Social Science Research 77 (2019) 79–87

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Modeling the micro-macro link: Understanding macro-level outcomes using randomization tests on micro-level data

Jacob Dijkstra[a,*], Loes Bouman[b], Dieko M. Bakker[a], Marcel A.L.M. van Assen[c]

[a] *University of Groningen, ICS Interuniversity Center for Social Science Theory and Methodology, The Netherlands*
[b] *University of Milano-Bicocca, Italy*
[c] *University of Tilburg and Utrecht University, ICS Interuniversity Center for Social Science Theory and Methodology, Netherlands*

ARTICLE INFO

Keywords:
Analytical sociology
Randomization tests
Micro-macro link
Social mechanisms

ABSTRACT

Analytical sociology explains macro-level outcomes by referring to micro-level behaviors, and its hypotheses thus take macro-level entities (e.g. groups) as their units of analysis. The statistical analysis of these macro-level units is problematic, since macro units are often few in number, leading to low statistical power. Additionally, micro-level processes take place within macro units, but tests on macro-level units cannot adequately deal with these processes. Consequently, much analytical sociology focuses on testing micro-level predictions. We propose a better alternative; a method to test macro hypotheses on micro data, using randomization tests. The advantages of our method are (i) increased statistical power, (ii) possibilities to control for micro covariates, and (iii) the possibility to test macro hypotheses without macro units. We provide a heuristic description of our method and illustrate it with data from a published study. Data and R-scripts for this paper are available in the *Open Science Framework* (https://osf.io/scfx3/).

1. Introduction

Analytical sociologists share James Coleman's outlook that "[t]he principal task of the social sciences lies in the explanation of social phenomena, not the behavior of single individuals" (1990, p.2). Looking at empirical research in sociology, however, it is clear that while "[s]ocial theory continues to be about the functioning of social systems of behavior, (…) empirical research is often concerned with explaining individual behavior" (Coleman, 1990, p.1). The analysis of individual behavior is *necessary* for a complete understanding of systemic outcomes (a position known as 'methodological individualism'). As analytical sociologists stress, however, it is not generally *sufficient* for this understanding (e.g., Boudon, 1979). The reason for the lack of sufficiency lies in the potentially complex ways in which individual behaviors interact to produce macro-level outcomes. The explanation of a macro-level outcome in terms of its generative interdependent individual actions, is variously known as the 'micro-macro link approach' or the 'social mechanisms approach'. In the words of Hedström (2005, p.24) "[t]he core idea behind the mechanism approach is that we explain not by evoking universal laws, or by identifying statistically relevant factors, but by specifying mechanisms that show how phenomena are brought about." Hedström offers a definition of mechanisms as consisting of "(…) a constellation of entities and activities that are organized such that they regularly bring about a particular type of outcome" (2005, p.25).

The potential complexity of social mechanisms has long been acknowledged (e.g. Boudon, 1977; Lindenberg, 1977). Coleman (1990, p.22), for instance, stresses that the macro-level outcomes to be analyzed frequently involve true *interdependencies* of individual behavior, going beyond mere "aggregation". Hedström (2005, p.26) similarly emphasizes that "(…) the same entities

[*] Corresponding author.
E-mail address: j.dijkstra@rug.nl (J. Dijkstra).

https://doi.org/10.1016/j.ssresearch.2018.08.013
Received 19 December 2017; Received in revised form 31 August 2018; Accepted 31 August 2018
Available online 04 September 2018

(individuals actors) strung together in different ways can be expected to regularly bring about different types of outcomes", suggesting that the "stringing together" (i.e., the way in which individual actions interact) is an important factor determining macro outcomes. Obviously, not all social mechanisms are complex in this sense, and 'easy cases' can be found where simple aggregation explains macro outcomes. Generally, however, it is safe to say that analytical sociology has a lot to offer when explaining macro-level outcomes produced by interdependent individual actions. For the explanation of such phenomena, we "(…) have to look at the *system of interaction* between individuals and their environment, that is, between individuals and other individuals or between individuals and the collectivity" (Schelling, 2006, p.14). To understand this system of interaction, analytical sociological theories of social mechanisms are comprised of at least two parts: (i) a theory of individual behavior (the "micro foundations", cf. Wittek et al., 2013, p.5), and (ii) a theory of how individual behaviors combine, under specific rules (i.e., norms, institutions), to produce macro-level outcomes (Coleman, 1990).

The focus on social mechanisms implies that in empirical research at least some of the hypotheses (and according to analytical sociologists the quintessentially "sociological" among them) take human collectivities such as groups, teams or societies as their units of analysis. This is true for most studies in the domain of 'experimental behavioral game theory' (e.g., Camerer, 2011), on which we focus in this paper and in which groups of participants make strategic choices. The interest of such games (e.g., public good games, coordination games, volunteer's dilemmas, etc.) lies not only in the behavior of individuals, but also in the group-level outcomes (e.g., whether the group's public good is successfully produced). Treating experimental groups as the unit of analysis for at least some hypotheses seems to imply that a great many participants must be recruited and tested, which is potentially very onerous, as statistical power of tests using macro-level outcomes is generally low with a small number of groups. Therefore, most studies from this domain focus on testing the micro-level hypotheses implied by their theories, which increases statistical power because of a larger number of observations.

Focusing on the explanation of individual behavior works at least to some extent in the simple cases in which the macro-level outcome is an aggregate (such as a count, a sum or a mean) of individual actions (or their consequences), for which a sampling distribution can be derived. For instance, in an experimental linear public goods game with several treatments (e.g. Fehr and Gachter, 2000), mean contributions to the public good by participants can be computed per treatment, and treatments can be statistically compared with regard to this mean. But even in this simple case the researcher has very little information about the sampling distributions of macro-level explananda (e.g., total contribution at the group level) for different treatments, due to the low number of groups.

From an analytical sociology perspective, testing only individual-level hypotheses fails altogether when the macro-level outcome depends on interdependent individual actions in more complex ways. In such cases, the sampling distribution of aggregated individual behavior (in the sense of counts, sums or means) does not equal the sampling distribution of the macro-level outcome. Moreover, the latter distribution (which is necessary for statistical inference concerning macro-level hypotheses) frequently cannot be analytically derived. For instance, even in a simple step-level public goods game (e.g., Kerr, 1992, Van de Kragt et al., 1983), where the public good is produced if and only if at least a certain minimum number of group members contribute, a higher number of contributors across all groups does not imply a higher number of 'successful' groups (the relevant macro-level outcome). The number of successful groups critically depends on the *distribution* of contributors across groups, and the aggregate sum of contributors contains limited information about this. Hence, one is again left with taking observed groups as the unit of analysis for inference about the macro-level outcome, resulting in low statistical power unless a considerable number of groups is available.

In this paper we offer an alternative to the standard solutions employed in experimental research of either exclusively focusing on individual behavior or settling for low power tests and propose and illustrate a methodology allowing statistical inference concerning the micro-macro link central to analytical sociology (Buskens et al., 2014; Hedström and Swedberg, 1998), *using micro-level data only*. Our method allows statistically more powerful inferences about social mechanisms by exploiting all the information about macro-level outcomes the data contain. Since we make all the data and R-scripts used for this research publicly available on the *Open Science Framework* (https://osf.io/scfx3/), our methodology can be employed by any researcher whose data have a certain structure and whose data and theory meet two assumptions.

The data structure to which our method applies is as follows. A set of N micro-level units (typically, (the behavior of) individual people) is distributed across T treatments, with each micro unit assigned to a single treatment (i.e., in a between-subjects design). In the remainder we will refer to such a distribution of micro units across treatments as a *permutation*. Within each treatment t the micro units are distributed across K_t groups, with each micro unit assigned to a unique group. In the remainder we will refer to such a set of groupings (one per treatment t) as a *partition*. The two core assumptions of our methodology are that (i) the micro-macro link or social mechanism is formally expressible as a function of the micro-level data ('individual behavior'), and (ii) the macro-level units ('groups') in the data are randomly formed and do not lead to statistically dependent individual behavior (i.e., individual behavior is statistically independent of group membership).[1]

Data from behavioral game theory experiments often satisfy these two assumptions. First, such research typically involves explicit 'rules' mapping individual behaviors to group outcomes and these rules thus codify the experiment's social mechanism. Meeting this

[1] This independence is potentially conditional on covariates such as 'treatment group' or 'order of experimental tasks'. Also, sometimes data on observational covariates are collected within an experiment. Representing all such experimental and observational covariates by the symbol Z, the second assumption can be formally expressed as $Y \perp X \mid Z$, where Y signifies individual behavior and X group membership. In words: individuals' behavior in each treatment is statistically independent, conditional on relevant other variables. These other variables may be demographic (sex, SES, education, etc.), personal characteristics (social preferences, intelligence, etc.) or behavior or choices in previous tasks of the experiment.

assumption forces the researcher to explicate the social mechanism at the core of her theory. Quite apart from statistical inference, this assumption is a *conditio sine qua non* for analytical sociology. Second, experimental macro units (such as groups or networks) are very often artificially created in the laboratory, and hence do not involve many of the statistical dependencies real-world groups have. This renders the second assumption more plausible. Note, however, that experiments involving communication between individuals before the behavior of interest almost inevitably violate the second assumption. This is not to say our method is principally unsuited to observational research, but the purest illustrations are likely found in experiments. Therefore, we draw our two examples from a published experimental paper in the field of experimental behavioral game theory. The paper used for illustration is the Dijkstra and Bakker (2017) publication on Step-level Public Goods (SPG).

In the next section we state the purpose of our method and give its outline in terms of six steps that have to be completed in order to apply it. The subsequent section illustrates with two examples. A conclusion and discussion section closes the paper.

2. Purpose and outline of our method

The purpose of our paper is to introduce and illustrate a statistical method enabling inferences on the social mechanism or micro-macro link in applications, by deriving the sampling distributions of relevant macro-level explananda using micro-level data in those applications. Our method is based on the principle of *randomization tests* (e.g., Edgington, 1995) to compare different conditions under which the social mechanism under consideration is predicted to yield different macro-level outcomes (we will consistently use the term “treatments” to refer to these conditions). We are thus concerned with research in which a number of human groups (such as societies, communities, teams, neighborhoods, etc.) are categorized into different treatments (in our illustrations experimentally, but potentially also observationally). A randomization test computes a statistic in the sample (here denoted S^*) and compares its value to the distribution of that statistic (S) assuming that the micro-level units of the treatments are exchangeable, i.e., the treatments are identical (the macro-level H0). The “micro-level units of the treatments are exchangeable” is implemented in the randomization test by repeatedly randomly allocating all micro-level units (of all treatments combined) to the treatments, i.e., by considering alternative random permutations. The exact sampling distribution under H0 is obtained by tabulating the statistic S as calculated for each possible permutation. Relevant p-values of the randomization test can then be computed.

Our test is based on the simple intuition that with random assignment of participants to treatments (i.e., random permutations), and with randomly formed groups within treatments (i.e., random partitions given the permutations), experimental groups are *arbitrary*. Thus, rather than testing macro-level hypotheses on these arbitrary groups, the test considers *all permutations and all partitions that could have occurred* given the H0 of exchangeable micro-level units and identical treatments. More specifically, six steps must be completed to apply our method. The contents of step 6 depend on whether a *homogeneous* or *heterogeneous* version of our test is employed. In a homogeneous test, there are no covariates in the data explaining individual behavior, except for design characteristics such as different roles agents may have in the social situation, and the treatment identifier. In a heterogeneous test, such covariates do exist and are used when modeling individual behavior. Examples 1 and Examples 2 below illustrate the homogeneous and heterogeneous cases, respectively. The six steps of our method are:

1. The macro-level explanandum (S) must be expressed as a function of micro-level behavior.
2. A (statistical) micro-level model must be estimated predicting individual behavior as a function of the treatment and potentially other covariates.
3. Based on the first two steps, compute the observed sample value (S^*) of the explanandum. Note how S* is based on the *observed permutation and partition* of micro-level units, in the data.

Then, for each of a fixed number N of iterations, take steps 4 through 6. Step 4 generates a new permutation of micro-level units, whereas steps 5 and 6 mirror steps 2 and 3, respectively:

4. Permute the micro-level units across the treatments, such that each micro unit is assigned to a unique treatment. It is important to note that each micro unit *together with all its values on the outcome variable (e.g. behavior choice), and all covariates (in the heterogeneous case)* is randomly assigned to a treatment, in this step.
5. Estimate the same micro-level model as specified in step 2, but this time on the permuted data, rather than on the original data. With the same micro-model we mean that the statistical micro-model has the same parameters as the model in step 2, but the parameters are once more estimated using the permuted data, that is *as if the current permutation were the true assignment of micro-level units (with all their variable scores) to treatments.*

If a homogeneous test is implemented, the micro-level predictions in step 5 (and step 2) will be identical for all micro-level units with the same role, assigned to the same treatment in a given permutation. Consequently, under a homogeneous test all possible partitions given a permutation yield the same value of the macro-level outcome S:

6 (Homogeneous). Compute S for this permutation.

If a heterogeneous test is implemented, the micro-level predictions in step 5 (and in step 2 as well) may differ among micro-level units assigned to the same treatment in a given permutation, since other covariates besides treatments are used to model individual behavior. Thus, not all possible groups of micro-level units from the same treatment in a given permutation necessarily yield the same macro-level outcome, and we also have to inspect the different possible partitions given each permutation. Step 6 then becomes:

6 (Heterogeneous). In the given permutation, consider K partitions. For each of these K partitions, compute S and tabulate the distribution of the K values of S, for this permutation.

In a typical case (as in our examples below), micro-level units are so numerous that it is impractical to exhaustively enumerate all N permutations mentioned at step 4 and all K partitions mentioned at step 6 (Heterogeneous), when applicable. In these cases, N permutations are randomly drawn from the universe of permutations, and steps 4–6 are repeated many times ($N = 10{,}000$, in our examples). Within each of these repetitions, step 6 (Heterogeneous), if applicable, is also repeated many times ($K = 500$ partitions, in Example 2).

Care must be taken that the permutations and partitions in steps 4 and 6 respect the original study design characteristics such as roles agents may have in the social situation. For instance, if each group within a treatment by design consists of fixed numbers of individuals with certain roles (such as one leader and three followers, or two parents and one child), then each partition in step 6 must yield groups of the same composition. Similarly, if a particular treatment in the original study contained 25 women and 13 men, and if the gender composition is fixed by design, then in each permutation this treatment must contain 25 women and 13 men.

3. Two examples

To facilitate the understanding of our method, we draw two examples from the Dijkstra and Bakker (2017) publication on Step-level Public Goods (SPG), from this journal.

Some theoretical and empirical background for our two examples

Many researchers regard the cooperative solution of adaptive problems (such as hunting or defense) as a key driver of human evolution (e.g., Cosmides and Tooby, 1992; Kiyonari et al., 2000; Fehr and Fischbacher, 2003; Nowak, 2006; Mesterton-Gibbons and Dugatkin, 1992; Trivers, 2006). However, cooperation is as much a *problem* as it is an adaptive solution. This conclusion follows from the analysis of the large number of social situations producing *social dilemmas* that are identified in the behavioral sciences (Kollock, 1998; Dawes, 1980). In social dilemmas individual interests are at odds with group interests, in that all group members would be better off if cooperation were wide-spread, but no individual group member has sufficient incentives to be the (sole) initiator of cooperation.

Ranking high among cooperation problems are collective action attempts aimed at the production of *public goods* (or *collective* goods; Ledyard, 1995; Olson, 1965). With respect to public goods, individuals participating in the joint action are said to *invest* in the (production of the) public good. Public goods are goods for which it is true that (i) consumption by one individual does not rival with consumption by another individual (jointness of supply), and (ii) it is (practically) infeasible to exclude individuals from consumption (non-excludability) (Offerman, 2013). The latter characteristic makes provision of the good through the market mechanism impossible, since consumption cannot be rationed by prices. The former characteristic renders the production of public goods rife with social dilemmas. In particular, if no single individual has sufficient interest in the public good and/or is sufficiently resourceful so as to be willing and able to produce the public good alone, various social dilemmas arise (Raub et al., 2015). The *social movements* literature in particular (e.g., McAdam and Diani, 2003; Goodwin and Jasper, 2014) is rife with examples of real-world public goods, as is the work of Elinor Ostrom (1999).

Coordination problems (Ochs, 1995) present an important class of social dilemmas associated with public good production. In coordination problems what is the best course of action for each individual depends on the behavior of others. Coordination problems in public good production typically arise when the good's “production technology” implies that the good is produced if and only if a sufficiently large or powerful subgroup of individuals invests. The coordination problem then amounts to identifying and motivating the coalition of individuals who should invest. Such problems typically involve a complex micro-macro link: the ‘right number’ of the ‘right kind of people’ should invest to make the group successful, implying that investments cannot simply be tallied or summed to determine group success. Thus, coordination problems in public good production are typically amenable to analysis with our method.

Several specific models have been proposed to analyze the coordination problem of finding the subset of necessary investors in public good coordination problems (Marwell and Oliver, 1993; Diekmann, 1985; Bramoullé and Kranton, 2007). A prominent theoretical model, from which we draw our two examples, is the *Step-Level Public Good* game (SPG; Van de Kragt et al., 1983; Kerr and Kaufman-Gilliland, 1994; Dijkstra and Bakker, 2017; Dijkstra and Oude Mulders, 2014). In an SPG a group of individuals can jointly produce a public good valuable to all. Investing in the public good is costly, but the value of the public good (when produced) exceeds this cost for each investor. Individuals may differ in the extent to which their investments have an impact on the production of the public good. The public good is produced if and only if the total impact of the investors exceeds a given threshold. If investments fall short of the threshold the public good is not produced, and any investors incur a net loss. If the public good is produced, all group members enjoy a net gain. However, the net gain for investors is less than the net gain for non-investors, as the latter do not incur the costs of investment. Hence, the SPG models a coordination problem in the context of the production of a public good, the problem being the quest for a set of sufficiently impactful investors.

We employ the data from Dijkstra and Bakker (2017). These data are from an SPG experimental game introduced in Dijkstra and Oude Mulders (2014). Individuals in the game are referred to as ‘players’, and this SPG has five of them. Each player has an endowment of 10 points and decides (anonymously and in isolation) whether to *invest* or *keep* her entire endowment. When all players have decided, the game ends.

Investing means a player loses her 10 points. In the event that the SPG is produced all players, regardless of their investments, receive 15 points. Thus, investors end up with a total of 15 points (for a net gain of 5) while non-investors end up with a total 25 points (for a net gain of 15). If the SPG is not produced no points are awarded. In that case, investors end up with 0 points (for a net loss of 10), whereas non-investors simply keep their endowments of 10 points (gain nor loss).

The rules determining SPG production are as follows. Each player is assigned a *share* between 1 and 50, modeling her impact. It is common knowledge that the shares of all five players sum to 100 and that no single player has a share greater than 50. The share distribution is such that one player has a share of 50, one player has a share of 2, and three players have a share of 16. The SPG is produced if and only if the shares of the investors sum to at least 51.

A successful group of investors must include the player with a share of 50. However, to put the sum of shares over the threshold, the share 50 player must be joined by at least one other player. The game in the complete information version has nine Nash equilibria (Dijkstra and Oude Mulders, 2014): one in which no player invests, four in which two players invest with certainty, and another four in which two players invest with high probability. In the equilibria where players invest, the investors are the share 50 player and *any one of the other players*.

Dijkstra and Bakker (2017) report three experimental studies using this experimental SPG, and we apply our method to all three studies separately. In all three experimental studies there were two *treatments*. In the *incomplete information treatment* (IIT) players knew only their own shares. In the *complete information treatment* (CIT) players knew the distribution of all five shares in their group. For a detailed description of the experimental design we refer to Dijkstra and Bakker (2017). Dijkstra and Bakker (2017) statistically tested hypotheses about the effects of their information treatment on individual investment decisions. Our method, however, allows drawing statistical inferences about *treatment differences in a group's probability of successfully producing the public good*.

Study 1 of Dijkstra and Bakker (2017) was a lab experiment with $N_1 = 120$ participants. Participants were randomly assigned to their player roles in groups of five, and groups were randomly assigned to the IIT and the CIT. Participants then played a one-shot SPG. The experiment was programmed in the Z-Tree software (Fischbacher, 2007).[2] Additionally, participants' *social value orientation* (SVO) was measured using the 9-item triple dominance measure of Van Lange (1999). The SPG was incentivized, with participants receiving 10 eurocents per point earned in the one-shot game.

Study 2 was an online experiment with $N_2 = 177$ participants, using the Qualtrics online survey software (Qualtrics, 2014). None of the participants in Study 2 had participated in Study 1. Participants were randomly assigned to their roles in either the IIT or the CIT, and played the same one-shot SPG as in Study 1.[3] Participants also completed the 9-item triple dominance SVO measure. Participants were randomly drawn to be paid for their investment decisions. The probability of being paid for at least one of their decisions was about 0.19. Participants selected for payment received 30 eurocents for each point earned in the experiment.

Study 3 was an online experiment with $N_3 = 315$ participants, using the Qualtrics online survey software (Qualtrics, 2014). None of the participants in Study 3 had participated in either of the other studies. The design and procedures of Study 3 were virtually identical to the design of Study 2 (some additional observational measures were taken after the SPG decisions had been taken). Importantly, note that in Studies 2 and 3 no real groups of participants were formed and no interaction between participants took place. Randomly selected participants were only grouped after data collection for payment purposes.

Example 1. A homogeneous test.

Step 1 of our method requires us to specify the micro-macro link. The macro-level explanandum in our test is *the group probability of success*, i.e. the probability that the public good is produced. Consider an arbitrary group *g*. If we let $p_{s,i,g,t}$ denote the probability of investment by a share *s* player *i* in group *g* of treatment *t*, and let $P_{t,g}$ denote group *g*'s probability of success in treatment *t*, then the rules of the SPG imply the following relation between these probabilities:

$$P_{t,g} = p_{50,i,g,t}[1 - (1 - p_{2,j,g,t})(1 - p_{16,k,g,t})(1 - p_{16,l,g,t})(1 - p_{16,m,g,t})] \tag{1}$$

Where *i*, *j*, *k*, *l*, and *m* index distinct players composing group *g*. Equation (1) reflects that all individuals with the same role within each treatment are exchangeable (i.e., all participants with the same share in the same treatment have exactly the same investment probabilities), which is characteristic of the homogeneous case. Equation (1) follows directly from the fact that players' decisions are statistically independent in the one-shot game. Note that the part in square brackets represents 1 minus the probability that none of the share 2 and share 16 players invest. $P_{t,g}$ is a macro-level variable determined by the micro-level decisions of group members. Thus, equation (1) explicitly models the micro-macro link of analytical sociology, and completes step 1 of our method.

Step 2 is the specification of a statistical model predicting individual behavior. Table 1 aggregates the micro-level decisions by study, share, and treatment. In each cell the proportions of invest/not invest decisions are given for participants with a given share, in a given study, in a given treatment. For instance, the top-left cell of the table shows that in the CIT of Study 1 a proportion of 0.25 of the share 2 players invested. In the simplest micro-level model this is also the predicted probability that a share 2 player in the CIT of Study 1 invests. Consequently, any group of the right composition (one share 2 player, one share 50 player, and three share 16 players) drawn from the same treatment, will have the same success probability (cf. Equation (1)).

Before proceeding, we note that the first core assumption of our method (i.e., the micro-macro link is a function of micro-level data) is met by the specification of Equation (1). Concerning the second core assumption (i.e., the micro-level data points are statistically independent), since the SPG game was one-shot, no repeated interactions took place between participants. Any groups in

[2] Random assignment of participants to groups and roles was handled by Z-tree. Random assignment of groups to treatments was implemented by randomly assigning *sessions* (comprising between 2 and 4 groups) to treatments.

[3] In studies 2 and 3, random assignment of participants to roles was handled by Qualtrics. For payment purposes the researchers randomly drew cases without replacement from the data file.

Table 1
Investment decisions in each study, broken down by information treatment (CIT = complete information treatment; IIT = incomplete information treatment) and shares, for step 2 of the homogeneous test; numbers in brackets are proportions of invest/not invest decisions.

Study	Treatment	Share 2		Share 16		Share 50		Tot
		Investment Decision						
		Invest	Keep	Invest	Keep	Invest	Keep	
Study 1	CIT	3 (.25)	9 (.75)	11 (.306)	25 (.694)	11 (.917)	1 (.083)	60
	IIT	2 (.167)	10 (.833)	15 (.417)	21 (.583)	9 (.75)	3 (.25)	60
	Totals	24		72		24		120
Study 2	CIT	11 (.379)	18 (.621)	12 (.444)	15 (.556)	24 (.774)	7 (.226)	87
	IIT	12 (.387)	19 (.613)	19 (.704)	8 (.296)	26 (.813)	6 (.188)	90
	Totals	60		54		63		177
Study 3	CIT	9 (.29)	22 (.71)	46 (.517)	43 (.483)	19 (.633)	11 (.367)	150
	IIT	7 (.226)	24 (.774)	48 (.48)	52 (.52)	30 (.882)	4 (.118)	165
	Totals	62		189		64		315

the data were randomly formed for payment purposes only, and this grouping did not affect participants' decisions. This is also true for Study 1, in which participants *did* come to the laboratory to participate in separate sessions. Session nor group had any relation to individual decisions in Study 1 (analyses not shown). Thus, also the second core assumption is satisfied.

Step 3 requires us to define the test statistic and compute its sample value. We will test the H0 $P_{IIT,g} \leq P_{CIT,g'}$ against the alternative that $P_{IIT,g} > P_{CIT,g'}$ (cf. equation (1)), for two arbitrary groups, g and g'.[4] For the homogeneous test $S = P_{IIT,g} - P_{CIT,g}$, where we calculate $P_{IIT,g}$ and $P_{CIT,g}$ using (1) for each permutation of the data. The *p*-values of the randomization tests equal $P(S \geq S^*)$. We approximate $P(S \geq S^*)$ and show the distribution of S for each study. Table 2 below reports the sample success probabilities for each treatment $P_{t,g}$ and the value of S^*, for each study separately (last column). The $P_{t,g}$ were computed based on Table 1 and Equation (1). Study 1 shows a negative difference of $S^* = -0.060$ running counter to our alternative hypothesis, and studies 2 and 3 show positive differences of $S^* = 0.108$ and $S^* = 0.204$, respectively.

Steps 4 through 6 were performed $N = 10{,}000$ times. Each time, we drew a random permutation (step 4), calculated the equivalent of Table 1 for that permutation (step 5) and computed the value of S (the equivalent of Table 2) for that permutation (step 6). The left column of Fig. 1 shows both the distribution of S and $P(S \geq S^*)$ for each study for the homogeneous test. The approximated *p*-values for Study 1, Study 2, and Study 3 are 0.6377, 0.1482, 0.0233, respectively, only suggesting a negative effect of complete information on the provision of the public good in Study 3.[5]

Comparison of our homogeneous test to a test on observed experimental groups

Although Dijkstra and Bakker (2017) do not perform a statistical test on the macro-level, their Study 1 does contain actual experimental groups that could be used for a standard test of differences in group success between treatments. Of the 12 groups in the IIT, 9 were successful and 3 failed. Of the 12 groups in the CIT, 8 were successful and 4 failed. A Fisher exact test on the difference in success between the observed groups reveals no significant difference (p = 1), which is the same conclusion our test reached. It is noteworthy that this classical Fisher test only uses 24 observations at the macro-level, whereas our test employs patterns in individual data. A classical test comparing the two treatments cannot be performed for Study 2 and Study 3, as no actual groups were formed in these studies. Hence classical methodology does not allow for testing the hypotheses, as opposed to the method proposed in this paper.

Example 2. A heterogeneous test.

Step 1 in the heterogeneous test is identical to step 1 in the homogeneous test, and the two core assumptions are also met, since we use the same theory and data as in the first example. In step 2 of the heterogeneous case we estimate the micro-level model by controlling for certain design variables (i.e., the order in which experimental tasks were completed, for details see Dijkstra and Bakker, 2017) and for SVO in addition to treatment and share, estimating each participant's investment probability using logistic regression. Table 3 shows the logistic regression results for each study separately.

The models control for the order of experimental tasks. In all three studies the SVO questionnaire was either taken before the SPG decision task or after (SVO first = 1 or 0, respectively). In Study 1 a repeated version of the SPG game was played (in a different randomly composed group) either before or after the one-shot SPG we analyze. In studies 2 and 3 a version of the one-shot SPG in which letter labels were used rather than numeric shares was either played before or after the shares treatment. We code the variable

[4] We use this directed alternative hypothesis, because in the original paper the authors show how investments in the CIT are lower for share 16 players than in the IIT, while investments of other players are not much different.

[5] The supplementary information on the *Open Science Framework* (https://osf.io/scfx3/) contains R-scripts demonstrating that the p-values of the sampling distributions our method generates are indeed uniformly distributed, Hence, the probability of a Type I error can effectively be controlled at the chosen alpha level.

Table 2
Success probabilities for each treatment for the homogeneous test, and the difference between treatments.

	Sample size	Success Probability Incomplete Info $P_{IIT,g}$	Success Probability Complete Info $P_{CIT,g}$	Difference $S^* = P_{IIT,g} - P_{CIT,g}$
Study 1	120	0.626	0.686	−0.060
Study 2	177	0.800	0.692	0.108
Study 3	315	0.786	0.583	0.204

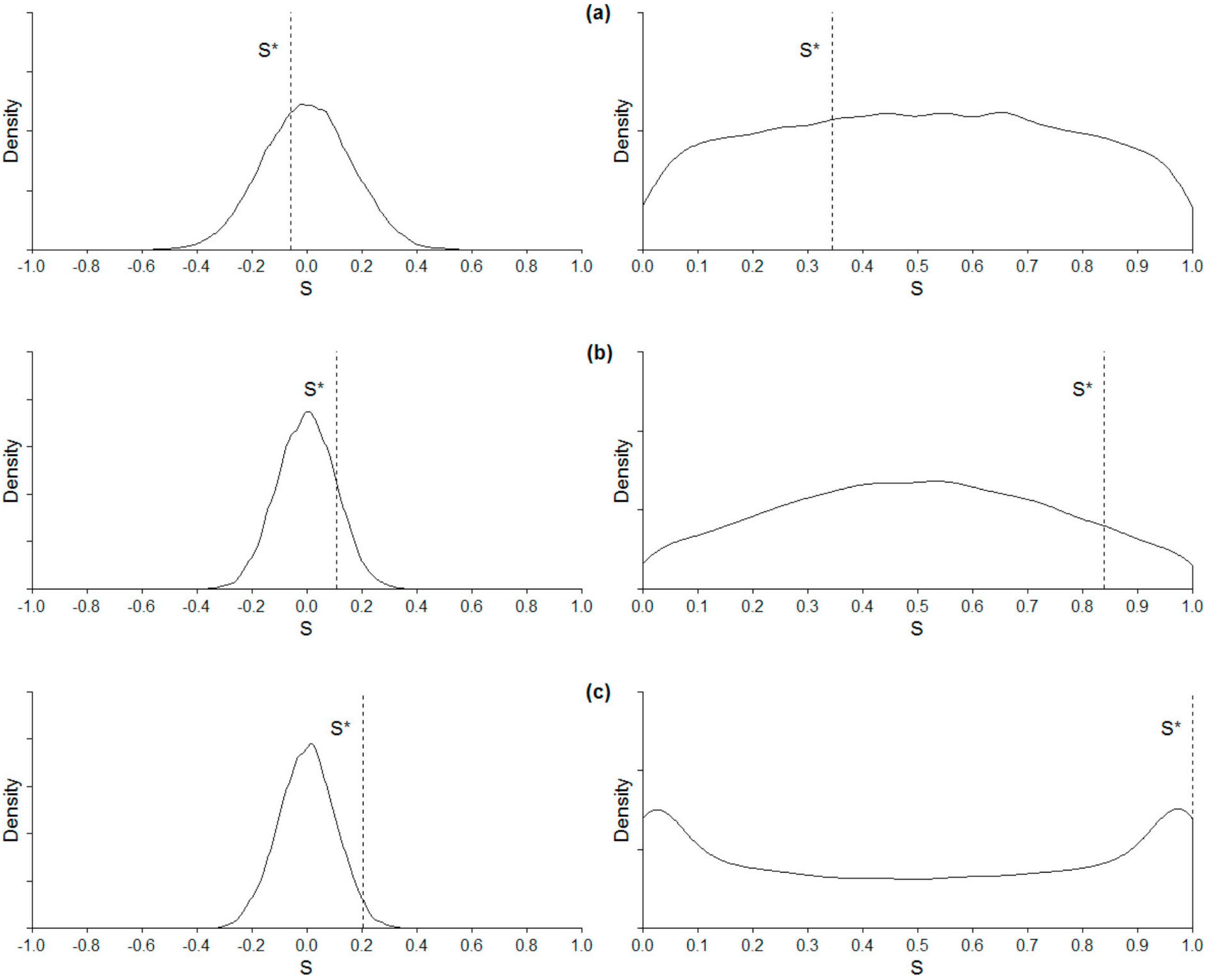

Fig. 1. Plots of approximate sampling distributions of $S = P_{IIT,g}$ - $P_{CIT,g}$ in the test without covariates (left column) and $S = \Pr[P_{IIT,g} > P_{CIT,g'}]$ in the test with covariates (right column), for Study 1 (a), Study 2 (b), and Study 3 (c). The value of S*, the value of the statistic for the actual data, is shown in all figures.

Game (1 = first, 0 = second) to indicate whether the one-shot SPG with shares we analyze in this paper was played first or second. Similar to the homogeneous test (Table 1) the models control for the share the participant was assigned to (share 16 is the reference category). Finally, the models control for the number of proself choices the participants made out of 9 (Proself). We estimate individual investment probabilities for each information treatment separately by including Info (1 = CIT, 0 = IIT) and all its interactions. The Hosmer-Lemeshow GOF tests reported in Table 3 suggest the logistic model is a good fit to the data (Lemeshow and Hosmer, 1982). Note how for each study, using the model in Table 3 to predict individual investment probabilities leads to (potentially) different predicted probabilities for two players with the same share in the same treatment, due to differences in the other covariates. Therefore, we will have to consider alternative partitions for each permutation, leading to a heterogeneous test.

Step 3 requires us to define the test statistic and compute its sample value. We test the same H0 as in the homogeneous case. For the heterogeneous test $P_{IIT,g}$ and $P_{CIT,g}$ are not constant across partitions given the same permutation, as the values of the covariates that affect the probability of contributing are not the same across partitions in the same permutation. Therefore, for the heterogeneous test we take $S = \Pr[P_{IIT,g} > P_{CIT,g'}]$ as our test statistic. To approximate S^* we apply step 6 (Heterogeneous) for 500 partitions for the *true* partition of participants across treatments, calculating $P_{IIT,g}$ and $P_{CIT,g}$ for each partition. This yields two distributions, one

Table 3
Logistic regression results for step 2 of the heterogeneous test. Dependent variable is the log odds of the probability to invest.

	Study 1 (N_1 = 120)		Study 2 (N_2 = 177)		Study 3 (N_3 = 315)	
	B	S.E.	B	S.E.	B	S.E.
Intercept	0.15	0.61	1.34*	0.64	0.24	0.39
SVO first	−0.66	0.60	0.15	0.50	−0.28	0.35
Game	0.85	0.62	−0.43	0.49	0.03	0.35
Share 2	−1.19	0.87	−1.19*	0.58	−1.18*	0.48
Share 50	1.83*	0.83	0.89	0.66	2.01***	0.58
Proself	−0.16	0.09	−0.12	0.07	−0.05	0.05
Info	0.24	0.88	−1.35	0.84	0.01	0.54
Info × SVO first	−0.28	0.90	0.27	0.69	0.54	0.48
Info × Game	−1.11	0.91	0.22	0.68	−0.27	0.49
Info × Share 2	0.80	1.20	0.92	0.80	0.19	0.66
Info × Share 50	2.59	1.61	0.53	0.88	−1.55*	0.73
Info × Proself	−0.14	0.17	0.02	0.09	−0.003	0.07
Hosmer – Lemeshow GOF test	$\chi^2(8) = 4.60$ p = 0.80		$\chi^2(8) = 5.61$ p = 0.69		$\chi^2(8) = 7.96$ p = 0.44	

Note: * p < 0.05, ** p < 0.01, *** p < 0.001.

for $P_{IIT,g}$ and one for $P_{CIT,g}$, from which we find $S^* = 0.344$ for Study 1, $S^* = 0.839$ for Study 2, and $S^* = 1$ for Study 3.

Steps 4 through 6 were again performed $N = 10,000$ times. Each time, we drew a random permutation (step 4), estimated individual investment probabilities by calculating the equivalent of Table 3 for that permutation (step 5) and computed the value of S for that permutation by randomly drawing 500 partitions (step 6). The right column of Fig. 1 shows the approximated distribution of $S = \Pr[P_{IIT,g} > P_{CIT,g'}]$ in the heterogeneous test. Adding the covariates did not change the p-values substantially compared to the homogeneous case; $p = 0.6764$ for Study 1, $p = 0.1013$ for Study 2, and $p = 0.0381$ for Study 3, again suggesting only an effect of information on the macro-level outcome in Study3.[6]

Note how no standard test is available to test for the success differences between observed groups in Study 1 (such as the Fisher exact test we employed when comparing to our homogeneous test), controlling for the covariates in Table 3. This reveals another advantage of our method over alternative standard tests; whereas traditional methodology cannot test hypotheses at the macro-level while adequately controlling for covariates at the level of individuals, our proposed methodology can.

4. Discussion and conclusion

Analytical sociology is about explaining macro-level phenomena or relations by referring to micro-level behavior and interactions. This implies that analytical sociology hypotheses take macro-level entities (such as groups, teams, organizations, or communities) as their units of analysis. The statistical analysis of these macro-level units, however, is problematic. In the first place, in empirical research macro units are often few in number because they are expensive to investigate in their entirety. Thus, statistical tests on these units have low power. In the second place, micro-level behavioral processes take place that affect macro-level outcomes. In fact, this is the basic tenet of analytical sociology. But tests on macro-level units cannot adequately deal with these micro-level processes. In response to these challenges, much analytical sociology focuses on testing micro-level (behavioral) predictions. Our method offers a better alternative; formally express the dependence of macro-level outcomes on micro-level data, and test macro-level hypotheses using randomization tests on the micro-level units.

Our method is suited for data that have a 'between-subjects' design: micro-level units (individuals) nested in macro-level units (groups) nested in conditions (treatments). The core assumptions of our method are that (i) the dependence of the macro-level phenomena of interest on the micro-level data is formally expressed, and (ii) the macro-level units ('groups') in the data are randomly formed and do not lead to statistically dependent individual behavior. Data from experiments are most likely to have this structure and meet these assumptions, but observational data are not principally excluded. The availability of our method (see https://osf.io/scfx3/) may actually provide experimental researchers with incentives to collect more data meeting our assumptions, such as "one-shot" designs, or series of "one-shot" experiments as in stranger or perfect-stranger matching designs.

We would like to draw attention to four strengths of our method. The first is that it forces researchers to explicate the micro-macro link or social mechanism of their theory. The second strength is that statistical tests of social mechanisms are statistically more powerful than standard tests that take the macro entities as their units of analysis. The third strength is that, contrary to what these standard tests have to offer, our method allows controlling for micro-level covariates when testing macro-level hypotheses (cf. the heterogeneous test in Example 2). The fourth strength is that our method facilitates the testing of macro-level hypotheses even in the absence of actual macro-level units (cf. studies 2 and 3 in our examples).

Within the boundary of its assumptions, our method is general. Both of our examples involved dichotomous behavior (keep/invest decisions), leading to logistic regression models of individual behavior in steps 2 and 5. However, our model can also be applied to

[6] In the online files (https://osf.io/scfx3/) we demonstrate that our randomization tests of our hypothesis have the required properties, i.e., they provide uniformly distributed p-values when the H0 is true.

ordinal or continuous individual-level behavior variables, just as randomization tests can be applied to variables of any scale (nominal, ordinal, interval, ratio). Simply replacing the binary logistic regression with the appropriate multinomial, ordinal, or OLS regressions suffices. On the other hand, like any statistical method, our method is limited by its assumptions. The main limitation is the assumption of (conditionally) independent micro-level observations. This assumption is constraining, since many real-world groups imply statistically dependent individual behavior, virtually limiting the applicability of our method to experimental research. A possible extension of our method to cases of dependent micro-level observations would be to include a covariance parameter in the models of steps 2 and 5 to reflect the dependence between micro observations from the same macro-level unit. The value of this covariance parameter could either be estimated from the data or, better yet, a range of values could be tried to investigate the sensitivity of the statistical results to degrees of dependence in the data. As an example of such a covariance parameter we would like to mention the intraclass coefficient, which is well-known from the context of multilevel analysis dealing with dependence of observations within macro-level units. Obviously, as permutations need to deal with this statistical dependency, an alternative procedure for random sampling of permutations needs to be developed to in that case.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ssresearch.2018.08.013.

References

Boudon, R., 1977. Effets pervers et ordre social, vol. 1 Presses universitaires de France.

Boudon, R., 1979. La logique du social. Hachette.

Bramoullé, Y., Kranton, R., 2007. Public goods in networks. J. Econ. Theor. 135 (1), 478–494. https://doi.org/10.1016/j.jet.2006.06.006.

Buskens, V., Raub, W., Van Assen, M. (Eds.), 2014. Micro-macro Links and Microfoundations in Sociology. Routledge.

Camerer, C.F., 2011. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.

Coleman, J., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, Mass.

Cosmides, L., Tooby, J., 1992. Cognitive adaptations for social exchange. Adap. Mind: Evol. Psychol. Generat. Cult. 163, 163–228.

Dawes, R.M., 1980. Social dilemmas. Annu. Rev. Psychol. 31 (1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125.

Diekmann, A., 1985. “Volunteer's dilemma. J. Conflict Resolut. 29, 605–610. https://doi.org/10.1177/0022002785029004003.

Dijkstra, J., Oude Mulders, J.A.A.P., 2014. Efficacy, beliefs, and investment in step-level public goods. J. Math. Sociol. 38 (4), 285–301. https://doi.org/10.1080/0022250X.2013.826214.

Dijkstra, J., Bakker, D.M., 2017. Relative power: material and contextual elements of efficacy in social dilemmas. Soc. Sci. Res. 62, 255–271. https://doi.org/10.1016/j.ssresearch.2016.08.011.

Edgington, E.S., 1995. Randomization Tests, third ed. Dekker, New York.

Fehr, E., Gachter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90 (4), 980–994. https://doi.org/10.1257/aer.90.4.980.

Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425 (6960), 785–791. https://doi.org/10.1038/nature02043.

Fischbacher, U., 2007. z-Tree: zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178. https://doi.org/10.1007/s10683-006-9159-4.

Goodwin, J., Jasper, J.M. (Eds.), 2014. The Social Movements Reader: Cases and Concepts. John Wiley & Sons.

Hedström, P., Swedberg, R. (Eds.), 1998. Social Mechanisms: an analytical approach to Social Theory. Cambridge University Press.

Hedström, P., 2005. Dissecting the Social: on the Principles of Analytical Sociology, vol. 10 Cambridge University Press, Cambridge.

Kerr, N.L., 1992. Efficacy as a Causal and Moderating Variable in Social Dilemmas.

Kerr, N.L., Kaufman-Gilliland, C.M., 1994. Communication, commitment, and cooperation in social dilemma. J. Pers. Soc. Psychol. 66 (3), 513. https://doi.org/10.1037/0022-3514.66.3.513.

Kiyonari, T., Tanida, S., Yamagishi, T., 2000. Social exchange and reciprocity: confusion or a heuristic? Evol. Hum. Behav. 21 (6), 411–427. https://doi.org/10.1016/S1090-5138(00)00055-6.

Kollock, P., 1998. Social dilemmas: the anatomy of cooperation. Annu. Rev. Sociol. 24 (1), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183.

Lemeshow, S., Hosmer Jr., D.W., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. 115 (1), 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113284.

Ledyard, J.O., 1995. Pubilc goods: a survey of experimental research. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 111–194.

Lindenberg, S., 1977. Individuelle Effekte, kollektive Phänomene und das Problem der Transformation. Probleme der Erklärung sozialen Verhaltens 1, 46–84.

Marwell, G., Oliver, P., 1993. The Critical Mass in Collective Action. Cambridge University Press.

McAdam, D., Diani, M., 2003. Social Movements and Networks: Relational Approaches to Collective Action (Comparative Politics).

Mesterton-Gibbons, M., Dugatkin, L.A., 1992. Cooperation among unrelated individuals: evolutionary factors. Q. Rev. Biol. 67 (3), 267–281. https://doi.org/10.1086/417658.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314 (5805), 1560–1563. https://doi.org/10.1126/science.1133755.

Ochs, J., 1995. Coordination problems. Handbook of experimental economics 195–252.

Offerman, T., 2013. Beliefs and Decision Rules in Public Good Games: Theory and Experiments. Springer Science & Business Media.

Olson, M., 1965. Logic of Collective Action: Public Goods and the Theory of Groups (Harvard Economic Studies. V. 124). Harvard University Press.

Ostrom, E., 1999. Coping with tragedies of the commons. Annu. Rev. Polit. Sci. 2 (1), 493–535.

Qualtrics, L.L.C., 2014. Online Survey Software Tools and Solutions: Qualtrics. (Retrieved from).

Raub, W., Buskens, V., Corten, R., 2015. Social dilemmas and cooperation. In: Braun, Norman, Saam, Nicole J. (Eds.), Handbuch Modellbildung und Simulation in den Sozialwissenschaften. Springer VS, Wiesbaden, pp. 597–626.

Schelling, T.C., 2006. Micromotives and Macrobehavior. WW Norton & Company.

Trivers, R., 2006. Reciprocal altruism: 30 years later. In: Cooperation in Primates and Ting Set as a Solution to Public Goods Problems, vol. 77. American Political Science Review, pp. 112–122. https://doi.org/10.2307/1956014. 01.

Van de Kragt, A.J., Orbell, J.M., Dawes, R.M., 1983. The minimal contributing set as a solution to public goods problems. Am. Polit. Sci. Rev. 77 (1), 112–122. https://doi.org/10.2307/1956014.

Van Lange, P.A., 1999. The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. J. Pers. Soc. Psychol. 77 (2), 337. https://doi.org/10.1037/0022-3514.77.2.337.

Wittek, R., Snijders, T.A., Nee, V. (Eds.), 2013. The Handbook of Rational Choice Social Research. Stanford University Press.