Strategic Misrepresentation in Personality Testing: An Experimental Study using the Public Goods Game (Report for Pre-registration)

Click here for the latest version

Daniel Woods*

October 10, 2023

Abstract

The use of 'psychometric' personality testing to help identify suitable candidates for job positions is commonplace in hiring practices. However, using these tests in this manner is logically self-defeating, as applicants have an incentive to misreport their answers to more closely match the job's ideal personality profile. This paper uses a lab experiment to explore the problem of strategic misrepresentation, and how it impacts subsequent teamwork. Both of these factors are of key interest to employers. The experiment uses the Public Goods Game (PGG) as a stand-in for a cooperative work environment. Prior to the PGG, a 'Big Five' personality test is conducted. Multiple studies have identified pro-social cooperation in this game to be correlated with the personality trait of 'Agreeableness'. Therefore, assigning subjects into groups based on Agreeableness should improve cooperation for high Agreeableness groups. The experiment varies the point at which this assignment rule is revealed (either before or after the personality test), as well as the assignment rule itself (ei-

*Department of Economics, University of Innsbruck, Universitaetsstrasse 15, 6020 Innsbruck, Austria. Email: daniel.j.woods@uibk.ac.at. Funding from the IFREE Small Grants Program is gratefully acknowledged. In the spirit of transparent and credible science, an exhaustive OSF project for this paper is available at `https://osf.io/mdb7a/?view_only=5d74405d3cd94a8eb99beb8a3e0d50aa` (OSF account required).

ther randomly or by Agreeableness). This design permits the study of the extent of strategic misrepresentation and its subsequent effect on cooperation.

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. Firm hiring processes are so intertwined with psychometric testing, it even pervades its dictionary definition.[1] 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, Tuzinski, & Raines, 2018). Psychometric tests are multi-billion dollar industry globally, with expenditure reaching 12.32 billion USD in 2021 and forecast to hit 23.28 billion in 2030 (Emergen Research, 2022). It remains an open question whether this expenditure is justified. Finally, it has been suggested that psychometric testing is unfair or discriminatory against minorities or those with disabilities (Weber & Dwoskin, 2014; Hawkins & Monroe, 2021; McGee & McGee, 2022a). All of these elements illustrate the economic importance of psychometric testing, and why it is crucial to understand their efficacy, impacts, and any unintended consequences.

The problem with using psychometric personality tests for hiring is that potential employees have a material incentive to misrepresent their responses on the tests in order to more closely align towards what they perceive an employer is looking for. When completing a job application, job-seekers know they are being evaluated at every step. Job-seekers tailor their cover letters, embellish their CVs, and curate their references, all to paint themselves in the best light possible. It is no stretch of the imagination that they also adjust their responses to personality questions. After all, no one would willingly state they 'insult people' or 'get irritated easily' during a job interview, so why would they in a job personality test? Therefore, using personality tests during the hiring process may be uninformative of the potential employee's true personality. Such tests introduce further frictions and inefficiencies into the labor market, which must function smoothly for the overall well-being of society and the economy. Therefore, it is important to study whether personality testing maintains its effectiveness given the incentive to misrepresent.

[1] *'Psychometric test: A test that is designed to show someone's personality, mental ability, opinions, etc., often used by companies when they are deciding whether or not to employ someone.'* (Cambridge Business English Dictionary, 2023)

There is some evidence that using psychometric personality testing is effective in job hiring processes (Autor & Scarborough, 2008; Hoffman, Kahn, & Li, 2018), but it is not clear why. One proposed channel is that it puts more weight on objective measures rather than subjective opinions, but that would still require job seekers to represent themselves honestly. Another channel is that misrepresentation on personality tests could be attenuated by a preference for honesty, meaning the signal-to-noise ratio may be good enough for it to remain a useful metric. The challenge of accurately identifying what an employer is seeking would also decrease the likelihood of misrepresentation. On this note, it is possible that testing might instead be capturing some measure of intelligence (Borghans, Duckworth, Heckman, & Weel, 2008). Those that are smart enough to identify what the employer wants could also be more effective workers based on that intelligence. More negatively, those that are cunning enough to shape their personality responses may be ruthless enough to achieve results. Such workers would not value the firm's reputation or sustainability, and potentially drive out more genuine workers. Therefore, selecting workers in this fashion may be beneficial in the short-term but problematic in the longer-term, something that previous studies have not considered. The effects of personality testing in job hiring are multifaceted and warrant further study. There are still open empirical questions on whether such tests are effective, why they are effective, and if there are unintended consequences arising from their use.

In this paper, I design a laboratory experiment to evaluate under what conditions personality testing is effective. Lab experiments are becoming a common method to help inform firm personnel and hiring processes. Experiments are a cost-effective tool to evaluate potential firm policies and why they work, without confounds like employee self-selection and while still retaining good external validity (Villeval, 2016). I design the experiment to closely mirror the important elements of hiring using personality testing, and the subsequent work output. 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, which are widely employed in both hiring procedures and academic research in economics.[2] For the

[2] The psychometric testing firms Big Five Assessments, Hogan Assessments, and SHL, among others, incorporate elements of the Big Five as part of their battery of psychometric testing services that they offer to firms. For a variety of examples of the Big Five in economics research, see (Bartling, Fehr, Maréchal, & Schunk, 2009; Fréchette, Schotter, & Trevino, 2017; Donato, Miller, Mohanan, Truskinovsky, & Vera-Hernández, 2017; Holmén, Holzmeister, Kirchler, Stefan, & Wengström, 2021).

cooperation task, I use the Public Goods Game (PGG) as an representation of a cooperative work environment. In the PGG subjects can make socially-optimal contributions to a public good, but face a personal incentive to free-ride and contribute less. I interpret contributions to the public good as effort at work, 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 research finds it positively impacts contributions in the PGG and other similar social dilemmas (Perugini, Tan, & Zizzo, 2010; Volk, Thöni, & Ruigrok, 2012; Kagel & McGee, 2014; Thielmann, Spadaro, & Balliet, 2020). I sort subjects 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 crucial 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, subjects have an incentive to misrepresent their personality in order to try and get into a more cooperative group. This situation is similar to the current status quo, where job seekers are aware they are being evaluated for the job by the test. The compression of Agreeableness scores, alongside any mistrust that might arise due to the potential for strategic misrepresentation, makes this a challenging environment for personality testing to be effective in increasing PGG contributions. Whereas in the *Never* Treatment, subjects are never informed about how groups are formed, and therefore have no material incentive to misrepresent their personality. Without strategic misrepresentation, forming groups by the elicited Agreeableness scores is more likely to be effective in increasing PGG contributions in high Agreeableness groups. Finally, in the *After* treatment, subjects also have no material incentive to misrepresent their personality as the group formation rule is only revealed directly after the personality test. If subjects know that they are in a group with similarly cooperative people, then they can be more confident of current and future cooperation. Combined with the absence of strategic misrepresentation, this scenario is the most favorable environment for personality testing to be effective. The situations represented by *After* and *Never* are not particularly realistic, but instead they address the question of what conditions are required for personality testing to be effective. They highlight a strength of economic experiments, in that it allows for an exploration of counterfactuals that would otherwise not occur.

The second treatment dimension is the group formation rule itself. Groups are typically randomly assigned in economics experiments, which makes a *Random* treatment a natural baseline for the *Agreeableness* group formation rule. The experiment is a 3x2 design, so subjects in the *Random* treatment also have the group formation rule revealed to them either *Before* or *After* the personality test, or *Never*. With this battery of treatments, I aim to address the following research questions:

Question 1 *Under what conditions are personality tests effective in encouraging cooperative behavior?*

Question 2 *To what extent do individuals misrepresent their personality when they have strategic reasons to do so?*

Question 3 *Does using personality tests in an unexpected way influence responses in subsequent tests?*

I answer Question 1 by comparing the impact of each treatment dimension on contributions in the PGG while holding the other dimension fixed. This approach allows me to isolate and identify the most significant empirical factors influencing behavior. I answer Question 2 by comparing the responses to the personality test between the treatment with *Agreeableness* group formation rule that is revealed *Before* to all other treatments, as strategic misrepresentation can only be present in the former. Question 3 considers the situation where an initially naive job-seeker represents their personality honestly, but comes to realize that their answers were used against them in some fashion. They would then likely misrepresent themselves in future tests, which reduces the effectiveness of personality tests as a hiring tool. Question 3 also addresses an important methodological question in experimental economics - whether the unexpected use of previous responses change how subjects behave in the future. I answer Question 3 by conducting another personality test after the PGG, and focus on subject responses in the *Agreeableness* group formation rule that is revealed *After*, as these subjects have had their personality responses used in an unannounced way.

1.1 Contribution to the Literature

I contribute to the voluminous literature on the PGG. When I refer to the PGG in this paper, I am using this as a shorthand reference for the commonly studied Linear Voluntary Contribution Mechanism, although it is worth noting other formulations exist (Ledyard, 1995). A typical pattern of behavior in the PGG starts out with average contributions to the public good of around 50%, which decays steadily over time (Ledyard, 1995; Chaudhuri, 2011; Villeval, 2020). The socially optimum contribution level is 100%, but subjects face an individual incentive to free-ride off the contributions of others. One specific focus has centered on mechanisms or interventions aimed at enhancing contributions in the PGG. Examples include allowing for punishment (Fehr & Gächter, 2000) or facilitating endogenous group formation (Ahn, Isaac, & Salmon, 2009). I contribute to this strand of the PGG literature by considering exogenous group formation through personality sorting. Prior studies on exogenous group formation have sorted subjects based on their previous contribution behavior in a PGG, and found that this type of sorting is effective (Burlando & Guala, 2005; Gächter & Thöni, 2005; Gunnthorsdottir, Houser, & McCabe, 2007; Ones & Putterman, 2007). Typically in these experiments the sorting rule is withheld from subjects, and in all cases the information given on the sorting rule is constant by treatment. I contribute to this literature by examining how knowledge of the sorting rule affects contributions in the PGG. Additionally, I contribute to this line of literature by exploring whether it is possible to effectively sort subjects by indirect measures of their contribution rate, namely their personality traits.

Another strand of the PGG literature considers the effects that individual characteristics have on contribution behavior in the PGG. Of particular interest to the current paper are studies that elicit Big Five personality characteristics.[3] The Big Five personality trait of Agreeableness has been found to be a significant predictor of contribution behavior in the PGG (Perugini et al., 2010; Volk et al., 2012). I aim to tackle the logical next question in this line of research: Given our understanding that Agreeableness influences contributions, how can we leverage this insight? Creating PGG groups by Agreeableness in order to improve

[3] Some other relevant papers on individual characteristics and the PGG are (Anderson, Mellor, & Milyo, 2004; Carpenter, Daniere, & Takahashi, 2004; Catola, D'Alessandro, Guarnieri, & Pizziol, 2021).

contributions is a natural next step, and is analogous to role of employers using personality testing to select well-suited employees.

Naturally, I also contribute to the literature on psychometric personality testing. Misrepresentation, regardless of the motivation behind it, has been a longstanding concern in psychology due to the threat it poses to the validity of psychometric testing.[4] The main limitation of psychology studies on misrepresentation is the absence of monetary incentives. Subjects are typically explicitly instructed to misrepresent themselves in a particular way, which effectively gives permission to lie. As a result, this type of research fails to capture the significant trade-off between honest representation and material gain. Furthermore, it is cognitively costly to determine which questions coincide with a specific personality trait. Without the incentive to do so, individuals will put less effort into this task. The use of incentives is a key difference between the fields of experimental psychology and experimental economics.

The most closely related paper in economics on personality testing using incentivized experimental methods is by McGee and McGee (2022b) (henceforth MM). In their experiment, they first elicit subjects' Big Five personality traits in an initial baseline session. In a follow-up session a week later, subjects complete a second Big Five assessment. Before taking the second personality test, subjects are informed that they will receive an extra payment if they are 'hired' for a hypothetical job. The hiring process is based in part on their Big Five characteristics as elicited in the second personality test. Subjects are given a job description that is designed to indicate that Big 5 personality trait of Extroversion would be ideal.[5] MM find that subjects misrepresent their personality in the presence of incentives.

I take a different approach from MM which makes a complementary but distinct contribution to the literature. Firstly, my focus is how misrepresentation impacts subsequent work behavior. An employee is likely to behave differently when it comes to cooperative team decisions if they suspect their colleagues are manipulative, dishonest, and ill-suited for their roles due to misrepresentation. MM focus on the magnitude of misrepresentation given the incentive of being hired for a job that is never undertaken. Whereas, I extend the hiring

[4] For select examples see (Braun & Gomez, 1966; Velicer & Weiner, 1975; Kroger & Wood, 1993)

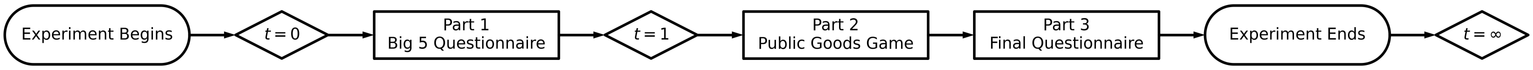
[5] MM also use a job description aimed at Introversion as well as a neutral description as robustness checks.

process analogy to include the ensuing job effort decisions, in order to focus on the effects of misrepresentation. My main goal is to uncover whether personality testing is effective in fostering cooperative environments, and under what conditions. Secondly, I consider the misrepresentation of personality traits in a between-subject design rather than within-subjects as in MM. In this regard, I follow the experimental literature on dishonesty, which emphasizes that dishonest behavior is difficult to observe at the individual level (Fischbacher & Föllmi-Heusi, 2013). Comparing individual responses across two personality tests introduces a potential confounding factor. Subjects might be concerned that any substantial misrepresentation will be detected by comparing the two tests, leading them to be more honest than they would otherwise be. Finally, this paper contributes by proving a conceptual replication of some of the elements in MM. Replication is not the primary purpose of this study, but the externality is a welcome one given the current credibility crisis in the social sciences (Butera, Grossman, Houser, List, & Villeval, 2020). Independent replications can greatly increase the likelihood that any detected effect is actually true (Maniadis, Tufano, & List, 2014).

2 Experimental Design

I first briefly describe the experiment and its treatments, so that the necessity of some of the finer design elements are more apparent. 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 *Random* (R) treatments, groups of three are formed randomly from all subjects in the session. In the *Agreeableness* (A) treatments, subjects are first randomly shuffled into silos of six. Within each silo, the three subjects with the highest Agreeableness scores (as elicited in Part 1) are assigned to one group, while the remaining three are assigned to another group. The second treatment dimension is the timing of when information about the group formation rule is provided. This is either *Before* Part 1 ($t = 0$), *After* Part 1 but before Part 2 ($t = 1$), or *Never* ($t = \infty$). An illustration of the timing of the experiment is presented in Figure 1. The experiment is a 3x2 design, meaning all combinations of the two treatment dimensions are considered, as summarized in Table 1. Henceforth, I denote each treatment with two characters as in Table 1, with the letter representing either the A(greeableness) or R(andom) group formation rule, and the number representing $t = 0$, $t = 1$, or $t = \infty$.

Figure 1: Timeline of Experiment



A diamond (⋄) represents a possible treatment point at which the group assignment rule for Part 2 is revealed.

Table 1: Treatments

	$t = 0$ (*Before*)	$t = 1$ (*After*)	$t = \infty$ (*Never*)
Agreeableness (A)	$A0$	$A1$	$A\infty$
Random (R)	$R0$	$R1$	$R\infty$

2.1 Part 1 - Big Five Elicitation

Part 1 consists of 50 questions designed to elicit the Big Five personality traits (McCrae & John, 1992). These traits are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. As the experiment is conducted in Austria, the questions (and all of the experiment) are in German. Each Big Five characteristic is elicited using the 30 question 'BFI-2-S Inventory' (Soto & John, 2017; Rammstedt, Danner, Soto, & John, 2020). 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 A treatments. Subjects are asked to answer the personality questions accurately, and enter their responses using a 5-point Likert scale (Likert, 1932). Agreeableness is calculated based on each subject's numerical responses on the relevant questions.[6]

[6] See Appendix D for details on the personality trait questions and their scoring.

In all treatments, subjects are given a short overview of the PGG in Part 2 before completing the Part 1 questions. If information about the group formation rule is provided (i.e. $t \leq 1$), it is provided either directly before or directly after subjects complete the 50 questions. The *Agreeableness* treatments have an in-depth description of the group formation rule, which outlines the Agreeableness trait, its positive relationship to cooperative decisions (with academic references), and that the three subjects with the highest Agreeableness scores from a random silo of six will be grouped together.[7] A high level of detail is provided to help subjects understand the specific personality trait that is being used in the group formation rule, and why it could be beneficial or desirable to be in the high Agreeableness group.

2.1.1 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 ($A0$). The second are those that know in advance that their Part 1 responses will not be used to form groups in Part 2 ($R0$). The final group are those that do not know in advance about the group formation rule in Part 2 ($t > 0$). 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 subjects are not aware of the purpose of the questionnaire. Subjects 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 2.

The experimental design permits a clean test of both incentives and suspicion, i.e. the two comparisons in Table 2 where only one channel is added or removed by moving between them. Both channels have the potential to influence Agreeableness. Incentives should increase the

[7] The instructions for Parts 1 and 3 are presented in Appendix C.

Table 2: Misrepresentation of Agreeableness - Treatment Comparisons

Treatment Comparison	Incentive	Suspicion
$A0$ to $R0$	$-$	0
$R0$ to $t > 0$	0	$+$
$A0$ to $t > 0$	$-$	$+$

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 $t > 0$ grouping includes all treatments except for $A0$ and $R0$.

reported Agreeableness scores, as subjects will prefer to be in H groups (or avoid L groups). I propose that suspicion would push responses towards being more socially-acceptable, and therefore increase reported Agreeableness scores. There are many possibilities of what a subject 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-acceptable responses. Hypotheses 1 and 2 formalizes the prediction that the reported Agreeableness scores in Part 1 in the presence of incentives or suspicion respectively.

Hypothesis 1 *Agreeableness scores are higher in* $A0$ *than in* $R0$

Hypothesis 2 *Agreeableness scores are higher in* $t > 0$ *treatments than in* $R0$

2.2 Part 2 - Public Goods Game

Part 2 consists of a PGG adapted from the version used by Lugovskyy, Puzzello, Sorensen, Walker, and Williams (2017). Groups of three are assigned from silos of six subjects 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 subject has 25 tokens they can allocate to either a Private account or a ‘Cooperation’ account.[8] Each token a subject allocates to the Private account earns that subject 10 points. Each token a subject allocates to the Cooperation account earns each of the three group members (i.e. including the subject in question) 4 points each. In other words, one token allocated to the Cooperation account earns the group 12 points overall. I refer to tokens allocated to the Cooperation

[8] Framing the PGG in terms of group cooperation is likely to increase contributions (Dufwenberg, Gächter, & Hennig-Schmidt, 2011). This is not an issue as all treatments are framed in the same way.

account as ‘contributions’. The marginal per-capita return (the ratio of the private benefit of one token to the Cooperation account to that of the opportunity cost of that token) is $MPCR = \frac{4}{10} = 0.4$. For $MPCR = \frac{4}{10} = 0.4$, it is a well-replicated result that groups’ average contribution rates typically start at around 50% and then decline steadily over time (Ledyard, 1995; Chaudhuri, 2011). As I do not anticipate full contributions in the baseline $R0$ treatment, there is plenty of room for a intervention to increase contributions without censoring. Subjects make their decision by deciding how many tokens to allocate to the Cooperation account, with the remainder being allocated to their Private account. After making their decision, subjects are reminded of their own contribution, and also given information on the total group contribution in that round. These two pieces of information are also available at any time during Part 2 in a history table that is displayed at the bottom of the screen.

2.2.1 Predictions: Efficiency

I define efficiency as the number of tokens allocated to the Cooperation account, as full contributions are the first-best social optimum (i.e. socially efficient). One important distinction to make is that in any treatment that sorts by Agreeableness (i.e. A treatments), one group will have higher Agreeableness than the other. The high group is likely 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.[9] 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. The Agreeableness (A) group formation rule should be effective in increasing contributions, as this personality trait is linked with cooperation and generosity. Hypothesis 3 tests this conjecture under each timing ($t = i$) condition .

Hypothesis 3 *The number of tokens contributed in AiH is greater than in Ri.*

[9] In the employment framing of this environment, the low group would simply not be hired. However, given the expectations of lab subjects this is not practical to implement.

The number of tokens contributed in Ri is greater than in AiL.

The number of tokens contributed in AiH is greater than in AiL.

However, the effectiveness of the Agreeableness group formation rule will differ depending on when information about the rule is revealed. Consider comparing $t = 0$ to $t = 1$, two treatments where subjects know the group formation rule before the PGG. In $t = 0$ the group formation rule is known prior to when Agreeableness is measured. Subjects 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 will 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 $t = 1$, the group formation rule is only revealed after the Agreeableness elicitation, precluding strategic misrepresentation. The Agreeableness group formation rule should be more effective in the absence of strategic misrepresentation. In terms of the Random group formation rule, I posit that t has no effect. Hypothesis 4 formalizes these conjectures.

Hypothesis 4 *The number of tokens contributed in $A0H$ is lower than in $A1H$.*

The number of tokens contributed in $R0$ is the same as in $R1$.

The number of tokens contributed in $A0L$ is higher than in $A1L$.

Now consider comparing $t = 1$ to $t = \infty$, two treatments that do not have strategic misrepresentation but differ in whether subjects know the group formation rule prior to the PGG. Knowing that the Agreeableness group formation rule is in effect means that subjects are aware they are grouped with similarly cooperative people. Such confidence will increase initial contributions if subjects are concerned about being taken advantage of by lower contributors. Higher initial contributions will have a flow-on effect if subjects are conditional cooperators. Therefore, Agreeableness group formation should be more effective when the rule is known in the absence of strategic misrepresentation. Hypothesis 5 summarizes these conjectures.

Hypothesis 5 *The number of tokens contributed in $A1H$ is higher than in $A\infty H$*

The number of tokens contributed in $R1$ is the same as in $R\infty$

The number of tokens contributed in $A1L$ is lower than in $A\infty L$

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

Table 3: Efficiency - Selected Treatment Comparisons

Treatment Comparison	Incentive to Misrepresent	Knowledge of group formation rule	Agreeableness group formation
$A0$ to $A1$	$-$	0	0
$A1$ to $A\infty$	0	$-$	0
$A0$ to $R0$	$-$	0	$-$
$A1$ to $R1$	0	0	$-$
$A\infty$ to $R\infty$	0	0	$-$

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.

2.3 Part 3 - Final Questionnaire

In Part 3, subjects are first informed that they are to complete a final survey, and that their final earnings for the experiment have already been set. Subjects then answer 16 personality questions using the same 5-point Likert scale format as the questions in Part 1, and a standard demographic questionnaire. The 16 questions elicit the three elements of the 'Dark Triad' (Paulhus & Williams, 2002), and 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). Machiavellianism is marked by a calculating, manipulative, and deceitful nature towards other people. Narcissism is defined as being egotistic and prideful with limited empathy for others. Psychopathy is characterized by selfishness, impulsiveness and a lack of remorse for ones actions. Honesty-Humility is a personality trait where people avoid manipulating others for personal gain, and feel little temptation to break rules.

Part 3 investigates an important methodological issue in experimental economics - whether omission of information leads to a loss of control over subjects' beliefs and expectations. There is a strong norm against using deception in economics experiments, which has existed from the inception of the field (Svorenčík, 2016). If a subject becomes aware they were deceived in an economics experiment, then they should not believe all of what they are told in experiments after that point in time. Subjects would adjust their responses to account for the fact that the underlying rules may suddenly change in a way that may be detrimental to them. Therefore, they would not reveal what they would actually do if the situation were exactly as described, resulting in a loss of experimental control. The current experiment does not use deception (it cannot - it is an economics experiment). Every piece of information provided to subjects, whether in the instructions or elsewhere, is technically correct. However, there are 'gray-areas' where full consensus among researchers about their acceptability has not yet been reached (Cason & Wu, 2019; Charness, Samek, & van de Ven, 2022). A relevant scenario is 'unexpected data use', when responses are used in a way not described or revealed to subjects. Charness et al. (2022) find that this technique is generally regarded by researchers as non-deceptive and is assessed as appropriate and useful. However, they also find that of the scenarios they consider, student subjects state that unexpected data use is the most likely to influence their future responses and perceive it as more deceptive than researchers do. If subjects do change their future responses based on unexpected data use, this is a methodological problem on a similar scale as outright deception, despite what researchers may believe.

Part 3 provides a very conservative test of whether unexpected data use affects subjects' subsequent responses. It is conservative as subjects 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 subjects have no material incentive to misrepresent their personality in their Part 3 responses. However in the $A1$ treatment, information about how the earlier Part 1 responses would be used in Part 2 was initially withheld and later disclosed to subjects. The unexpected data use from Part 1 may cause subjects to change their Part 3 responses in anticipation of additional unexpected data use, despite explicit statements to the contrary. It would be concerning if subjects in the $A1$ 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 subject anticipates unexpected data use, then they would misrepresent themselves towards what is more socially acceptable. I propose two channels that would influence a subject's beliefs that their Part 3 responses will be used to affect something in the experiment. The first channel is whether subjects are aware that the data from personality questions have been used for something in the experiment. These are subjects in the $A0$ and $A1$ treatments, as they know the group formation rule in Part 2 was based on their Agreeableness score from Part 1. The subjects in the other treatments remain $Unaware$ that personality responses could be used in other parts of the experiment. Subjects 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. Subjects in $A0$ expected this data use when completing Part 1, as they were told of the Agreeableness group formation rule in advance. Whereas, subjects in $A1$ did not expect it, but found out about it after completing Part 1. Subjects in the $A1$ treatment have a justified belief that their Part 3 responses may 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 4 describes which channels are present between each group of treatments.

Table 4: Misrepresentation in Part 3 - Treatment Comparisons

Treatment Comparison	Unexpected Data Use Revealed	Knowledge of Personality Data Use
$A0$ to $A1$	$+$	0
$A0$ to $Unaware$	0	$-$
$A1$ to $Unaware$	$-$	$-$

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 $A0$ and $A1$.

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* $A1$ *than in* $A0$

Hypothesis 7 *Reported Positive Perception is higher in* $A1$ *than in Unaware treatments*

Hypothesis 8 *Reported Positive Perception is higher in* $A0$ *than in Unaware treatments*

2.4 Planned Procedures

The experiment will be conducted in the EconLab at the University of Innsbruck.[10] Subjects will be recruited using the online database hroot (Bock, Baetge, & Nicklisch, 2014), where UIBK students who are interested in participating in economics experiments can sign up. The experiment is computerized using oTree (Chen, Schonger, & Wickens, 2016), so further details about the implementation can be obtained by looking at the code attached to the OSF project.

I will collect observations from 432 subjects, i.e., 144 groups of three. Each R treatment will have 16 groups, and each A treatment will have 32 groups. I collect a different number of groups as the A treatment is split between L and H groups. A power analysis given 432 subjects is conducted and reported for each statistical test in Section 4. The power analyses are provided in the code for the statistical analysis. A session consists of 6, 12, 18, or 24 subjects (depending on how many show up for a session), as multiples of six are required for the A treatments.[11] All subjects within a session face the same treatment. I assign treatments to sessions by randomly shuffling a list of the treatments, and then sampling the shuffled list without replacement. The list of treatments includes two entries for each A treatment as twice as many groups are required for this treatment. Such a process seeks to avoid the A treatments being disproportionately represented in the latter part of the data collection. Once each item in the list is assigned to one session, the list is repopulated, randomly shuffled, and the process repeats. The code used for this randomization is attached

[10] The plan is to complete the entire data collection in the Winter Semester 2023/2024 at UIBK (02.10.2023 - 03.02.2024). It is possible that the UIBK EconLab subject pool could be exhausted or the lab becomes unavailable due to some unforeseen event. In these events, data collection will either be delayed into the subsequent semester, or another German-speaking lab will be used for the remainder of the sessions.

[11] R sessions will also use multiples of six for consistency even though only multiples of three are needed.

to this pre-registration, as well as the resulting order that I will use in sequential sessions. Once a treatment reaches its required number of observations, I will manually remove (i.e. cross it out) from the printed shuffled session list. I will then follow the same order, but only using the treatments that are not yet crossed out.

3 Simulations

As the PGG is finitely repeated, the Nash equilibrium can be solved by backwards induction, and results in all subjects contributing zero to the public good. The Nash equilibrium in the PGG has been resoundingly refuted by a large body of evidence. Therefore, I adapt the utility function used in Arifovic and Ledyard (2012), which incorporates other-regarding preferences alongside a desire to not be taken advantage of. Their learning model is able to capture various empirical regularities of the PGG while remaining relatively simple. The simulation exercise that I undertake is not designed to accurately predict the magnitude of contributions, rather it is to explore how the treatment effects described in Section 2 could operate. I also use it to simulate a sample data-set on which the statistical analysis in Section 4 is applied, in order to demonstrate the functionality of that code prior to data collection.

The base utility function is $U_i(c) = \pi_i(c) + \beta\bar{\pi}(c) - \gamma_i max\{0, \bar{\pi}(c) - \pi_i(c)\}$, where c denotes the total contribution of the group, $\pi_i(c) = 4c + 10(25 - c_i)$, $\bar{\pi}(c) = \frac{1}{3}\sum_{i=1,2,3}\pi_i(c)$, and β_i and γ_i are individual weights on the total average payoff and being taken advantage of respectively. I incorporate the Agreeableness of an individual $A_i \in [0, 1]$, as well as the the Agreeableness of each group member $A_{j\neq i}$ by adjusting the weights on each element present in the utility function: $U_i(c) = (1 - A_i)\pi_i(c) + \frac{A_1+A_2+A_3}{3}\bar{\pi}(c) - (1 - A_i)max\{0, \bar{\pi}(c) - \pi_i(c)\}$. I propose that a higher level of Agreeableness reduces the weight on an individual's own payoff as well as their envy disutility. A higher level of Agreeableness also increases the weight placed on the group's average payoffs, but to a lesser extent as it also depends on the Agreeableness of the other two group members.

I conduct simulations based on this utility function that incorporates Agreeableness. I assume that $A_i \sim U[0, 1]$. In all treatments, I assume that individuals have an estimate of

A_j as reported in Part 1.[12] However, this estimate is distorted depending on the treatment. In the $A0$ treatment, the reported Agreeableness and true Agreeableness diverge due to misrepresentation. I model this misrepresentation as $\tilde{A}_i = max\{A_i + U[0,1], 1\}$, as the ability to misrepresent is likely heterogeneous across individuals. However, as a result of this misrepresentation, the reported $\tilde{A}_i$ are not believed, and the estimate is weighted down towards 0.50, the average expected A_i given the uniform prior of $U[0,1]$. The weighting in $A0$ is 0.75 on the uniform prior and 0.25 on the Part 1 report. In the $A1$ treatment, I assume that individuals do not misrepresent their personality, and therefore they believe the reports from Part 1 are accurate (i.e. there is zero weight on the uniform prior). In the Random R treatments and in $A\infty$, I assume Agreeableness is somewhat obfuscated as it is not known that it is important. In particular, I assume that equal weights are given to the reported A_i and the uniform prior.

I assume that individuals estimate $\bar{\pi}(c)$ from the other group member's contributions in the previous round. In round 1, I assume they expect $c = c_i + \sum_{i \neq j}(25A_j)$. Finally, I assume that decisions are made probabilistically using a logistic / quantal response decision rule with parameter λ that increases round on round to reflect the effect of experience. I present the results from 10,000 simulations in Figure 2. Figure 2 suggests comparative statics in the directions posited in Section 2, providing an example of how treatment differences could emerge.

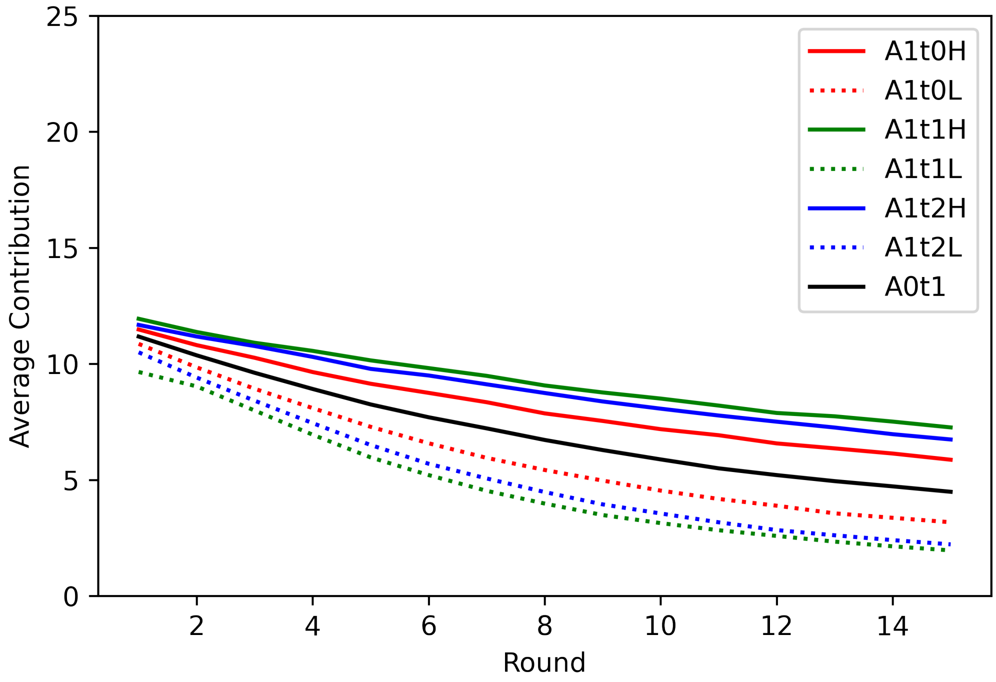
4 Planned Statistical Analysis

The statistical analysis is conducted using Python and Stata in a Jupyter Notebook, and this code is attached to the OSF project. I simulate a data-set using oTree bots based on Section 3, which has the advantage of outputting realistic data in the same file format that the actual experiment will have. On the simulated data-set, I conduct the statistical analysis in the same manner as I will on the actual data-set. As the statistical analysis is already programmed, I can credibly commit to this registered analysis.

I describe statistical results as strongly statistically significant when $p < 0.01$ (***),

[12] This is only for the simulation, subjects in the experiment are never told Agreeableness scores.

Figure 2: Simulation Results by Round



statistically significant when $p < 0.05$ (**), and weakly statistically significant when $p < 0.10$ (*). I report all statistical results using conservative two-sided tests regardless of whether the associated hypothesis is directional or not.

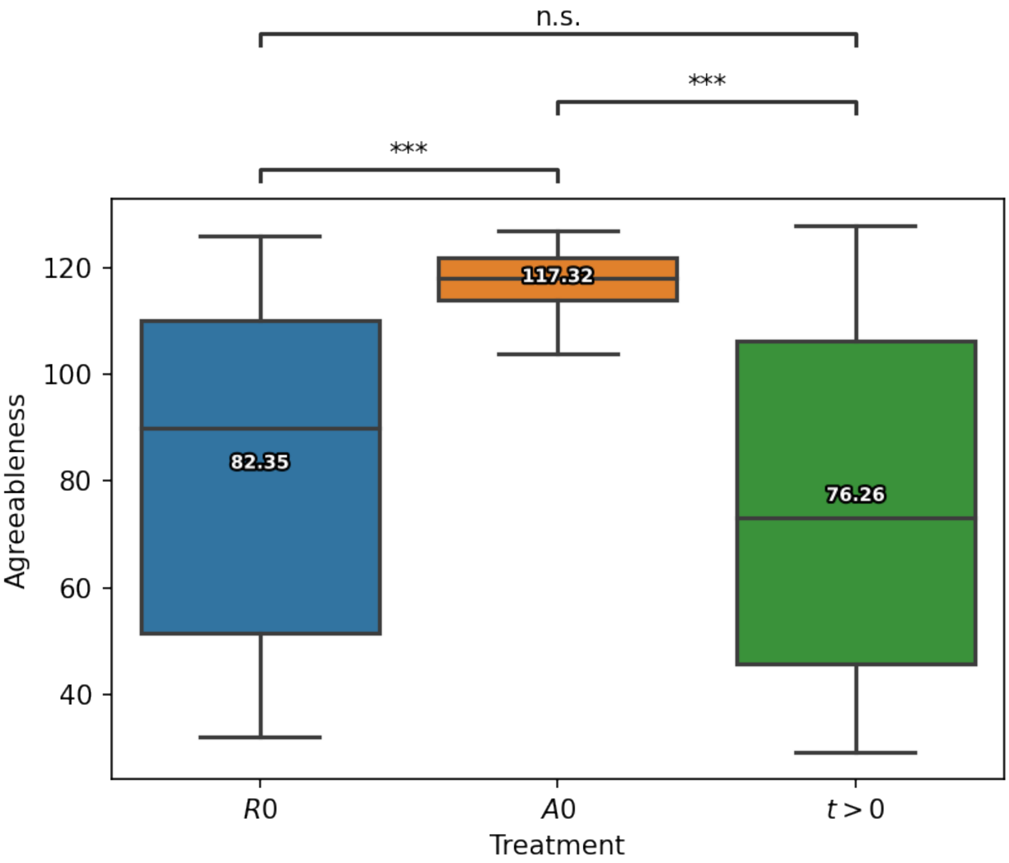
For each statistical test, I conduct a simulation-based power analysis in order to determine the minimum detectable effect size that attains 80% power given a $\alpha = 0.05$ rejection threshold. The power analysis is conducted in the same code as the statistical analysis. I report the minimum effect size for each test at the relevant points of the following statistical analysis.

4.1 Primary Analysis

4.1.1 Part 1: Strategic Misrepresentation of Agreeableness

As described in Section 2.1.1, there are three major groups of treatments: $A0$, $R0$, and all $t > 0$ treatments, because treatment differences can only impact Part 1 responses if they occur before Part 1. The outcome of interest is each subject's Agreeableness score, calculated from their responses to the Part 1 questions. Each of the three groups of treatments is tested using a Mann-Whitney test, with each subject being an independent observation.[13] The comparison between $R0$ and $A0$ tests Hypothesis 1, and the comparison between $R0$ and all $t > 0$ treatments tests Hypothesis 2. I report these results alongside summary statistics in Figure 3.

Figure 3: Agreeableness by Treatment



Mean Agreeableness overlaid. Statistical results are based on a two-sided Mann Whitney test. ***=$p < 0.01$, **=$p < 0.05$, and *=$p < 0.10$.

Figure 3 shows that Agreeableness scores do not differ between $R0$ and $t > 0$, which is evidence against Hypothesis 2 and suggests that any suspicion from the omission of the Part 2 group formation rule has a minimal impact in Part 1. Figure 3 also shows strong evidence in support of Hypotheses 1, as Agreeableness scores are much higher in the $A0$ treatment.

[13] The power analysis for the $R0$ to $A0$ comparison suggests a minimum detectable effect size of 7.7 units, and for the $A0$ *Others* comparison it is 5.1. These minimum detectable effect sizes are reasonable given they represent an average change of one or two out of the 26 Agreeableness questions being flipped from 1 to 5.

There is strong evidence of strategic misrepresentation when there is an incentive to report a higher level of Agreeableness.[14]

I also consider whether strategic misrepresentation is sophisticated or not. If misrepresentation is sophisticated then subjects only misrepresent the relevant trait of Agreeableness. However, if misrepresentation is unsophisticated, then responses will change for all of the elicited Big 5 characteristics. Table 5 reports the analysis for all Big 5 characteristics, and shows evidence for sophisticated misrepresentation as only Agreeableness is substantially different in the $A0$ treatment. When subjects have an incentive to misrepresent Agreeableness, they misrepresent Agreeableness but no other Big 5 traits, suggesting subjects are able to identify Agreeableness questions from other personality questions.

Table 5: All Big 5 characteristics by Treatment

Characteristic	$R0$	$A0$	$t > 0$	p-values
Agreeableness	82.35	117.32	76.26	0.00, 0.26, 0.00
Open Mindedness	18.21	17.76	18.08	0.42, 0.75, 0.46
Negative Emotionality	17.79	17.38	17.97	0.38, 0.90, 0.13
Extraversion	18.44	17.95	17.89	0.46, 0.40, 0.97
Conscientiousness	17.50	18.24	18.08	0.30, 0.21, 1.00

The treatment columns report the average score of the given personality trait. Agreeableness $\in [26, 130]$ and all other personality traits $\in [6, 30]$. The p-values column reports the results from Mann-Whitney tests on the pairs: $R0$ to $A0$; $R0$ to $t > 0$; and $A0$ to $t > 0$ respectively.

4.1.2 Part 2: PGG Contributions

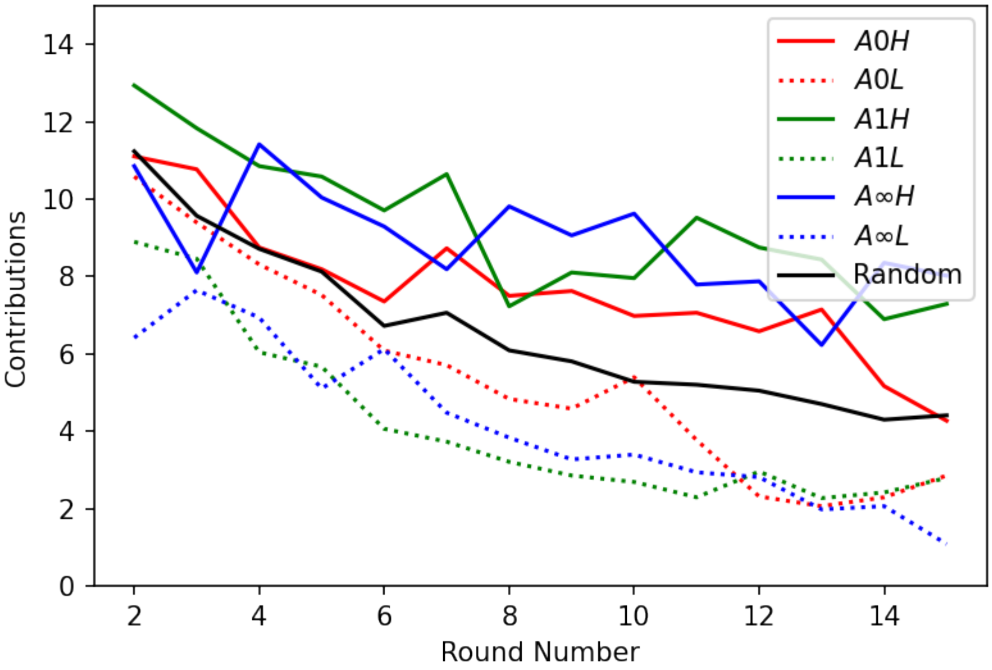
Figure 4 summarizes average contributions over time by treatment and group type. Table 6 reports the average contribution by each group member to the public good by treatment and group type. Both Table 6 and Figure 4 suggest that $H > L$, which justifies considering these two types of groups separately.

Table 6: Average Individual Contributions by Treatment

	$t = 0$	$t = 1$	$t = \infty$
AH	22.45	27.54	26.38
R	18.90	20.12	18.63
AL	15.85	12.00	11.95

[14] The reported 'results' in this paper are currently from simulated data using Section 3.

Figure 4: Average Contributions by Round



Instead of using an ultra conservative test where each group is a single independent observation and their contributions are averaged over all periods (Clark, 2002; Harrison, 2007), I conduct a more sophisticated statistical analysis that uses a panel data approach in order to utilize more of the data while accounting for the underlying dependencies. I use the group's average contribution in a period as the dependent variable, and a treatment dummy alongside the period for the independent variables. I use a panel-data Tobit regression for the possible censoring that occurs at 0 and 25 tokens for upper and lower limits respectively.[15] For each relevant comparison between two treatments (or group types within a treatment), I run the regression using only data from the pair that is being considered. Table 7 summarizes the results from the comparisons that are relevant for testing the proposed Hypotheses. Table 7 provides support for all tested hypotheses except for Hypotheses $5H$.

[15] A simulation based power analysis suggests a minimum detectable effect size of 1.7 tokens.

Table 7: Efficiency - Regressions

Pairwise Comparison	Coefficient	Hypothesis
Within $t = 0$		
$A0H$ - $R0$	1.18**	H3 +
$A0L$ - $R0$	-1.02**	H3 $-$
$A0H$ - $A0L$	2.21***	H3 +
Within $t = 1$		
$A1H$ - $R1$	2.47***	H3 +
$A1L$ - $R1$	-2.74***	H3 $-$
$A1H$ - $A1L$	5.21***	H3 +
Within $t = \infty$		
$A\infty H$ - $R\infty$	2.58***	H3 +
$A\infty L$ - $R\infty$	-2.25***	H3 $-$
$A\infty H$ - $A\infty L$	4.84***	H3 +
Within A_1		
$A0H$ - $A1H$	-1.70**	H4 $-$
$A0L$ - $A1L$	1.30***	H4 +
$A1H$ - $A\infty H$	0.39	H5 +
$A1L$ - $A\infty L$	0.02	H5 $\sim$
Within A_0		
$R0L$ - $R1L$	-0.41	H4 $\sim$
$R1L$ - $R\infty L$	0.50	H5 $\sim$

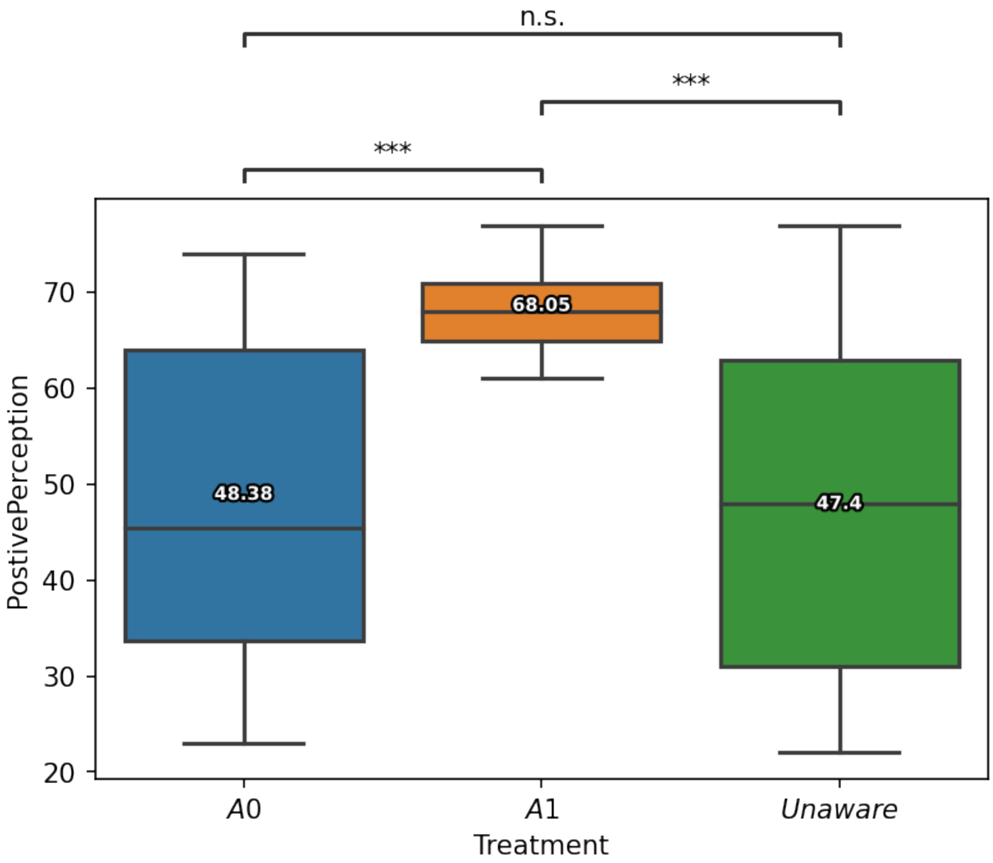
Second group in the pair is the omitted dummy. *Within X* are the groups of hypotheses holding X fixed. +, $-$, and $\sim$ indicate a positive, negative, or neutral predicted effect respectively. ***=$p < 0.01$, **=$p < 0.05$, and *=$p < 0.10$.

4.1.3 Part 3: Other Personality Measures

The main test in Part 3 is to detect whether 'Data Use' of questionnaires to sort groups affects future responses. There are three groups, *'Expected'* Data Use (of their Part 1 personality responses) ($A0$), *'Unexpected'* Data Use ($A1$), and all treatments where subjects are $Unaware$ of Data Use. I combine all of the characteristics elicited in Part 3 into one measure based on how likely it is they would be positively perceived by an observer. That is, I reverse code the Dark Triad as these traits are negative, and leave the coding for Honest-Humility as it is. I call this combined measure 'Positive Perception'. As there are 16 questions elicited on a 5-point Likert scale, it can take a minimum value of 16, and a maximum of 80. I use a Mann-Whitney test to test for differences between the three relevant subject groups.[16]

[16] A power analysis suggests minimum detectable effect sizes of 6.3 and 5.2 respectively.

Figure 5: Positive Perception by Treatment



Positive Perception mean overlaid. Statistical results are based on a two-sided Mann Whitney test. ***=$p <$ 0.01, **=$p < 0.05$, and *=$p < 0.10$.

Figure 5 summarizes the results of these comparisons. I find no evidence for Hypothesis 8, which suggests that the knowledge alone that personality traits can be used for group formation is not enough to influence responses on later personality tests. Whereas, I find strong evidence in support of Hypothesis 6 and 7, suggesting that data use needs to be both known and unexpected in order to contaminate later responses.

4.2 Exploratory Analysis

Naturally, the empirical findings of the experiment may suggest additional analysis that is unanticipated and thus not described above. It would be remiss to not follow the data where interesting results lie. These results can be used to generate new hypotheses, but should be viewed with an appropriate amount of skepticism until they are replicated in a future experiment. Any exploratory analysis will be clearly distinguished from the pre-registered analysis described above.

5 Conclusion

Using psychometric personality testing in the context of job hiring is a complex and sometimes controversial topic. These tests have become integral to modern hiring processes, helping firms to evaluate potential employees. However, a challenge is the incentive for job-seekers to tailor their responses to align with their beliefs of the employers' expectations. The incentive to strategically misrepresent one's preferences undermines the validity of such tests and their usefulness for job hiring decisions.

To shed light on this issue, I design and conduct an incentivized laboratory experiment that mirrors real-world hiring scenarios. I first elicit Big 5 characteristics through a questionnaire, much like what job-seekers have to fill out at some stage during the hiring process. I then use a standard PGG to represent a cooperative work environment. The Big 5 characteristic of Agreeableness has been found to positively impact contributions in previous studies, so sorting (or hiring) based on this trait makes sense. By changing the timing of the revelation of the sorting rule to before or after the initial questionnaire, I am able to quantify the level of misrepresentation and evaluate its subsequent impact on cooperative behavior.

I find that ... *(this paragraph and concluding statements will depend on the results).*

References

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

Anderson, L. R., Mellor, J. M., & Milyo, J. (2004, 4). Social Capital and Contributions in a Public-Goods Experiment. *American Economic Review: Papers & Proceedings*, *94*(2), 373–376. doi: 10.1257/0002828041302082

Arifovic, J., & Ledyard, J. (2012, 10). Individual evolutionary learning, other-regarding preferences, and the voluntary contributions mechanism. *Journal of Public Economics*, *96*(9-10), 808–823. doi: 10.1016/j.jpubeco.2012.05.013

Ashton, M., & Lee, K. (2009, 7). The HEXACO-60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment*, *91*(4), 340–345. doi: 10.1080/00223890902935878

Autor, D. H., & Scarborough, D. (2008, 2). Does Job Testing Harm Minority Workers? Evidence from Retail Establishments. *Quarterly Journal of Economics*, *123*(1), 219–277. doi: 10.1162/qjec.2008.123.1.219

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

Bartling, B., Fehr, E., Maréchal, M. A., & Schunk, D. (2009, 4). Egalitarianism and Competitiveness. *American Economic Review*, *99*(2), 93–98. doi: 10.1257/aer.99.2.93

Bock, O., Baetge, I., & Nicklisch, A. (2014, 10). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, *71*, 117–120. doi: 10.1016/j.euroecorev.2014.07.003

Borghans, L., Duckworth, A. L., Heckman, J. J., & Weel, B. t. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, *43*(4), 972–1059. doi: 10.3368/jhr.43.4.972

Braun, J. R., & Gomez, B. J. (1966, 10). Effects of Faking Instructions on the Eysenck Personality Inventory. *Psychological Reports*, *19*(2), 388–390. doi: 10.2466/pr0.1966

.19.2.388

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

Butera, L., Grossman, P. J., Houser, D., List, J. A., & Villeval, M. C. (2020). *A New Mechanism to Alleviate the Crises of Confidence in Science With An Application to the Public Goods Game.*

Cambridge Business English Dictionary. (2023, 7). *Psychometric Test - English Meaning - Cambridge Dictionary.* Retrieved from `https://dictionary.cambridge.org/dictionary/english/psychometric-test`

Carpenter, J. P., Daniere, A. G., & Takahashi, L. M. (2004, 12). Cooperation, trust, and social capital in Southeast Asian urban slums. *Journal of Economic Behavior & Organization*, *55*(4), 533–551. doi: 10.1016/j.jebo.2003.11.007

Cason, T. N., & Wu, S. Y. (2019, 7). Subject Pools and Deception in Agricultural and Resource Economics Experiments. *Environmental and Resource Economics*, *73*(3), 743–758. doi: 10.1007/s10640-018-0289-x

Catola, M., D'Alessandro, S., Guarnieri, P., & Pizziol, V. (2021, 10). Personal norms in the online public good game. *Economics Letters*, *207*, 110024. doi: 10.1016/j.econlet.2021.110024

Charness, G., Samek, A., & van de Ven, J. (2022, 4). What is considered deception in experimental economics? *Experimental Economics*, *25*(2), 385–412. doi: 10.1007/s10683-021-09726-7

Chaudhuri, A. (2011, 3). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47–83. Retrieved from `http://link.springer.com/10.1007/s10683-010-9257-1` doi: 10.1007/s10683-010-9257-1

Chen, D. L., Schonger, M., & Wickens, C. (2016, 3). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2214635016000101?via%3Dihub` doi: 10.1016/J.JBEF.2015.12.001

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism.* Elsevier. doi: 10.1016/

C2013-0-10497-7

Clark, J. (2002). House Money Effects in Public Good Experiments. *Experimental Economics*, *5*(3), 223–231. Retrieved from `http://link.springer.com/10.1023/A:1020832203804` doi: 10.1023/A:1020832203804

Donato, K., Miller, G., Mohanan, M., Truskinovsky, Y., & Vera-Hernández, M. (2017, 5). Personality Traits and Performance Contracts: Evidence from a Field Experiment among Maternity Care Providers in India. *American Economic Review: Papers & Proceedings*, *107*(5), 506–510. doi: 10.1257/aer.p20171105

Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011, 11). The framing of games and the psychology of play. *Games and Economic Behavior*, *73*(2), 459–478. doi: 10.1016/j.geb.2011.02.003

Dylman, A. S., & Zakrisson, I. (2023, 3). The effect of language and cultural context on the BIG-5 personality inventory in bilinguals. *Journal of Multilingual and Multicultural Development*, 1–14. doi: 10.1080/01434632.2023.2186414

Emergen Research. (2022, 2). *Assessment Services Market, By Product Type (Psychometric Test, Aptitude Tests, Coding Tests), By Service Type, By Medium (Online, Offline), By Sectors (K-12, Higher Education, Corporate, Government), and By Region Forecast to 2030* (Tech. Rep.).

Fehr, E., & Gächter, S. (2000, 9). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, *90*(4), 980–994. Retrieved from `http://pubs.aeaweb.org/doi/10.1257/aer.90.4.980` doi: 10.1257/aer.90.4.980

Fischbacher, U., & Föllmi-Heusi, F. (2013, 6). Lies in Disguise - An Experimental Study on Cheating. *Journal of the European Economic Association*, *11*(3), 525–547. doi: 10.1111/jeea.12014

Fréchette, G. R., Schotter, A., & Trevino, I. (2017, 7). Personality, Information Acquisition, and Choice Under Uncertainty: An Experimental Study. *Economic Inquiry*, *55*(3), 1468–1488. doi: 10.1111/ecin.12438

Gächter, S., & Thöni, C. (2005, 5). Social Learning and Voluntary Cooperation among like-Minded People. *Journal of the European Economic Association*, *3*(2-3), 303–314. doi: 10.1162/jeea.2005.3.2-3.303

Goldberg, L. (2002). *Big Five Factor Markers.* Retrieved from `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, 2). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. doi: 10.1016/j.jrp.2005.08.007

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

Hare, R. D. (1985, 2). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, *53*(1), 7–16. doi: 10.1037/0022-006X.53.1.7

Harrison, G. W. (2007, 11). House money effects in public good experiments: Comment. *Experimental Economics*, *10*(4), 429–437. Retrieved from `http://link.springer.com/10.1007/s10683-006-9145-x` doi: 10.1007/s10683-006-9145-x

Hawkins, T., & Monroe, M. (2021, 3). *Persona: The Dark Truth Behind Personality Tests.* HBO Max. Retrieved from `https://www.imdb.com/title/tt14173880/`

Hoffman, M., Kahn, L. B., & Li, D. (2018, 5). Discretion in Hiring. *The Quarterly Journal of Economics*, *133*(2), 765–800. doi: 10.1093/qje/qjx042

Holmén, M., Holzmeister, F., Kirchler, M., Stefan, M., & Wengström, E. (2021). *Economic Preferences and Personality Traits Among Finance Professionals and the General Population.* Innsbruck.

Jonason, P. K., & Webster, G. D. (2010, 6). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420–432. doi: 10.1037/a0019265

Kagel, J., & McGee, P. (2014, 8). Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters*, *124*(2), 274–277. doi: 10.1016/j.econlet.2014.05.034

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

Kroger, R. O., & Wood, L. A. (1993, 12). Reification, "faking," and the Big Five. *American Psychologist*, *48*(12), 1297–1298. doi: 10.1037/0003-066X.48.12.1297

Küfner, A. C. P., Dufner, M., & Back, M. D. (2015, 1). Das Dreckige Dutzend und die Niederträchtigen Neun. *Diagnostica*, *61*(2), 76–91. doi: 10.1026/0012-1924/a000124

Ledyard, J. (1995). Public Goods: A Survey of Experimental Evidence. In J. Kagel & A. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton: Princeton University Press.

Lee, K., & Ashton, M. C. (2009). *HEXACO-PI-R Materials for Researchers.* Retrieved from `https://hexaco.org/hexaco-inventory`

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, 3). An experimental study of finitely and infinitely repeated linear public goods games. *Games and Economic Behavior*, *102*, 286–302. doi: 10.1016/j.geb.2017.01.004

Maniadis, Z., Tufano, F., & List, J. A. (2014, 1). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, *104*(1), 277–290. doi: 10.1257/aer.104.1.277

McCrae, R. R., & John, O. P. (1992, 6). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

McGee, A., & McGee, P. (2022a). *Gender and race differences on incentivized personality measures.* Retrieved from `http://www.hivereview.org/project/35`

McGee, A., & McGee, P. (2022b). Whoever You Want Me to Be: Personality and Incentives. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4163677

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

Paulhus, D. L., & Williams, K. M. (2002, 12). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. doi: 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.

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2020, 1). 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. doi: 10.1027/1015-5759/a000481

Raskin, R. N., & Hall, C. S. (1979, 10). A Narcissistic Personality Inventory. *Psychological Reports*, *45*(2), 590–590. doi: 10.2466/pr0.1979.45.2.590

Soto, C. J., & John, O. P. (2017, 6). 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. doi: 10.1016/j.jrp.2017.02.004

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

Svorenčík, A. (2016, 12). The Sidney Siegel Tradition: The Divergence of Behavioral and Experimental Economics at the End of the 1980s. *History of Political Economy*, *48*(suppl_1), 270–294. doi: 10.1215/00182702-3619310

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

Velicer, W. F., & Weiner, B. J. (1975, 8). Effects of Sophistication and Faking Sets on the Eysenck Personality Inventory. *Psychological Reports*, *37*(1), 71–73. doi: 10.2466/pr0.1975.37.1.71

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

Villeval, M. C. (2020). Public goods, norms and cooperation. In C. M. Capra, R. Croson, M. Rigdon, & T. Rosenblat (Eds.), *Handbook of experimental game theory* (chap. 7). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.

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

Weber, L., & Dwoskin, E. (2014, 9). *Are Workplace Personality Tests Fair?* Retrieved from https://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

A Deviations from the Pre-registration Document

Any deviations from what was pre-registered will be documented here. They are expected to be minor or technical details, as any exploratory analysis will be already clearly labeled as such in the main body of the text.

B Conceptual Replication Results

B.1 Personality tests and incentives

The current experimental design permits a conceptual replication of some elements of McGee and McGee (2022b). In particular, Research Questions 1 and 2 from that paper can be partially answered.

MM Research Question 1: *How important are incentives when measuring personality?*

The incentives in McGee and McGee (2022b) were a direct lump-sum payment if selected for a job. Whereas in the current study the incentive is indirect, as it is membership in the more cooperative H group that could increase earnings in the PGG. Research Question 1 is addressed by Hypothesis 1 and the comparisons in Figure 3. As there is strong evidence for Hypothesis 1, I conclude that indirect incentives are also important when measuring personality. This conceptual replication increases our confidence in, as well as the generalizability of, the results reported in McGee and McGee (2022b) that address their Research Question 1.

MM Research Question 2: *Are incentivized personality measures influenced by traits other than personality?*

McGee and McGee (2022b) posit that traits such as intelligence, Machiavellianism, self-deception, optimism, acceptability of lying, risk aversion, and locus of control could be correlated with misrepresentation. They find that most of these characteristics are uncorrelated with misrepresentation in all treatments of their experiment.

In particular, McGee and McGee (2022b) find no evidence that Machiavellianism is correlated with misrepresentation in any of their treatments. This is an interesting result, given that people high in Machiavellianism tend to be manipulative and strategically self-serving in their words and actions. In this paper, Part 3 elicits Machiavellianism using a different set of questions, and its relationship to misrepresentation of Agreeableness in Part 1 can be explored. I test this relationship with a Tobit regression censored at 26 and 130[17] of the following form $Agreeableness_i = \beta_0 + \beta_1 Machiavellianism + \beta_2 Machiavellianism \times A0 + \epsilon_i$. The coefficient β_1 represents the correlation between Agreeableness and Machiavellianism,

[17] The minimum and maximum value that Agreeableness can take in this experiment.

Table A1: Personality Traits and Agreeableness Misrepresentation

Trait	β_1	β_2
Honesty Humility	2.71***	2.43***
Machiavellianism	-4.91***	3.90***
Narcissism	-4.68***	3.83***
Psychopathy	-4.51***	3.80***

β_1 represents the correlation between the trait and Agreeableness, and β_2 represents the correlation between the trait and misrepresentation of Agreeableness.***=$p < 0.01$, **=$p < 0.05$, and *=$p < 0.10$.

and β_2 represents the increase (if > 0) in reported Agreeableness when there is an incentive to misrepresent (i.e. in $A0$).[18] However, such a regression would only be appropriate in the event that there is misrepresentation in Part 1 but not in Part 3, i.e. if there is support for Hypothesis 1 but not for Hypothesis 6.

Alongside Machiavellianism, all of the traits elicited in Part 3 could also address this research question using the same statistical test. I expect that Honesty-Humility will be correlated with Agreeableness, but not with misrepresentation of Agreeableness. Psychopathy likely follows a similar pattern of behavior as Machiavellianism. The effect that Narcissim would have is unclear, and I do not offer any predictions.

Table A1 summarizes and shows that Agreeableness is increasing in Honesty Humility and decreasing in the Dark Triad traits, as expected, and that misrepresentation is increasing in all personality traits.[19]

B.2 Individual characteristics and contributions in the PGG

A variety of papers have considered the impact of Big Five personality traits (Perugini et al., 2010; Volk et al., 2012) and/or other individual characteristics (Anderson et al., 2004; Carpenter et al., 2004; Catola et al., 2021) on contributions in the PGG. There are also papers that consider the impact of the Big Five on other pro-social actions in other strategic games (Kagel & McGee, 2014; Thielmann et al., 2020). This line of literature can be broadly summarized in by the following research question.

Research Question: *How do individual characteristics impact pro-social behavior?*

[18] A simulation-based power analysis suggests a minimum detectable effect size of around 1 unit.

[19] The simulated data didn't model this in detail, so these results are currently a little unintuitive.

The current paper can address this research question using the elicited individual characteristics in Parts 1 and 3. I use a mixed-effects panel tobit regression (censored at 0 and 25) clustered at the individual and group levels. The regression includes all of the elicited personality characteristics and demographics alongside treatment dummies and the average group contribution by others in the previous period.[20] I exclude the $A0$ treatment data from this regression, as I anticipate misrepresentation in Agreeableness in this treatment.[21] Table A2 lists the relevant coefficients from this regression, and suggests that the personality traits of Agreeableness, Negative Emotionality, and Honesty Humility all positively effect an individual's contributions. The other personality traits and demographics do not appear to substantially impact an individual's contribution rate.

[20] Following Bardsley and Moffatt (2007), the initial lagged contribution in period 1 is found using a grid search.

[21] If misrepresentation is detected in other treatments in Part 1 or 3, then as a robustness check I will also conduct this regression excluding those treatments.

Table A2: Individual Characteristics on Contributions

Independent Variable	Coefficient
Lagged Avg. Group Cont.	0.16***
Agreeableness	0.03***
Open Mindedness	-0.01
Negative Emotionality	0.06**
Extraversion	0.04
Conscientiousness	0.05*
Honesty Humility	0.17***
Machiavellianism	-0.04
Narcissism	-0.03
Psychopathy	-0.01
Female	-0.16
2nd Year at Uni.	0.14
3rd Year at Uni.	0.09
4th+ Year at Uni.	-0.20
Grad. Student	-0.12
GPA	-0.06
Black	-0.46
Caucasian	-0.45
Hispanic	-0.25
Other Race	-0.34
Economics	-0.45
Arts and Humanities	-0.20
Natural Sciences	-0.52*
Education	-0.02

An individual's contribution to the public good is the dependent variable. Results are from a mixed-effects panel tobit regression (censored at 0 and 25) clustered at the individual and group levels. Controls for Treatment and Period are included in the regression but not listed. ***=$p < 0.01$, **=$p < 0.05$, and *=$p < 0.10$.

C Instructions

Full Instructions are provided in the OSF project and/or replication packet - either in the oTree code or a .docx file for the paper instructions. Selected parts of the Instructions that are particularly relevant for the understanding of the experiment are presented below.

C.1 Part 1

C.1.1 First Screen

This experiment will have two parts.

Part 1 will be a set of questions about yourself. We ask that you answer these questions accurately.

Part 2 has 15 decision rounds. A brief summary of one decision round follows:

- Subjects are in groups of 3
- Each subject has 25 tokens that they divide between their Private Account or a Cooperation Account
- Each token placed in their Private Account earns that subject 10 points.
- Each token placed in the Cooperation Account earns the entire group 12 points.
- Everyone in the group receives an equal portion of the earnings from the Cooperation account, that is, they earn 12*1/3=4 points per token in the Cooperation Account.

This is only a basic outline of Part 2. More instructions will be provided before starting Part 2.

[$t > 0$ Treatments:]

We will now start with Part 1 - the set of questions about yourself.

C.1.2 Second Screen

[Second Screen only in $t = 0$ Treatments]

[Random Treatments:] For Part 2, you will be assigned to a group of three **randomly**.

[*A*greeableness Treatments:] 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 states:

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.**

[All $t = 0$ treatments:] Each group of three will remain together for all 15 decisions in Part 2.

We will now proceed with Part 1 - the set of questions about yourself.

C.2 Part 2

[$t \leq 1$ Treatments:]

[*A*greeableness treatments:] 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.

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.**

[*R*andom treatments:] For Part 2, you will be assigned to a group of three **randomly**.

C.3 Part 3

Parts 1 and 2 of the experiment are now complete.

We ask you to fill out a final short survey, before your final earnings are displayed. Your final earnings have already been calculated and set.

There are no further parts to the experiment after this final survey.

D Personality Questions

The 50 questions in Part 1 are taken from the 30 question 'BFI-2-S Inventory' (Soto & John, 2017), and the 20 questions on Agreeableness from the International Personality Item Pool's (IPIP) '100-Item Lexical Big-Five Factor Markers' (Goldberg, 2002; Goldberg et al., 2006). The 16 questions in Part 3 are taken from the 'Dirty Dozen' (Jonason & Webster, 2010) and four Honesty-Humility questions from HEXACO's 60-item version (Lee & Ashton, 2009). Subjects 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. Subjects face blocks of five questions on a page, and all questions are presented in a random order that differs across subjects.[22] Personality traits are scored based on each subject's numerical (i.e. 1-5) responses by the following formula: $Trait = \sum_{i \in Q}(LikertValue_{+veKey} + (6 - LikertValue_{-veKey}))$, where Q is the set of relevant questions to that trait. Appendices D.1 and D.2 report which questions are related to each trait and whether the questions are positively or negatively keyed.

As the experiment is conducted in Austria the experiment is conducted in German. While the majority of the university students that make up the subject pool are fluent in English, it is important to conduct personality tests in their native language. Firstly, there will be heterogeneity in subjects' confidence or ability in using English. Secondly, there is a literature that suggests that elicited personality traits are different in bilingual speakers depending on what language is being used (see Dylman and Zakrisson (2023) for examples). I would rather observe a subject's 'regular' personality rather than one that is shaped by a foreign language. Strategic misrepresentation of personality is already likely to be difficult enough as it is, let alone with an additional levels of complexity on top of that. The question sets used in Parts 1 and 3 all have pre-existing German translations. Rammstedt et al. (2020) translate the BFI-2-S. Streib and Wiedmaier (2001) translate the 100-Item IPIP. Küfner, Dufner, and Back (2015) translate the Dirty Dozen. A translation for HEXACO is provided by Lee and Ashton (2009). A list of the questions and their translations are

[22] Technically it is possible that two subjects face exactly the same ordering, however this is unlikely as the probability of that occurring in Part 1 is $p = \frac{1}{50!}$ and in Part 3 $p = \frac{1}{16!}$.

provided in Appendices D.1 and D.2.

Some questions were changed or removed. In Part 1, a question was changed slightly to avoid excessive repetition, from *'I am compassionate and soft-hearted'* to *'I am compassionate'*, as another question is *'I have a soft heart'*. Two less relevant questions from Honesty-Humility were removed in order to maintain an equal number of questions between each trait in Part 3. The removed questions are *'I'd be tempted to use counterfeit money, if I were sure I could get away with it.'* and *'If I knew that I could never get caught, I would be willing to steal a million dollars.'*

Some of the pre-existing translations were changed based on feedback from native German speakers. The question *'I have a soft heart'* was changed from *'Ich habe ein weiches Herz'* to *'Ich bin gutherzig'*. The question *'I have a good word for everyone'* was changed from *'Ich habe ein gutes Wort für jeden'* to *'Ich rede gut über andere'*. These two changes were implemented as the original translations were considered ambiguous and a little too literal. The question *'I make people feel at ease'* was changed from *'Ich mache andere Leute ungezwungen'* to *'Ich kann andere beruhigen'*. Ungezwungen can be interpreted as being unhinged rather than calm, and may also be grammatically incorrect. The question *'Ich habe getäuscht oder gelogen, um meinen Willen durchzusetzen'* was changed to *'Ich neige dazu, zu täuschen oder zu lügen, um meinen Willen durchzusetzen'*, and similarly the question *'Ich habe Schmeicheleien genutzt, um meinen Willen durchzusetzen'* to *'Ich neige dazu, Schmeicheleien zu benutzen, um meinen Willen durchzusetzen'*. The other questions in the Dirty Dozen all have *'Ich neige dazu'* (I have the tendency to), and I was concerned that the question about lying could be interpreted as whether they have been deceitful in the current experiment, rather than a tendency in general.

D.1 Part 1 Questions and Translations

English Questions	**German Questions**
Agreeableness Positively Keyed	
I am interested in people.	Ich interessiere mich für Leute.
I sympathize with other's feelings.	Ich kann die Gefühle anderer nachempfinden.
I have a soft heart.	Ich bin gutherzig.
I take time out for others.	Ich nehme mir Zeit für andere.
I feel other's emotions	Ich kann die Gefühle anderer nachfühlen.
I make people feel at ease.	Ich mache andere Leute ungezwungen.
I inquire about other's well-being.	Ich erkundige mich nach dem Wohlbefinden anderer.
I know how to comfort others.	Ich weiß wie ich andere trösten kann.
I love children.	Ich liebe Kinder.
I am on good terms with nearly everyone.	Ich komme mit fast jedem gut aus.
I have a good word for everyone.	Ich rede gut über andere.
I show my gratitude.	Ich zeige meine Dankbarkeit.
I think of others first.	Ich denke zuerst an andere.
I love to help others.	Ich liebe es anderen zu helfen.
I am compassionate.	Ich bin einfühlsam.
I assume the best about people.	Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.
I am respectful and treat others with respect.	Ich begegne anderen mit Respekt.
Agreeableness Negatively Keyed	
I insult people.	Ich beleidige Leute.
I am not interested in other people's problems.	Ich interessiere mich nicht für die Probleme anderer Leute.
I feel little concern for others.	Andere Menschen kümmern mich wenig.
I am not really interested in others.	Ich interessiere mich nicht wirklich für andere.
I am hard to get to know.	Mich kennenzulernen ist schwer.
I am indifferent to the feelings of others.	Ich bin den Gefühlen anderer gegenüber gleichgültig.
I am sometimes rude to others.	Ich bin manchmal unhöflich und schroff.
I can be cold and uncaring.	Andere sind mir eher gleichgültig, egal.
I tend to find fault with others.	Ich neige dazu, andere zu kritisieren.
Extraversion Positively Keyed	
I am dominant and act as a leader.	Ich neige dazu, die Führung zu übernehmen.
I am full of energy.	Ich bin voller Energie und Tatendrang.
I am outgoing and sociable.	Ich gehe aus mir heraus, bin gesellig.
Extraversion Negatively Keyed	
I tend to be quiet.	Ich bin eher ruhig.
I prefer to have others take charge.	In einer Gruppe überlasse ich lieber anderen die Entscheidung.
I am less active than other people.	Ich bin weniger aktiv und unternehmungslustig als andere.

Table A3: Part 1 Questions 1-32

English Questions	**German Questions**
Conscientiousness	*Positively Keyed*
I am reliable and can always be counted on.	Ich bin verlässlich, auf mich kann man zählen.
I keep things neat and tidy.	Ich mag es sauber und aufgeräumt.
I am persistent and work until a task is finished.	Ich bleibe an einer Aufgabe dran, bis sie erledigt ist.
Conscientiousness	*Negatively Keyed*
I tend to be disorganized.	Ich bin eher unordentlich.
I have difficulty getting started on tasks.	Ich neige dazu, Aufgaben vor mir herzuschieben.
I can be somewhat careless.	Ich bin manchmal ziemlich nachlässig.
Negative Emotionality	*Positively Keyed*
I worry a lot.	Ich mache mir oft Sorgen.
I tend to feel depressed and blue.	Ich bin oft deprimiert, niedergeschlagen.
I am temperamental and get emotional easily.	Ich reagiere schnell gereizt oder genervt.
Negative Emotionality	*Negatively Keyed*
I am emotionally stable and not easily upset.	Ich bin ausgeglichen, nicht leicht aus der Ruhe zu bringen.
I am relaxed and handle stress well.	Ich bleibe auch in stressigen Situationen gelassen.
I feel secure and comfortable with myself.	Ich bin selbstsicher, mit mir zufrieden.
Open-mindedness	*Positively Keyed*
I am fascinated by art, music, or literature.	Ich kann mich für Kunst, Musik und Literatur begeistern.
I am original and come up with new ideas.	Ich bin originell, entwickle neue Ideen.
I am complex and a deep thinker.	Es macht mir Spaß, gründlich über komplexe Dinge nachzudenken und sie zu verstehen.
Open-mindedness	*Negatively Keyed*
I have little interest in abstract ideas.	Mich interessieren abstrakte Überlegungen wenig.
I have few artistic interests.	Ich bin nicht sonderlich kunstinteressiert.
I have little creativity.	Ich bin nicht besonders einfallsreich.

Table A4: Part 2 Questions 33-50

D.2 Part 3 Questions and Translations

English Questions	**German Questions**
Narcissism Positively Keyed	
I tend to want others to admire me.	Ich neige dazu, von anderen bewundert werden zu wollen.
I tend to want others to pay attention to me.	Ich neige dazu, von anderen beachtet werden zu wollen.
I tend to expect special favors from others.	Ich neige dazu, besondere Gefälligkeiten von anderen zu erwarten.
I tend to seek prestige or status.	Ich neige dazu, nach Ansehen oder Status zu streben.
Psychopathy Positively Keyed	
I tend to lack remorse.	Ich neige dazu, keine Gewissensbisse zu haben.
I tend to be callous or insensitive.	Ich neige dazu, gefühllos oder unsensibel zu sein
I tend to not be too concerned with morality or the morality of my actions.	Ich neige dazu, mich nicht um die Moral meiner Handlungen zu kümmern.
I tend to be cynical.	Ich neige dazu, zynisch zu sein.
Machiavellianism Positively Keyed	
I have used deceit or lied to get my way.	Ich neige dazu, zu täuschen oder zu lügen, um meinen Willen durchzusetzen.
I tend to manipulate others to get my way.	Ich neige dazu, andere zu manipulieren, um meinen Willen durchzusetzen.
I have used flattery to get my way.	Ich neige dazu, Schmeicheleien zu benutzen, um meinen Willen durchzusetzen.
I tend to exploit others towards my own end.	Ich neige dazu, andere für meine Zwecke auszunutzen.
Honesty-Humility Positively Keyed	
I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed.	Ich würde keine Schmeicheleien benutzen, um eine Gehaltserhöhung zu bekommen oder befördert zu werden, auch wenn ich wüsste, dass es erfolgreich wäre.
I wouldn't pretend to like someone just to get that person to do favors for me.	Ich würde nicht vortäuschen, jemanden zu mögen, nur um diese Person dazu zu bringen, mir Gefälligkeiten zu erweisen.
I would never accept a bribe, even if it were very large.	Ich würde niemals Bestechungsgeld annehmen, auch wenn es sehr viel wäre.
Honesty-Humility Negatively Keyed	
If I want something from someone, I will laugh at that person's worst jokes.	Wenn ich von jemandem etwas will, lache ich auch noch über dessen schlechteste Witze.

Table A5: Part 3 Questions