Papers). http://hdl.handle.net/10419/180475

Fellner, G., & Lünser, G. K. (2014). Cooperation in local and global groups. *Journal of Economic Behavior and Organization*, *108*, 364–373. https://doi.org/10.1016/j.jebo.2014.02.007

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*(1), 541–556.

Fischbacher, U., Schudy, S., & Teyssier, S. (2014). Heterogeneous reactions to heterogeneity in returns from public goods. *Social Choice and Welfare*, *43*(1), 195–217. https://doi.org/10.1007/s00355-013-0763-x

Flache, A., & Macy, M. W. (2006). Learning and framing in social exchange. In D. Fechtenhauer, A. Flache, A. Buunk, & S. M. Lindenberg (Eds.), *Solidarity and prosocial behavior: An integration of sociological and psychological perspectives* (pp. 45–60). Springer. https://doi.org/10.1007/0-387-28032-4_3

Francis, H. (1985). The law, oral tradition and the mining community. *Journal of Law and Society*, *12*(3), 267–271.

Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing “conditional cooperation” in a field experiment. *American Economic Review*, *94*(5), 1717–1722. https://doi.org/10.1257/0002828043052187

Frey, U. J. (2017). Cooperative strategies outside the laboratory — evidence from a long-term large-N-study in five countries. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2016.07.006

Frey, U. J. (2019). Long-term evidence on cooperation and cultural differences in public goods dilemmas. *Biology Letters*, *15*(4). https://doi.org/10.1098/rsbl.2019.0143

Frey, U. J., & Rusch, H. (2012). An evolutionary perspective on the long-term efficiency of costly punishment. *Biology and Philosophy*, *27*(6), 811–831. https://doi.org/10.1007/s10539-012-9327-1

Friedman, B. D., & Oprea, R. (2012). A continuous dilemma. *American Economic Review*, *102*(1), 337–363. https://doi.org/10.1257/aer.102.1.337

Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *78*(2), 1–8. https://doi.org/10.1103/PhysRevE.78.026117

Gächter, S. (2007). Conditional cooperation. Behavioral regularities from the lab and the field and their policy implications. In B. Frey & A. Stutzer (Eds.), *Economics and psychology. A promising new cross-disciplinary field. CESifo Seminar Series*. MIT Press.

Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior and Organization*, *55*(4 SPEC.ISS.), 505–531. https://doi.org/10.1016/j.jebo.2003.11.006

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2651–2661. https://doi.org/10.1098/rstb.2010.0135

Gächter, S., Mengel, F., Tsakas, E., & Vostroknutov, A. (2017). Growth and inequality in public good provision. *Journal of Public Economics*, *150*, 1–13. https://doi.org/10.1016/j.jpubeco.2017.03.002

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510. https://doi.org/10.1126/science.1164744

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common ingroup identity model*. Psychology Press.

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games : A systematic lab-field study. *Management Science*, *65*(3), 976–1002.

Gallier, C., Goeschl, T., Kesternich, M., Lohse, J., Reif, C., & Römer, D. (2019). Leveling up? An inter-neighborhood experiment on parochialism and the efficiency of multi-level public goods provision. *Journal of Economic Behavior and Organization*, *164*, 500–517. https://doi.org/10.1016/j.jebo.2019.05.028

Gallus, J. (2017). Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia. *Management Science*, *63*(12), 3999–4015. https://doi.org/10.1287/mnsc.2016.2540

Gangadharan, L., Nikiforakis, N., & Villeval, M. C. (2017). Normative conflict and the limits of self-governance in heterogeneous populations. *European Economic Review*, *100*, 143–156. https://doi.org/10.1016/j.euroecorev.2017.07.013

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068–6073. https://doi.org/10.1073/pnas.1703857114

Gelcich, S., Guzman, R., Rodríguez-Sickert, C., Castilla, J. C., & Cárdenas, J. C. (2013). Exploring external validity of common pool resource experiments: Insights from artisanal benthic fisheries in Chile. *Ecology and Society*, *18*(3). https://doi.org/10.5751/ES-05598-180302

Gereke, J., Schaub, M., & Baldassarri, D. (2021). Immigration, integration and cooperation:

experimental evidence from a public goods game in Italy. *Journal of Ethnic and Migration Studies*, *0*(0), 1–28. https://doi.org/10.1080/1369183X.2021.1949269

Ghidoni, R., Cleave, B. L., & Suetens, S. (2019). Perfect and imperfect strangers in social dilemmas. *European Economic Review*, *116*, 148–159. https://doi.org/10.1016/j.euroecorev.2019.04.002

Ghimire, R., & Skinner, J. (2019). The relationship between length of community tenure and residents' volunteering at community events: Results from the Metro Atlanta Speaks Survey. *Journal of Rural Social Sciences*, *34*(2), 3.

Gneezy, U., Leibbrandt, A., & List, J. A. (2016). Ode to the sea: Workplace organizations and norms of cooperation. *Economic Journal*, *126*(595), 1856–1883. https://doi.org/10.1111/ecoj.12209

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*(3), 481–510. https://doi.org/10.1086/228311

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Grimson, D., Knowles, S., & Stahlmann-Brown, P. (2020). How close to home does charity begin? *Applied Economics*, *52*(34), 3700–3708. https://doi.org/10.1080/00036846.2020.1720906

Grossman, G., & Baldassarri, D. (2012). The impact of elections on cooperation: Evidence from a lab-in-the-field experiment in uganda. *American Journal of Political Science*, *56*(4), 964–985. https://doi.org/10.1111/j.1540-5907.2012.00596.x

Grund, C., Harbring, C., & Thommes, K. (2015). Public good provision in blended groups of partners and strangers. *Economics Letters*, *134*, 41–44. https://doi.org/10.1016/j.econlet.2015.06.008

Grund, C., Harbring, C., & Thommes, K. (2018). Group (re-) formation in public good games: The tale of the bad apple? *Journal of Economic Behavior and Organization*, *145*, 306–319. https://doi.org/10.1016/j.jebo.2017.11.009

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–59. https://doi.org/10.1017/S0140525X11001142

Guala, F., & Mittone, L. (2010). How history and convention create norms: An experimental study. *Journal of Economic Psychology*, *31*(4), 749–756. https://doi.org/10.1016/j.joep.2010.05.009

Guido, A., Robbett, A., & Romaniuc, R. (2019). Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence. *Journal of Economic Behavior and Organization*, *159*, 192–209. https://doi.org/10.1016/j.jebo.2019.02.009

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, *120*, 220–230. https://doi.org/10.1016/j.jpubeco.2014.10.001

Gurven, M. (2004). To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences*, *27*(4), 543–560. https://doi.org/10.1017/s0140525x04000123

Habyarimana, J., Humphreys, M., Posner, D. N., & Weinstein, J. M. (2009). *Coethnicity:*

diversity and the dilemmas of collective action. Russell Sage Foundation.

Hainmueller, J., Hangartner, D., & Lawrence, D. (2016). When lives are put on hold: Lengthy asylum processes decrease employment among refugees. *Science Advances*, *2*(8), 1–8. https://doi.org/10.1126/sciadv.1600432

Hainmueller, J., & Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, *17*, 225–249. https://doi.org/10.1146/annurev-polisci-102512-194818

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, *111*(3), 465–497. https://doi.org/10.1086/374182

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*(1), 1–16. https://doi.org/10.1016/0022-5193(64)90038-4

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Hargreaves Heap, S. P., Ramalingam, A., & Stoddard, B. V. (2016). Endowment inequality in public goods games: A re-examination. *Economics Letters*, *146*, 4–7. https://doi.org/10.1016/j.econlet.2016.07.015

Hauge, K. E. (2015). Moral opinions are conditional on the behavior of others. *Review of Social Economy*, *73*(2), 154–175. https://doi.org/10.1080/00346764.2015.1035910

Hauser, O. P., Hendriks, A., Rand, D. G., & Nowak, M. A. (2016). Think global, act local: Preserving the global commons. *Scientific Reports*, *6*, 1–7. https://doi.org/10.1038/srep36079

Hauser, O. P., Hilbe, C., Chatterjee, K., & Nowak, M. A. (2019). Social dilemmas among unequals. *Nature*, *572*, 524–527. https://doi.org/10.1038/s41586-019-1488-5

Hawkes, C., Altman, J., Beckerman, S., Grinker, R. R., Harpending, H., Jeske, J., Peterson, N., Smith, E. A., Wenzel, G. W., & Yellen, J. E. (1993). Why hunter gatherers work: An ancient version of the problem of public goods. *Current Anthropology*, *34*(4), 341–361. https://doi.org/10.1086/204182

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

Henrich, J. (2015). *The secret of our success*. Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78. https://doi.org/10.1257/aer.91.2.73

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815. https://doi.org/10.1017/S0140525X05000142

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?

Behavioral and Brain Sciences, *33*, 61–135.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardaroas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesoronol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Hergueux, J., Algan, Y., Benkler, Y., & Fuster Morell, M. (2015). Cooperation in a peer production economy: Experimental evidence from Wikipedia. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2843518

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*(4), 311–335. https://doi.org/10.1002/ejsp.2420200404

Heyblom, O. (2020). Racisme in ons taalgebruik: 'Witte Nederlanders moeten elkaar corrigeren.' *Het Parool*. https://www.parool.nl/nederland/racisme-in-ons-taalgebruik-witte-nederlanders-moeten-elkaar-corrigeren~b9cc690b/

Hill, K. (2002). Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Human Nature*, *13*(1), 105–128. https://doi.org/10.1007/s12110-002-1016-3

Hill, L. (2006). Low voter turnout in the united states: Is compulsory voting a viable solution? *Journal of Theoretical Politics*, *18*(2), 207–232. https://doi.org/10.1177/0951629806061868

Hobbes, T. (1968 [1651]). *Leviathan*. Penguin.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), 36–44. https://doi.org/10.1038/s41562-019-0734-z

Huckman, R. S., Staats, B. R., & Upton, D. M. (2009). Team familiarity, role experience, and performance: Evidence from Indian software services. *Management Science*, *55*(1), 85–100. https://doi.org/10.1287/mnsc.1080.0921

Hume, D. (1952 [1739]). *A treatise of human nature*. J.M. Dent.

Igami, M. (2017). Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *Journal of Political Economy*, *125*(3), 798–847. https://doi.org/10.1086/691524

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., Ivory, J. D., & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*(11), 1092–1094. https://doi.org/10.1038/s41562-020-00990-w

Ikariam fandom. (2020). *Donations to the shared island resources*. https://web.archive.org/web/20200926104810/https://ikariam.fandom.com/wiki/Island

Ikariam leecher checker. (2011). https://web.archive.org/web/20110824065054/http://olimario.110mb.com/ikariam_leecher_checker.php

Innocenti, A. (2017). Virtual reality experiments in economics. *Journal of Behavioral and Experimental Economics* , *69*, 71–77. https://doi.org/10.1016/j.socec.2017.06.001

Isaac, R., & Walker, J. (1988a). Communication and free riding behavior: the voluntary contributions mechanism. *Economic Inquiry*, *26*(4), 585–608.

Isaac, R., & Walker, J. M. (1988b). Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, *103*(1), 179–199. https://doi.org/10.2307/1882648

Israel, S., Weisel, O., Ebstein, R. P., & Bornstein, G. (2012). Oxytocin, but not vasopressin, increases both parochial and universal altruism. *Psychoneuroendocrinology*, *37*(8), 1341–1344. https://doi.org/10.1016/j.psyneuen.2012.02.001

Jackson, M., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology*, *39*, 27–49. https://doi.org/10.1146/annurev-soc-071811-145443

Jackson, R. B., Friedlingstein, P., Andrew, R. M., Canadell, J. G., Le Quéré, C., & Peters, G. P. (2019). Persistent fossil fuel growth threatens the Paris Agreement and planetary health. *Environmental Research Letters*, *14*(12). https://doi.org/10.1088/1748-9326/ab57b3

Jetten, J., Spears, R., & Manstead, A. S. R. (1996). Intergroup norms and intergroup discrimination : Distinctive self-categorization and social identity effects. *Journal of Personality and Social Psychology*, *71*(6), 1222–1233. https://doi.org/10.1037/0022-3514.71.6.1222

Kallhoff, A. (2014). Why societies need public goods. *Critical Review of International Social and Political Philosophy*, *17*(6), 635–651.

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405

Keser, C., & Winden, F. van. (2000). Conditional cooperation and voluntary contributions to public goods. *The Scandinavian Journal of Economics*, *102*(1), 23–39.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608–638. https://doi.org/10.1111/jeea.12152

Kingsley, D. C. (2016). Endowment heterogeneity and peer punishment in a public good experiment : Cooperation and normative conflict. *Journal of Behavioral and Experimental Economics*, *60*, 49–61. https://doi.org/10.1016/j.socec.2015.12.002

Klarenbeek, L. M. (2021). Reconceptualising 'integration as a two-way process.' *Migration Studies*, *9*(3), 902–921. https://doi.org/10.1093/migration/mnz033

Klein, J. T. (2017). Typologies of interdisciplinarity. In R. Frodeman, J. T. Klein, & R. C. D. S. Pacheco (Eds.), *The Oxford handbook of interdisciplinarity* (pp. 21–34). Oxford University Press.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*, *112*(4), 1251–1288. https://doi.org/10.1162/003355300555475

Kocher, M. G., Martinsson, P., & Visser, M. (2008). Does stake size matter for cooperation and punishment? *Economics Letters*, *99*(3), 508–511. https://doi.org/10.1016/j.econlet.2007.09.048

Kölle, F. (2015). Heterogeneity and cooperation: The role of capability and valuation on public goods provision. *Journal of Economic Behavior and Organization*, *109*, 120–134.

https://doi.org/10.1016/j.jebo.2014.11.009

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*, 182–214. https://doi.org/10.1146/annurev.soc.24.1.183

Koopmans, R., & Veit, S. (2014). Cooperation in ethnically diverse neighborhoods: A lost-letter experiment. *Political Psychology*, *35*(3), 379–400. https://doi.org/10.1111/pops.12037

Kraft-Todd, G., Yoeli, E., Bhanot, S., & Rand, D. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, *3*, 96–101. https://doi.org/10.1016/j.cobeha.2015.02.006

Kruissink, M., & Essers, A. A. M. (2004). *Zelfgerapporteerde jeugdcriminaliteit in de periode 1990-2001* (Report).

Krupka, E. L., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), 307–320. https://doi.org/10.1016/j.joep.2008.11.005

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006

Kube, S., Schaube, S., Schildberg-Hörisch, H., & Khachatryan, E. (2015). Institution formation and cooperation with heterogeneous agents. *European Economic Review*, *78*, 248–268. https://doi.org/10.1016/j.euroecorev.2015.06.004

Kulczycki, A., & Windle, S. (2011). Honor killings in the Middle East and North Africa: A systematic review of the literature. *Violence Against Women*, *17*(11), 1442–1464. https://doi.org/10.1177/1077801211434127

Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences*, *102*(5), 1803–1807. https://doi.org/10.1073/pnas.0408759102

Lai, C. K., & Lisnek, J. A. (2023). The impact of implicit-bias-oriented diversity training on police officers' beliefs, motivations, and actions. *Psychological Science*. https://doi.org/10.1177/09567976221150617

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, *90*, 375–402. https://doi.org/10.1016/j.euroecorev.2015.11.011

Lange, A. (2006). The impact of equity-preferences on the stability of international environmental agreements. *Environmental & Resource Economics*, *34*(2), 247–267. https://doi.org/10.1007/s10640-005-0006-4

Lange, A., Löschel, A., Vogt, C., & Ziegler, A. (2010). On the self-interested use of equity in international climate negotiations. *European Economic Review*, *54*(3), 359–375. https://doi.org/10.1016/j.euroecorev.2009.08.006

Lange, A., Schmitz, J., & Schwirplies, C. (2022). Inequality, role reversal and cooperation in multiple group membership settings. *Experimental Economics*, *25*(1), 68–110. https://doi.org/10.1007/s10683-021-09705-y

Laury, S. K., & Taylor, L. O. (2008). Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *Journal of Economic Behavior and Organization*, *65*(1), 9–29. https://doi.org/10.1016/j.jebo.2005.05.011

Lazega, E., Snijders, T. A., & Wittek, R. P. (Eds.). (2022). *A research agenda for social networks and social resilience*. Edward Elgar Publishing.

Leach, C. W., Van Zomeren, M., Zebel, S., Vliek, M. L. W., Pennekamp, S. F., Doosje, B., Ouwerkerk, J. W., & Spears, R. (2008). Group-Level self-definition and self-investment: A hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology*, *95*(1), 144–165. https://doi.org/10.1037/0022-3514.95.1.144

Ledyard, J. O. (1994). Public goods: A survey of experimental research. In J. Kagel & A. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 111–194). Princeton University Press. https://doi.org/10.1002/em.10197

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the Stereotype Content Model. *International Journal of Intercultural Relations*, *30*(6), 751–768. https://doi.org/10.1016/j.ijintrel.2006.06.005

Lefkowitz, M., Blake, R. R., & Mouton, J. S. (1955). Status factors in pedestrian violation of traffic signals. *Journal of Abnormal and Social Psychology*, *51*(3), 704–706. https://doi.org/10.1037/h0042000

Leszczensky, L., & Wolbring, T. (2022). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. *Sociological Methods and Research*, *51*(2), 837–865. https://doi.org/10.1177/0049124119882473

Levine, J. M., & Moreland, R. L. (1994). Group socialization: Theory and research. *European Review of Social Psychology*, *5*(1), 305–336. https://doi.org/10.1080/14792779543000093

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

Lindenberg, S. (1992). The method of decreasing abstraction. In J. S. Coleman & T. J. Fararo (Eds.), *Rational choice theory. Advocacy and critique* (pp. 3–20). Sage.

Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, *147*(2), 228–242. https://doi.org/10.1037/xge0000365

Lubbers, M., Coenders, M., & Scheepers, P. (2006). Objections to asylum seeker centres: Individual and contextual determinants of resistance to small and large centres in the Netherlands. *European Sociological Review*, *22*(3), 243–257. https://doi.org/10.1093/esr/jci055

Macanovic, A., & Przepiorka, W. (2022). *A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from feedback texts and tweets* (Working Paper).

Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American Sociological Review*, *61*(6), 999–1017.

Maier-Rigaud, F. P., Martinsson, P., & Staffiero, G. (2010). Ostracism and the provision of a public good : experimental evidence. *Journal of Economic Behavior and Organization*, *73*(3), 387–395. https://doi.org/10.1016/j.jebo.2009.11.001

Malkin, J., & Wildavsky, A. (1991). Why the traditional distinction between public and private goods should be abandoned. *Journal of Theoretical Politics*, *3*(4), 355–378. https://doi.org/10.1177/0951692891003004001

Marlowe, F. W. (2010). *The Hazda: Hunter-gatherers of Tanzania*. University of California Press.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, *93*(1), 366–380. https://doi.org/10.1257/000282803321455359

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, *108*(28), 11375–11380. https://doi.org/10.1073/pnas.1105604108

McCarter, M. W., & Sheremeta, R. M. (2013). You can't put old wine in new bottles: The effect of newcomers on coordination in groups. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0055058

McGinnis, M., & Ostrom, E. (2008). Will lessons from small-scale social dilemmas scale up? In A. Biel, D. Eek, T. Gärling, & M. Gustafsson (Eds.), *New issues and paradigms in research on social dilemmas* (pp. 189–211). Springer.

Melamed, D., Harrell, A., & Simpson, B. (2018). Cooperation, clustering, and assortative mixing in dynamic networks. *Proceedings of the National Academy of Sciences*, *115*(5), 951–956. https://doi.org/10.1073/pnas.1715357115

Milgram, S., Liberty, H. J., Toledo, R., & Wackenhut, J. (1986). Response to intrusion into waiting lines. *Journal of Personality and Social Psychology*, *51*(4), 683–689. https://doi.org/10.1037/0022-3514.51.4.683

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48. https://doi.org/10.1037/0021-843X.110.1.40

Molenberghs, P., & Louis, W. R. (2018). Insights from fMRI studies into ingroup bias. *Frontiers in Psychology*, *9*(OCT), 1–12. https://doi.org/10.3389/fpsyg.2018.01868

Moreland, R. L. (1985). Social categorization and the assimilation of "new" group members. *Journal of Personality and Social Psychology*, *48*(5), 1173–1190. https://doi.org/10.1037/0022-3514.48.5.1173

Mujcic, R., & Frijters, P. (2020). The colour of a free ride. *The Economic Journal*, *131*, 970–999.

Mullen, B., Copper, C., & Driskell, J. E. (1990). Jaywalking as a function of model behavior. *Personality and Social Psychology Bulletin*, *16*(2), 320–330. https://doi.org/10.1177/0146167290162012

Muller, E. N., & Opp, K.-D. (1986). Rational choice and rebellious collective action. *American Political Science Review*, *80*(2), 471–487. https://doi.org/10.2307/1958269

Naylor, R. (1989). Strikes, free riders, and social customs. *The Quarterly Journal of Economics*, *104*(4), 771–785. https://doi.org/10.2307/2937867

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, *96*(9–10), 797–807. https://doi.org/10.1016/j.jpubeco.2012.05.014

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. grave le B., Rohrer, J. M., Romero, F., Scheel,

A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosenzo, D., Quercia, S., & Sefton, M. (2015). Cooperation in small groups: the effect of group size. *Experimental Economics*, *18*(1), 4–14. https://doi.org/10.1007/s10683-013-9382-8

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Number of daily travelers at Dutch train stations. (2022). https://web.archive.org/web/20220920070234/https://dashboards.nsjaarverslag.nl/reizigersgedrag/

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.

Olsson, L., Jerneck, A., Thoren, H., Persson, J., & O'Byrne, D. (2015). Why resilience is unappealing to social science: Theoretical and empirical investigations of the scientific use of resilience. *Science Advances*, *1*(4), 1–12. https://doi.org/10.1126/sciadv.1400217

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Oprea, R., Charness, G., & Friedman, D. (2014). Continuous time and communication in a public-goods experiment. *Journal of Economic Behavior and Organization*, *108*, 212–223. https://doi.org/10.1016/j.jebo.2014.09.012

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158. https://doi.org/10.1257/jep.14.3.137

Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2020). Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels. *Scientific Reports*, *16702*, 1–12. https://doi.org/10.1038/s41598-020-73314-7

Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2021). Cooperation between newcomers and incumbents: The role of normative disagreements. *Journal of Economic Psychology*, *87*, 102448. https://doi.org/10.1016/j.joep.2021.102448

Otten, K., Frey, U. J., Buskens, V., Przepiorka, W., & Ellemers, N. (2022). Human cooperation in changing groups in a large-scale public goods game. *Nature Communications*, *13*(1), 1–11. https://doi.org/10.1038/s41467-022-34160-5

Ouwerkerk, J. W., Ellemers, N., & Kortekaas, P. (1999). Self-categorisation, commitment to the group and group self-esteem as related but distinct aspects of social identity. *European Journal of Social Psychology*, *29*(2–3), 371–389. https://doi.org/10.1002/(SICI)1099-0992(199903/05)29:2/3<371::AID-EJSP932>3.0.CO;2-U

Ozono, H., Kamijo, Y., & Shimizu, K. (2020). The role of peer reward and punishment for public goods problems in a localized society. *Scientific Reports*, *10*(1), 1–6. https://doi.org/10.1038/s41598-020-64930-4

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, *34*, 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

Pereda, M., Tamarit, I., Antonioni, A., Cuesta, J. A., Hernández, P., & Sánchez, A. (2019). Large scale and information effects on cooperation in public good games. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-50964-w

Pettigrew, T. F. (1991). Normative theory in intergroup relations: Explaining both harmony and conflict. *Psychology and Developing Societies*, *3*(1), 3–16. https://doi.org/10.1177/097133369100300102

Pettigrew, T. F. (2020). Intergroup attribution. *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.326

Peysakhovich, A., & Rand, D. G. (2013). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647. https://doi.org/10.1021/la00030a006

Piskorski, M. J., & Gorbatai, A. D. (2017). Testing Coleman's social-norm enforcement mechanism: Evidence from Wikipedia. *American Journal of Sociology*, *122*(4), 1183–1222. https://doi.org/10.1086/689816

Polzer, J. T., Milton, L. P., & Gruenfeld, D. H. (2009). *Asymmetric subgroup communication in nested social dilemmas* (Working Paper).

Polzer, J. T., Stewart, K. J., & Simmons, J. L. (1999). A social categorization explanation for framing effects in nested social dilemmas. *Organizational Behavior and Human Decision Processes*, *79*(2), 154–178. https://doi.org/10.1006/obhd.1999.2842

Pratsinakis, M. (2018). Established and outsider nationals: Immigrant–native relations and the everyday politics of national belonging. *Ethnicities*, *18*(1), 3–22. https://doi.org/10.1177/1468796817692838

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S., Walleczek, J., & Schooler, J. W. (2020). *High replicability of newly-discovered social-behavioral findings is achievable* (Working Paper).

Przepiorka, W., & Berger, J. (2016). The sanctioning dilemma: A quasi-experiment on social norm enforcement in the train. *European Sociological Review*, *32*(3), 439–451. https://doi.org/10.1093/esr/jcw014

Przepiorka, W., & Diekmann, A. (2018). Heterogeneous groups overcome the diffusion of responsibility problem in social norm enforcement. *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0208129

Przepiorka, W., & Diekmann, A. (2021). Parochial cooperation and the emergence of signalling norms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 533–547. https://doi.org/10.1098/rstb.2020.0294

Przepiorka, W., Szekely, A., Andrighetto, G., Diekmann, A., & Tummolini, L. (2022). How norms emerge from conventions (and change). *Socius*, *8*. https://doi.org/10.1177/23780231221124556

Quervain, D. J.-F. de, Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology*, *32*, 299–328.

https://doi.org/10.1146/annurev.soc.32.061604.123132

Ramalingam, A., Morales, A. J., & Walker, J. M. (2018). Varying experimental instructions to improve comprehension: Punishment in public goods games. *Journal of Behavioral and Experimental Economics*, *73*(66–73). https://doi.org/10.1016/j.socec.2018.01.008

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, *108*(48), 19193–19198. https://doi.org/10.1073/pnas.1108243108

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272–1275. https://doi.org/10.1126/science.1177418

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Rand, D. G., Nowak, M. A., Fowler, J. H., & Christakis, N. A. (2014). Static network structure can stabilize human cooperation. *Proceedings of the National Academy of Sciences*, *111*(48), 17093–17098. https://doi.org/10.1073/pnas.1400406111

Ranehill, E., Schneider, F., & Weber, R. A. (2014). *Growing groups, cooperation, and the rate of entry* (No. 4719; CESifo Working Paper Series). https://ssrn.com/abstract=2425583

Rathbun, B. C., Powers, K. E., & Anders, T. (2019). Moral hazard: German public opinion on the Greek debt crisis. *Political Psychology*, *40*(3), 523–541. https://doi.org/10.1111/pops.12522

Raub, W., & Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, *96*(3), 626–654. https://doi.org/10.1086/229574

Rauhut, H., & Winter, F. (2017). Types of normative conflicts and the effectiveness of punishment. In W. Przepiorka & B. Jann (Eds.), *Social dilemmas, institutions, and the evolution of cooperation* (pp. 239–256).

Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, *77*(1), 122–137. https://doi.org/10.1016/j.geb.2012.10.001

Reuben, E., Riedl, A., & Bernard, M. (2015). *Fairness and coordination: The role of fairness principles in coordination failure and success* (Working Paper). https://sites.google.com/site/markbernard1984/research-papers

Riedl, A., & Ule, A. (2002). *Exclusion and cooperation in social network experiments* (Working Paper).

Rink, F., Kane, A. A., Ellemers, N., & Van der Vegt, G. (2013). Team receptivity to newcomers: Five decades of evidence and future research themes. *Academy of Management Annals*, *7*(1), 247–293. https://doi.org/10.1080/19416520.2013.766405

Ritchie, H., & Roser, M. (2019). *Age structure*. Our World in Data. https://ourworldindata.org/age-structure

Robbett, A. (2016). Sustaining cooperation in heterogeneous groups. *Journal of Economic Behavior and Organization*, *132*, 121–138. https://doi.org/10.1016/j.jebo.2016.09.012

Romano, A., Spadaro, G., Balliet, D., Joireman, J., Lissa, C. Van, Jin, S., Agostini, M., Bélanger, J. J., Gützkow, B., Kreienkamp, J., Leander, N. P., Abakoumkin, G., Khaiyom, J. H. A., Ahmedi, V., Akkas, H., Almenara, C. A., Atta, M., Bagci, S. C., Basel, S., …

Zúñiga, C. (2021). Cooperation and trust across societies during the COVID-19 pandemic. *Journal of Cross-Cultural Psychology*, *52*(7), 622–642. https://doi.org/10.1177/0022022120988913

Romano, A., Sutter, M., Liu, J. H., Yamagishi, T., & Balliet, D. (2021). National parochialism is ubiquitous across 42 nations around the world. *Nature Communications*, *12*(1), 1–8. https://doi.org/10.1038/s41467-021-24787-1

Rösler, I. K. (2022). *Hear me out: How to create an open mind towards moral criticism*. Utrecht University.

Ruist, J. (2015). The fiscal cost of refugee immigration: The example of Sweden. *Population and Development Review*, *41*(4), 567–581. https://doi.org/10.1111/j.1728-4457.2015.00085.x

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, *330*, 961–965. https://doi.org/10.1126/science.1193649

Salmon, T. C., & Weber, R. A. (2017). Maintaining efficiency while integrating entrants from lower performing groups: an experimental study. *Economic Journal*, *127*(600), 417–444. https://doi.org/10.1111/ecoj.12308

Samuelson, P. A. (1954). The pure theory of public expenditure. *The Review of Economics and Statistics*, *36*(4), 387–389.

Sanders, K. (2009). Cooperative behaviours in organizations. In R. Morrison & S. Wright (Eds.), *Friends and Enemies in Organizations: A Work Psychology Perspective* (pp. 101–121). London: Palgrave Macmillan. https://doi.org/10.1057/9780230248359

Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006). Cooperation prevails when individuals adjust their social ties. *PLoS Computational Biology*, *2*(10), 1284–1291. https://doi.org/10.1371/journal.pcbi.0020140

Santos, F. P., Santos, F. C., & Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, *555*(7695), 242–245. https://doi.org/10.1038/nature25763

Scheepers, D., & Derks, B. (2016). Revisiting social identity theory from a neuroscience perspective. *Current Opinion in Psychology*, *11*, 74–78. https://doi.org/10.1016/j.copsyc.2016.06.006

Scheffer, M., Van Bavel, B., Van de Leemput, I. A., & Van Nes, E. H. (2017). Inequality in nature and society. *Proceedings of the National Academy of Sciences*, *114*(50), 13154–13157. https://doi.org/10.1073/pnas.1706412114

Schmitt, B. H., Dubé, L., & Leclerc, F. (1992). Intrusions into waiting lines: Does the queue constitute a social system? *Journal of Personality and Social Psychology*, *63*(5), 806–815. https://doi.org/10.1037/0022-3514.63.5.806

Schultz, C. (1996). Polarization and inefficient policies. *Review of Economic Studies*, *63*(2), 331–344. https://doi.org/10.2307/2297855

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. Harper.

SCOOP. (2019). *Sustainable Cooperation annual report 2019: Roadmaps to a resilient society*. https://www.scoop-program.org/images/pdf/SCOOP_annual-report_2019.pdf

Scrimpshire, A. J., Stone, T. H., Kisamore, J. L., & Jawahar, I. M. (2017). Do birds of a feather cheat together? How personality and relationships affect student cheating. *Journal of*

Academic Ethics, *15*(1), 1–22. https://doi.org/10.1007/s10805-016-9267-5

Shaw, K., & Lazear, E. P. (2008). Tenure and output. *Labour Economics*, *15*(4), 704–723. https://doi.org/10.1016/j.labeco.2007.08.004

Sherif, M. (1966). *In common predicament: social psychology of intergroup conflict and cooperation*. Houghton Miffl.

Simkins, P. (1988). *Kitchener's army: The raising of the new armies*. Manchester United Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology 26*, 4–7. https://doi.org/10.2139/ssrn.2160588

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Smerdon, D., Offerman, T., & Gneezy, U. (2019). 'Everybody's doing it': on the persistence of bad social norms. *Experimental Economics*, *0123456789*. https://doi.org/10.1007/s10683-019-09616-z

Smith, K. M., Larroucau, T., Mabulla, I. A., & Apicella, C. L. (2018). Hunter-Gatherers maintain assortativity in cooperation despite high levels of residential change and mixing. *Current Biology*, *28*(19), 3152-3157.e4. https://doi.org/10.1016/j.cub.2018.07.064

Smith, S. R. (2017). Modelling "the expanding circle" of cooperation towards a sustainable future. *European Journal of Sustainable Development*, *6*(4), 341–352. https://doi.org/10.14207/ejsd.2017.v6n4p341

Sonnemans, J., Schram, A., & Offerman, T. (1999). Strategic behavior in public good games: When partners drift apart. *Economics Letters*, *62*(1), 35–41. https://doi.org/10.1016/s0165-1765(98)00203-1

Spadaro, G., Graf, C., Jin, S., Arai, S., Inoue, Y., Lieberman, E., Rinderu, M. I., Yuan, M., Lissa, C. J. V., & Balliet, D. (2022). Cross-cultural variation in coperation: A meta-analysis. *Journal of Personality and Social Psychology*, *123*(5), 1024–1088. https://doi.org/10.1037/pspi0000389

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Teije, A. ten, Team, C., & Balliet, D. (2022). The Cooperation Databank: Machine-readable science accelerates research synthesis. *Perspectives on Psychological Science*, 1–18. https://doi.org/10.1177/17456916211053319

Spiller, J., Ufert, A., Vetter, P., & Will, U. (2016). Norms in an asymmetric public good experiment. *Economics Letters*, *142*, 35–44. https://doi.org/10.1016/j.econlet.2016.01.014

Stagnaro, M. N., Arechar, A. A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition*, *167*, 212–254. https://doi.org/10.1016/j.cognition.2017.01.017

Steinert-Threlkeld, Z. C. (2017). Spontaneous collective action: Peripheral mobilization during the arab spring. *American Political Science Review*, *111*(2), 379–403. https://doi.org/10.1017/S0003055416000769

Stoop, J., Noussair, C. N., & Van Soest, D. (2012). From the lab to the field: Cooperation among fishermen. *Journal of Political Economy*, *120*(6), 1027–1056. https://doi.org/10.1086/669253

Storm, I., Sobolewska, M., & Ford, R. (2017). Is ethnic prejudice declining in Britain? Change in social distance attitudes among ethnic majority and minority Britons. *British Journal of Sociology*, *68*(3), 410–434. https://doi.org/10.1111/1468-4446.12250

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of Intergroup Relations. Key readings in social psychology* (pp. 7–24). Nelson-Hall.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cohen, D. (2012). Teams are changing: Are research and practice evolving fast enough? *Industrial and Organizational Psychology*, *5*(1), 2–24. https://doi.org/10.1111/j.1754-9434.2011.01396.x

Tavoni, A., Dannenberg, A., Kallis, G., & Loschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, *108*(29), 11825–11829. https://doi.org/10.1073/pnas.1102493108

Thielemann, E. (2018). Why refugee burden-sharing initiatives fail: Public goods, free-riding and symbolic solidarity in the EU. *Journal of Common Market Studies*, *56*(1), 63–82. https://doi.org/10.1111/jcms.12662

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.19004

Thijssen, L., Coenders, M., & Lancee, B. (2020). Ethnic discrimination in the Dutch labor market: Differences between ethnic minority groups and the role of personal information about job applicants—Evidence from a field experiment. *Journal of International Migration and Integration*. https://doi.org/10.1007/s12134-020-00795-w

Thöni, C., & Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, *171*, 37–40. https://doi.org/10.1016/j.econlet.2018.06.022

Thravalou, E., Martinovic, B., & Verkuyten, M. (2020). Humanitarian assistance and permanent settlement of asylum seekers in Greece : The role of sympathy, perceived threat, and perceived contribution. *International Migration Review*, 1–27. https://doi.org/10.1177/0197918320949826

Tietzel, M., & Weber, M. (1994). The economics of the Iron Curtain and the Berlin Wall. *Rationality and Society*, *6*(1), 58–78. https://doi.org/10.1177/1043463194006001005

Titlestad, K., Snijders, T. A. B., Durrheim, K., Quayle, M., & Postmes, T. (2019). The dynamic emergence of cooperative norms in a social dilemma. *Journal of Experimental Social Psychology*, *84*(103799). https://doi.org/10.1016/j.jesp.2019.03.010

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*(1), 231–255. https://doi.org/10.1146/annurev-psych-113011-143812

Trainer, H. M., Jones, J. M., Pendergraft, J. G., Maupin, C. K., & Carter, D. R. (2020). Team membership change “events”: A review and reconceptualization. *Group & Organization Management*, *45*(2), 219–251. https://doi.org/10.1177/1059601120910848

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Tullock, G. (1971). Public decisions as public goods. *Journal of Political Economy*, *79*(4), 913–918. https://doi.org/10.1086/259799

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987).

Rediscovering the Social Group: A Self-Categorization Theory. Oxford: Basil Blackwell. https://doi.org/10.2307/2073157

Ullmann-Margalit, E. (1977). PD Norms. In *The Emergence of Norms* (pp. 18–73). Oxford Univ. Press.

Van Den Assem, M. J., Van Dolder, D., & Thaler, R. H. (2012). Split or steal? Cooperative behavior when the stakes are large. *Management Science*, *58*(1), 2–20. https://doi.org/10.1287/mnsc.1110.1413

Van der Laan, J., De Jonge, E., Das, M., Te Riele, S., & Emery, T. (2022). A whole population network and its application for the social sciences. *European Sociological Review*, *October*, 1–15. https://doi.org/10.1093/esr/jcac026

Van Dijk, E., & De Dreu, C. K. W. (2021). Experimental games and social decision making. *Annual Review of Psychology*, *72*, 415–438. https://doi.org/10.1146/annurev-psych-081420-110718

Van Gerwen, N., Buskens, V., & Van der Lippe, T. (2018). Individual training and employees' cooperative behavior: Evidence from a contextualized laboratory experiment. *Rationality and Society*, *30*(4), 432–462. https://doi.org/10.1177/1043463118771428

Van Lange, P. A. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141. https://doi.org/10.1016/j.obhdp.2012.11.003

Van Miltenburg, N., Buskens, V., Barrera, D., & Raub, W. (2014). Implementing punishment and reward in the public goods game: The effect of individual and collective decision rules. *International Journal of the Commons*, *8*(1), 47–78. https://doi.org/10.18352/ijc.426

Van Tubergen, F. (2020). *Introduction to sociology*. Routledge.

Van Vugt, M., Snyder, M., Tyler, T. R., & Biel, A. (Eds.). (2000). *Cooperation in modern society: Promoting the welfare of communities, states and organizations*. Routledge.

VanderWeele, T. J., & An, W. (2013). Social networks and causal inference. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 353–374). Springer. http://www.springer.com/series/6055

Voss, T. (2001). Game-theoretical perspectives on the emergence of social norms. In M. Hechter & K. D. Opp (Eds.), *Social Norms*. Russell Sage.

Wageman, R. (1995). Interdependence and group effectiveness. *Administrative Science Quarterly*, *40*(1), 145–180. https://doi.org/10.2307/2393703

Wang, J., Suri, S., & Watts, D. J. (2012). Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, *109*(36), 14363–14368. https://doi.org/10.1073/pnas.1120867109

Weber, R. (2006). Managing growth to achieve efficient coordination in large groups: Theory and experimental evidence. *American Economic Review*, *96*(1), 114–126. https://doi.org/10.1257/000282806776157588

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415. https://doi.org/10.1287/mnsc.49.4.400.14430

Weimann, J., Brosig-Koch, J., Heinrich, T., Hennig-Schmidt, H., & Keser, C. (2019). Public good provision by large groups – the logic of collective action revisited. *European*

Economic Review, *118*, 348–363. https://doi.org/10.1016/j.euroecorev.2019.05.019

Wenzel, M. (2004). An analysis of norm processes in tax compliance. *Journal of Economic Psychology*, *25*(2), 213–228. https://doi.org/10.1016/S0167-4870(02)00168-X

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, *11*(3), 1–22. https://doi.org/10.1371/journal.pone.0152719

Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi Bushmen: A case of strong reciprocity? *Human Nature*, *16*(2), 115–145. https://doi.org/10.1007/s12110-005-1000-9

Wike, R., Stokes, B., & Simmons, K. (2016). *Europeans fear wave of refugees will mean more terrorism, fewer jobs*. 2–5. http://www.pewglobal.org/2016/07/11/europeans-fear-wave-of-refugees-will-mean-more-terrorism-fewer-jobs/

Winter, F., Rauhut, H., & Helbing, D. (2012). How norms can generate conflict: An experiment on the failure of cooperative micro-motives on the macro-level. *Social Forces*, *90*(3), 919–948. https://doi.org/10.1093/sf/sor028

Winter, F., Rauhut, H., & Miller, L. (2018). Dynamic bargaining and normative conflict. *Journal of Behavioral and Experimental Economics*, *74*, 112–126. https://doi.org/10.1016/j.socec.2018.03.003

Winter, F., & Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. *Proceedings of the National Academy of Sciences*, *115*(11), 2722–2727. https://doi.org/10.1073/pnas.1718309115

Wit, A. P., & Kerr, N. L. (2002). "Me versus just us versus us all" categorization and cooperation in nested social dilemmas. *Journal of Personality and Social Psychology*, *83*(3), 616–637. https://doi.org/10.1037/0022-3514.83.3.616

Wolbring, T., Bozoyan, C., & Langner, D. (2013). "Walk left, stand right!": A field experiment on the enforcement of informal norms on escalators. *Zeitschrift für Soziologie*, *42*(3), 239–258. https://doi.org/10.1515/zfsoz-2013-0305

Xiao, E., & Houser, D. (2011). Punish in public. *Journal of Public Economics*, *95*(7–8), 1006–1017. https://doi.org/10.1016/j.jpubeco.2010.11.021

Yang, W., Liu, W., Viña, A., Tuanmu, M. N., He, G., Dietz, T., & Liu, J. (2013). Nonlinear effects of group size on collective action and resource outcomes. *Proceedings of the National Academy of Sciences*, *110*(27), 10916–10921. https://doi.org/10.1073/pnas.1301733110

Yasarcan, H. (2009). Improving understanding, learning, and performances of novices in dynamic managerial simulation games. *Complexity*, *15*(4), 31–42. https://doi.org/10.1002/cplx.20292

Ye, M., Zheng, J., Nikolov, P., & Asher, S. (2020). One step at a time: Does gradualism build coordination? *Management Science*, *66*(1), 113–129.

Ye, Y., Zhang, Q., Wei, X., Cao, Z., Yuan, H. Y., & Zeng, D. D. (2022). Equitable access to COVID-19 vaccines makes a life-saving difference to all countries. *Nature Human Behaviour*, *6*(2), 207–216. https://doi.org/10.1038/s41562-022-01289-8

Young, H. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*(3), 299–310. https://doi.org/10.1023/A:1026277420119

Zhang, N., Aidenberger, A., Rauhut, H., & Winter, F. (2019). Prosocial behaviour in interethnic encounters: Evidence from a field experiment with high-and low-status immigrants. *European Sociological Review*, *35*(4), 582–597. https://doi.org/10.1093/esr/jcz030

Zhang, N., Gereke, J., & Baldassarri, D. (2022). Everyday discrimination in public spaces: A field experiment in the Milan Metro. *European Sociological Review*, *38*(5), 679–693. https://doi.org/10.1093/esr/jcac008

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. https://doi.org/10.1007/s10683-009-9230-z

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies*, *42*(7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279

Acknowledgments

It is a pleasure to acknowledge the contributions of many people to this book. First and foremost, I want to express my sincere gratitude to my supervisors Vincent Buskens, Naomi Ellemers, and Wojtek Przepiorka. Vincent was one of the main reasons to apply for a PhD position. I only applied to his PhD project, and had I not been selected for this project, I am not sure I would have applied for others. His kindness, sharp intellect, and commitment to his students were already clear to me as a master student, and I was not disappointed during the subsequent PhD project. The combination of being a renowned expert in your field and still being so intellectually humble is rare, but so valuable as a scientist and mentor. I am very thankful I could benefit from it with Vincent as my daily supervisor. I also much appreciated the freedom I got. Apart from the project having to do with newcomers, cooperation, and experiments, it was pretty much open for my own input. The freedom to pursue my own ideas was rewarding and required confidence from the supervisor. It also means the first results may take somewhat longer than usual – I conducted my first experiment for this project only in the second year. Yet, Vincent remained patient and trustful with me. His academic work shows that being trustful can pay off, and I hope the quality of my work does not challenge that conclusion.

Next to Vincent's expertise in theoretical sociology, I was lucky to benefit from Naomi's expertise in social psychology. Although these two fields may seem similar, I learned that they are in fact very different. Naomi helped me to appreciate the differences, see the benefits of the approach taken in social psychology, and the even larger benefits in combining and transcending the different approaches. Many of our first meetings involved efforts to move from two disciplinary languages to one interdisciplinary language. We had insightful discussions about fundamental assumptions underlying the different disciplines. In designing our first experiment, for example, we extensively discussed incentives; what are they, do they have to be monetary, how are people engaged by incentives, and is 'incentive' even the right word? Although sometimes difficult, these discussions with Naomi were very helpful for me in learning that there is no one way to do science. I feel lucky to have received training as an interdisciplinary scientist under her guidance. I was also impressed by how many activities and responsibilities she takes on simultaneously – both scientific and societal.

Although the PhD project started with Vincent and Naomi as supervisors, Wojtek soon joined the team. One might think that adding another sociologist to the supervising team shifts the balance in favor of sociology over social psychology. However, this was not the case. Wojtek is an interdisciplinary researcher par excellence, and his addition to the team served only to expand and improve the project's approach. Wojtek is super committed to the work of his students and is actively involved in each step of the research cycle. He replies very fast (even in the late hours) and not only provides comments, but also detailed edits for the text. I am very grateful that he joined the team. I enjoyed working together and learned a lot from him. My fellow PhD students often say that their different supervisors provide similar comments on similar issues. In my case, I found the feedback of my three supervisors to be almost always complementary rather than substitutionary. Sure, this may mean three times the work whenever I received feedback, but it always helped to improve the papers considerably. Altogether, I have been very lucky to receive expert guidance from an all-star team of supervisors.

Next to my supervisors, I had the chance to work with some other outstanding researchers on this project. In the second year of my project, I reached out via email to Ulrich Frey after reading two of his papers about an online multiplayer game *Ikariam*. Ulrich was very kind in his reply, and after a short Skype call, we decided to work together on a joint paper. That paper became Chapter 4 in this thesis, on which Ulrich is the second author. Ulrich not only provided the data, but he also prepared it extensively for data analysis, helped with conceptualizing the research, and provided many valuable comments and insights for the paper. With the lab work lying largely dormant due to COVID-19 restrictions, the large-scale and complex data from Ikariam gave me a useful and enjoyable way to spend my working hours. Also, I started playing the game myself to better understand the data. I convinced myself that playing the game is a useful way to spend my time, and even Vincent played the game for a bit. Vincent wisely stopped after a few weeks, but I am still playing. I even met some online ‘friends’ who took me under their wing and gave me first-hand experience of cooperation in Ikariam. Incidentally, Ikariam takes place in a Greek archipelago, and I spent a significant share of the COVID-19 lockdown in the Greek island of Samos, next to the island of Ikaria.

Through a collaboration between Utrecht University and the Hebrew University of Jerusalem, I came into contact with Boaz Cherki and Salomon Israel, two smart and interesting (neuro)psychologists. Initially we had planned multiple visits between our two universities, but due to COVID-19 restrictions the visits had to be canceled. Still, we managed to collaborate on two experiments. The first was conducted at Utrecht University and became Chapter 5 in this thesis. The second is currently being conducted at the Hebrew University of Jerusalem and is an extension of the first. Despite not being able to meet physically, the collaboration with Boaz and Salomon was smooth and enjoyable.

I was also very fortunate to share a workplace with many nice colleagues. I have shared more than four years with my office roommate Philipp, and he contributed greatly to making the time at the office enjoyable. Not only did he help in numerous informal chats about papers, with Philipp I could also vent about all the bad things in academia, share the good things of academia, play table tennis and table football, feed the crows from our office window, have some beers or even ouzo, go the gym, and so on. I am happy that an office mate became a good friend. I also want to thank the rest of my sociology year-group friends: Sanjana, Kim, and Anne. Sanjana was very kind and supportive, always open for a chat, and her table tennis skills noticeably improved during the four years we spent together. Kim and Anne I already met during our time as master students, and it was very nice to enter into this PhD-trajectory together. Thanks also to the PhD students from the other year groups at our sociology department, in particular Hendrik, Tara, Rita, Ana, Ece, Christian, Nick, Lian, Jos, Kevin, Jan-Willem, Paula, and Amina, for making time at the office enjoyable. Members of the support staff such as Ellen, Marianne, Babs, Pim, Marjet, Sofie, Aleida, Marjolein, and Karin also helped me with several practical matters, for which I am grateful.

I would also like to thank all the other colleagues at the Interuniversity Center for Social Science Theory and Methodology (ICS), too many to mention by name, for feedback and chats during the forum days and other events, both at Utrecht University and at Groningen, Nijmegen, and Amsterdam. Next to the ICS, I met many smart and kind colleagues as part of the SCOOP program. In particular, I would like to thank my fellow PhD students Aliona, Anne, Carlos, Chris, Damion, Esmee, Ewout, Francisca, Hendrik, Jonas, José Luis, Julian, Klara, Larisa, Leonie, Louisa, Maikel, Marlou, Maud, Piet, Sofie, and Thomas, and the training from Liesbet, Geetha, Daniela, Saskia, and Rafael. Thanks also go to members of the Cooperative Relations group for their feedback on my papers and presentations. Among them, I would like to acknowledge Rense, Jiamin, and Werner, who have not yet been mentioned.

I further want to express my gratitude to the ten young actors who participated in the role of fare dodgers for the field experiment in Chapter 6: Rumi Ünal, Kevin Christiaans, Ayad Murad, Rens Heilema, Alexander van Reeken, Bas Nillesen, Tarik Rahali, Oussama Dib, Silvan Zomerdijk, and Feyyaz Bademoglu. They bravely approached travelers who are about to go through check-in gates at Dutch train stations with the request to follow them without checking in themselves. This request was not always met with kindness by the travelers, let alone the NS police, but all actors persevered. Even more so, they made the data collection fun, and we always had enjoyable snack bar breaks. Some of my supervisors were quite hesitant about letting me do a field experiment given the considerable time investment it takes and the many ways it can go wrong. I now understand their hesitance; it really did take a lot of time and effort to pull it off, and I could not have done it without these actors who were all so motivated to make it work.

Many thanks also to the promotion committee for taking the time to read and evaluate my thesis. I would also like to thank my new colleagues at the WODC, and in particular members of the ASMI team, for giving me a warm welcome and allowing me to extend my research on newcomers to Ukrainian refugees, Syrian refugees, and migration in general. I moved from Utrecht University to the WODC with Vardan, and it has been great fun to work together with such a positive and kind person. I am sure I have forgotten to name several important colleagues who also contributed to my work. Those who know me well, are aware that I sometimes complain about a bad memory, so I hope they will understand and forgive me when I forget to mention them.

Before the number of mentioned people reaches Dunbar’s (empirically unsupported) number, I want to mention an important nonhuman contributor to this book – luck. I have been incredibly lucky to be born in an affluent country, to be able to spend so much time on my education, to have met people who encouraged me to go to university, enroll in a research master’s program, and start a PhD project. Several deciding factors that made this trajectory possible were largely beyond my control, and I fully realize that many people more capable than me have not had a fraction of the chances I had. Despite this large element of luck, finishing the PhD project is handsomely rewarded with a prestigious title, scientific publications, and a book to parade around. These rewards sometimes make me feel uneasy; there are so many jobs that contribute as much, if not more, to society and are rarely rewarded. A PhD project has its place in society and is hard work, but it is not more important or more work than a nurse caring for the sick, a garbage collector keeping our streets clean, a musician providing joy and comfort, and so on.

My luck extends to my social circle. I have been very fortunate to have the support from my friends, family, and family-in-law. They contributed to this book precisely by providing a distraction from it. The enjoyment and comfort that they bring reminds me there is much more to life than academia and recharges me, thereby giving me a clearer headspace for when I do work. I am especially grateful for the love of my partner Elli, mom Marja, dad Martin, brother Ivar, and sister Esmée.

About the author

Kasper Otten was born in Enschede, the Netherlands, on October 9, 1993. In 2016, he obtained his bachelor's degree in Sociology combined with the Von Humboldt honors program at Utrecht University. In 2018, he completed the research master Sociology and Social Research cum laude and was awarded the Best Student Prize. During his studies, he did an internship at the Centraal Bureau voor de Statistiek (CBS) and worked as a part-time data manager at Data Archiving and Networked Services (DANS). In 2018, he started working as a Ph.D. candidate at the Interuniversity Centre for Social Science Theory and Methodology (ICS), the Department of Sociology, and the research group Cooperative Relations at Utrecht University. He wrote his dissertation under the supervision of prof. dr. ir. Vincent Buskens (Department of Sociology, Utrecht University), prof. dr. Naomi Ellemers (Department of Social, Health and Organisational Psychology, Utrecht University), and dr. Wojtek Przepiorka (Department of Sociology, Utrecht University). His research revolves around cooperation, competition, norms, experiments, game theory, and newcomers. He currently works as a scientific researcher at the Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC).

Peer-reviewed publications

Otten, K., & te Riele, S. (2015). Latrelaties in Nederland. *Bevolkingstrends*, *14*, 1-25.

Otten, K. (2020). When upward social mobility leads to frustration: Boudon's game-theoretic model of relative deprivation and experimental evidence. *Research in Social Stratification and Mobility*, *65*.

Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2020). Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels. *Scientific reports*, *10*(1), 16702.

Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2021). Cooperation between newcomers and incumbents: the role of normative disagreements. *Journal of Economic Psychology*, *87*, 102448.

Otten, K. (2022). Human competition is not lower if competing is socially wasteful instead of socially beneficial. *Scientific Reports*, *12*(1), 10740.

Otten, K., Frey, U. J., Buskens, V., Przepiorka, W., & Ellemers, N. (2022). Human cooperation in changing groups in a large-scale public goods game. *Nature Communications*, *13*(1), 6399.

Otten, K. (2022). The logic of relative frustration revisited: theoretical revision of Boudon's competition model and experimental evidence. *European Sociological Review*.

Otten, K., & Mandemakers, J. J. (2023). Partners in health: Investigating social genetic effects among married and cohabiting couples. *Behavior Genetics*, *53*, 348–358.

Working papers

Otten, K., Buskens, V., Przepiorka, W., Cherki, B., & Israel, S. (2022). Cooperation, punishment, and group change in multilevel public goods experiments.

Otten, K., Buskens, V., Przepiorka, W., & Ellemers, N. (2023). Double standards in facilitating norm violations: A field experiment on attempted free-riding in public transport.

Otten, K. (2023). Ingroup cooperators are outgroup cooperators in international prisoner's dilemmas.

ICS dissertation series

The ICS series presents dissertations of the Interuniversity Center for Social Science Theory and Methodology. Each of these studies aims at integrating explicit theory formation with state-of-the-art empirical research or at the development of advanced methods for empirical research. The ICS was founded in 1986 as a cooperative effort of the universities of Groningen and Utrecht. Since 1992, the ICS has expanded to the University of Nijmegen and, since 2017, to the University of Amsterdam. Most of the projects are financed by the participating universities or by the Dutch Research Council. The international composition of the ICS graduate students is mirrored in the increasing international orientation of the projects and thus of the ICS series itself.

1. Kees van Liere (1990), *"Lastige leerlingen: Een empirisch onderzoek naar sociale oorzaken van probleemgedrag op basisscholen."* Amsterdam: Thesis Publishers
2. Marco van Leeuwen (1990), *"Bijstand in Amsterdam, ca. 1800-1850: Armenzorg als beheersings en overlevingsstrategie."* ICS dissertation, Utrecht
3. Ineke Maas (1990), *"Deelname aan podiumkunsten via de podia, de media en actieve beoefening: Substitutie of leereffecten?"* Amsterdam: Thesis Publishers
4. Marjolein Broese van Groenou (1991), *"Gescheiden netwerken: De relaties met vrienden en verwanten na echtscheiding."* Amsterdam: Thesis Publishers
5. Jan van den Bos (1991), *"Dutch EC policy making: A model guided approach to coordination and negotiation."* Amsterdam: Thesis Publishers
6. Karin Sanders (1991), *"Vrouwelijke pioniers: Vrouwen en mannen met een 'mannelijke' hogere beroepsopleiding aan het begin van hun loopbaan."* Amsterdam: Thesis Publishers
7. Sjerp de Vries (1991), *"Egoism, altruism, and social justice: Theory and experiments on cooperation in social dilemmas."* Amsterdam: Thesis Publishers
8. Ronald Batenburg (1991), *"Automatisering in bedrijf."* Amsterdam: Thesis Publishers
9. Rudi Wielers (1991), *"Selectie en allocatie op de arbeidsmarkt. Een uitwerking voor de informele en geïnstitutionaliseerde kinderopvang."* Amsterdam: Thesis Publishers
10. Gert Westert (1991), *"Verschillen in ziekenhuisgebruik."* ICS dissertation, Groningen
11. Hanneke Hermsen (1992), *"Votes and policy preferences: Equilibria in party systems."* Amsterdam: Thesis Publishers
12. Cora Maas (1992), *"Probleemleerlingen in het basisonderwijs."* Amsterdam: Thesis Publishers
13. Ed Boxman (1992), *"Contacten en carrière : Een empirisch theoretisch onderzoek naar de relatie tussen sociale netwerken en arbeidsmarktposities"* Amsterdam: Thesis Publishers
14. Conny Taes (1992), *"Kijken naar banen: Een onderzoek naar de inschatting van arbeidsmarktkansen bij schoolverlaters uit het middelbaar beroepsonderwijs."* Amsterdam: Thesis Publishers
15. Peter van Roozendaal (1992), *"Cabinets in multi party democracies: The effect of dominant and central parties on cabinet composition and durability."* Amsterdam: Thesis Publishers
16. Marcel van Dam (1992), *"Regio zonder regie: Verschillen in en effectiviteit van gemeentelijk arbeidsmarktbeleid."* Amsterdam: Thesis Publishers
17. Tanja van der Lippe (1993), *"Arbeidsverdeling tussen mannen en vrouwen."* Amsterdam: Thesis Publishers
18. Marc Jacobs (1993), *"Software: Kopen of kopiëren? Een sociaal wetenschappelijk onderzoek onder PC gebruikers."* Amsterdam: Thesis Publishers
19. Peter van der Meer (1993), *"Verdringing op de Nederlandse arbeidsmarkt: Sector- en sekseverschillen."* Amsterdam: Thesis Publishers
20. Gerbert Kraaykamp (1993), *"Over lezen gesproken: Een studie naar sociale differentiatie in leesgedrag."* Amsterdam: Thesis Publishers
21. Evelien Zeggelink (1993), *"Strangers into friends: The evolution of friendship networks using an individual oriented modeling approach."* Amsterdam: Thesis Publishers
22. Jaco Berveling (1994), *"Het stempel op de besluitvorming: Macht, invloed en besluitvorming op twee Amsterdamse beleidsterreinen."* Amsterdam: Thesis Publishers

23. Wim Bernasco (1994), *“Coupled careers: The effects of spouse's resources on success at work.”* Amsterdam: Thesis Publishers
24. Liset van Dijk (1994), *“Choices in child care: The distribution of child care among mothers, fathers and non parental care providers.”* Amsterdam: Thesis Publishers
25. Jos de Haan (1994), *“Research groups in Dutch sociology.”* Amsterdam: Thesis Publishers
26. Kwasi Boahene (1995), *“Innovation adoption as a socioeconomic process: The case of the Ghanaian cocoa industry.”* Amsterdam: Thesis Publishers
27. Paul Ligthart (1995), *“solidarity in economic transactions: An experimental study of framing effects in bargaining and contracting.”* Amsterdam: Thesis Publishers
28. Roger Leenders (1995), *“Structure and influence: Statistical models for the dynamics of actor attributes, network structure, and their interdependence.”* Amsterdam: Thesis Publishers
29. Beate Volker (1995), *“Should auld acquaintance be forgot...? Institutions of communism, the transition to capitalism and personal networks: The case of East Germany.”* Amsterdam: Thesis Publishers
30. Anneke Cancrinus-Matthijsse (1995), *“Tussen hulpverlening en ondernemerschap: Beroepsuitoefening en taakopvattingen van openbare apothekers in een aantal West-Europese landen.”* Amsterdam: Thesis Publishers
31. Nardi Steverink (1996), *“Zo lang mogelijk zelfstandig: Naar een verklaring van verschillen in oriëntatie ten aanzien van opname in een verzorgingstehuis onder fysiek kwetsbare ouderen.”* Amsterdam: Thesis Publishers
32. Ellen Lindeman (1996), *“Participatie in vrijwilligerswerk.”* Amsterdam: Thesis Publishers
33. Chris Snijders (1996), *“Trust and commitments.”* Amsterdam: Thesis Publishers
34. Koos Postma (1996), *“Changing prejudice in Hungary. A study on the collapse of state socialism and its impact on prejudice against Gypsies and Jews.”* Amsterdam: Thesis Publishers
35. Jooske van Busschbach (1996), *“Uit het oog, uit het hart? Stabiliteit en verandering in persoonlijke relaties.”* Amsterdam: Thesis Publishers
36. René Torenvlied (1996), *“Besluiten in uitvoering: Theorieën over beleidsuitvoering modelmatig getoetst op sociale vernieuwing in drie gemeenten.”* Amsterdam: Thesis Publishers
37. Andreas Flache (1996), *“The double edge of networks: An analysis of the effect of informal networks on cooperation in social dilemmas.”* Amsterdam: Thesis Publishers
38. Kees van Veen (1997), *“Inside an internal labor market: Formal rules, flexibility and career lines in a Dutch manufacturing company.”* Amsterdam: Thesis Publishers
39. Lucienne van Eijk (1997), *“Activity and wellbeing in the elderly.”* Amsterdam: Thesis Publishers
40. Róbert Gál (1997), *“Unreliability: Contract discipline and contract governance under economic transition.”* Amsterdam: Thesis Publishers
41. Anne-Geerte van de Goor (1997), *“Effects of regulation on disability duration.”* ICS dissertation, Utrecht
42. Boris Blumberg (1997), *“Das Management von Technologiekooperationen: Partnersuche und Verhandlungen mit dem Partner aus Empirisch Theoretischer Perspektive.”* ICS dissertation, Utrecht
43. Marijke von Bergh (1997), *“Loopbanen van oudere werknemers.”* Amsterdam: Thesis Publishers
44. Anna Petra Nieboer (1997), *“Life events and well-being: A prospective study on changes in well-being of elderly people due to a serious illness event or death of the spouse.”* Amsterdam: Thesis Publishers
45. Jacques Niehof (1997), *“Resources and social reproduction: The effects of cultural and material resources on educational and occupational careers in industrial nations at the end of the twentieth century.”* ICS dissertation, Nijmegen
46. Ariana Need (1997), *“The kindred vote: Individual and family effects of social class and religion on electoral change in the Netherlands, 1956 1994.”* ICS dissertation, Nijmegen
47. Jim Allen (1997), *“Sector composition and the effect of education on Wages: An international comparison.”* Amsterdam: Thesis Publishers
48. Jack Hutten (1998), *“Workload and provision of care in general practice: An empirical study of the relation between workload of Dutch general practitioners and the content and quality of their care.”* ICS dissertation, Utrecht
49. Per Kropp (1998), *“Berufserfolg im Transformationsprozeß: Eine theoretisch empirische Studie über die Gewinner und Verlierer der Wende in Ostdeutschland.”* ICS dissertation, Utrecht

50. Maarten Wolbers (1998), *"Diploma-inflatie en verdringing op de arbeidsmarkt: Een studie naar ontwikkelingen in de opbrengsten van diploma's in Nederland."* ICS dissertation, Nijmegen
51. Wilma Smeenk (1998), *"Opportunity and marriage: The impact of individual resources and marriage market structure on first marriage timing and partner choice in the Netherlands."* ICS dissertation, Nijmegen
52. Marinus Spreen (1999), *"Sampling personal network structures: Statistical inference in ego graphs."* ICS dissertation, Groningen
53. Vincent Buskens (1999), *"Social networks and trust."* ICS dissertation, Utrecht
54. Susanne Rijken (1999), *"Educational expansion and status attainment: A cross-national and over-time comparison."* ICS dissertation, Utrecht
55. Mérove Gijsberts (1999), *"The legitimation of inequality in state-socialist and market societies, 1987-1996."* ICS dissertation, Utrecht
56. Gerhard van de Bunt (1999), *"Friends by choice: An actor-oriented statistical network model for friendship networks through time."* ICS dissertation, Groningen
57. Robert Thomson (1999), *"The party mandate: Election pledges and government actions in the Netherlands, 1986 1998."* Amsterdam: Thela Thesis
58. Corine Baarda (1999), *"Politieke besluiten en boeren beslissingen: Het draagvlak van het mestbeleid tot 2000."* ICS dissertation, Groningen
59. Rafael Wittek (1999), *"Interdependence and informal control in organizations."* ICS dissertation, Groningen
60. Diane Payne (1999), *"Policy Making in the European Union: An analysis of the impact of the reform of the structural funds in Ireland."* ICS dissertation, Groningen
61. René Veenstra (1999), *"Leerlingen, Klassen, Scholen: Prestaties en vorderingen van leerlingen in het voortgezet onderwijs."* Amsterdam, Thela Thesis
62. Marjolein Achterkamp (1999), *"Influence strategies in collective decision making: A comparison of two models."* ICS dissertation, Groningen
63. Peter Mühlau (2000), *"The governance of the employment relation: Arelational signaling perspective."* ICS dissertation, Groningen
64. Agnes Akkerman (2000), *"Verdeelde vakbeweging en stakingen: Concurrentie om leden."* ICS dissertation, Groningen
65. Sandra van Thiel (2000), *"Quangocratization: Trends, causes and consequences."* ICS dissertation, Utrecht
66. Sylvia Peacock-Korupp (2000), *"Mothers and the process of social stratification."* ICS dissertation, Utrecht
67. Rudi Turksema (2000), *"Supply of day care."* ICS dissertation, Utrecht
68. Bernard Nijstad (2000), *"How the group affects the mind: Effects of communication in idea generating groups."* ICS dissertation, Utrecht
69. Inge de Wolf (2000), *"Opleidingsspecialisatie en arbeidsmarktsucces van sociale wetenschappers."* ICS dissertation, Utrecht
70. Jan Kratzer (2001), *"Communication and performance: An empirical study in innovation teams."* ICS dissertation, Groningen
71. Madelon Kroneman (2001), *"Healthcare systems and hospital bed use."* ICS/NIVEL-dissertation, Utrecht
72. Herman van de Werfhorst (2001), *"Field of study and social inequality: Four types of educational resources in the process of stratification in the Netherlands."* ICS dissertation, Nijmegen
73. Tamás Bartus (2001), *"Social capital and earnings inequalities: The role of informal job search in Hungary."* ICS dissertation Groningen
74. Hester Moerbeek (2001), *"Friends and foes in the occupational career: The influence of sweet and sour social capital on the labour market."* ICS dissertation, Nijmegen
75. Marcel van Assen (2001), *"Essays on actor perspectives in exchange networks and social dilemmas."* ICS dissertation, Groningen
76. Inge Sieben (2001), *"Sibling similarities and social stratification: The impact of family background across countries and cohorts."* ICS dissertation, Nijmegen
77. Alinda van Bruggen (2001), *"Individual production of social well-being: An exploratory study."* ICS dissertation, Groningen

78. Marcel Coenders (2001), *"Nationalistic attitudes and ethnic exclusionism in a comparative perspective: An empirical study of attitudes toward the country and ethnic immigrants in 22 countries."* ICS dissertation, Nijmegen
79. Marcel Lubbers (2001), *"Exclusionistic electorates: Extreme right-wing voting in Western Europe."* ICS dissertation, Nijmegen
80. Uwe Matzat (2001), *"Social networks and cooperation in electronic communities: A theoretical-empirical analysis of academic communication and internet discussion groups."* ICS dissertation, Groningen
81. Jacques Janssen (2002), *"Do opposites attract divorce? Dimensions of mixed marriage and the risk of divorce in the Netherlands."* ICS dissertation, Nijmegen
82. Miranda Jansen (2002), *"Waardenoriëntaties en partnerrelaties: Een panelstudie naar wederzijdse invloeden."* ICS dissertation, Utrecht
83. Anne-Rigt Poortman (2002), *"Socioeconomic causes and consequences of divorce."* ICS dissertation, Utrecht
84. Alexander Gattig (2002), *"Intertemporal decision making."* ICS dissertation, Groningen
85. Gerrit Rooks (2000), *"Contract en conflict: Strategisch management van inkooptransacties."* ICS dissertation, Utrecht
86. Károly Takács (2002), *"Social networks and intergroup conflict."* ICS dissertation, Groningen
87. Thomas Gautschi (2002), *"Trust and exchange, effects of temporal embeddedness and network embeddedness on providing and dividing a surplus."* ICS dissertation, Utrecht
88. Hilde Bras (2002), *"Zeeuwse meiden: Dienen in de levensloop van vrouwen, ca. 1850 – 1950."* Aksant Academic Publishers, Amsterdam
89. Merijn Rengers (2002), *"Economic lives of artists: Studies into careers and the labour market in the cultural sector."* ICS dissertation, Utrecht
90. Annelies Kassenberg (2002), *"Wat scholieren bindt: Sociale gemeenschap in scholen."* ICS dissertation, Groningen
91. Marc Verboord (2003), *"Moet de meester dalen of de leerling klimmen? De invloed van literatuuronderwijs en ouders op het lezen van boeken tussen 1975 en 2000."* ICS dissertation, Utrecht
92. Marcel van Egmond (2003), *"Rain falls on all of us (but some manage to get more wet than others): Political context and electoral participation."* ICS dissertation, Nijmegen
93. Justine Horgan (2003), *"High-performance human resource management in Ireland and the Netherlands: Adoption and effectiveness."* ICS dissertation, Groningen
94. Corine Hoeben (2003), *"LETS' be a community: Community in local exchange trading systems."* ICS dissertation, Groningen
95. Christian Steglich (2003), *"The framing of decision situations: Automatic goal selection and rational goal pursuit."* ICS dissertation, Groningen
96. Johan van Wilsem (2003), *"Crime and context: The impact of individual, neighborhood, city and country characteristics on victimization."* ICS dissertation, Nijmegen
97. Christiaan Monden (2003), *"Education, inequality and health: The impact of partners and life course."* ICS dissertation, Nijmegen
98. Evelyn Hello (2003), *"Educational attainment and ethnic attitudes: How to explain their relationship."* ICS dissertation, Nijmegen
99. Marnix Croes en Peter Tammes (2004). *Gif laten wij niet voortbestaan: Een onderzoek naar de overlevingskansen van Joden in de Nederlandse gemeenten, 1940-1945.* Aksant Academic Publishers, Amsterdam.
100. Ineke Nagel (2004), *"Cultuurdeelname in de levensloop."* ICS dissertation, Utrecht
101. Marieke van der Wal (2004), *"Competencies to participate in life: Measurement and the impact of school."* ICS dissertation, Groningen
102. Vivian Meertens (2004), *"Depressive symptoms in the general population: A multifactorial social approach."* ICS dissertation, Nijmegen
103. Hanneke Schuurmans (2004), *"Promoting well-being in frail elderly people: Theory and intervention."* ICS dissertation, Groningen

104. Javier Arregui Moreno (2004), *"Negotiation in legislative decision-making in the European Union."* ICS dissertation, Groningen
105. Tamar Fischer (2004), *"Parental divorce, conflict and resources: The effects on children's behaviour problems, socioeconomic attainment, and transitions in the demographic career."* ICS dissertation, Nijmegen
106. René Bekkers (2004), *"Giving and volunteering in the Netherlands: Sociological and psychological perspectives."* ICS dissertation, Utrecht
107. Renée van der Hulst (2004), *"Gender differences in workplace authority: An empirical study on social networks."* ICS dissertation, Groningen
108. Rita Smaniotto (2004), *"'You scratch my back and I scratch yours' versus 'love thy neighbour': Two proximate mechanisms of reciprocal altruism."* ICS dissertation, Groningen
109. Maurice Gesthuizen (2004), *"The life course of the low-educated in the Netherlands: Social and economic risks."* ICS dissertation, Nijmegen
110. Carlijne Philips (2005), *"Vakantiegemeenschappen: Kwalitatief en kwantitatief onderzoek naar gelegenheid en refreshergemeenschap tijdens de vakantie."* ICS dissertation, Groningen
111. Esther de Ruijter (2005), *"Household outsourcing."* ICS dissertation, Utrecht
112. Frank van Tubergen (2005), *"The integration of immigrants in cross-national perspective: Origin, destination, and community effects."* ICS dissertation, Utrecht
113. Ferry Koster (2005), *"For the time being: Accounting for inconclusive findings concerning the effects of temporary employment relationships on solidary behavior of employees."* ICS dissertation, Groningen
114. Carolien Klein Haarhuis (2005), *"Promoting anti-corruption reforms: Evaluating the implementation of a World Bank anti-corruption program in seven African countries (1999-2001)."* ICS dissertation, Utrecht
115. Martin van der Gaag (2005), *"Measurement of individual social capital."* ICS dissertation, Groningen
116. Johan Hansen (2005), *"Shaping careers of men and women in organizational contexts."* ICS dissertation, Utrecht
117. Davide Barrera (2005), *"Trust in embedded settings."* ICS dissertation, Utrecht
118. Mattijs Lambooij (2005), *"Promoting cooperation: Studies into the effects of long-term and short-term rewards on cooperation of employees."* ICS dissertation, Utrecht
119. Lotte Vermeij (2006), *"What's cooking? Cultural boundaries among Dutch teenagers of different ethnic origins in the context of school."* ICS dissertation, Utrecht
120. Mathilde Strating (2006), *"Facing the challenge of rheumatoid arthritis: A 13-year prospective study among patients and cross-sectional study among their partners."* ICS dissertation, Groningen
121. Jannes de Vries (2006), *"Measurement error in family background variables: The bias in the intergenerational transmission of status, cultural consumption, party preference, and religiosity."* ICS dissertation, Nijmegen
122. Stefan Thau (2006), *"Workplace deviance: Four studies on employee motives and self-regulation."* ICS dissertation, Groningen
123. Mirjam Plantinga (2006), *"Employee motivation and employee performance in child care: The effects of the introduction of market forces on employees in the Dutch child-care sector."* ICS dissertation, Groningen
124. Helga de Valk (2006), *"Pathways into adulthood: A comparative study on family life transitions among migrant and Dutch youth."* ICS dissertation, Utrecht
125. Henrike Elzen (2006), *"Self-Management for chronically ill older people."* ICS dissertation, Groningen
126. Ayse Güveli (2007), *"New social classes within the service class in the Netherlands and Britain: Adjusting the EGP class schema for the technocrats and the social and cultural specialists."* ICS dissertation, Nijmegen
127. Willem-Jan Verhoeven (2007), *"Income attainment in post-communist societies."* ICS dissertation, Utrecht
128. Marieke Voorpostel (2007), *"Sibling support: The exchange of help among brothers and sisters in the Netherlands."* ICS dissertation, Utrecht
129. Jacob Dijkstra (2007), *"The effects of externalities on partner choice and payoffs in exchange networks."* ICS dissertation, Groningen
130. Patricia van Echtelt (2007), *"Time-greedy employment relationships: Four studies on the time claims of post-Fordist work."* ICS dissertation, Groningen

131. Sonja Vogt (2007), *"Heterogeneity in social dilemmas: The case of social support."* ICS dissertation, Utrecht
132. Michael Schweinberger (2007), *"Statistical methods for studying the evolution of networks and behavior."* ICS dissertation, Groningen
133. István Henrik Back (2007), *"Commitment and evolution: Connecting emotion and reason in long-term relationships."* ICS dissertation, Groningen
134. Ruben van Gaalen (2007), *"Solidarity and ambivalence in parent-child relationships."* ICS dissertation, Utrecht
135. Jan Reitsma (2007), *"Religiosity and solidarity: Dsimensions and relationships disentangled and tested."* ICS dissertation, Nijmegen
136. Jan Kornelis Dijkstra (2007), *"Status and affection among (pre)adolescents and their relation with antisocial and prosocial behavior."* ICS dissertation, Groningen
137. Wouter van Gils (2007), *"Full-time working couples in the Netherlands: Causes and consequences."* ICS dissertation, Nijmegen
138. Djamila Schans (2007), *"Ethnic diversity in intergenerational solidarity."* ICS dissertation, Utrecht
139. Ruud van der Meulen (2007), *"Brug over woelig water: Lidmaatschap van sportverenigingen, vriendschappen, kennissenkringen en veralgemeend vertrouwen."* ICS dissertation, Nijmegen
140. Andrea Knecht (2008), *"Friendship selection and friends' influence: Dynamics of networks and actor attributes in early adolescence."* ICS dissertation, Utrecht
141. Ingrid Doorten (2008), *"The division of unpaid work in the household: A stubborn pattern?"* ICS dissertation, Utrecht
142. Stijn Ruiter (2008), *"Association in context and association as context: Causes and consequences of voluntary association involvement."* ICS dissertation, Nijmegen
143. Janneke Joly (2008), *"People on our minds: When humanized contexts activate social norms."* ICS dissertation, Groningen
144. Margreet Frieling (2008), *"'Joint production' als motor voor actief burgerschap in de buurt."* ICS dissertation, Groningen
145. Ellen Verbakel (2008), *"The partner as resource or restriction? Labour market careers of husbands and wives and the consequences for inequality between couples."* ICS dissertation, Nijmegen
146. Gijs van Houten (2008), *"Beleidsuitvoering in gelaagde stelsels: De doorwerking van aanbevelingen van de Stichting van de Arbeid in het CAO-overleg."* ICS dissertation, Utrecht
147. Eva Jaspers (2008), *"Intolerance over time: Macro- and micro-level questions on attitudes towards euthanasia, homosexuality and ethnic minorities."* ICS dissertation, Nijmegen
148. Gijs Weijters (2008), *"Youth delinquency in Dutch cities and schools: A multilevel approach."* ICS dissertation, Nijmegen
149. Jessica Nooij (2009), *"The self in social rejection."* ICS dissertation, Groningen
150. Gerald Mollenhorst (2009), *"Networks in contexts: How meeting opportunities affect personal relationships."* ICS dissertation, Utrecht
151. Tom van der Meer (2009), *"States of freely associating citizens: Comparative studies into the impact of state institutions on social, civic and political participation."* ICS dissertation, Nijmegen
152. Manuela Vieth (2009), *"Commitments and reciprocity in trust situations: Experimental studies on obligation, indignation, and self-consistency."* ICS dissertation, Utrecht
153. Rense Corten (2009), *"Co-evolution of social networks and behavior in social dilemmas: Theoretical and empirical perspectives."* ICS dissertation, Utrecht
154. Arieke Rijken (2009), *"Happy families, high fertility? Childbearing choices in the context of family and partner relationships."* ICS dissertation, Utrecht
155. Jochem Tolsma (2009), *"Ethnic hostility among ethnic majority and minority groups in the Netherlands: An investigation into the impact of social mobility experiences, the local living environment and educational attainment on ethnic hostility."* ICS dissertation, Nijmegen
156. Freek Bucx (2009), *"Linked lives: Young adults' life course and relations with parents."* ICS dissertation, Utrecht

157. Philip Wotschack (2009), *"Household governance and time allocation: Four studies on the combination of work and care."* ICS dissertation, Groningen
158. Nienke Moor (2009), *"Explaining worldwide religious diversity: The relationship between subsistence technologies and ideas about the unknown in pre-industrial and (post-)industrial societies."* ICS dissertation, Nijmegen
159. Lieke ten Brummelhuis (2009), *"Family matters at work: Depleting and enriching effects of employees' family lives on work outcomes."* ICS dissertation, Utrecht
160. Renske Keizer (2010), *"Remaining childless: Causes and consequences from a life-course perspective."* ICS dissertation, Utrecht
161. Miranda Sentse (2010), *"Bridging contexts: The interplay between family, child, and peers in explaining problem behavior in early adolescence."* ICS dissertation, Groningen
162. Nicole Tieben (2010), *"Transitions, tracks and transformations: Social inequality in transitions into, through and out of secondary education in the Netherlands for cohorts born between 1914 and 1985."* ICS dissertation, Nijmegen
163. Birgit Pauksztat (2010), *"Speaking up in organizations: Four studies on employee voice."* ICS dissertation, Groningen
164. Richard Zijdeman (2010), *"Status attainment in the Netherlands, 1811-1941: Spatial and temporal variation before and during industrialization."* ICS dissertation, Utrecht
165. Rianne Kloosterman (2010), *"Social background and children's educational careers: The primary and secondary effects of social background over transitions and over time in the Netherlands."* ICS dissertation, Nijmegen
166. Olav Aarts (2010), *"Religious diversity and religious involvement: A study of religious markets in Western societies at the end of the twentieth century."* ICS dissertation, Nijmegen
167. Stephanie Wiesmann (2010), *"24/7 Negotiation in couples' transition to parenthood."* ICS dissertation, Utrecht
168. Borja Martinovic (2010), *"Interethnic contacts: A dynamic analysis of interaction between immigrants and natives in Western countries."* ICS dissertation, Utrecht
169. Anne Roeters (2010), *"Family life under pressure? Parents' paid work and the quantity and quality of parent-child and family time."* ICS dissertation, Utrecht
170. Jelle Sijtsema (2010), *"Adolescent aggressive behavior: Status and stimulation goals in relation to the peer context."* ICS dissertation, Groningen
171. Kees Keizer (2010), *"The spreading of disorder."* ICS dissertation, Groningen
172. Michael Mäs (2010), *"The diversity puzzle. explaining clustering and polarization of opinions."* ICS dissertation, Groningen
173. Marie-Louise Damen (2010), *"Cultuurdeelname en CKV: Studies naar effecten van kunsteducatie op de cultuurdeelname van leerlingen tijdens en na het voortgezet onderwijs."* ICS dissertation, Utrecht
174. Marieke van de Rakt (2011), *"Two generations of crime: The intergenerational transmission of convictions over the life course."* ICS dissertation, Nijmegen
175. Willem Huijnk (2011), *"Family life and ethnic attitudes: The role of the family for attitudes towards intermarriage and acculturation among minority and majority groups."* ICS dissertation, Utrecht
176. Tim Huijts (2011), *"Social ties and health in Europe: Individual associations, cross-national variations, and contextual explanations."* ICS dissertation, Nijmegen
177. Wouter Steenbeek (2011), *"Social and physical disorder: How community, business presence and entrepreneurs influence disorder in Dutch neighborhoods."* ICS dissertation, Utrecht
178. Miranda Vervoort (2011), *"Living together apart? Ethnic concentration in the neighborhood and ethnic minorities' social contacts and language practices."* ICS dissertation, Utrecht
179. Agnieszka Kanas (2011), *"The economic performance of immigrants: The role of human and social capital."* ICS dissertation, Utrecht
180. Lea Ellwardt (2011), *"Gossip in organizations: A social network study."* ICS dissertation, Groningen
181. Annemarije Oosterwaal (2011), *"The gap between decision and implementation: Decision making, delegation and compliance in governmental and organizational settings."* ICS dissertation, Utrecht

182. Natascha Notten (2011), *"Parents and the media: Causes and consequences of parental media socialization."* ICS dissertation, Nijmegen
183. Tobias Stark (2011), *"Integration in schools: A process perspective on students' interethnic attitudes and interpersonal relationships."* ICS dissertation, Groningen
184. Giedo Jansen (2011), *"Social cleavages and political choices: Large-scale comparisons of social class, religion and voting behavior in Western democracies."* ICS dissertation, Nijmegen
185. Ruud van der Horst (2011), *"Network effects on treatment results in a closed forensic psychiatric setting."* ICS dissertation, Groningen
186. Mark Levels (2011), *"Abortion laws in European Countries between 1960 and 2010: Legislative developments and their consequences for women's reproductive decision making."* ICS dissertation, Nijmegen
187. Marieke van Londen (2012), *"Exclusion of ethnic minorities in the Netherlands: The effects of individual and situational characteristics on opposition to ethnic policy and ethnically mixed neighbourhoods."* ICS dissertation, Nijmegen
188. Sigrid Mohnen (2012), *"Neighborhood context and health: How neighborhood social capital affects individual health."* ICS dissertation, Utrecht
189. Asya Zhelyazkova (2012), *"Compliance under controversy: Analysis of the transposition of European directives and their provisions."* ICS dissertation, Utrecht
190. Valeska Korff (2012), *"Between cause and control: Management in a humanitarian organization."* ICS dissertation, Groningen
191. Maike Gieling (2012), *"Dealing with diversity: Adolescents' support for civil liberties and immigrant rights."* ICS dissertation, Utrecht
192. Katya Ivanova (2012), *"From parents to partners: The impact of family on romantic relationships in adolescence and emerging adulthood."* ICS dissertation, Groningen
193. Jelmer Schalk (2012), *"The performance of public corporate actors: Essays on effects of institutional and network embeddedness in supranational, national, and local collaborative contexts."* ICS dissertation, Utrecht
194. Alona Labun (2012), *"Social networks and informal power in organizations."* ICS dissertation, Groningen
195. Michal Bojanowski (2012), *"Essays on social network formation in heterogeneous populations: Models, methods, and empirical analyses."* ICS dissertation, Utrecht
196. Anca Minescu (2012), *"Relative group position and intergroup attitudes in Russia."* ICS dissertation, Utrecht
197. Marieke van Schellen (2012), *"Marriage and crime over the life course: The criminal careers of convicts and their spouses."* ICS dissertation, Utrecht
198. Mieke Maliepaard (2012), *"Religious trends and social integration: Muslim minorities in the Netherlands."* ICS dissertation, Utrecht
199. Fransje Smits (2012), *"Turks and Moroccans in the Low Countries around the year 2000: Determinants of religiosity, trend in religiosity and determinants of the trend."* ICS dissertation, Nijmegen
200. Roderick Sluiter (2012), *"The diffusion of morality policies among Western European countries between 1960 and 2010: A comparison of temporal and spatial diffusion patterns of six morality and eleven non-morality policies."* ICS dissertation, Nijmegen
201. Nicoletta Balbo (2012), *"Family, friends and fertility."* ICS dissertation, Groningen
202. Anke Munniksma (2013), *"Crossing ethnic boundaries: Parental resistance to and consequences of adolescents' cross-ethnic peer relations"* ICS dissertation, Groningen
203. Anja-Kristin Abendroth (2013), *"Working women in Europe: How the country, workplace, and family context matter."* ICS dissertation, Utrecht
204. Katia Begall (2013), *"Occupational hazard? The relationship between working conditions and fertility."* ICS dissertation, Groningen
205. Hidde Bekhuis (2013), *"The popularity of domestic cultural products: Cross-national differences and the relation to globalization."* ICS dissertation, Utrecht

206. Lieselotte Blommaert (2013), *"Are Joris and Renske more employable than Rashid and Samira? A study on the prevalence and sources of ethnic discrimination in recruitment in the Netherlands using experimental and survey data."* ICS dissertation, Utrecht
207. Wiebke Schulz (2013), *"Careers of men and women in the 19th and 20th centuries."* ICS dissertation, Utrecht
208. Ozan Aksoy (2013), *"Essays on social preferences and beliefs in non-embedded social dilemmas."* ICS dissertation, Utrecht
209. Dominik Morbitzer (2013), *"Limited farsightedness in network formation."* ICS dissertation, Utrecht
210. Thomas de Vroome (2013), *"Earning your place: The relation between immigrants' economic and psychological integration in the Netherlands."* ICS dissertation, Utrecht
211. Marloes de Lange (2013), *"Causes and consequences of employment flexibility among young people: Recent developments in the Netherlands and Europe."* ICS dissertation, Nijmegen
212. Roza Meuleman (2014), *"Consuming the Nation: Domestic cultural consumption: Its stratification and relation with nationalist attitudes."* ICS dissertation, Utrecht
213. Esther Havekes (2014), *"Putting interethnic attitudes in context: The relationship between neighbourhood characteristics, interethnic attitudes and residential behaviour."* ICS dissertation, Utrecht
214. Zoltán Lippényi (2014), *"Transitions toward an open society? Intergenerational occupational mobility in Hungary in the 19th and 20th centuries."* ICS dissertation, Utrecht
215. Anouk Smeekes (2014), *"The presence of the past: Historical rooting of national identity and current group dynamics."* ICS dissertation, Utrecht
216. Michael Savelkoul (2014), *"Ethnic diversity and social capital: Testing underlying explanations derived from conflict and contact theories in Europe and the United States."* ICS dissertation, Nijmegen
217. Martijn Hogerbrugge (2014), *"Misfortune and family: How negative events, family ties, and lives are linked."* ICS dissertation, Utrecht
218. Gina-Felicia Potarca (2014), *"Modern love: Comparative insights in online dating preferences and assortative mating."* ICS dissertation, Groningen
219. Mariska van der Horst (2014), *"Gender, aspirations, and achievements: Relating work and family aspirations to occupational outcomes."* ICS dissertation, Utrecht
220. Gijs Huitsing (2014), *"A social network perspective on bullying"* ICS dissertation, Groningen
221. Thomas Kowalewski (2015), *"Personal growth in organizational contexts."* ICS dissertation, Groningen
222. Manuel Muñoz-Herrera (2015), *"The impact of individual differences on network relations: Social exclusion and inequality in productive exchange and coordination games."* ICS dissertation, Groningen
223. Tim Immerzeel (2015), *"Voting for a change: The democratic lure of populist radical right parties in voting behavior."* ICS dissertation, Utrecht
224. Fernando Nieto Morales (2015), *"The control imperative: Studies on reorganization in the public and private sectors."* ICS dissertation, Groningen
225. Jellie Sierksma (2015), *"Bounded helping: How morality and intergroup relations shape children's reasoning about helping."* ICS dissertation, Utrecht
226. Tinka Veldhuis (2015), *"Captivated by fear: An evaluation of terrorism detention policy."* ICS dissertation, Groningen
227. Miranda Visser (2015), *"Loyality in humanity: Turnover among expatriate humanitarian aid workers."* ICS dissertation, Groningen
228. Sarah Westphal (2015), *"Are the kids alright? Essays on postdivorce residence arrangements and children's well-being."* ICS dissertation, Utrecht
229. Britta Rüschoff (2015), *"Peers in careers: Peer relationships in the transition from school to work."* ICS dissertation, Groningen
230. Nynke van Miltenburg (2015), *"Cooperation under peer sanctioning institutions: Collective decisions, noise, and endogenous implementation."* ICS dissertation, Utrecht
231. Antonie Knigge (2015), *"Sources of sibling similarity: Status attainment in the Netherlands during modernization."* ICS dissertation, Utrecht
232. Sanne Smith (2015), *"Ethnic segregation in friendship networks: Studies of its determinants in English, German, Dutch, and Swedish school classes."* ICS dissertation, Utrecht

233. Patrick Präg (2015), *"Social stratification and health: Four essays on social determinants of health and wellbeing."* ICS dissertation, Groningen
234. Wike Been (2015), *"European top managers' support for work-life arrangements"* ICS dissertation, Utrecht
235. André Grow (2016), *"Status differentiation: New insights from agent-based modeling and social network analysis."* ICS dissertation, Groningen
236. Jesper Rözer (2016), *"Family and personal networks: How a partner and children affect social relationships."* ICS dissertation, Utrecht
237. Kim Pattiselanno (2016), *"At your own risk: The importance of group dynamics and peer processes in adolescent peer groups for adolescents' involvement in risk behaviors."* ICS- dissertation, Groningen
238. Vincenz Frey (2016), *"Network formation and trust."* ICS dissertation, Utrecht
239. Rozemarijn van der Ploeg (2016), *"Be a buddy, not a bully? Four studies on social and emotional processes related to bullying, defending, and victimization."* ICS dissertation, Groningen
240. Tali Spiegel (2016), *"Identity, career trajectories and wellbeing: A closer look at individuals with degenerative eye conditions."* ICS- dissertation, Groningen
241. Felix Christian Tropf (2016), *"Social science genetics and fertility."* ICS dissertation, Groningen
242. Sara Geven (2016), *"Adolescent problem behavior in school: The role of peer networks."* ICS dissertation, Utrecht
243. Josja Rokven (2016), *"The victimization-offending relationship from a longitudinal perspective."* ICS dissertation, Nijmegen
244. Maja Djundeva (2016), *"Healthy ageing in context: Family welfare state and the life course."* ICS dissertation, Groningen
245. Mark Visser (2017), *"Inequality between older workers and older couples in the Netherlands: A dynamic life course perspective on educational and social class differences in the late career."* ICS dissertation, Nijmegen
246. Beau Oldenburg (2017), *"Bullying in schools: The role of teachers and classmates."* ICS dissertation, Groningen
247. Tatang Muttaqin (2017), *"The education divide in Indonesia: Four essays on determinants of unequal access to and quality of education."* ICS dissertation, Groningen
248. Margriet van Hek (2017), *"Gender inequality in educational attainment and reading performance: A contextual approach."* ICS dissertation, Nijmegen
249. Melissa Verhoef-van Dorp (2017), *"Work schedules, childcare and well-being: Essays on the associations between modern-day job characteristics, childcare arrangements and the well-being of parents and children."* ICS dissertation, Utrecht
250. Timo Septer (2017), *"Goal priorities, cognition and conflict: Analyses of cognitive maps concerning organizational change."* ICS dissertation, Groningen
251. Bas Hofstra (2017), *"Online social networks: Essays on membership, privacy, and structure."* ICS dissertation, Utrecht
252. Yassine Khoudja (2018), *"Women's labor market participation across ethnic groups: The role of household conditions, gender role attitudes, and religiosity in different national contexts."* ICS dissertation, Utrecht
253. Joran Laméris (2018), *"Living together in diversity: Whether, why and where ethnic diversity affects social cohesion."* ICS dissertation, Nijmegen
254. Maaike van der Vleuten (2018), *"Gendered Choices: Fields of study of adolescents in the Netherlands."* ICS dissertation, Utrecht
255. Mala Sondang Silitonga (2018), *"Corruption in Indonesia: The impact of institutional change, norms, and networks."* ICS dissertation, Groningen
256. Manja Coopmans (2018), *"Rituals of the past in the context of the present: The role of Remembrance Day and Liberation Day in Dutch society."* ICS dissertation, Utrecht
257. Paul Hindriks (2018), *"The struggle for power: Attitudes towards the political participation of ethnic minorities."* ICS dissertation, Utrecht
258. Nynke Niezink (2018), *"Modeling the dynamics of networks and continuous behavior."* ICS dissertation, Groningen

259. Simon de Bruijn (2018), *"Reaching agreement after divorce and separation: Essays on the effectiveness of parenting plans and divorce mediation."* ICS dissertation, Utrecht
260. Susanne van 't Hoff-de Goede (2018), *"While you were locked up: An empirical study on the characteristics, social surroundings and wellbeing of partners of prisoners in The Netherlands."* ICS dissertation, Utrecht
261. Loes van Rijsewijk (2018), *"Antecedents and consequences of helping among adolescents."* ICS dissertation, Groningen
262. Mariola Gremmen (2018), *"Social network processes and academic functioning: The role of peers in students' school well-being, academic engagement, and academic achievement."* ICS dissertation, Groningen
263. Jeanette Renema (2018), *"Immigrants' support for welfare spending: The causes and consequences of welfare usage and welfare knowledgeability."* ICS dissertation, Nijmegen
264. Suwatin Miharti (2018), *"Community health centers in Indonesia in the era of decentralization: The impact of structure, staff composition and management on health outcomes."* ICS dissertation, Groningen
265. Chaïm la Roi (2019), *"Stigma and stress: Studies on attitudes towards sexual minority orientations and the association between sexual orientation and mental health."* ICS dissertation, Groningen
266. Jelle Lössbroek (2019), *"Turning grey into gold: Employer-employee interplay in an ageing workforce."* ICS dissertation, Utrecht
267. Nikki van Gerwen (2019), *"Employee cooperation through training: A multi-method approach."* ICS dissertation, Utrecht
268. Paula Thijs (2019), *"Trends in cultural conservatism: The role of educational expansion, secularisation, and changing national contexts."* ICS dissertation, Nijmegen
269. Renske Verweij (2019), *"Understanding childlessness: Unravelling the link with genes and the socio-environment."* ICS dissertation, Groningen
270. Niels Blom (2019), *"Partner relationship quality under pressing work conditions: Longitudinal and cross-national investigation."* ICS dissertation, Nijmegen
271. Müge Simsek (2019), *"The dynamics of religion among native and immigrant youth in Western Europe."* ICS dissertation, Utrecht
272. Leonie van Breeschoten (2019), *"Combining a career and childcare: The use and usefulness of work-family policies in European organizations."* ICS dissertation, Utrecht
273. Roos van der Zwan (2019), "*The political representation of ethnic minorities and their vote choice."* ICS dissertation, Nijmegen
274. Ashwin Rambaran (2019), "*The classroom as context for bullying: A social network approach*." ICS dissertation, Groningen
275. Dieko Bakker (2019), *"Cooperation and social control: Effects of preferences, institutions, and social structure."* ICS dissertation, Groningen
276. Femke van der Werf (2019), "*Shadow of a rainbow? National and ethnic belonging in Mauritius*." ICS dissertation, Utrecht
277. Robert Krause (2019), *"Multiple imputation for missing network data."* ICS dissertation, Groningen
278. Take Sipma (2020), *"Economic insecurity and populist radical right voting."* ICS dissertation, Nijmegen
279. Mathijs Kros (2020), *"The nature of negative contact: Studies on interethnic relations in Western societies."* ICS dissertation, Utrecht
280. Lonneke van den Berg (2020), *"Time to leave: Individual and contextual explanations for the timing of leaving home."* ICS dissertation, Amsterdam
281. Marianne Hooijsma (2020), "*Clashrooms: Interethnic peer relationships in schools."* ICS dissertation, Groningen
282. Marina Tulin (2020*), "Blind spots in social resource theory: Essays on the creation, maintenance and returns of social capital."* ICS dissertation, Amsterdam
283. Tessa Kaufman (2020), "*Toward tailored interventions: Explaining, assessing, and preventing persistent victimization of bullying."* ICS dissertation, Groningen
284. Lex Thijssen (2020), *"Racial and ethnic discrimination in western labor markets: Empirical evidence from field experiments."* ICS dissertation, Utrecht

285. Lukas Norbutas (2020), *"Trust on the dark web: An analysis of illegal online drug markets."* ICS dissertation, Utrecht
286. Tomáš Diviák (2020), *"Criminal networks: Actors, mechanisms, and structures."* ICS dissertation, Groningen
287. Tery Setiawan (2020), "*Support for interreligious conflict in Indonesia*." ICS dissertation, Nijmegen
288. Vera de Bel (2020), "*The ripple effect in family networks: Relational structures and well-being in divorced and non-divorced families*." ICS dissertation, Groningen
289. Diego Palacios (2020), *"How context and the perception of peers' behaviors shape relationships in adolescence: A multiplex social network perspective."* ICS dissertation, Groningen
290. Saskia Glas (2020), *"Where are the Muslim Feminists? Religiosity and support for gender equality in the Arab region*." ICS dissertation, Nijmegen
291. Tomas Turner-Zwinkels (2020), *"A new macro-micro approach to the study of political careers: Theoretical, Methodological and empirical challenges and solutions."* ICS dissertation, Groningen
292. Lotte Scheeren (2020), "*Not on the same track? Tracking age and gender inequality in education."* ICS dissertation, Amsterdam
293. Joris Broere (2020), "*Essays on how social network structure affects asymmetric coordination and trust*." ICS dissertation, Utrecht
294. Marcus Kristiansen (2021*), "Contact with benefits: How social networks affect benefit receipt dynamics in the Netherlands."* ICS dissertation, Utrecht
295. Judith Kas (2021), "*Trust and reputation in the peer-to-peer platform economy*." ICS dissertation, Utrecht
296. Andrea Forster (2021), "*Navigating educational institutions: Mechanisms of educational inequality and social mobility in different educational systems."* ICS dissertation, Amsterdam
297. Jannes ten Berge (2021), *"Technological change and work: The relation between technology implementation within organizations and changes in workers' employment."* ICS dissertation, Utrecht
298. Jolien Geerlings (2021), "*Teaching in culturally diverse classrooms: The importance of dyadic relations between teachers and children*." ICS dissertation, Utrecht
299. Kirsten van Houdt (2021), "*Stepfamilies in adulthood: Solidarity between parents and adult children*." ICS dissertation, Amsterdam
300. Suzanne de Leeuw (2021), "*The intergenerational transmission of educational attainment after divorce and remarriage."* ICS dissertation, Amsterdam
301. Fleur Goedkoop (2021), "*Involvement in bottom-up energy transitions: The role of local and contextual embeddedness*." ICS dissertation, Groningen
302. Eva Vriens (2021), "*Mutualism in the 21st century: The why, when, and how behind successful risk-sharing institutions*." ICS dissertation, Utrecht
303. Ardita Muja (2021), "*From school to work: The role of the vocational specificity of education in young people's labor market integration*." ICS dissertation, Nijmegen
304. Siyang Kong (2021), "*Women and work in contemporary China: The effect of market transition on women's employment, earnings, and status attainment*." ICS dissertation, Utrecht
305. Marijn Keijzer (2022), *"Opinion dynamics in online social media."* ICS dissertation, Groningen
306. Sander Kunst (2022), *"The educational divide in openness towards globalisation in Western Europe."* ICS dissertation, Amsterdam
307. Nella Geurts (2022), "*Puzzling pathways: The integration paradox among migrants in Western Europe."* ICS dissertation, Nijmegen
308. Dragana Stojmenovska (2022), *"Men's place: The incomplete integration of women in workplace authority."* ICS dissertation, Amsterdam
309. Bram Hogendoorn (2022), *"Divorce and inequality: Stratification in the risk and consequences of union dissolution."* ICS dissertation, Amsterdam
310. Tom Nijs (2022), *"This place is ours: Collective psychological ownership and its social consequences."* ICS dissertation, Utrecht
311. Nora Storz (2022), *"'This land is ours—but is it also theirs?' Collective ownership beliefs and reconciliation in territorial conflict regions."* ICS dissertation, Utrecht
312. Tara Koster (2022), *"Parenting and fairness in diverse families."* ICS dissertation, Utrecht

313. Danelien van Aalst (2022), "*Elements Contributing to Teachers' Role in Bullying."* ICS dissertation, Groningen
314. Wybren Nooitgedagt (2022), *"Who owns the country? Collective psychological ownership and intergroup relations in settler societies"* ICS dissertation, Utrecht
315. Marija Dangubić (2022), *"Rejecting Muslim minority practices: Principles and prejudices"* ICS dissertation, Utrecht
316. Sara Cvetkovska (2022), *"Lines in the shifting sand: The implications of being tolerated"* ICS dissertation, Utrecht
317. Maaike Hornstra (2022), "*Going beyond the dyad: Adult intergenerational closeness after divorce and remarriage"* ICS dissertation, Amsterdam
318. Wouter Kiekens (2022), *"Sexual and gender minority youth's mental health and substance use: Disparities, mechanisms, and protective factors"* ICS dissertation, Groningen
319. Carlijn Bussemakers (2022), *"Adversity and educational inequality: The interplay between adverse experiences and parental resources for children's educational attainment."* ICS dissertation, Nijmegen
320. Evi Velthuis (2022), " '*To tolerate or not to tolerate?' Reasons for tolerance of minority group practices among majority members in the Netherlands and Germany."* ICS dissertation, Utrecht
321. Hendrik Nunner (2023), *"Co-evolution of social networks and infectious diseases."* ICS dissertation, Utrecht
322. Carly van Mensvoort (2023), *"Inspiring leaders: An empirical study of female supervisors at work."* ICS dissertation, Nijmegen
323. Anne van der Put (2023), *"Healthy at work: The role of the work environment in worksite health promotion."* ICS dissertation, Utrecht
324. Rowan ten Kate (2023), "*Understanding loneliness among older migrants*" ICS dissertation, Groningen
325. Sanne Kellij (2023), "*I see, I see what you don't see: Neural and behavioral social-cognitive processes underlying (persistent) victimization* " ICS dissertation, Groningen
326. Inge Hendriks (2023), "*Changes and contrasts in attitudes towards ethnic minorities*" ICS dissertation, Nijmegen
327. Eleonora Marucci (2023), "*Antecedents and consequences of teacher attunement in primary and secondary school classrooms"* ICS dissertation, Groningen
328. Kasper Otten (2023), "*Cooperation in changing groups: How newcomers and norms shape public good provision in the lab, online games, and the field.*" ICS dissertation, Utrecht

How people cooperate to provide public goods is a major scientific and societal question. Most existing research shows how people cooperate in stable groups. However, many real-world public good problems occur in groups that gradually change due to old members leaving and new members arriving. We have a limited understanding of how cooperation can be sustained when newcomers arrive whose views and needs might differ from those of the existing members. In this thesis, we examine this issue with lab experiments, a large-scale online game, and a field experiment. We find that newcomers initially contribute differently to public goods than existing members. Over time, newcomers largely conform to existing members. Newcomers are punished more for norm deviations than existing members and have less influence on how public goods are shaped. Altogether, the thesis reveals that cooperation dynamics in changing groups are different from stable groups.

Kasper Otten obtained his master's degree in Sociology and Social Research at Utrecht University. He conducted the present study as part of his Ph.D. research at the Interuniversity Center for Social Science Theory and Methodology (ICS), the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP), and the Department of Sociology at Utrecht University.

SCOOP

SUSTAINABLE COOPERATION