Judgment and Decision Making (2026), **xx**:xx 1–**??**

CAMBRIDGE UNIVERSITY PRESS

Empirical Article

Strategic Misrepresentation in Personality Testing: An Experimental Study using the Public Goods Game

Daniel Woods[1]

[1]Macquarie Business School, Macquarie University, Sydney, 2109, NSW, Australia.

Keywords: Personality Testing, Public Goods Game, Agreeableness, Big Five

Abstract

Personality tests are commonly used to hire suitable employees but this process is susceptible to strategic misrepresentation by job-seekers. This paper uses a lab experiment as an analogy of such a hiring process by using a repeated Public Goods Game (PGG) as a proxy for a cooperative work environment. Participants first complete a Big Five personality test, focusing on the trait of "Agreeableness", which some previous studies have associated with prosocial cooperation in the PGG. Two groups are formed: a high Agreeableness group and a low Agreeableness group. The experiment manipulates the timing of revealing the group formation rule, as knowing the rule before the personality test allows for misrepresentation of Agreeableness. I find no evidence of substantial misrepresentation when the group formation rule is revealed before the personality test. I do find that Agreeableness group formation increases contributions for both high and low groups, but only when it is described to participants before the PGG. I find no evidence that Agreeableness is related to contributions in the PGG.

1. Introduction

Psychometric tests, designed to measure a person's personality or other latent aspects that cannot be directly observed, are an established standard in many firms' hiring procedures. Psychometric tests are used on approximately 60 to 70% of US job-seekers (Weber & Dwoskin, 2014), and 75% of international firms either use or plan to use them in the future (Kantrowitz et al., 2018). However, strategic misrepresentation in these tests has been a long-standing concern both in industry and in academia (Morgeson et al., 2007; Tett & Simonet, 2021; Viswesvaran & Ones, 1999). It has also been suggested that psychometric testing is unfair or discriminatory against minorities or those with disabilities (Hawkins & Monroe, 2021; McGee & McGee, 2025; Weber & Dwoskin, 2014).

In this paper, I design a laboratory experiment to evaluate under what conditions personality testing may be effective. I design the lab experiment to be analogous to 'hiring' using personality testing, and the subsequent 'work effort', at least in so far as a lab experiment permits. The experiment consists of two main parts, a personality test followed by a cooperation task. For the personality test, I elicit the 'Big Five' personality traits, and for the cooperation task, I use a repeated Public Goods Game (PGG). In the PGG participants can make socially-optimal contributions to a public good, but face a personal incentive to free-ride and contribute less (for the exact implementation, see Section 3.2.2). I interpret contributions to the public good as 'work effort', which is something an employer would like to encourage. I focus on the Big Five personality trait of 'Agreeableness', the tendency to act in a cooperative, unselfish manner, as some research finds it is positively associated with contributions in the PGG and other similar social dilemmas. I sort participants into groups for the PGG based on their Agreeableness score, to mimic the role of an employer hiring based on personality tests in an attempt to maximize their firm's success.

The use of Agreeableness was motivated in the pre-registration by its relationship with behavior in the PGG citing the following papers: (Kagel & McGee, 2014; Perugini et al., 2010; Thielmann et al., 2020; Volk et al., 2012). I caveat this motivation using the meta-analysis of Thielmann et al. (2020) rather than the individual studies. While this meta-analysis reports Agreeableness is associated with prosocial behavior both over all games and all PGGs they consider, they find no evidence of a relationship in social dilemmas with repeated interaction, i.e. of the type used in this study.[1] I used a repeated PGG to be analogous to a cooperative teamwork environment, which is rarely one-shot or without feedback. While the following was not pre-registered, using Agreeableness is realistic as it has been recommended for team selection (Morgeson et al., 2005), is a strong predictor of team field performance (Bell, 2007), and two of the key themes of its impact are teamwork and work investment (Wilmot & Ones, 2022).[2]

The key treatment dimension in the experiment is the timing of information about the purpose of the initial personality questionnaire, i.e., the group formation rule for the PGG. There are three treatments on the time dimension, *Before* the personality test, *After* the personality test (but before the PGG), and *Never*. In the *Before* treatment, participants have an incentive to misrepresent their personality to try and get into a more cooperative group. Whereas in the *Never* treatment, participants are never informed about how groups are formed, and therefore have no material incentive to misrepresent their personality. Finally, in the *After* treatment, participants also have no material incentive to misrepresent their personality as the group formation rule is only revealed directly after the personality test. Strategic misrepresentation could reduce the effectiveness of the group formation rule in increasing contributions in the PGG due to the compression of Agreeableness scores and potential mistrust. Whereas, knowledge of the rule could increase its effectiveness by reducing uncertainty about group members' cooperativeness, suggesting *After* > *Never* > *Before* in terms of effectiveness. The second treatment dimension is the group formation rule itself. Groups are typically randomly assigned in economics experiments, which makes a *Random* treatment (henceforth *RAND*) a natural baseline for the *Agreeableness* group formation rule (henceforth *AGRE*). The experiment is a 3x2 design, so participants in the *RAND* treatment also have the group formation rule revealed to them either *Before* or *After* the personality test, or *Never*, meaning they are clean comparisons for their respective *AGRE* treatment.

1.1. Related Literature

One relevant area of research in the PGG has centered on mechanisms or interventions aimed at increasing contributions in the PGG. Examples include allowing for punishment (Fehr & Gächter, 2000), endogenous group formation (Ahn et al., 2009; Charness & Yang, 2014), inducing a group identity with a team-building task (Charness et al., 2014), or priming with words related to cooperation (Drouvelis et al., 2015). Another strand of PGG research sorts groups based on previous contribution behavior, and found that this type of sorting is effective (Burlando & Guala, 2005; Gächter & Thöni, 2005; Gunnthorsdottir et al., 2007; Ones & Putterman, 2007). I contribute by studying an intervention that sorts using a personality test. I vary the timing of information about the sorting rule, whereas previous studies on sorting hold this constant, and typically do not reveal it.

Another relevant strand of literature considers how individual characteristics are related to behavior in social dilemma games like the PGG. Of particular interest are studies that elicit the Big Five personality trait of Agreeableness. In terms of individual studies in economics, Volk et al. (2012) and Perugini et al. (2010) find Agreeableness to be correlated with contribution behavior in the one-shot and repeated PGG respectively. Whereas in the related repeated Prisoner's Dilemma, Kagel and McGee (2014) find a positive correlation of Agreeableness with cooperation, while Proto et al. (2019) observe this only in

[1] Thielmann et al. (2020) report $\hat{\rho} = 0.12$, $p < 0.001$ for all games (pg. 51, Table 7); $\hat{\rho} = 0.07$, $p < 0.05$ for all social dilemmas (pg. 54, Table 8); $\hat{\rho} = 0.09$, $p = 0.001$ for all PGGs (Table S10); but $\hat{\rho} = 0.03$, $p > 0.05$ (Table S14) for repeated PGGs. The last (lack of) result was not noticed until after the study was complete.

[2] In terms of actual job performance, successful traits naturally vary by job type, although Conscientiousness is a consistently strong predictor (Sackett et al., 2022; Zell & Lesick, 2022).

early periods. Additionally, Gill and Rosokha (2024) find that the trust facet of Agreeableness is related to cooperation through learning. In terms of comparable studies in psychology (i.e. incentivized with no deception), Corr et al. (2015) finds a positive relationship between Agreeableness and cooperation in a one-shot PGG, while Glöckner and Hilbig (2012) finds no relationship in a one-shot Prisoner's Dilemma. A meta-analysis over all fields finds a positive relationship between Agreeableness and prosocial actions in one-shot social dilemma games, but not when they are repeated (Thielmann et al., 2020).

This paper also contributes to the literature on faking personality tests in general, and the implications for using them in job-hiring. These studies can be grouped into two categories: 'fake-good' (or bad) studies, where participants are explicitly instructed to misrepresent themselves, and 'applicant-incumbent' studies, where job-seekers' responses are compared to those already in a similar job. In fake-good studies, meta-analyses by Viswesvaran and Ones (1999) and Walker et al. (2022) suggest that people can misrepresent their personality in Big Five and Dark Triad personality tests respectively. On applicant-incumbent studies, meta-analyses suggest job-seekers do misrepresent themselves on Big Five personality tests (Birkeland et al., 2006; Hu & Connelly, 2021). I contribute to this literature by: indirectly incentivizing misrepresentation through PGG group type, considering whether suspected misrepresentation contaminates subsequent work cooperation, and exploring the counterfactual where job-hiring is through a personality test but misrepresentation is ruled out (*AGRE After*).

The closest related paper is by McGee and McGee (2024). In their experiment, participants' Big Five traits were measured in a baseline session, then again a week later. Before the second test, participants were offered an extra payment if 'hired' for a hypothetical job based partly on the second Big Five test, using a description that implied a specific trait was ideal. They find that participants misrepresent their personality in the presence of incentives. My approach is complementary but distinct, as I consider how misrepresentation could impact subsequent behavior, and test misrepresentation in a between-participant design as is common in the dishonesty literature (Fischbacher & Föllmi-Heusi, 2013).

Two other relevant papers are by Drouvelis and Georgantzis (2019) and Cartwright et al. (2023). Drouvelis and Georgantzis (2019) have participants complete a Big Five personality test and then play a Dictator Game and a two-player one-shot PGG. They pair participants by their Agreeableness (High, Low, or Mixed), and tell this to the pair, except in a baseline no info treatment. They find contributions are higher when participants are told their pair is High compared to Low or Mixed. However, they find no difference in contributions by Agreeableness group type when no info is given. Cartwright et al. (2023) study a repeated four-player PGG that takes place after a Social Value Orientation (SVO) task (Murphy et al., 2011). SVO overlaps with Agreeableness and is related to cooperative behavior (Pletzer et al., 2018). They vary by treatment what info is given about other group members' SVO, and only find differences in contributions between pro-social and pro-self types when at least some information about SVO is given. My main additional contribution is permitting strategic misrepresentation of Agreeableness - both of these designs are somewhat similar to my *After* or *Never* treatments. Another difference is that I do not provide info on specific composition of a participant's group (i.e. high or low Agreeableness).

My experiment also tests whether 'unexpected data use' (Charness et al., 2022) influences participants' future decisions. Unexpected data use is when responses are used in a way not described to participants when they provided that data. This could cause problems for a similar reason as to why deception is not used in economics experiments - a loss of control over participants' beliefs and expectations (Cason & Wu, 2019; Cooper, 2014). If a participant does not believe all of what they are told in experiments, then they would not always reveal what they would do if the situation was exactly as described. For example, previous unexpected data use may cause participants to behave more pro-socially, as they anticipate their decision could have some additional future consequence, like being revealed to others. The experimenter would then be unable to observe a participant's true pro-social preferences for the specified decision environment. Charness et al. (2022) find that researchers consider unexpected data use as useful and not deceptive, but that student participants' views differ, making it important to examine whether participants change their behavior after its use.

2. Experiment

All aspects of this experiment and the statistical analysis were pre-registered unless otherwise stated. The experiment consists of three parts that are common to all treatments. Part 1 is a Big Five questionnaire, Part 2 is a PGG, and Part 3 is a short questionnaire that elicits four other personality traits. The first treatment dimension is how groups are formed in Part 2, the PGG. In the *RAND* treatments, groups of three are formed randomly from all participants in the session. In the *AGRE* treatments, participants are first randomly shuffled into silos of six. Within each silo, the three participants with the highest Agreeableness scores (as elicited in Part 1) are assigned to one group (henceforth the *H* group), while the remaining three are assigned to another group (the *L* group). If information on the Agreeableness group formation rule is provided, then participants are told the definition of Agreeableness, and how being in a high Agreeableness group could be beneficial for the Part 2 PGG (see Section 3.3). The second treatment dimension is the timing of when information about the group formation rule is provided. This is either *Before* Part 1, *After* Part 1 but before Part 2, or *Never*. A full 3x2 factorial design is conducted. The main outcome variables of interest are the Agreeableness scores as reported in Part 1, the Part 2 PGG contributions, and 'Positive Perception', an aggregation of socially-desirable responses (defined in Section 3.2.3) in the Part 3 personality questions.

2.1. Hypotheses

A list of all pre-registered Hypotheses is presented in Table 1. I now briefly describe the possible channels for each Hypothesis, and provide a more detailed description in Appendix B.

Table 1: Summary of Hypotheses

	Outcome Variable	Prediction
H1	Agreeableness	*AGRE Before* > *RAND Before*
H2	Agreeableness	*After & Never* > *RAND Before*
H3A	PGG Contribution	*AGRE H* > *RAND*
H3B	PGG Contribution	*AGRE L* < *RAND*
H3C	PGG Contribution	*AGRE H* > *AGRE L*
H4A	PGG Contribution	*AGRE Before H* < *AGRE After H*
H4B	PGG Contribution	*RAND Before* = *RAND After*
H4C	PGG Contribution	*AGRE Before L* > *AGRE After L*
H5A	PGG Contribution	*AGRE After H* > *AGRE Never H*
H5B	PGG Contribution	*RAND After* = *RAND Never*
H5C	PGG Contribution	*AGRE After L* < *AGRE Never L*
H6	Positive Perception	*AGRE After* > *AGRE Before*
H7	Positive Perception	*AGRE After* > *Unaware*
H8	Positive Perception	*AGRE Before* > *Unaware*

Notes: Treatment Groups: *After & Never* (effectively the same treatment during Part 1) = { *AGRE After*, *AGRE Never*, *RAND After*, *RAND Never* }, *Unaware* (that personality responses were/could be used in Part 2) = { *AGRE Never*, *RAND Before*, *RAND After*, *RAND Never* }

In the Part 1 questionnaire, there are effectively three treatment groups, *AGRE Before*, *RAND Before*, and *After & Never*, as only participants in the *Before* treatments have seen anything different. In *AGRE Before*, participants have an incentive to misreport and increase their Agreeableness scores if they want to be in (avoid) *H* (*L*) PGG groups in Part 2 (H1). Whereas, participants in *After & Never* know there is a Part 2 PGG, but have not been told about group formation. If they respond differently in Part 1, it would be due to the suspicion (which is sometimes correct) that their responses would be used in some way in Part 2, and I posit most anticipated uses would likely be in the direction of what is socially desirable (H2).

In the Part 2 PGG, I first test whether the Agreeableness group formation rule is effective through creating groups of different levels of Agreeableness. I compare PGG contributions between *H*, *L*, and *RAND* groups within their respective timing conditions separately, to avoid the possible timing confounds described below. As groups with higher Agreeableness are predicted to contribute more, then *H* groups should contribute more than *RAND*, and they in turn more than *L* (H3). I then test the same type of group (i.e. *H*, *L*, or *RAND*) across adjacent timing conditions. Comparing *Before* and *After*, I predict *H* groups in *Before* will contribute less due to mistrust from possible Agreeableness misrepresentation in Part 1 and due to compressed Agreeableness scores reducing the group formation rule's effectiveness by making groups more similar to *RAND* (H4), with the latter effect also increasing contributions in *L* groups. Comparing *After* and *Never*, I predict that those in *H* (*L*) groups in *After* will contribute more (less) due to (dis-)encouragement they are with similarly (dis-)Agreeable people, as they will believe they are less (more) likely to be taken advantage of (H5).[3]

In the Part 3 questionnaire there are three comparison groups of interest in terms of previously disclosed 'data use' of personality responses. The first group is participants in *AGRE Before*, the second group is participants in *AGRE After*, and the third group is all other participants, as they remain *Unaware* of potential data use of personality responses. The variable of interest for Part 3 is 'Positive Perception', a combination of all personality questions in Part 3 based on how positively they might be perceived by others (exact details in Section 3.2.3). Participants in *AGRE After* learn their Part 1 personality responses were used in an initially unannounced way. As a result, they may also believe that their Part 3 personality responses could also be used in some unannounced way (despite statements to the contrary), and report more socially-desirable responses than they otherwise would have (H6, H7). Whereas those in *AGRE After* had their Part 1 responses used in an announced way and thus would be more trusting of the experimenter that Part 3 responses will be used as stated (H6). However, knowing personality responses were used in some relevant way could still result in increased reported Positive Perception (H8).

3. Methods

3.1. Participants

There were 432 participants (182 male, 246 female, 4 other, $M_{age} = 23.2$, $SD_{age} = 4.7$, see Table C4 for full demographics by treatment), i.e., 144 groups of three. Each *RAND* treatment has observations from 16 groups of three, and each *AGRE* treatment has observations from 32 groups of three. I collected a different number of groups by treatment as observations in the *AGRE* treatments are split between *L* and *H* groups. I determined the number of participants based on what was feasible given the total research budget and then considered whether this number was appropriate using simulation based power analyses. I based this on whether the minimum detectable treatment effect at the 5% level with 80% power seemed reasonable given the number of participants in each treatment (see Appendix SE for more details). No definition of exactly what was 'reasonable' was pre-registered, only that I considered it so at that time. The simulated datasets imply the treatment comparisons that have the lowest power would have a minimum detectable effect size of $d = 0.52$, $d = 0.49$, and $d = 0.43$ for Agreeableness, Average PGG Group Contributions, and Positive Perception respectively. These are all about medium effect sizes ($d \approx 0.5$) assuming the assumptions were correct.[4]

Participants were German-speakers in either Innsbruck or Vienna, who had signed up to participate in economics experiments at the EconLab at the University of Innsbruck (UIBK) or the Vienna Center for Experimental Economics (VCEE) at the University of Vienna. 20 sessions (366 participants) were run at the UIBK EconLab, and 4 sessions (66 participants) were run at VCEE. Participants were recruited for

[3]The proposed channels for H4 and H5 can be similarly described as being through the fear of exploitation being shifted by beliefs of others' pro-social behavior (Hilbig et al., 2018; Tan et al., 2025).

[4]Given the results appear closer to a small effect size ($d \approx 0.2$) if anything, power in the pre-registered tests could be a problem in retrospect. Exploratory analysis that pools data based on the results in order to improve power are considered in Section 4.3. The pre-registration assessed 'reasonableness' based on absolute treatment effect sizes and not Cohen's d.

sessions using the online database hroot (Bock et al., 2014) at the UIBK EconLab, and ORSEE at VCEE (Greiner, 2015). Participants were paid based on their cumulative earned points over 15 rounds of the PGG described in Section 3.2.2. Points were converted at a rate of 1000 points = €3, and participants received a show-up fee of €4. The experiment was 45-60 minutes long with average earnings of €16.03.

3.2. Materials and Measures

3.2.1. Part 1 - Big Five Elicitation

Part 1 consists of 50 questions to elicit the Big Five personality traits (McCrae & John, 1992). These traits are Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each Big Five characteristic is elicited using the 30 question 'BFI-2-S Inventory' (Rammstedt et al., 2020; Soto & John, 2017). The remaining 20 questions are all on Agreeableness, and sourced from the International Personality Item Pool's (IPIP) '100-Item Lexical Big-Five Factor Markers' (Goldberg, 2002; Goldberg et al., 2006; Streib & Wiedmaier, 2001). The Agreeableness trait is disproportionately weighted (26/50) as it is of primary interest and used for group formation in Part 2 in the *AGRE* treatments. Participants are asked how much they agree each statement applies to them using a 5-point Likert scale (Likert, 1932). The 5 points are labeled: 1 = Disagree strongly, 2 = Disagree a little, 3 = Neither agree nor disagree, 4 = Agree a little, and 5 = Agree strongly. They are presented using horizontal radio buttons. Participants face blocks of five questions on a page, and all questions are presented in a random order drawn independently across participants. Personality traits are scored based on each participant's numerical (i.e. 1-5) responses by the following formula: $Trait = \frac{\sum_{i \in Q} Response \times I(+veKey_i) + (6 - Response) \times I(-veKey_i))}{n}$, where Q is the set of relevant questions to that trait and n is the size of set Q. See Appendix SB for further details on the personality trait questions and their scoring.

3.2.2. Part 2 - Public Goods Game

Part 2 consists of a PGG adapted from the version used by Lugovskyy et al. (2017). Groups of three are assigned by the group formation rule (i.e. randomly or by Agreeableness). Each group of three remains together for 15 'group cooperation decisions'. In each decision, each participant has 25 tokens they can allocate to either a Private account or a 'Cooperation' account. Each token a participant allocates to the Private account earns that participant 10 points. Each token a participant allocates to the Cooperation account earns each of the three group members (i.e. including the participant in question) 4 points each (i.e. 12 points to the group overall). I refer to tokens allocated to the Cooperation account as 'contributions'. Participants decide how many tokens to allocate to the Cooperation account, with the rest being allocated to their Private account. After making their decision, participants are given a summary of their own contribution and the total group contribution in that round. These round summaries are also available at any time during Part 2 in a history table that is displayed at the bottom of the screen.

3.2.3. Part 3 - Final Questionnaire

In Part 3, participants are first told they are to complete a final survey, and that their final earnings for the experiment have already been set. Participants then answer 16 personality questions and a demographic questionnaire. The 16 questions elicit the Dark Triad (Paulhus & Williams, 2002), and the Sincerity and Fairness facets of the Honesty-Humility trait from HEXACO (Ashton & Lee, 2009). The three Dark Triad measures are Machiavellianism (Christie & Geis, 1970), Narcissism (Raskin & Hall, 1979), and Psychopathy (Hare, 1985). The questions were elicited in the same format as in Part 1, with the exception that only 4 questions were shown on each page given there were 16 total questions. I combine all of the personality traits elicited in Part 3 into one measure based on how likely it is they would be positively perceived by others. I take the average of each participant's 16 responses to the Part 3 questions, where Dark Triad traits are reversed ($6 - Response$) and Honesty-Humility is left unchanged (after the initial positively- or negatively-keyed adjustment is made). I call this combined measure 'Positive Perception'.

After the personality questions, participants filled in their age, gender, field of study, GPA, years at uni, number of previous economics experiments, and their enjoyment of the experiment.

3.3. Procedures

All participants within a session faced the same treatment, which was randomly assigned.[5] In all treatments, participants were given a short overview of the PGG in Part 2 (see Appendix SD) before completing the Part 1 questions. If information about the group formation rule is provided (i.e. in *Before* or *After*), it is provided either directly before or directly after participants complete the Part 1 questions. In the *AGRE Before* and *After* treatments, participants are presented with the following message:

> For Part 2, you will be assigned to a group of three **based on your 'Agreeableness' score. Your Agreeableness score is determined by your responses to particular questions in Part 1.**
>
> Agreeableness is a personality trait where people high in Agreeableness are often described as *selfless, trusting, good-natured, generous, and forgiving.* (Costa, McCrae, & Dembroski, 1989)
>
> In scientific studies, **a high level of Agreeableness has been found to have a positive effect on group cooperation decisions** similar to the type in Part 2. [References button with pop-up window that listed the following info: Perugini, Tan, & Zizzo in Economic Issues, Volume 15, Part 1, 2010. Volk, Thöni, & Ruigrok in the Journal of Economic Behavior & Organization, Volume 81, Issue 2, 2012. Kagel & McGee in Economics Letters, Volume 124, Issue 2, 2014. Thielmann, Spadaro, & Balliet in Psychological Bulletin, Volume 146, Issue 1, 2020.]
>
> Each group of three is formed from six randomly selected subjects. **The three subjects with the highest Agreeableness scores will be assigned to one group, and the remaining three subjects to the other group.**

After all participants completed the Part 1 questions the computer formed groups by the group formation rule, and they made their Part 2 PGG contribution decisions, followed by the Part 3 questions.

4. Results

All analyses were conducted in Python using Stata 17 (`tobit`, `xttobit`, `metobit`) (StataCorp, 2021), SciPy (`mannwhitneyu`) (Virtanen et al., 2020), or Pingouin (`compute_effsize`, `compute_esci`, `cronbach_alpha`, `rcorr`, `pairwise_corr`) (Vallat, 2018).

4.1. Descriptive Results

A summary of the dependent variables by treatment and high or low Agreeableness group type (where applicable) is presented in Table 2. Table 3 reports the means, correlations, and Cronbach's alpha between all elicited personality traits and individual contributions in the PGG. Figure 1 displays a distributional summary of the two personality dependent variables, separated into their relevant treatment groups. Figure 1 suggests any treatment effect, if present, is quite small.

Figure 2 graphically illustrates average contributions over time by treatment. Figure 2 displays declining contributions over time, which is typical in this type of PGG. Figure 2 visually suggests that *AGRE Before* and *AGRE After* increases contributions by about 2-4 tokens on average when compared to *RAND*, and that this level shift is sustained over time. However, this increase is observed regardless of whether the group is of *H* or *L* Agreeableness. In addition, *AGRE Never* appears to be ineffective as it is quite close to *RAND*. From the patterns in Figure 2, it follows that the Agreeableness group formation rule does not increase contributions by creating groups with high levels of Agreeableness.

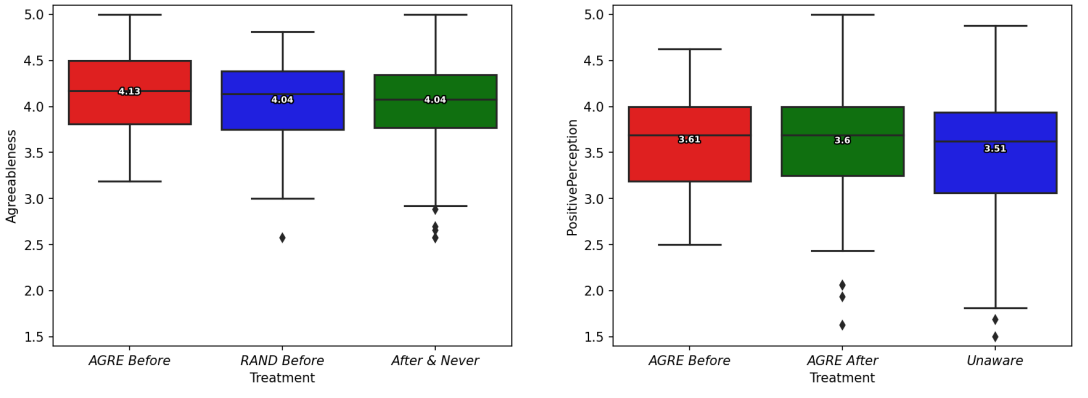
[5] Detailed procedures for a session are described in Appendix SC. The experiment was computerized using oTree (D. L. Chen et al., 2016).

Table 2: Summary Statistics by Treatment and Agreeableness Group Type

	RAND Before	*RAND After*	*RAND Never*	*AGRE Before*	*AGRE After*	*AGRE Never*
Part 2. Avg. PGG Group Cont. [Low Agr. High Agr.]	6.91 (6.62)	8.00 (7.70)	5.65 (5.34)	11.25 (7.86)	8.79 (7.93)	7.21 (7.21)
				10.35 (8.09)	9.37 (7.64)	6.81 (6.00)
Part 1. Agreeableness	4.04 (0.47)	4.09 (0.38)	4.01 (0.53)	3.81 (0.32)	3.62 (0.35)	3.81 (0.35)
				4.45 (0.24)	4.33 (0.28)	4.38 (0.28)
Part 3. Positive Perception	3.53 (0.60)	3.45 (0.65)	3.58 (0.54)	3.49 (0.53)	3.44 (0.63)	3.44 (0.63)
				3.72 (0.49)	3.77 (0.55)	3.56 (0.58)
Num. Participants	48	48	48	48	48	48
				48	48	48

Note: Mean given, then standard deviation in parentheses. Low and High Agreeableness groups are reported separately for *AGRE* treatments, with Low Agreeableness groups reported first (on the top) in a cell, and High Agreeableness groups reported second (bottom). Avg. PGG Group Cont. = Average PGG Group Contribution $\in [0, 25]$ and reported at the group per round level, meaning $(48 \div 3) \times 15 = 240$ observations.

Figure 1: Personality Dependent Variables by Treatment Group



(a) Part 1 Agreeableness

(b) Part 3 Positive Perception

Mean value overlaid. The three lines in the box are the 75%, 50% and 25% quartile when going from top to bottom, the top (bottom) whisker is the largest (smallest) value that is below (above) 1.5 times the difference between the 75% and 25% quartiles, and values outside this range are diamonds.

4.2. Pre-registered Analysis

This section presents all analysis that was in the main body of the pre-registration, except for unsophisticated misrepresentation which is now in Appendix SA.3.1. This section also presents the analysis of individual traits on PGG contributions which was in the appendix of the pre-registration.

Table 3: Means and inter-correlations between personality traits and PGG contributions

	M (SD)	1	2	3	4	5	6	7	8	9
1. Agreeabl.	4.06 (0.45)	*0.88*								
2. Openness	3.66 (0.70)	0.13***	*0.69*							
3. Neurot.	2.63 (0.80)	-0.11**	-0.09*	*0.80*						
4. Extrav.	3.43 (0.73)	0.16***	0.20***	-0.33***	*0.76*					
5. Conscie.	3.59 (0.69)	0.20***	0.08*	-0.22***	0.20***	*0.74*				
6. Hon. Hum.	3.16 (0.80)	0.13***	0.10**	0.06	-0.15***	0.07	*0.47*			
7. Machiav.	2.06 (0.87)	-0.27***	-0.00	-0.10**	0.23***	-0.07	-0.38***	*0.79*		
8. Psychop.	2.02 (0.77)	-0.48***	-0.06	-0.07	0.06	-0.09**	-0.14***	0.54***	*0.64*	
9. Narcissism	2.87 (0.90)	-0.05	0.09*	0.08*	0.18***	-0.03	-0.19***	0.42***	0.27***	*0.74*
PGG Cont.	8.26 (6.38)	0.03	0.18***	-0.02	-0.02	-0.13***	0.05	-0.03	-0.04	-0.01

Note: Cronbach's alpha given in italics in the diagonal. Pearson correlation coefficient in the off diagonal, with ***=$p < 0.01$, **=$p < 0.05$, *=$p < 0.10$. Pooled observations over all treatments ($n = 432$). PGG Cont. is averaged at the individual level. Abbreviations: Agreeabl. = Agreeableness, Openness = Openness to experience, Neurot. = Neuroticism, Extrav. = Extraversion, Conscie = Conscientiousness, Hon. Hum. = Honesty Humility, Machiav. = Machiavellianism, Psychop. = Psychopathy, PGG Cont. = PGG Contributions.

4.2.1. Part 1: Strategic Misrepresentation of Agreeableness

For misrepresentation of Agreeableness, there are three comparison groups: *AGRE Before*, *RAND Before*, and all *After* & *Never* treatments, because treatments can only impact Part 1 responses if they differ before Part 1. I find that Agreeableness in *RAND Before* ($M = 4.04$, $SD = 0.47$) is not statistically significantly different than in *AGRE Before* ($M = 4.13$, $SD = 0.43$, $U_{48,96} = 2090.5$, $p = 0.37$, $d = 0.21$ 95% $CI = [-0.14, 0.56]$), suggesting people are not substantially misrepresenting their Agreeableness when they know PGG groups will be formed based on this trait. I also find that Agreeableness in *RAND Before* ($M = 4.04$, $SD = 0.47$) is very similar to *After & Never* ($M = 4.04$, $SD = 0.46$, $U_{48,288} = 7062$, $p = 0.81$, $d < 0.01$ 95% $CI = [-0.31, 0.31]$), suggesting people do not change their answers in response to suspicion about how they might be used in future parts. Finally, the comparison between *AGRE Before* and *After & Never* also finds no statistically significant difference ($p = 0.10$, $d = 0.20$, $U_{96,288} = 15369$, 95% $CI = [-0.03, 0.44]$). In Appendix SA.3.1 I also consider 'unsophisticated' misrepresentation of the other Big Five traits, but find no differences in most traits.

4.2.2. Part 2: PGG Contributions

To test for treatment differences between two comparison groups, I first drop all observations except for those two comparison groups, and create a dummy variable called Treatment that is 1 if the observation belongs to one of the comparison groups and 0 otherwise. I then conduct a panel Tobit regression over PGG groups and rounds, with random effects for PGG groups and a linear time trend. The dependent variable is the average number of tokens contributed by the 3 PGG group members in a given round, which is possibly censored at 0 and 25, hence the use of the Tobit. The regression specification is as follows: $Average\ PGG\ Contribution_{g,r} = \beta_0 + \beta_1 Treatment_g + \beta_2 Round + u_g + \epsilon_{g,r}$.

Figure 2: Average Contributions by Round

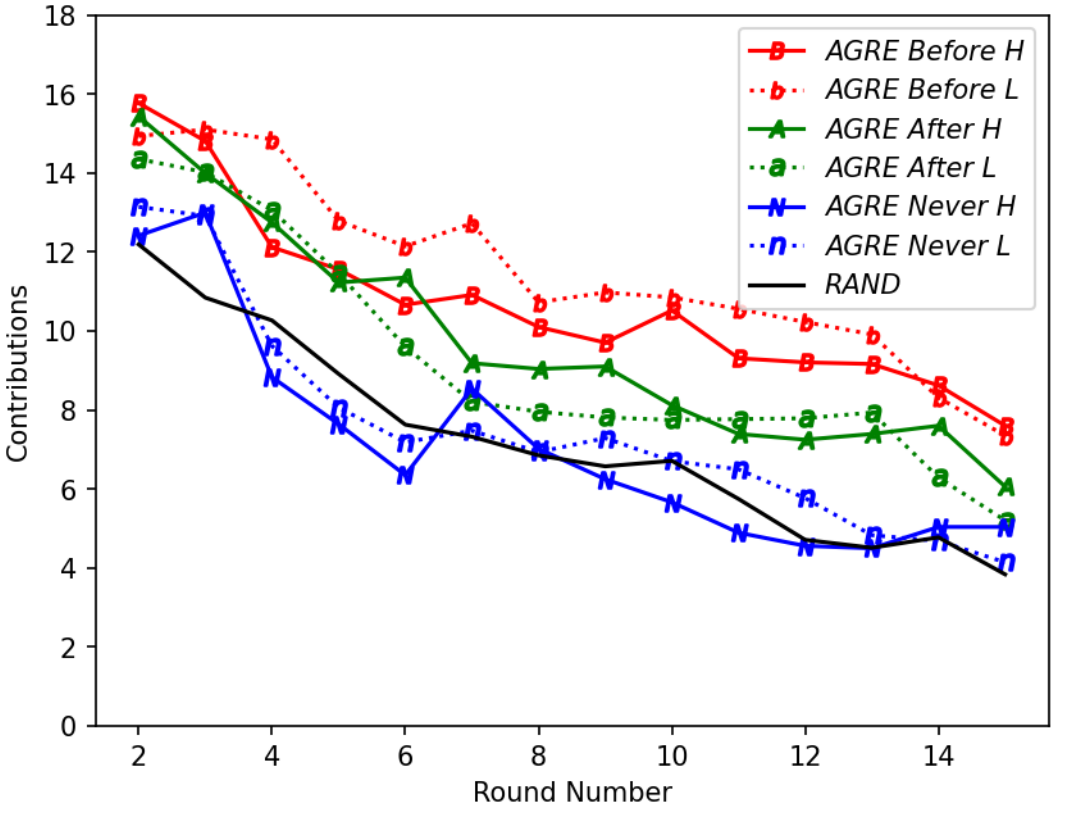


Table 4 summarizes the results from tests of the numbered Hypotheses on PGG contributions. The only statistically significant results are between the *RAND Before* treatment and those in either the *H* group or the *L* group of the *AGRE Before* treatment. Table 4 provides no support for any of H3, H4, or H5, suggesting three things. Firstly, group formation by Agreeableness is generally ineffective in changing contributions in the PGG - both *H* and *L* Agreeableness groups contribute similarly within each of the *Before*, *After*, and *Never* treatments (H3). Secondly, the possibility of strategic misrepresentation of personality in Part 1 due to Part 2 group incentives does not affect contributions in the PGG - *H*(*L*) groups in *AGRE Before* do not contribute less (more) than *H*(*L*) groups in *AGRE After* (H4). Thirdly, there is no (dis-)encouragement effect of knowing the group has similarly high (low) Agreeableness - *H*(*L*) groups in *AGRE After* do not contribute more (less) compared to *H*(*L*) groups in *AGRE Never* (H5).

4.2.3. Part 3: Positive Perception

There are three relevant comparison groups, 'expected' or pre-announced data use (of Part 1 personality responses) (*AGRE Before*), unexpected data use (*AGRE After*), and all treatments where participants remain *Unaware* of data use. I find no difference between Positive Perception in *AGRE Before* ($M = 3.60, SD = 0.52$) and *AGRE After* ($M = 3.61$, $SD = 0.61$, $U_{96,96} = 4511.5$, $p = 0.80$, $d = 0.01$ 95% $CI = [-0.28, 0.29]$), suggesting that unexpected data use does not affect responses to the Part 3 personality questions when it was known that Part 1 responses affected group composition in Part 2. I also find no evidence that Positive Perception in *AGRE After* is higher than in *Unaware* ($M = 3.51$, $SD = 0.60$, $U_{96,240} = 12657.5$, $p = 0.25$, $d = 0.16$ 95% $CI = [-0.08, 0.40]$), which means it is

Table 4: Efficiency - Regressions

Pairwise Comparison	H	B	SE	z	p	d
AGRE Before H - RAND Before	H3 +	4.34	2.53	1.72	0.09	0.53
AGRE Before L - RAND Before	H3 −	5.19	2.41	2.15	0.03	0.65
AGRE Before H - AGRE Before L	H3 +	-0.81	2.79	-0.29	0.77	-0.09
AGRE After H - RAND After	H3 +	2.18	2.54	0.86	0.39	0.26
AGRE After L - RAND After	H3 −	1.36	2.61	0.52	0.60	0.16
AGRE After H - AGRE After L	H3 +	0.85	2.54	0.34	0.74	0.10
AGRE Never H - RAND Never	H3 +	1.65	1.63	1.01	0.31	0.29
AGRE Never L - RAND Never	H3 −	1.59	2.01	0.79	0.43	0.23
AGRE Never H - AGRE Never L	H3 +	0.16	1.93	0.08	0.93	0.02
AGRE Before H - AGRE After H	H4 −	1.26	2.69	0.47	0.64	0.15
AGRE Before L - AGRE After L	H4 +	2.97	2.64	1.12	0.26	0.35
AGRE After H - AGRE Never H	H5 +	2.78	2.06	1.35	0.18	0.40
AGRE After L - AGRE Never L	H5 −	2.19	2.45	0.89	0.37	0.27
RAND Before - RAND After	H4 ∼	-0.92	2.37	-0.39	0.70	-0.12
RAND After - RAND Never	H5 ∼	2.42	2.20	1.10	0.27	0.33

Note: For brevity, only the statistics of the comparison dummy variable are reported. Second group in the pair is the omitted dummy. Each block groups the comparisons within the *Before*, *After*, *Never*, *AGRE*, and *RAND* treatments respectively. For (H)ypotheses: +, −, and ∼ indicate a positive, negative, or neutral predicted effect respectively. Full regression output reported in Appendix SF.1.1.

unlikely that unexpected data use affects Part 3 responses compared to the case where it is unknown if/how Part 1 responses is used in Part 2. Finally, I find no statistically significant difference between Positive Perception in *AGRE Before* and in *Unaware* ($U_{96,240} = 12435.5$, $p = 0.25$, $d = 0.16, 95\%\ CI = [-0.08, 0.39]$), implying that the knowledge that Part 1 responses were used in creating Part 2 groups does not change responses to Part 3 questions when that usage was expected (i.e. revealed before Part 1).

4.2.4. Agreeableness and PGG Contributions

The results suggest that the Agreeableness group formation rule is not effective in increasing contributions by creating groups with higher levels of Agreeableness. This follows from the observation that there is no difference between *H* and *L* groups, and that the *AGRE Never H* group contributes similarly to those in *RAND*. Creating higher contribution groups through sorting by Agreeableness relies on the assumption that individuals with higher Agreeableness contribute more.

To see whether this is the case, I regress an individual's personality traits and demographics on their contributions in the PGG. The regression is a panel Tobit regression on individual PGG contributions in each round, censored at 0 and 25 (the min and max possible contribution), with individual- and group-level random effects. It has dummies for each treatment (with *RAND Before* as the omitted dummy) and fits a linear trend over rounds. It includes the lagged average contribution of the other two group members, where the first round value is found by which integer value maximizes the log-likelihood (Bardsley & Moffatt, 2007). It then includes all elicited Part 1 and Part 3 personality scores, and dummy variables for gender (male and other as omitted variable), year at uni (first year as the omitted dummy), and main subject of study (Mgmt./Business as the omitted dummy). Finally, all observations from *AGRE Before* are excluded, as misrepresentation of Agreeableness could be present. Table 5 reports the output from this regression, and suggests that an individual's Agreeableness is uncorrelated with their contribution behavior. This lack of correlation is robust to a variety of alternative specifications, like the pairwise correlation test reported in Table 3, and additional tests reported in Appendix SA.2. This result is in line with the meta-analysis of Thielmann et al. (2020), who find no evidence of a relationship between Agreeableness and pro-social behavior in social dilemma games when play is repeated, with the latter

proving to be an important caveat. With this in mind, it is unsurprising then that the Agreeableness group formation rule by itself proved ineffective in increasing group contributions in the repeated PGG.

Table 5: Individual Characteristics on Contributions

Variable	B	SE	z	p
RAND After	0.81	2.98	0.27	0.78
RAND Never	-2.65	2.97	-0.89	0.37
AGRE After H	2.18	2.98	0.73	0.46
AGRE Never H	0.16	2.98	0.05	0.96
AGRE After L	0.88	3.02	0.29	0.77
AGRE Never L	-0.06	2.97	-0.02	0.98
Round***	-0.93	0.04	-24.99	< 0.01
Lagged Avg. Group Cont.***	0.55	0.05	11.10	< 0.01
Agreeableness	0.10	1.13	0.09	0.93
Openness to experience***	1.72	0.55	3.12	< 0.01
Neuroticism	0.67	0.53	1.26	0.21
Extroversion	0.78	0.56	1.39	0.16
Conscientiousness***	-1.59	0.54	-2.93	< 0.01
Honesty Humility	-0.27	0.50	-0.55	0.58
Machiavellianism	-0.37	0.61	-0.61	0.54
Narcissism	-0.52	0.46	-1.12	0.26
Psychopathy	-0.34	0.61	-0.55	0.58
Female	-0.59	0.82	-0.72	0.47
2nd Year at Uni.	0.25	1.42	0.17	0.86
3rd Year at Uni.*	-2.34	1.40	-1.67	0.09
4th+ Year at Uni.	-1.67	1.41	-1.18	0.24
Grad. Student	-2.26	1.53	-1.48	0.14
GPA	0.27	0.39	0.68	0.49
Economics	-0.07	1.48	-0.04	0.96
Arts and Humanities	1.78	1.97	0.90	0.37
Natural Sciences*	3.29	1.70	1.94	0.05
Education	3.23	2.26	1.42	0.15
Engineering	4.08	2.68	1.52	0.13
Law	0.09	2.31	0.04	0.97
Social Sciences	0.63	1.85	0.34	0.73
Medicine	0.18	2.36	0.08	0.94
Other	1.05	1.64	0.64	0.52

Note: Results are from a multilevel panel Tobit regression (censored at 0 and 25) with individual and group-level random effects. An individual's contribution to the public good per round is the dependent variable. All observations from *AGRE Before* excluded. Full regression output and details are reported in Appendix SF.1.2. ***=$p < 0.01$, **=$p < 0.05$, and *=$p < 0.10$.

In terms of the other personality traits, Table 5 suggests that higher levels of Openness to experience are associated with higher contributions. These findings should be interpreted through the lens of the broader literature. Openness to experience has been proposed as a possible factor in cooperative prosociality by Lawn et al. (2022), although they found only modest correlations in one-shot PGGs, and while not directly reported in the meta-analysis of Thielmann et al. (2020), their dataset suggests little to no relationship in repeated social dilemmas. Table 5 also suggests that Conscientiousness is negatively correlated with contributions. While Proto et al. (2019) find a similar effect in a repeated Prisoner's Dilemma and propose this could be driven by the Cautiousness sub-facet, the meta-analysis of Thielmann

et al. (2020) suggests no relationship between Conscientiousness and behavior in social dilemmas. In the absence of pre-registered theoretical predictions these results are likely false positives, although they do contribute to future meta-analyses as there is a relative dearth of studies considering personality and repeated social dilemmas (Thielmann et al., 2020). Despite these results, individual differences do play an important role in cooperation, as (not pre-registered) likelihood ratio tests of the specification in Table 5 strongly support including individual level random effects ($\chi^2_1 \geq 233.15$, $p < 0.01$).

4.3. Exploratory Analysis

4.3.1. Agreeableness Misrepresentation

The results imply that there may be a small amount of misrepresentation ($d \approx 0.20$) by those in *AGRE Before*. However, the ex-ante power analysis suggests that only $d \geq 0.49$ is sufficiently powered. Ex-post there is little that can be done to increase power. However, based on the results, pooling observations from *RAND Before* and *After & Never* seems appropriate as Agreeableness is very similar between these two groups ($d < 0.01$). This yields a marginally insignificant two-sided result ($U_{96,336} = 17886.5$, $p = 0.10$, $d = 0.20$ 95% $CI = [-0.02, 0.43]$), while additionally conducting the better-powered t-test yields a significant result at the 10% level ($t_{430} = 1.75$, $p = 0.08$). Given H1 was directional, it could also be reasonable to use a one-sided p-value, which would then be significant at the 5% level.

Agreeableness was elicited using two different question sets: 20 from the IPIP's 100-Item Lexical Big-Five Factor Markers and 6 from the BFI-2-S which are then combined (see Section 3.2.1). However, as these inventories were not explicitly designed to be combined in such a manner, I also consider them separately. Using the same tests as were pre-registered, I find no evidence of misrepresentation when using the IPIP scale (e.g. the lowest p-value is comparing *AGRE Before* ($M = 4.14$, $SD = 0.44$) to *After & Never* ($M = 4.07$, $SD = 0.47$, $U_{96,288} = 14999$, $p = 0.21$, $d = 0.16$ 95% $CI = [-0.07, 0.40]$)). For the BFI-2-S Agreeableness scale, there is evidence that participants in the *AGRE Before* ($M = 4.10$, $SD = 0.48$) report higher Agreeableness than those in *RAND Before* ($M = 3.89$, $SD = 0.54$, $U_{48,96} = 1834$, $p = 0.05$, $d = 0.43$ 95% $CI = [0.07, 0.78]$)), and those in *After & Never* ($M = 3.96$, $SD = 0.54$, $U_{96,288} = 15874$, $p = 0.03$, $d = 0.28$ 95% $CI = [0.05, 0.51]$)).

In summary, there may be some evidence to suggest strategic misrepresentation of Agreeableness, and the null result from the pre-registered analysis may be due to low power or the specific Agreeableness metric used. However, even if there is misrepresentation, its effect size is usually small ($d \approx 0.20$) and may not be empirically relevant. Also, comparing multiple metrics for Agreeableness could yield false positives due to the multiple comparisons problem. For example, the comparison of *RAND Before* and *After & Never* for the BFI-2-S Agreeableness scale is no longer statistically significant at the 5% level when correcting for a family of three comparisons (i.e. one for each tested metric) by either the Bonferroni-Holm (Holm, 1979) or Benjamini and Hochberg (1995) procedures. Furthermore, while the analysis in the above two paragraphs may seem reasonable, they were not pre-registered and could be seen as an example of how one could unintentionally be drawn to analysis that yields $p < 0.05$.

4.3.2. Increased Contributions in Agreeableness Before and After

The results show that the Agreeableness group formation rule does not improve contributions by grouping people with high Agreeableness together, as was initially proposed. Rather, as both *H* and *L* groups contribute similarly at increased levels, it must be something on the additional screen (described in Section 3.2.2) that increases contributions. To test whether this is the case, I pool observations from both *H* and *L* groups, in *AGRE Before* and *AGRE After*, and compare them to participants in all other treatments. Using the same statistical test as in Section 4.2.2, I find contributions are higher in *AGRE Before & After* than in the other treatments ($B = 3.74$, $SE = 1.13$, $Z = 3.31$, $p =< 0.01$, $d = 0.48$). This result is robust at the 10% significance level to different aggregations of *H* and *L* groups such

as considering each timing condition separately (see Appendix SF.2.1), except when comparing only *AGRE After* to only *RAND After* ($B = 1.76$, $SE = 2.22$, $Z = 0.79$, $p = 0.43$, $d = 0.21$).

5. Discussion

5.1. Possible Operative Channels

One finding of this experiment is the lack of substantial misrepresentation of Agreeableness in Part 1 when it is known Part 2 PGG groups will be formed based on this trait. There are multiple possible explanations as to why this is the case. Firstly, Section 4.3.1 suggests low power may be behind this result. In addition, Figure 1 illustrates that average Agreeableness is already quite high relative to its maximum value, meaning low power could be exasperated by a ceiling effect. Secondly, a preference for honesty (see Abeler et al. (2019) for a meta-study) could outweigh the indirect benefits of being in an *H* group. Thirdly, it could be that participants were unable to misrepresent their personality. This seems unlikely given the previous literature, e.g. McGee and McGee (2024) find misrepresentation in a more difficult task of inferring the personality trait in question from a job description.

The results show strong evidence of increased PGG contributions in the *AGRE Before* treatment and some evidence of increased contributions in *AGRE After* treatment, regardless of whether the group was of high or low Agreeableness. Alongside the lack of effect in *AGRE Never*, where participants were not told they were grouped by Agreeableness, the increased contributions must be driven by the additional text those in the *AGRE Before* and *AGRE After* treatments see (described in Section 3.3). This text states group formation is done by Agreeableness calculated from Part 1 responses, and defines people with high Agreeableness as being *'selfless, trusting, good-natured, generous, and forgiving'*. It then states scientific studies suggest a positive relationship between Agreeableness and group cooperation decisions, and that high and low Agreeableness groups will be created from subsets of 6 people. A variety of explanations could explain why this text increases contributions in the PGG. Being grouped with those with similar personality scores could induce a group identity (Y. Chen & Li, 2009; Eckel & Grossman, 2005), or the test be seen as a team-building exercise (Charness et al., 2014). The positive words used to describe Agreeableness could prime cooperation (Drouvelis et al., 2015), or as the positive words are linked to contributions, people may contribute to maintain or build a positive self-image (Bénabou & Tirole, 2006, 2011). Another possibility is an experimenter demand effect (Zizzo, 2010) through framing the PGG as being about cooperative decisions. While such framing is likely to increase contributions (Dufwenberg et al., 2011), all treatments have the public good framed as a 'cooperation account', while the additional text only repeats this link. Finally, it could be that most participants thought they were in the *H* group, as this experiment does not reveal this type of info, while previous studies only found differences when such info was given (Cartwright et al., 2023; Drouvelis & Georgantzis, 2019).

These potential operative channels have implications for using personality testing in hiring decisions, assuming external validity. Firstly, a preference for honesty (or against lying) would limit the threat that strategic misrepresentation poses. Getting a job can be very high-stakes for an individual, which would suggest a higher levels of dishonest behavior (e.g. Gneezy and Kajackaite, 2020; Hilbig and Thielmann, 2017; Kajackaite and Gneezy, 2017) and thus more misrepresentation. However, there is also evidence that the 'costs' of lying increase when considering the distance between the truth and what someone actually states (Abeler et al., 2019; Hilbig & Hessler, 2013), suggesting the job seeker may misrepresent their personality somewhat, but not to a large extent. Secondly, the increased contributions of those who knew the personality test determined their group, regardless of *H* or *L*, suggests that using personality tests in hiring may have an additional positive channel beyond any predictive power it may have for job suitability. Having to do the personality test may induce a group identity, or a common belief that coworkers either share similar personalities or should be well-suited to their jobs.

5.2. Limitations and Future Research

The most important limitation is the use of Agreeableness as the personality trait for PGG group formation. While a meta-analysis (Thielmann et al., 2020) shows Agreeableness is positively related with pro-social actions overall and in one-shot social dilemmas, it provides no evidence that it affects behavior in repeated social dilemmas like the repeated PGG used in this paper. Future research could consider if sorting by other personality traits increases contributions in a repeated PGG. For a repeated social dilemma, better candidates could be Pro-environmentalism (Schultz et al., 2004) or the HEXACO version of Agreeableness (Ashton & Lee, 2009). These traits have the strongest correlations in Figure 7 of Thielmann et al. (2020), although with the caveat that there are relatively few studies that assess those traits in a repeated setting, and that the PGG would likely need to be framed for Pro-environmentalism.

Another limitation of this research is that it was not practical to decompose all possible channels, foreseen or otherwise. For example, the posited compression of Agreeableness scores due to incentives in *AGRE Before* could have also been achieved by adding a random number to scores prior to sorting (removing malicious intent), or each of element of the additional text in *AGRE Before* and *AGRE After* could be included or excluded. A full decomposition would be ruinous in terms of sample size constraints, particularly given power in the current study seems a bit low ex-post.

Finally, while this study imitates the important parts of the work hiring process using personality tests followed by a cooperative job, it is a lab experiment. Lab experiments can be a cost-effective tool to evaluate firm personnel policies in an environment without real-world confounds or complexities (Villeval, 2016). However, they also have limitations like limited external validity, abstract settings (Schram, 2005), relatively low stakes (Harrison, 1989), experimenter demand effects, and self-selected, homogeneous student participant pools (Henrich et al., 2010; Henry, 2008; Kaźmierczak et al., 2023).

6. Conclusion

Using psychometric personality testing when there may be incentives to misrepresent personality is a complex topic. To shed light on this issue, I design and conduct an incentivized laboratory experiment. I first elicit Big Five characteristics, and then conduct a repeated PGG. I create groups for the PGG based on either on the Big Five trait of Agreeableness, or randomly. By changing the timing of the revelation of the sorting rule to before or after the initial questionnaire, I am able to explore misrepresentation and evaluate any subsequent impact on cooperative behavior.

I find that participants do not substantially misrepresent their personality to get into groups with higher levels of Agreeableness. I also find that the Agreeableness group formation rule increases contributions in the *AGRE Before* and *AGRE After* treatments, but not in the *AGRE Never* treatment. In addition, the increase in contributions occurs for both *H* and *L* groups. Lastly, I find that using personality tests in an unannounced way does not affect subsequent personality tests. Withholding information about the group formation rule did not influence later behavior, suggesting that experimental control was maintained.

Acknowledgments. A previous version of this paper circulated under the title ‘Personality Testing and the Public Goods Game’. I thank Annika Kieninger and Matthias Waldauf for excellent research assistance. Helpful comments and feedback were received from the editor Isabel Thielmann and anonymous referees, as well as Tim Cason, Raphael Epperson, David Gill, Elisabeth Gsottbauer, Stanton Hudja, Christian König-Kersting, and Andrew McGee.

Funding Statement. Funding from the IFREE Small Grants Program is gratefully acknowledged.

Competing Interests. None.

Data Availability Statement. The pre-registration is available at https://doi.org/10.17605/OSF.IO/YWM64 and the data and analysis code is available at http://doi.org/10.17605/OSF.IO/MDB7A.

Ethical Standards. This research received ethical approval from the University of Innsbruck (54/2023).

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, *87*(4), 1115–1153. https://doi.org/10.3982/ECTA14673

Ahn, T. K., Isaac, R. M., & Salmon, T. C. (2009). Coming and going: Experiments on endogenous group sizes for excludable public goods. *Journal of Public Economics*, *93*(1-2), 336–351. https://doi.org/10.1016/j.jpubeco.2008.06.007

Ashton, M., & Lee, K. (2009). The HEXACO-60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment*, *91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Bardsley, N., & Moffatt, P. G. (2007). The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, *62*(2), 161–193. https://doi.org/10.1007/s11238-006-9013-3

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology*, *92*(3), 595–615. https://doi.org/10.1037/0021-9010.92.3.595

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5).

Bénabou, R., & Tirole, J. (2011). Identity, Morals, and Taboos: Beliefs as Assets. *The Quarterly Journal of Economics*, *126*(2), 805–855. https://doi.org/10.1093/qje/qjr002

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://www.jstor.org/stable/2346101

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, *71*, 117–120. https://doi.org/10.1016/j.euroecorev.2014.07.003

Burlando, R. M., & Guala, F. (2005). Heterogeneous Agents in Public Goods Experiments. *Experimental Economics*, *8*(1), 35–54. https://doi.org/10.1007/s10683-005-0436-4

Cartwright, E., Chai, Y., & Xue, L. (2023). *Are My Team Members Pro-Social? Information About Social Value Orientation Influences Cooperation in Public Good Games*. https://doi.org/10.2139/ssrn.4628488

Cason, T. N., & Wu, S. Y. (2019). Subject Pools and Deception in Agricultural and Resource Economics Experiments. *Environmental and Resource Economics*, *73*(3), 743–758. https://doi.org/10.1007/s10640-018-0289-x

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2014). Identities, selection, and contributions in a public-goods game. *Games and Economic Behavior*, *87*, 322–338. https://doi.org/10.1016/j.geb.2014.05.002

Charness, G., Samek, A., & van de Ven, J. (2022). What is considered deception in experimental economics? *Experimental Economics*, *25*(2), 385–412. https://doi.org/10.1007/s10683-021-09726-7

Charness, G., & Yang, C.-L. (2014). Starting small toward voluntary formation of efficient large groups in public goods provision. *Journal of Economic Behavior & Organization*, *102*, 119–132. https://doi.org/10.1016/j.jebo.2014.03.005

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.1016/J.JBEF.2015.12.001

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–457. https://doi.org/10.1257/aer.99.1.431

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. Elsevier. https://doi.org/10.1016/C2013-0-10497-7

Cooper, D. J. (2014). A Note on Deception in Economic Experiments. *Journal of Wine Economics*, *9*(2), 111–114. https://doi.org/10.1017/jwe.2014.18

Corr, P. J., Hargreaves Heap, S. P., Seger, C. R., & Tsutsui, K. (2015). An experiment on individual 'parochial altruism' revealing no connection between individual 'altruism' and individual 'parochialism'. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01261

Drouvelis, M., & Georgantzis, N. (2019). Does revealing personality data affect prosocial behaviour? *Journal of Economic Behavior & Organization*, *159*, 409–420. https://doi.org/10.1016/j.jebo.2019.02.019

Drouvelis, M., Metcalfe, R., & Powdthavee, N. (2015). Can priming cooperation increase public good contributions? *Theory and Decision*, *79*(3), 479–492. https://doi.org/10.1007/s11238-015-9481-4

Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, *73*(2), 459–478. https://doi.org/10.1016/j.geb.2011.02.003

Eckel, C. C., & Grossman, P. J. (2005). Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, *58*(3), 371–392. https://doi.org/10.1016/j.jebo.2004.01.003

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

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in Disguise - An Experimental Study on Cheating. *Journal of the European Economic Association*, *11*(3), 525–547. https://doi.org/10.1111/jeea.12014

Gächter, S., & Thöni, C. (2005). Social Learning and Voluntary Cooperation among like-Minded People. *Journal of the European Economic Association*, *3*(2-3), 303–314. https://doi.org/10.1162/jeea.2005.3.2-3.303

Gill, D., & Rosokha, Y. (2024). Beliefs, Learning, and Personality in the Indefinitely Repeated Prisoner's Dilemma. *American Economic Journal: Microeconomics*, *16*(3), 259–283. https://doi.org/10.1257/mic.20210336

Glöckner, A., & Hilbig, B. E. (2012). Risk is relative: Risk aversion yields cooperation rather than defection in cooperation-friendly environments. *Psychonomic Bulletin & Review*, *19*(3), 546–553. https://doi.org/10.3758/s13423-012-0224-z

Gneezy, U., & Kajackaite, A. (2020). Externalities, stakes, and lying. *Journal of Economic Behavior & Organization*, *178*, 629–643. https://doi.org/10.1016/j.jebo.2020.08.020

Goldberg, L. (2002). Big Five Factor Markers. https://ipip.ori.org/newBigFive5broadKey.htm

Goldberg, L., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

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

Gunnthorsdottir, A., Houser, D., & McCabe, K. (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, *62*(2), 304–315. https://doi.org/10.1016/j.jebo.2005.03.008

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, *53*(1), 7–16. https://doi.org/10.1037/0022-006X.53.1.7

Harrison, G. W. (1989). Theory and Misbehavior of First-Price Auctions. *The American Economic Review*, *79*(4), 749–762. https://www.jstor.org/stable/1827930

Hawkins, T., & Monroe, M. (2021, February). Persona: The Dark Truth Behind Personality Tests. https://www.imdb.com/title/tt14173880/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Henry, P. J. (2008). College Sophomores in the Laboratory Redux: Influences of a Narrow Data Base on Social Psychology's View of the Nature of Prejudice. *Psychological Inquiry*, *19*(2), 49–71. https://doi.org/10.1080/10478400802049936

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, *49*(2), 263–266. https://doi.org/10.1016/j.jesp.2012.11.010

Hilbig, B. E., Kieslich, P. J., Henninger, F., Thielmann, I., & Zettler, I. (2018). Lead Us (Not) into Temptation: Testing the Motivational Mechanisms Linking Honesty–Humility to Cooperation. *European Journal of Personality*, *32*(2), 116–127. https://doi.org/10.1002/per.2149

Hilbig, B. E., & Thielmann, I. (2017). Does everyone have a price? On the role of payoff magnitude for ethical decision making. *Cognition*, *163*, 15–25. https://doi.org/10.1016/j.cognition.2017.02.011

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment*, *29*(3-4), 412–426. https://doi.org/10.1111/ijsa.12338

Kagel, J., & McGee, P. (2014). Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters*, *124*(2), 274–277. https://doi.org/10.1016/j.econlet.2014.05.034

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, *102*, 433–444. https://doi.org/10.1016/j.geb.2017.01.015

Kantrowitz, T. M., Tuzinski, K. A., & Raines, J. M. (2018). 2018 Global Assessment Trends Report.

Kaźmierczak, I., Zajenkowska, A., Rogoza, R., Jonason, P. K., & Ścigała, D. (2023). Self-selection biases in psychological studies: Personality and affective disorders are prevalent among participants. *PLOS ONE*, *18*(3), e0281046. https://doi.org/10.1371/journal.pone.0281046

Lawn, E. C. R., Zhao, K., Laham, S. M., & Smillie, L. D. (2022). Prosociality Beyond Big Five Agreeableness and HEXACO Honesty-Humility: Is Openness/Intellect Associated With Cooperativeness in the Public Goods Game? *European Journal of Personality*, *36*(6), 901–925. https://doi.org/10.1177/08902070211028104

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*.

Lugovskyy, V., Puzzello, D., Sorensen, A., Walker, J., & Williams, A. (2017). An experimental study of finitely and infinitely repeated linear public goods games. *Games and Economic Behavior*, *102*, 286–302. https://doi.org/10.1016/j.geb.2017.01.004

McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McGee, A., & McGee, P. (2024). Whoever you want me to be: Personality and incentives. *Economic Inquiry*, *62*(3), 1268–1291. https://doi.org/10.1111/ecin.13220

McGee, A., & McGee, P. (2025). Gender and race differences on incentivized personality measures. *Frontiers in Behavioral Economics*, *4*. https://doi.org/10.3389/frbhe.2025.1499464

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the Use of Personality Tests in Personnel Selection Contexts. *Personnel Psychology*, *60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting Individuals in Team Settings: The Importance of Social Skills, Personality Characteristics, and Teamwork Knowledge. *Personnel Psychology*, *58*(3), 583–611. https://doi.org/10.1111/j.1744-6570.2005.655.x

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.1017/S1930297500004204

Ones, U., & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior & Organization*, *62*(4), 495–521. https://doi.org/10.1016/j.jebo.2005.04.018

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Perugini, M., Tan, J. H. W., & Zizzo, D. J. (2010). Which is the More Predictable Gender? Public Good Contribution and Personality. *Economic Issues*, *15*(1), 83–110.

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., & Van Lange, P. A. (2018). Social Value Orientation, Expectations, and Cooperation in Social Dilemmas: A Meta–Analysis. *European Journal of Personality*, *32*(1), 62–83. https://doi.org/10.1002/per.2139

Proto, E., Rustichini, A., & Sofianos, A. (2019). Intelligence, Personality, and Gains from Cooperation in Repeated Interactions. *Journal of Political Economy*, *127*(3), 1351–1390. https://doi.org/10.1086/701355

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2020). Validation of the Short and Extra-Short Forms of the Big Five Inventory-2 (BFI-2) and Their German Adaptations. *European Journal of Psychological Assessment*, *36*(1), 149–161. https://doi.org/10.1027/1015-5759/a000481

Raskin, R. N., & Hall, C. S. (1979). A Narcissistic Personality Inventory. *Psychological Reports*, *45*(2), 590. https://doi.org/10.2466/pr0.1979.45.2.590

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225–237. https://doi.org/10.1080/13501780500086081

Schultz, P., Shriver, C., Tabanico, J. J., & Khazian, A. M. (2004). Implicit connections with nature. *Journal of Environmental Psychology*, *24*(1), 31–42. https://doi.org/10.1016/S0272-4944(03)00022-7

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

StataCorp. (2021, April). Stata Statistical Software: Release 17.

Streib, H., & Wiedmaier, M. (2001). German Translation of the 100-Item Lexical Big-Five Factor Markers. https://ipip.ori.org/German100-ItemBig-FiveFactorMarkers.htm

Tan, N. P.-J., Hilbig, B., Moshagen, M., Zettler, I., Payer, S., & Thielmann, I. (2025). Strangers in the dark: assumed similarity in judgments of unknown others on aversive personality. *Judgment and Decision Making*, *20*, e38. https://doi.org/10.1017/jdm.2025.10013

Tett, R., & Simonet, D. (2021). Applicant Faking on Personality Tests: Good or Bad and Why Should We Care? *Personnel Assessment and Decisions*, *7*(1). https://doi.org/10.25035/pad.2021.01.002

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. https://doi.org/10.1037/bul0000217

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Villeval, M. C. (2016). Can lab experiments help design personnel policies? *IZA World of Labor*. https://doi.org/10.15185/izawol.318

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Viswesvaran, C., & Ones, D. S. (1999). Meta-Analyses of Fakability Estimates: Implications for Personality Measurement. *Educational and Psychological Measurement*, *59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Volk, S., Thöni, C., & Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization*, *81*(2), 664–676. https://doi.org/10.1016/j.jebo.2011.10.006

Walker, S. A., Double, K. S., Birney, D. P., & MacCann, C. (2022). How much can people fake on the dark triad? A meta-analysis and systematic review of instructed faking. *Personality and Individual Differences*, *193*, 111622. https://doi.org/10.1016/j.paid.2022.111622

Weber, L., & Dwoskin, E. (2014, February). Are Workplace Personality Tests Fair? https://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

Wilmot, M. P., & Ones, D. S. (2022). Agreeableness and Its Consequences: A Quantitative Review of Meta-Analytic Findings. *Personality and Social Psychology Review*, *26*(3), 242–280. https://doi.org/10.1177/10888683211073007

Zell, E., & Lesick, T. L. (2022). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, *90*(4), 559–573. https://doi.org/10.1111/jopy.12683

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

Appendices

A. Deviations from the Pre-registration Document

This paper closely follows the pre-registered motivation and statistical tests or explicitly states otherwise. The original intention was to strictly adhere to using the text in the pre-registration as the final paper as far as practically possible. However, a substantial rewrite was required, in order to meet the expectations and norms of a different field. Instead, I leave the last version prior to this rewrite available at https://woodsd42.github.io/files/PTPGG.pdf and the original pre-registration at http://doi.org/10.17605/OSF.IO/MDB7A.

The random ordering of treatments was not able to be followed perfectly near the end of the data collection, as the number of participants that showed up could differ from what the next treatment required. For example, 18 participants might show up for a session that only requires 12 more observations. In this case, the next treatment that required 18 participants was conducted in that session, while the original assigned treatment was conducted in the next session that had 12 participants. The treatments that were assigned to VCEE were done so as soon as the decision to employ them was made, but show-up rates also necessitated some changes. Given the large majority of sessions followed the random ordering, then the treatments that were in these later sessions were also random. Based on this, there is no reasonable threat to the randomization procedure.

B. Predictions, Hypotheses, and Proposed Behavioral Channels

B.1. Predictions: Misrepresentation

When it comes to strategic misrepresentation in the Big Five questionnaire of Part 1, there are three treatment groups of interest. The first are those that know in advance that their Part 1 responses will be used to form groups in Part 2 (*AGRE Before*). The second are those that know in advance that their Part 1 responses will not be used to form groups in Part 2 (*RAND Before*). The final group are those that do not know in advance about the group formation rule in Part 2 (*After & Never*). The first two groups are aware of how their Part 1 responses affect Part 2 while answering Part 1, while the third group is unaware of this while answering Part 1.

I propose two behavioral channels that could influence Part 1 responses: the incentive to misrepresent Agreeableness, and the suspicion that Part 1 answers may be used in some way for Part 2. An incentive to misrepresent Agreeableness exists when it is known groups will be formed based on this trait. Suspicion occurs only when participants are not aware of the purpose of the questionnaire. Participants may believe (sometimes correctly) that the questionnaire will be used in some relevant way in the future, as they know there will be a following Part 2. Each of the comparisons between the relevant groups and the differences in operative channels between them are summarized in Table B1.

Table B1: Misrepresentation of Agreeableness - Treatment Comparisons

Treatment Comparison	Incentive	Suspicion
AGRE Before to *RAND Before*	−	0
RAND Before to *After & Never*	0	+
AGRE Before to *After & Never*	−	+

Going from the first treatment to the second, + indicates that channel has been added, 0 indicates no change, and − indicates that channel has been taken away. The *After & Never* grouping includes all treatments except for *AGRE Before* and *RAND Before*.

Table B1 demonstrates comparisons that isolate either channel: incentives through comparing *AGRE Before* to *RAND Before*, and suspicion through comparing *RAND Before* to *After & Never*. Both channels have the potential to influence Agreeableness. Incentives should increase the reported Agreeableness

scores, as participants will prefer to be in *H* groups (or avoid *L* groups). I propose that suspicion leads to more socially desirable responses, thereby increasing reported Agreeableness scores. There are many possibilities of what a participant might be suspicious of, but the obvious candidates of group formation or having answers revealed to others in Part 2 would both suggest a tendency towards more socially desirable responses.[6] Hypotheses 1 and 2 formalizes the conjecture that the reported Agreeableness scores in Part 1 in the presence of incentives or suspicion respectively.

Hypothesis 1. Agreeableness scores are higher in *AGRE Before* than in *RAND Before*

Hypothesis 2. Agreeableness scores are higher in *After* & *Never* treatments than in *RAND Before*

B.2. Predictions: Contributions

One important distinction to make is that in the *AGRE* treatments, one group will have higher Agreeableness than the other. The high group is predicted to have higher contributions than the low group. I therefore consider these two types of groups separately, as I would like to observe the positive effects of personality sorting.[7] I denote the two types of groups *H* and *L* for high and low Agreeableness respectively. In the following discussion, I take the viewpoint of the *H* group when describing potential effects.

I conjecture that there are three main factors at play here: the group formation rule itself, strategic misrepresentation of Agreeableness, and knowledge of the group formation rule. If Agreeableness is linked with cooperation and generosity, then the Agreeableness group formation rule could be effective in increasing contributions.[8] Hypothesis 3 tests this conjecture under each timing condition.

Hypothesis 3. The number of tokens contributed in *AGRE H* is greater than in *RAND*
The number of tokens contributed in *RAND* is greater than in *AGRE L*
The number of tokens contributed in *AGRE H* is greater than in *AGRE L*

However, the effectiveness of the Agreeableness group formation rule could differ depending on when information about the rule is revealed. Consider comparing *Before* to *After*, two treatments where participants know the group formation rule before the PGG. In *Before* the group formation rule is known prior to when Agreeableness is measured. Participants have an incentive to misrepresent themselves in the Agreeableness elicitation to try and be placed in the *H* group (or to avoid the *L* group). Agreeableness scores would be compressed and the end result would be more similar to Random group formation in terms of each group's true level of Agreeableness. Whereas in *After*, the group formation rule is only revealed after the Agreeableness elicitation, precluding strategic misrepresentation. The Agreeableness group formation rule should be more effective for *H* groups in the absence of strategic misrepresentation. By the logic of compressed Agreeableness scores, the effect on *L* groups should be the opposite. The *L* groups in *After* In terms of the Random group formation rule, I posit that the timing has no effect. Hypothesis 4 formalizes these conjectures.

Hypothesis 4. The number of tokens contributed in *AGRE Before H* is lower than in *AGRE After H*.
The number of tokens contributed in *RAND Before* is the same as in *RAND After*.
The number of tokens contributed in *AGRE Before L* is higher than in *AGRE After L*.

Now consider comparing *After* to *Never*, two treatments that do not have strategic misrepresentation but differ in whether participants know the group formation rule prior to the PGG. Knowing that the

[6] Both incentives and suspicion also have the potential to influence the other personality traits. Suspicion because it is not known which traits could be used, and incentives if misrepresentation is unsophisticated. Section 4.2.1 also tests the other personality traits.

[7] In the employment analogy, the low group would simply not be hired. However, given the expectations of lab participants this is not practical to implement.

[8] It should be noted that Thielmann et al. (2020) only find a statistically significant relationship between Agreeableness and cooperation in one-shot social dilemma games, not repeated games like the current environment.

Agreeableness group formation rule is in effect means that participants in the *After* treatment are aware they are grouped with similarly cooperative people. This would increase initial contributions if participants are concerned about being taken advantage of by lower contributors. Higher initial contributions would have a flow-on effect if participants are conditional cooperators. Therefore, Agreeableness group formation should be more effective for *H* groups when the rule is known in the absence of strategic misrepresentation. Whereas, for *L* groups, the effect is reversed - if they are aware they are with other potential low cooperators, the concern of being taken advantage of would decrease their initial contribution, and again have a flow-on effect if there are conditional cooperators. Hypothesis 5 formalizes these conjectures.

Hypothesis 5. The number of tokens contributed in *AGRE After H* is higher than in *AGRE Never H*
The number of tokens contributed in *RAND After* is the same as in *RAND Never*
The number of tokens contributed in *AGRE After L* is lower than in *AGRE Never L*

Table B2 presents especially interesting treatment comparisons that isolate the impact of a particular effect while holding other factors constant. This assumes effects are additively separable, but potential interactions means the full 3x2 design is prudent.

Table B2: Efficiency - Selected Treatment Comparisons

Treatment Comparison	Incentive to misrepresent	Knowledge of group formation rule	Agreeableness group formation
AGRE Before to *AGRE After*	–	0	0
AGRE After to *AGRE Never*	0	–	0
AGRE Before to *RAND Before*	–	0	–
AGRE After to *RAND After*	0	0	–
AGRE Never to *RAND Never*	0	0	–

Note: Going from the first treatment to the second, 0 indicates no change, and – indicates that channel has been taken away.

B.3. Predictions - Positive Perception

Part 3 provides a very conservative test of whether unexpected data use affects participants' subsequent responses. It is conservative as participants are explicitly informed that Part 3 is the last part of the experiment and that their final payments are already set. If this statement is taken seriously, then participants have no material incentive to misrepresent their personality in their Part 3 responses. However, in the *AGRE After* treatment, information about how the earlier Part 1 responses would be used in Part 2 was initially withheld and then later disclosed to participants. The unexpected data use from Part 1 may cause participants to change their Part 3 responses in anticipation of additional unexpected data use, despite explicit statements to the contrary. It would be concerning if participants in the *AGRE After* treatment responded in a different fashion than those in the other treatments, as it would imply a loss of experimental control. Such a finding would raise strong objections about using unexpected data use as a design feature in economics experiments going forward.

The traits elicited in Part 3 all have a clear direction in terms of social desirability. Narcissism, Machiavellianism, and Psychopathy are clearly negative traits from the perspective of society, while Honesty/Humility is considered a positive trait. I propose that if a participant anticipates unexpected data use, then they would misrepresent themselves towards what is more socially desirable. I propose two channels that would influence a participant's beliefs that their Part 3 responses will be used to affect something in the experiment. The first channel is whether participants are aware that the data from personality questions have been used for something in the experiment. These are participants in the

AGRE Before and *AGRE After* treatments, as they know the group formation rule in Part 2 was based on their Agreeableness score from Part 1. The participants in the other treatments remain *Unaware* that personality responses could be used in other parts of the experiment. Participants that know their personality questions in Part 1 were used in Part 2 could suspect that their personality responses in Part 3 are also used in some fashion, and misrepresent themselves accordingly. The second channel is whether the use of the personality data was unexpected. Participants in *AGRE Before* expected this data use when completing Part 1, as they were told of the Agreeableness group formation rule in advance. Whereas, participants in *AGRE After* did not expect it, but found out about it after completing Part 1. Participants in *AGRE After* may think that their Part 3 responses will be used in some way that has not yet been revealed, and thus would be the most likely to misrepresent themselves in Part 3. Table B3 describes which channels are present between each group of treatments.

Table B3: Misrepresentation in Part 3 - Treatment Comparisons

Treatment Comparison	Unexpected Data Use Revealed	Knowledge of Personality Data Use
AGRE Before to *AGRE After*	+	0
AGRE Before to *Unaware*	0	−
AGRE After to *Unaware*	−	−

Note: Going from the first treatment to the second, + indicates that channel has been added, 0 indicates no change, and − indicates that channel has been taken away. The *Unaware* grouping includes all treatments except for *AGRE Before* and *AGRE After*.

I aggregate each individual into one measure of 'Positive Perception', which positively weights Honesty/Humility and negatively weights the Dark Triad traits. Based on my previous reasoning, I posit the following Hypotheses about Positive Perception:

Hypothesis 6. Reported Positive Perception is higher in *AGRE After* than in *AGRE Before*

Hypothesis 7. Reported Positive Perception is higher in *AGRE After* than in *Unaware* treatments

Hypothesis 8. Reported Positive Perception is higher in *AGRE Before* than in *Unaware* treatments

C. Summary Statistics

Table C4 Notes: Standard deviations in parentheses. Low and High Agreeableness groups reported separately for *AGRE* treatments. Categorical variables are presented as is for summary purposes only and are otherwise unadjusted unless specifically noted. Female = 1 if reported gender was female, 0 if reported male, and not included otherwise (for this specific statistic only). Age is reported age in years. Num. Prev. Exp. is the number of previous experiments the participant reported they had participated in. Year At Uni.: 1 = First year, 2 = Second year, 3 = Third year, 4 = Fourth year +, 5 = Graduate student. GPA: 1 = [1, 1.5], 2 = [1.51,2.5], 3 = [2.51, 3.5], 4 = [3.51, ∞), 5 = N/A. Enjoyment: 1 = Disliked the experiment a lot, 2 = Disliked ... a little, 3 = Did not like or dislike, 4 = Liked ... a little, 5 = Liked ... a lot. Num. Fail CompQ 1 is the number of times that answers with at least one error or missing value was submitted over all questions on the first page of comprehension questions, similarly for Num. Fail CompQ 2 but for the second page. Abbreviations: Tot. = Total, Contr. = Contributions, Num. = Number, Prev. = Previous, Exp. = Experiments, Uni. = University, GPA = Grade Point Average, CompQ = Comprehension Questions.

Table C4: Summary Statistics

	RAND Before	*RAND After*	*RAND Never*	*AGRE Before*	*AGRE After*	*AGRE Never*
Openness to experience [Low Agre. High Agre.]	3.61 (0.78)	3.53 (0.58)	3.45 (0.71)	3.76 (0.72) 3.84 (0.70)	3.71 (0.75) 3.90 (0.68)	3.50 (0.60) 3.60 (0.63)
Neuroticism	2.75 (0.90)	2.71 (0.80)	2.56 (0.76)	2.72 (0.85) 2.50 (0.86)	2.64 (0.63) 2.55 (0.76)	2.61 (0.84) 2.64 (0.82)
Extroversion	3.24 (0.68)	3.35 (0.89)	3.48 (0.70)	3.29 (0.78) 3.39 (0.73)	3.44 (0.76) 3.49 (0.71)	3.47 (0.60) 3.67 (0.64)
Conscientiousness	3.71 (0.65)	3.34 (0.69)	3.66 (0.77)	3.56 (0.68) 3.76 (0.60)	3.21 (0.75) 3.66 (0.65)	3.62 (0.69) 3.76 (0.56)
Honesty Humility	3.15 (0.85)	3.02 (0.80)	3.04 (0.69)	3.14 (0.81) 3.36 (0.77)	3.14 (0.77) 3.25 (0.75)	3.21 (0.85) 3.16 (0.87)
Machiavellianism	2.01 (0.81)	2.24 (0.98)	2.05 (0.78)	2.17 (0.89) 1.89 (0.89)	2.24 (0.84) 1.73 (0.76)	2.18 (0.96) 2.02 (0.80)
Psychopathy	2.10 (0.73)	2.01 (0.89)	1.95 (0.77)	2.12 (0.67) 1.80 (0.69)	2.32 (0.83) 1.69 (0.57)	2.36 (0.81) 1.84 (0.66)
Narcissism	2.93 (0.87)	2.96 (0.86)	2.73 (1.02)	2.90 (0.86) 2.81 (0.92)	2.82 (0.85) 2.73 (0.88)	2.90 (0.93) 3.05 (0.91)
Female	0.60 (0.50)	0.70 (0.46)	0.54 (0.50)	0.46 (0.50) 0.70 (0.46)	0.44 (0.50) 0.64 (0.49)	0.44 (0.50) 0.67 (0.48)
Age	22.88 (3.42)	22.44 (2.58)	22.38 (2.73)	23.98 (4.61) 24.15 (4.42)	24.67 (9.89) 22.65 (2.91)	23.33 (4.11) 22.27 (2.52)
Num. Prev. Exp.	5.35 (6.31)	5.88 (8.05)	5.38 (4.51)	6.52 (7.16) 6.38 (6.86)	5.73 (7.37) 3.94 (4.33)	7.04 (7.40) 5.54 (5.32)
Year At Uni.	2.98 (1.47)	2.77 (1.28)	2.81 (1.33)	2.96 (1.35) 2.79 (1.27)	2.90 (1.53) 2.62 (1.39)	3.15 (1.38) 2.88 (1.44)
GPA	2.56 (1.32)	2.71 (1.25)	2.79 (1.35)	2.79 (1.34) 2.48 (1.37)	2.88 (1.45) 2.94 (1.62)	2.62 (1.30) 2.69 (1.49)
Enjoyment	3.81 (0.91)	3.73 (0.94)	3.73 (1.07)	3.90 (0.88) 3.77 (0.93)	3.77 (1.06) 3.90 (0.86)	3.71 (1.09) 3.98 (0.89)
Num. Fail CompQ 1	1.40 (3.04)	0.73 (1.09)	1.44 (3.21)	2.02 (4.39) 1.00 (2.13)	0.90 (1.57) 1.31 (2.27)	1.04 (1.70) 1.35 (2.35)
Num. Fail CompQ 2	1.65 (3.35)	1.50 (1.83)	1.56 (2.52)	1.40 (2.22) 0.98 (1.79)	1.38 (2.10) 1.06 (1.29)	1.60 (2.35) 1.25 (1.59)