Group-based Biases Influence Learning about Individual Trustworthiness

Marieke Vermue[1], Charles R. Seger[1], & Alan G. Sanfey[2, 3]

[1]School of Psychology, University of East Anglia, Norwich, United Kingdom

[2] Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands

[3]Behavioural Science Institute, Radboud University Nijmegen, the Netherlands

Corresponding author:

Marieke Vermue (m.vermue@uea.ac.uk)

Author Note

The authors would like to thank Natalie Wyer and Piers Fleming for their comments on an earlier version of this manuscript, and Will Penny for his assistance with sample size and statistical analyses.

Abstract

People often have generalised expectations of trustworthiness about ingroup and outgroup members, based on previous direct and indirect experience with these groups. How do these prior biases interact with new experiences when learning about individual group members' trustworthiness? These three studies are the first to examine the effect of group-level biases on learning about individuals' trustworthiness. Participants from the Netherlands and the United Kingdom played iterated Trust Games with trustworthy and untrustworthy members of both ingroups and outgroups. We show that the influence of group membership on trust decisions depended on the valence of the interactions with individual group members. When interacting with trustworthy partners, people displayed outgroup favouritism throughout the game, investing higher in outgroup members than ingroup members. However, for untrustworthy partners, initial outgroup favouritism disappeared, and ingroup and outgroup members were equally distrusted by the end of the game. Our work suggests that when individual experience is integrated with group-based biases, group membership influences trust decisions over time, but mostly when experiences are positive. These findings are discussed in relation to complexity-extremity theory and previous work on learning in the Trust Game.

Keywords: INTERGROUP BIAS, TRUST, INGROUP FAVORITISM, LEARNING, TRUST GAME

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Group-based Biases Influence Learning about Individual Trustworthiness

Imagine you are at a bar and see one person wearing a New England Patriots jersey, another sporting a Republican Party badge, and a third with a German accent talking to their friend. What would you infer about their personality and their attitudes? Which of these people would you choose to ask for a favour or trust to look after your bag? We quickly categorize others in terms of their group membership (Bargh, 1999; Willis & Todorov, 2006) and the social categories to which others belong are vitally important cues for making decisions about how we then act towards them (Balliet, Wu, & De Dreu, 2014). In three studies, we investigate how social category biases interact with individual experiences in forming decisions to trust.

Feelings of trust are essential for successful cooperation, particularly when the other person is relatively unknown to you, and you cannot therefore rely on previous experiences with the person (Balliet & Van Lange, 2013). In this situation, feelings of trust come from external cues such as a person's physical features (e.g. Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Todorov, Pakrashi, & Oosterhof, 2009) and particularly their group membership (Williams, 2001). Generally, people exhibit more trust, cooperation and positive reciprocation towards ingroup members than outgroup members (Balliet et al., 2014). This so-called ingroup bias for trust has been extensively observed using well-validated economic games, such as the Trust Game (Berg, Dickhaut, & McCabe, 1995). In this game, a trustor is given an endowment that he/she can invest in a trustee. If the trustor invests his/her endowment, the amount is multiplied and given to the trustee. The trustee then has the choice to reciprocate trust by returning some of the received amount to the trustor, but he/she does not have to. Both players can end the game with more money than they started out with, but only if they both cooperate. Ingroup favouritism in these cooperative settings has been found with many types of naturally occurring groups, such as race (e.g. Burns, 2006), nationality

(e.g. Stoddard & Leibbrandt, 2014), or religion (e.g. Rotella, Richeson, Chiao, & Bean, 2013) as well as in a minimal-group setting (e.g. Buchan, Johnson, & Croson, 2006).

However, people do not always prefer the ingroup or individual ingroup members. A considerable amount of research shows that outgroup preferences can exist when that group is perceived as high status (Jost, Pelham, & Carvallo, 2002; Trifiletti & Capozza, 2011), or high in warmth and competence (Cuddy, Fiske, & Glick, 2008). Even ethnic majorities can occasionally show outgroup preferences towards minority groups (Jussim, Coleman, & Lerch, 1987). One theory that accounts for how individual ingroup members can be viewed less favourably than outgroup members, particularly once some learning occurs, is the *Black Sheep Effect* (BSE; Marques, Yzerbyt, & Leyens, 1988). In the BSE, people punish deviating ingroup members more strongly than deviating outgroup members. This has the purpose of maintaining a positive image of the ingroup, which is vital for maintaining a positive social identity (Tajfel & Turner, 1979). The BSE predicts that extreme ingroup devaluations occur when the deviating member is relevant to one's social identity, and identification with the ingroup is strong.

Another theory that highlights differences between how ingroup and outgroup members are represented is *Complexity-Extremity Theory* (CET; Linville, 1982). CET accounts for situations in which ingroup members (or the ingroup as a whole) can be rated as less favourable than outgroup members, and describes situations in which negativity towards outgroups may be exacerbated. According to Linville, people's representations of outgroups are less complex than for ingroups, which leads to more extreme evaluations of outgroup members than ingroup members for both positive and negative information. Therefore, each piece of information about an outgroup member changes the evaluation more than when similar information is provided about an ingroup member. This can therefore lead to outgroup favouritism (Jussim et al., 1987).

Thus, these two theories both provide cases where the ingroup would not be favoured over the outgroup, but different patterns are predicted. According to BSE, ingroup members are generally favoured over the outgroup in positive situations, but more strongly devalued when behaving negatively. Complexity-extremity theory, however, would predict that the outgroup is evaluated more extremely positive than the ingroup when both are presented in a positive situation, but more extremely negative when negative information about these groups is learned.

Updating Trustworthiness Impressions

Previous studies have focused on *initial* trust reactions towards ingroup or outgroup members. These studies employed one-shot Trust Games whereby players interact only a single time with another person. However, this is not analogous to real-world settings, which require interaction over some course of time. We are interested in how group biases influence judgments about individuals' trustworthiness when *experience* becomes available. Group membership is a useful piece of information when having to decide on an initial response. Once you gain experience with an individual, group-based expectations should be integrated with the information you have learned. The aim of the current paper is to examine whether, and how, group information influences decisions to trust in these iterated settings, when people have to integrate experience with initial group-based biases.

The influence of group-level biases on learning about individuals' trustworthiness has not been examined before. However, studies utilising an iterated Trust Game show that people are able to learn about the behaviour of individual partners over multiple interactions, and adjust their trust decisions accordingly (Chang et al., 2010; Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; Fouragnan et al., 2013). In these studies, information about characteristics related to the partner's trustworthiness, as well as the amounts that the partner returned (reciprocity behaviour), were manipulated. Chang and colleagues (2010)

found initial beliefs based on facial trustworthiness of the partner influenced initial trust decisions, and remained important throughout the entire game. In the last round of the iterated game, participants invested more in the trustworthy appearing partners that acted trustworthy than in the untrustworthy appearing partners that showed similar trustworthy behaviour. However, investments were lower for the trustworthy-appearing partners that did not reciprocate trust than for untrustworthy appearing partners that did not reciprocate. In the current study, we examine the influence of group-based biases on investment decisions, instead of the individual-based biases of facial trustworthiness.

Overview of Studies and Hypotheses

The present research consists of three studies in two different European countries, with group membership manipulated through nationality. We adopted an iterated Trust Game paradigm, where participants played multiple rounds with several purported individuals from the ingroup and outgroup. Trustworthiness of behaviour was manipulated by pre-programming the return behaviour of the partner to be high or low. Study 1 and 2 explored ingroup and outgroup trust in the iterated Trust Game in two different European countries, the Netherlands and the United Kingdom. The outgroup consisted of people from different European foreign nationalities. Study 3 dived deeper into the underlying processes and examined how perceptions of trustworthiness, expectations of return, and affective feelings towards the partners are related to changes in investments over time. Moreover, in Study 3 the outgroup was restricted to one outgroup nationality to control for possible stereotype perceptions of the different countries.

We predicted that, based on the research described above, players should learn to distinguish between trustworthy and untrustworthy partners based on the game experiences over trials. Secondly, based on the literature on ingroup favouritism in cooperation (Balliet et al., 2014), and Chang and colleagues work, we hypothesised that, should initial ingroup

favouritism occur, trustworthy ingroup members would receive higher investments than trustworthy outgroup members across repeated interactions.

Thirdly, and most interestingly, both the BSE and Chang et al.'s (2010) study would predict that responses to untrustworthy ingroup members should be more negative than responses to untrustworthy outgroup members, as untrustworthy ingroup members defy the positive image and expectations of the ingroup, but this does not apply to the outgroup. However, complexity-extremity theory predicts that responses to outgroup members are more extreme for both positive and negative reciprocity, due to a low complexity of the group representation. From this theory, we would expect investments in trustworthy outgroup members to be higher than for trustworthy ingroup members, and investments in untrustworthy outgroup members to be lower than for untrustworthy ingroup members.

Study 1

Our first experiment was conducted in the Netherlands, with Dutch participants playing repeated Trust Games with (pre-programmed response) partners who were supposedly Dutch (ingroup) or from another Western European country (outgroup). In addition to the Trust Game, we measured expectations that participants had about these partners before playing the game, and the certainty of those expectations. Partners were also rated individually on trustworthiness, likeability and generosity after the game. Ingroup (Dutch) identification was additionally measured.

In accordance with past research (Ashraf, Bohnet, & Piankov, 2006; Buchan & Croson, 2004), we predicted that expectations would be related to any biases found in investment behaviour. Based on the results of Chang et al. (2010), partner ratings following the game were expected to reflect both learning from the game, with higher ratings for trustworthy than untrustworthy partners, and congruency with any initial group-level biases.

Ingroup identification may be related to ingroup bias (Kenworthy & Jones, 2009; Voci, 2006), with people that identify stronger with the ingroup showing more differentiation between ingroup and outgroup investments than people that identify to a lesser extent.

Methods

Participants and design. Participants (N = 40 students[1]) were recruited partly via an online database and through verbal recruitment. Four participants were removed from analyses due to double or different nationalities. The remaining 36 participants (75% female; M_{age} = 24.1 years, SD_{age} = 7.8 years) were Dutch. Participants received either a standard payment of €10 or course credit for their time. Ten participants were selected at random to receive a bonus monetary amount, which consisted of their average earnings in the game, converted to euros (one token = 40 euro cent).

A 2 (group: ingroup vs outgroup) x 2 (reciprocity behaviour: low vs high) within subject design was employed, with trial number (1 to 15) as continuous predictor. The main dependent variable was the investment transferred to the partner, which could range from zero to ten tokens. In addition to the Trust Game data, expectations of return and partner ratings were examined as dependent variables.

Materials and procedure. The experiment consisted of three parts, was performed on a computer and run on the software programme PsychoPy (Peirce, 2007). Initially, participants completed demographic questions and were presented with their eight game

[1] Sample size was guided by a power analysis performed in GPower 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) for a repeated measures F test. The power analysis was based on a medium effect size *f* of .15, a power of .80, 4 groups (Group x Reciprocity), and 15 measurements (individual trial number), producing a required sample size of 32.

partners. Participants were informed that these partners had played the game previously, and that their answers had been saved and would be matched with the investments that the participant made in the game. In reality, all the partner behaviours in the game were pre-programmed, and all information provided about these partners was predetermined. To introduce all game partners, participants were presented with a card with information for each partner. On this card, information about the nationality, age, gender, and academic major of the partner was presented (all predetermined). This information was provided to heighten the (false) sense of the partners being real people, as people can behave differently when they believe they are playing with a computer versus with a person (Chang et al., 2010). This information was only shown before the game began. Additionally, a figure was shown with the colours of the flag of the country, and the initials and age of the partner (see Figure 1), which would represent the partner throughout the Trust Game. Four of these partners were Dutch (ingroup), and four were foreigners (outgroup), consisting of two Germans, one Italian, and one French partner. The group membership manipulation was implemented through the coloured figures that were presented to the participant throughout the game.

Figure 1. Examples of figures used to indicate the partners. The figure on the left portrays and 20-year-old ingroup member, D.V. The figure on the right, figure O.B. is a 26-year-old German outgroup member.

Repeated Trust Game. Following this, the Trust Game was played. After six practice trials, participants played 15 rounds of the repeated Trust Game with each of the eight individual partners, thereby playing 120 rounds in total (see Figure 2). The order of the 120 rounds was fully randomised for each participant. In each round, participants played only the

role of trustor, where they decided how many of their 10 tokens to invest in the current partner. At the start of each round, participants were shown their current partner for three seconds (see Figure 1), then they were prompted to decide how many of the ten tokens to invest in that partner. The image of the partner remained on the screen throughout the trial. Any investment was tripled by the experimenter and sent to the partner; this multiplied amount was presented on the screen for three seconds. Last, participants were shown how many tokens the partner had transferred back to the participant on that round. This reciprocation rate was predetermined such that half of the partners always transferred a high amount in each round (45% to 70% of the received amount on each trial), and half of the partners always transferred a low amount (0% to 35% of the received amount). See Figure 2 for a visual representation of this procedure.

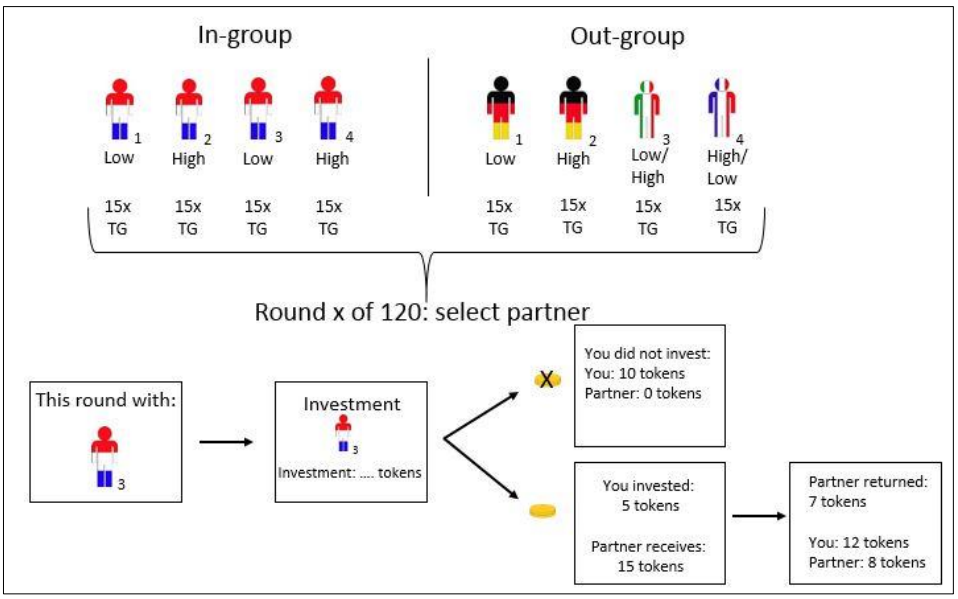


Figure 2. Visual representation of the design of the Trust Game and example of one round of the Trust Game. From the pool of eight partners (four ingroup, four outgroup, four high trustworthy, four low trustworthy), one partner is selected per round of the Trust Game.

Before starting with the game, but after receiving instructions, participants indicated their expectations of how much they thought each of the partners would return, as indicated

by percentages of the amount they would receive, on a scale ranging from 0% to 100%. Next, they were asked for their confidence in this expectation, also on a scale from 0% to 100%.

Following the Trust Game, participants rated each of the eight partners on generosity displayed during the game as well as general characteristic of kindness and trustworthiness, on a scale from zero to 100. This created a reliable index of partner ratings (α = .97). Finally, participants completed a 14 item ingroup identification questionnaire (Leach et al., 2008). Higher numbers indicate stronger identification with Dutch people (α = 0.89). This procedure was approved by a research committee before data collection commenced.

Data analysis. Data were analysed using linear multilevel models (also called mixed-effects models, e.g. Baayen, Davidson, & Bates, 2008), whereby per-subject random intercepts were added to incorporate the within-subject design. Per-subject random slopes of the relevant variables were included in the various models. Models of increasing complexity were compared, where the variables of interest (group[2], reciprocity behaviour and trial number) and their interactions were added in sequential steps[3]. Regression coefficients and bootstrapped 95% confidence intervals are reported alongside inferential statistics as simple effect sizes (Baguley, 2009).

[2] The effect of nationality of the partner (Dutch, German, Italian, or French) was analysed additionally. The results were mostly similar to the group based results, showing that the expectations, investments, and ratings of especially the French partners resembled the outgroup pattern. Expectations and ratings of German and Italian partners were also higher than expectations and ratings of the Dutch partners.

[3] Moreover, several exploratory analyses were performed on the effect of age of the partner (range 18 – 30 years) on investments in the Trust Game. Results showed higher trust in older partners. This effect did not confound the effects of group, reciprocity behaviour and sequential trial number on investments in the Trust Game.

Results

Trust Game. Based on a visual inspection of the data (see Figure 3), we decided to compare a linear model of the data to a non-linear square root model. This model consisted of the same main effects and interactions between group, reciprocity behaviour, and trial number, but included the square root effect of trial number instead of the linear effect. This square root model had a better fit to the data (*AIC* = 19573) than the linear model (*AIC* = 19648), χ^2 (0) = 74.53, p <.001. The square root effect of trial number indicated that the speed of change in investments decreased over time, with a steeper change in the first rounds of the game, and a flattening out during later rounds[4]. The fixed effects of this model alone explained 37% of the variance in the data. The addition of the random effects per subject increased the explained variance to 66%. The results of the square root model will be reported here.

Table 1

Mean (Standard Deviations) Investments in the Trust Game in Study 1, for Group, Reciprocity Behaviour and Condition (Group x Reciprocity Behaviour Interaction)

Group	**Investment**
Ingroup	5.09 tokens (1.82)
Outgroup	5.54 tokens (1.49)
difference	t(35) = -2.43, p = .020, d = 0.27

[4] Not only is this non-linear square root model a more accurate fit to the data, it is in accordance with general models of learning over time (e.g. Sloan & Ostrom, 1974).

Reciprocity behaviour	
High reciprocity	7.47 tokens (1.98)
Low reciprocity	3.16 tokens (1.89)
difference	$t(35) = 11.36$, $p < .001$, $d = 2.22$
Condition (Group x Reciprocity)	
Ingroup-High	6.96 tokens (2.27)
Outgroup-High	7.97 tokens (1.95)
difference	$t(35) = -4.11$, $p < .001$, $d = 0.48$
Ingroup-Low	3.21 tokens (2.17)
Outgroup-Low	3.11 tokens (1.85)
difference	$t(35) = 0.43$, $p = .668$, $d = .05$

In the full square root model, significant main effects of reciprocity behaviour, $F(1, 98) = 5.41$, $p = .022$, $B = 1.14$, 95% CI [0.22, 2.06], and group, $F(1, 575) = 9.69$, $p = .002$, $B = 1.13$, 95% CI [0.42, 1.81], were found. As expected, participants invested more in partners that returned high amounts than in partners that returned low amounts. Surprisingly, across all rounds of the Trust Game, the outgroup received higher investments than the ingroup (see Table 1 for descriptive statistics). Higher investments for the outgroup were already observed in the first round of the game, $t(35) = -2.52$, $p = .017$, $d = 0.37$ ($M_{ingroup} = 4.61$, $SD_{ingroup} =$

2.11; $M_{outgroup}$ = 5.39, $SD_{outgroup}$ = 2.05). The 3-way interaction of Group x Reciprocity x Trial was of most interest to our hypotheses. This 3-way interaction was marginally significant, $F(1, 4172) = 3.64$, $p = .056$, $B = -0.30$, 95% CI [-0.62, 0.01] (see Figure 3). Post-hoc analyses of the slopes indicated that only the slopes in the low reciprocity condition differed between the ingroup and outgroup, $\chi^2(1) = 9.85$, $p = .003$, where the slope of the outgroup was steeper ($B_{outgroup}$ = -0.29, $SE_{outgroup}$ = 0.02) than the slope of the ingroup ($B_{ingroup}$ = -0.20, $SE_{ingroup}$ = 0.02). The slopes of the ingroup ($B_{ingroup}$ = 0.28, $SE_{ingroup}$ = 0.02) and outgroup ($B_{outgroup}$ = 0.26, $SE_{outgroup}$ = 0.02) did not differ significantly in the high reciprocity condition, $\chi^2(1) = 0.16$, $p = .691$.

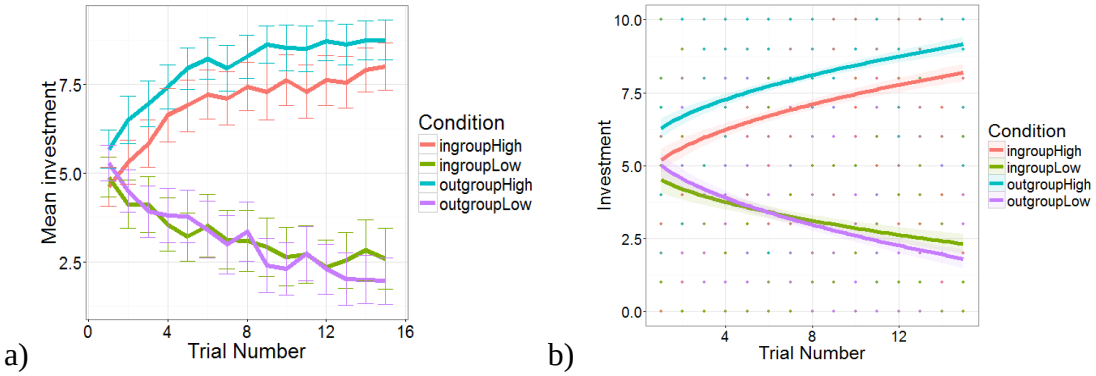


Figure 3. Investments in the Trust Game in Study 1 over sequential trials with a partner (1 to 15), for the different conditions: ingroup-high (red), outgroup-high (blue), ingroup-low (green), outgroup-low (purple). Figure 3a indicates the mean investment scores for each round and each condition. Error bars represent 95% confidence intervals. Figure 3b shows the data points with regression lines for the square root effect of time, with separate lines for each condition.

Additionally, we compared investments during the last rounds of the game, to examine possible biases still present at the end of the game. Paired t-tests demonstrated that, for high reciprocating partners, outgroup partners still received higher investments in the last round of the game (M = 8.81, SD = 2.29) than ingroup members (M = 7.72, SD = 2.79), $t(35)$ = -2.91, p = .006, d = 0.38. However, for low reciprocating partners, there was no difference

between last round investments between ingroup (M = 2.94, SD = 3.81) and outgroup partners (M = 2.06, SD = 2.87), $t(35) = 1.35$, $p = .19$, $d = 0.14$.

Ingroup identification and the Trust Game. No significant effects of ingroup identification on investment behaviour in the Trust Game were observed. The full model and all inferential statistics can be found in the online supplementary materials.

Expectations of return. A significant effect of group was found in the model predicting expectations of return, $F(1, 249) = 22.78$, $p < .001$, $B = 8.06$, 95% CI [4.90, 11.29]. Expectations of return were higher for the outgroup ($M_{outgroup}$ = 47.31, $SD_{outgroup}$ = 17.47) than for the ingroup ($M_{ingroup}$ = 39.79, $SD_{ingroup}$ = 18.34). The effect of group on the certainty of the expectations was not significant ($M_{ingroup}$ = 39.09, $SD_{ingroup}$ = 22.36; $M_{outgroup}$ = 37.48, $SD_{outgroup}$ = 21.50), $F(1, 248) = 0.68$, $p = .42$, $B = -0.96$, 95% CI [-3.30, 1.29].

Ratings of individual partners. The partner-rating index was used as the dependent variable in the linear multilevel model with group and reciprocity behaviour as within-subject factors. Results showed a significant main effect of reciprocity behaviour on partner ratings, $F(1, 246) = 147.34$, $p < .001$, $B = 1.27$, 95% CI [1.08, 1.49], and a significant Group x Reciprocity interaction, $F(1, 246) = 7.34$, $p = .007$, $B = 0.40$, 95% CI [0.09, 0.69]. See Figure 4 and supplementary materials for descriptive statistics. We did not observe a significant main effect of group, $F(1, 246) = 1.03$, $p = .310$, $B = -0.11$, 95% CI [-0.32, 0.12]. Post-hoc paired t-tests indicate that only in the high reciprocity condition, the outgroup was rated more positively than the ingroup, $t(34) = -3.34$, $p = .002$, $d = 0.57$. In the low reciprocity conditions, partner ratings did not differ significantly between the groups, $t(33) = 1.13$, $p = .269$, $d = 0.19$. This is in line with the investment behaviour in the last rounds of the game.

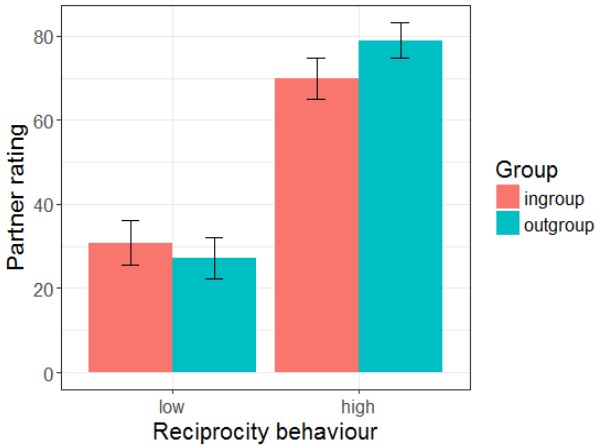


Figure 4. Mean partner ratings after the game for Study 1, separate for ingroup (red) and outgroup (blue) partners that showed high or low reciprocity behaviour. Error bars represent 95% confidence intervals.

Discussion

In Study 1, we investigated how group membership of game partners influences learning about their trustworthiness over repeated interactions. Surprisingly, people demonstrated a preference to trust outgroup members, which affected trust amounts throughout the game, and even influenced player ratings after the game. Such outgroup preferences do occasionally occur (e.g. Jussim et al., 1987). Relatedly, and likely the motivation behind the subsequent trust decisions, before the game people indicated higher expectations for the outgroup than the ingroup.

However, our main research question related to learning for trustworthy and untrustworthy group members over time. We observed that with initial trust being higher for the outgroup than the ingroup, trustworthy outgroup partners kept receiving higher investments than trustworthy ingroup partners throughout the game. However, for untrustworthy outgroup partners, investments decreased significantly faster than investments in untrustworthy ingroup partners, so that there was no difference in the ratings between groups after the game. These results are in line with complexity-extremity theory (CET;

Linville, 1982), showing that the investments for outgroup members are more extreme than for ingroup members. However, as investments in the last round were similar for untrustworthy ingroup and outgroup partners, the faster decrease in investments could also be explained by the initial bias towards the outgroup. Participants had to make a larger adjustment for outgroup untrustworthy partners to come to the same endpoint as the ingroup. Such an adjustment was not made in the trustworthy condition.

We did not find any evidence for the Black Sheep Effect (BSE; Marques et al., 1988) in this experiment. We did not observe ingroup favouritism, there were no differences between ingroup and outgroup investments in untrustworthy partners towards the end of the game, and we found no effects of ingroup identification.

Initial expectations in this study suggests that the participants viewed the outgroup as more trustworthy. This may be concordant with the general liberal attitudes demonstrated by university students in the Netherlands (van Leeuwen & Park, 2009). Perhaps it also could be related to ingroup stereotypes of the Dutch as more frugal than other cultures (Mlicki & Ellemers, 1996). Would this pattern hold in a different country with different national perceptions? Study 2 examines this possibility and attempts to replicate the extremity effects with a different sample.

Study 2

Study 2 was a replication of Study 1, conducted in the United Kingdom. We expected to replicate the general pattern in the initial study, with more extreme investment behaviour emerging towards the outgroup than towards the ingroup. As an additional investigation, if the outgroup preferences found in Study 1 were caused by factors unique to our Dutch sample, we would expect to find ingroup preferences, particularly for those high in ingroup identification, as was originally hypothesised. However, other motives might also be at play.

Social desirability and Motivation to Control Prejudice (MCP) scales were added to the study to explore the possibility that outgroup preferences may have been caused by social desirability concerns.

Method

Participants and design. Participants (N = 73 students[5]) were recruited via an online participant database. The data of 14 participants was removed from analysis due to double or foreign nationalities. The remaining 59 participants (86% female; M_{age} = 20.30 years, SD_{age} = 3.70 years) were of British nationality. Participants received course credit for their time and had a chance to win their average earnings in the game, converted to pounds (one token = 50 pence). Each participant had a 1 in 6 chance to win the bonus, based on a dice roll. The design of Study 2 was similar to Study 1, with the addition of the social desirability and MCP scales as continuous variables predicting investments alongside group and reciprocity behaviour.

Materials and procedure. Only the differences with Study 1 will be described here. First, the eight partners that the participants played the game with consisted of four British partners and four foreign partners with the following nationalities: Dutch, Italian, Belgian and Austrian. Second, the information about the age of the partner, which was shown next to the figure that represented the partner in Study 1, was only provided initially due to unexpected

[5] The sample size for Study 2 was increased from Study 1 due to the addition of the individual difference measures of social desirability and MCP. We performed another power analysis in GPower 3.1, which now included eight groups: 2 (group) x 2 (reciprocity behaviour) x 2 (social desirability/MCP). All other parameters were set the same as the power analysis for Study 1. This resulted in a sample size of 72. Due to more exclusions than anticipated, our final sample size was 59.

effects of age of the partner found in Study 1[2]. The figures representing each individual partner now only consisted of the figure with the colours of the flag and the initials.

Third, participants were presented with the social desirability and MCP scales at the end of the study. The social desirability scale consisted of sixteen items and was adapted from Stöber (2001). Higher scores indicated increased social desirability ($\alpha = 0.67$). Next, participants completed the 10-item Motivation to Control Prejudice (MCP) scale (Plant & Devine, 1998). Higher scores indicated a greater motivation to control prejudice ($\alpha = 0.72$). The full procedure was again approved by a research ethics committee before data collection commenced.

Data analysis. The same analyses were performed on the data of the Trust Game as in Study 1[6]. The standardised scores from the social desirability questionnaire and the MCP scale were added to the full model of investments in the Trust Game.

Results

Trust Game. We again compared a linear and square root model, and the square root model resulted in a better fit (*AIC* = 31605) to the data than the linear model (*AIC* = 31642), $\chi^2(0) = 36.56$, $p < .001$. The fixed effects of this model alone explained 15% of the variance in the data. The random effects per subject increased the explained variance to 59%. We will report the results of the model including a square root effect of trial.

[6] The effect of nationality of the partner (British, Italian, Dutch, Belgian, and Austrian) was again analysed. The results were mostly similar to the group based results, showing that especially the investments in the Belgian partner differed from the British partners, and expectations and ratings of the Italian partner differed from the British partners. Participants' behaviour towards each national outgroup pointed in the same direction, and was different from behaviour towards British partners.

In the full model, significant main effects of reciprocity behaviour, $F(1, 272) = 5.23$, $p = .023$, $B = 0.75$, 95% CI [0.10, 1.39], and trial number, $F(1, 136) = 55.77$, $p < .001$, $B = 0.70$, 95% CI [0.52, 0.87], were found. Participants invested more in partners that returned high amounts than in partners that returned low amounts (see Table 2 for descriptive statistics). Furthermore, there was a linear, positive trend in investments over trials. As part of the full model, the main effect of group was not significant, $F(1, 783) = 1.04$, $p = .307$, $B = 0.28$, 95% CI [-0.25, 0.77] (see Table 2), although in the first round people again invested more in the outgroup ($M = 5.17$, $SD = 2.60$) than the ingroup ($M = 4.63$, $SD = 2.41$), $t(58) = 2.04$, $p = .046$, $d = 0.22$.

Table 2

Mean (Standard Deviations) Investments in the Trust Game in Study 2, for Group, Reciprocity Behaviour and Condition (Group x Reciprocity Behaviour Interaction)

Group	**Investment**
Ingroup	5.23 tokens (2.11)
Outgroup	5.42 tokens (1.95)
difference	$t(58) = -1.23$, $p = .225$, $d = 0.09$
Reciprocity behaviour	
High reciprocity	6.53 tokens (2.31)
Low reciprocity	4.12 tokens (1.95)
difference	$t(58) = -10.55$, $p < .001$, $d = 1.13$

Condition (Group x Reciprocity)	
Ingroup-High	6.36 tokens (3.18)
Outgroup-High	6.71 tokens (2.86)
difference	$t(58) = -1.80$, $p = .077$, $d = 0.14$
Ingroup-Low	4.11 tokens (3.26)
Outgroup-Low	4.13 tokens (3.11)
difference	$t(58) = -0.11$, $p = .912$, $d = 0.01$

The 3-way Group x Reciprocity x Trial interaction between was of most interest for our hypotheses. This 3-way interaction was marginally significant, $F(1, 6840) = 3.67$, $p = .055$, $B = -0.23$, 95% CI [-0.47, -0.00]. Post-hoc analyses of the slopes indicated that, again, only the slopes in the low reciprocity condition differed between the ingroup and outgroup, $\chi^2(1) = 6.27$, $p = .025$, where the slope of the outgroup ($B_{outgroup} = -0.19$, $SE_{outgroup} = 0.021$) was steeper than the slope of the ingroup ($B_{ingroup} = -0.12$, $SE_{ingroup} = 0.014$). The slopes of the ingroup ($B_{ingroup} = 0.21$, $SE\text{-}_{ingroup} = 0.016$) and outgroup ($B_{outgroup} = 0.21$, $SE_{outgroup} = 0.015$) did not differ significantly in the high reciprocity condition, $\chi^2(1) = 0.07$, $p = .789$ (see Figure 5).

Again, we also compared investments during the last rounds of the game. The paired t-tests showed a marginally significant difference between ingroup ($M = 7.32$, $SD = 3.28$) and outgroup ($M = 7.78$, $SD = 2.88$) investments for high reciprocating partners, $t(58) = -1.68$, $p =$

.099, d = 0.22, but no significant difference for low reciprocating partners, $t(58)$ = -0.40, p = .69, d = 0.05 ($M_{outgroup}$ = 3.59, $SD_{outgroup}$ = 3.28; $M_{ingroup}$ = 3.44, $SD_{ingroup}$ = 3.45).

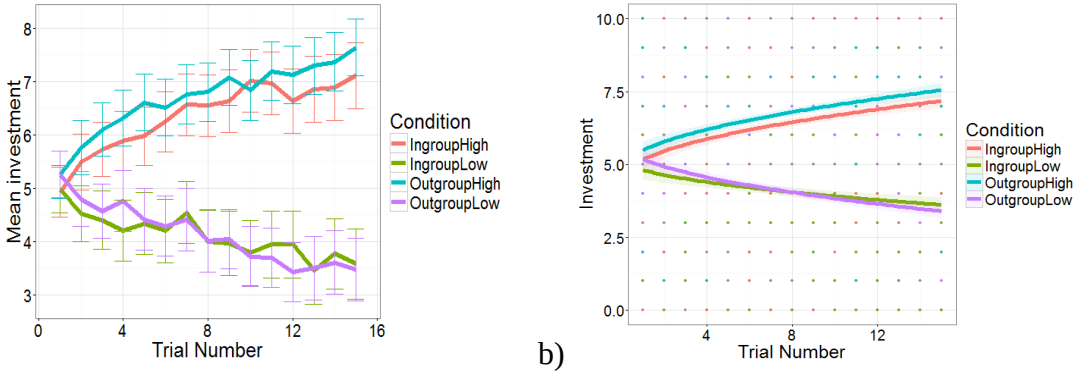


Figure 5. Investments in the Trust Game in Study 2 over sequential trials with a partner (1 to 15), for the different conditions: ingroup-high (red), outgroup-high (blue), ingroup-low (green), outgroup-low (purple). Figure 5a indicates the mean investment scores for each round and each condition. Error bars represent 95% confidence intervals. Figure 5b shows the data points with regression lines for the square root effect of time, with separate lines for each condition.

Group identification and the Trust Game. We observed a significant main effect of identification, indicating that participants with higher reported identification with the ingroup made higher investments overall, $F(1, 47)$ = 5.09, p = .029, B = 0.77, 95% CI [0.14, 1.41]. More interestingly, the Group x Identification interaction was significant, $F(1, 83)$ = 16.40, p < .001, B = -0.58, 95% CI [-0.87, -0.29]. The slope of the ingroup ($B_{ingroup}$ = 0.30, $SE_{ingroup}$ = 0.09) was steeper than the outgroup ($B_{outgroup}$ = 0.11, $SE_{outgroup}$ = 0.09). Low identifiers showed higher investments in the outgroup than the ingroup, while high identifiers showed the classic pattern of ingroup bias, with higher investments in the ingroup than the outgroup.

Expectations of return. Participants did not differ in the expectations for partners from the ingroup or outgroup, $F(1, 411)$ = 2.27, p = .133, B = 2.04, 95% CI [-0.61, 4.70]. However, participants were more certain about their expectations about the ingroup, ($M_{ingroup}$ = 47.34 % certain, $SD_{ingroup}$ = 25.11; $F(1, 412)$ = 26.44, p < .001, B = -6.15, 95% CI [-8.58, -

3.71]), than about the outgroup, ($M_{outgroup}$ = 41.18 % certain, $SD_{outgroup}$ = 23.71).

We observed a significant Identification x Group interaction, $F(1, 411) = 12.93$, $p < .001$, $B = -4.88$, 95% CI [-7.36, -2.25]. This interaction indicated that low identifiers expected more from outgroup members than ingroup members (as was the general pattern in Study 1), while high identifiers had higher expectations for ingroup members than outgroup members. The slope for the ingroup was slightly positive ($B_{ingroup}$ = .06, $SE_{ingroup}$ = 0.11), while the slope for the outgroup was negative ($B_{outgroup}$ = -0.19, $SE_{outgroup}$ = 0.10).

Ratings of individual partners. There were significant main effects of both group, $F(1, 410) = 6.79$, $p = .010$, $B = 0.27$, 95% CI [0.07, 0.48], and reciprocity behaviour, $F(1, 410) = 97.80$, $p < .001$, $B = -1.02$, 95% CI [-1.22, -0.81], on the partner ratings. Ratings were more positive for high reciprocity partners than for low reciprocity partners, and higher for outgroup members than for ingroup members (see Figure 6, and supplementary materials for descriptive statistics). Paired t-tests indicated that the difference in rating between ingroup and outgroup partners was significant for high reciprocating partners, $t(58) = -2.94$, $p = .005$, $d = 0.38$, but not for low reciprocating partners, $t(58) = -1.05$, $p = .296$, $d = 0.14$. This is in line with the last round of investments for partners from different conditions. The Group x Reciprocity interaction was not significant, $F(1, 410) = 1.60$, $p = .206$, $B = -0.18$, 95% CI [-0.48, 0.09].

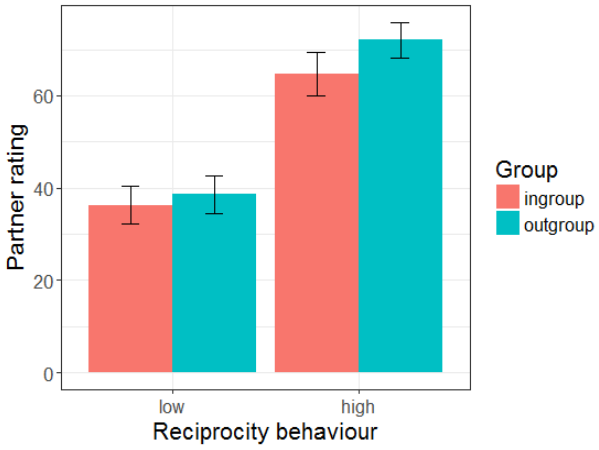


Figure 6. Mean partner ratings after the game for Study 2, separate for ingroup (red) and outgroup (blue) partners that showed high or low reciprocity behaviour. Error bars represent 95% confidence intervals.

Social desirability and the Trust Game. No significant effects of social desirability were found on investments in the Trust Game. The full description of the performed analyses and inferential statistics can be found in the supplementary materials.

Motivation to control prejudice and the Trust Game. We only observed a significant main effect of MCP on investments in the Trust Game, $F(1, 72) = 4.09$, $p = .047$, $B = -0.70$, 95% CI [-1.38, -0.06]. There was an overall tendency for people with higher MCP scores to invest less. No significant MCP x Group interaction was found, $F(1, 99) = 2.82$, $p = .096$, $B = -0.26$, 95% CI [-0.56, 0.04], nor MCP x Reciprocity, $F(1, 73) = 1.68$, $p = .20$, $B = 0.30$, 95% CI [-0.13, 0.77].

Discussion

The results of Study 2 generally replicate the results of Study 1. As expected, there was a strong effect of the behaviour of a partner (i.e. reciprocity behaviour) on the amount of trust placed in him/her, and a small overall effect of the partners' group membership, showing slight outgroup favouritism. Importantly, we again observed that investments in the outgroup remained significantly higher than investments in the ingroup for trustworthy partners, but decreased more quickly for untrustworthy partners. This led to similar levels of (low) investment in both untrustworthy ingroup and outgroup members by the end of the game. Again, this could be due to trust behaviour having an outsized effect for outgroup members, or because of initial outgroup favouritism. The latter would lead to a larger adjustment of investments towards untrustworthy outgroup partners to arrive at the same point as untrustworthy ingroup partners.

Together, the results of Study 1 and Study 2 show that people initially trust outgroup partners more than ingroup partners. This outgroup bias continues when learning about trustworthy partners, but investments in untrustworthy outgroup partners decrease more quickly so that no group bias is present for untrustworthy partners at the end of the game. This pattern of investments is also reflected in ratings of partners after the game. Unlike Study 1, here ingroup identification was related to patterns of ingroup and outgroup investments and expectations of return. Low identifiers displayed outgroup bias and high identifiers displayed ingroup bias. However, the effect of ingroup identification did not influence the relative effects of investment behaviour. There was no evidence that this effect is caused by social desirability concerns.

Framing the outgroup as people from different foreign nationalities might be argued to be problematic. These different nationalities might not have made a very cohesive out-group, but instead created multiple out-groups, which can lead to multiple categorization (for a review, see Crisp & Hewstone, 2007). Moreover, it is possible that people held stereotypical beliefs related to trustworthiness and generosity about the different nationalities. Therefore, Study 3 focussed on one particular nationality to represent the outgroup, namely Austrian. This nationality was pre-tested to be perceived as similar to the British on the aforementioned traits (see supplementary materials). The change in target outgroup to a relatively unknown nationality was predicted to reduce initial bias. This reduction would demonstrate whether initial outgroup favouritism is driving the magnitude of change in investments over time for trustworthy and untrustworthy ingroup and outgroup members. Otherwise we would expect that, due to low complexity of the outgroup, we again observe more extreme investment behaviour towards outgroup partners.

Study 3

Study 3 was conducted to examine how group membership can influence learning about trustworthiness when an initial bias is reduced, and further to examine the underlying processes that could explain the findings of the first two studies. While previously we could infer learning from behavioural responses to both trustworthy and untrustworthy partners, this could be explained by a variety of mechanisms. Therefore, we included here measures of perceptions of trustworthiness, expectations, and affective responses towards the partners in the game, as these are relevant theoretical components of trust that could help clarify the responsible processes (Lewicki, Tomlinson, & Gillespie, 2006; Mayer, Davis, & Schoorman, 1995; McAllister, 1995). These three measures target both impressions of the person (affect and trustworthiness), as well as expectations about behaviour, which could be perceived as a separate construct from person perceptions of trustworthiness (Newman & Uleman, 1993). These ratings were measured at three different time points: before the start of the Trust Game, during the game itself, and after completing the Trust Game. Using this design, we could examine how changes in investments throughout the game are related to changes in trustworthiness perceptions, expectations, and affect.

We predicted that if initial outgroup favouritism is no longer present, then two scenarios are possible. Firstly, if changes in investment behaviour can be explained by complexity-extremity theory (CET; Linville, 1982), we would expect similar findings as in Study 1 and 2. Investments in trustworthy outgroup members would increase quicker than investments in trustworthy ingroup members, and investments in untrustworthy outgroup members would decrease quicker than investments in untrustworthy ingroup members. Secondly, if adjustments after initial outgroup favouritism are driving differences in slopes, but not changes in the perceptions of trustworthiness, then we would expect that investments in ingroup and outgroup members would increase or decrease at similar rates.

With regard to the underlying processes, we predict that perceptions of trustworthiness, expectations, and affective responses towards the partners will change along with the investment behaviour in the game. However, as each of these factors may play a significant role, we do not specify a directional hypothesis regarding which would be the optimal predictor of the changes in investment amount.

Methods

Participants and design. Participants (N = 134 students[7]) were recruited via an online participant database. The data of 27 participants was removed from analysis due to foreign or double nationalities. The remaining 107 participants (86% female; M_{age} = 20.30 years, SD_{age} = 3.70 years) were of British nationality. Participants received course credit for their time, and had a chance to win their average earnings in the game, converted to pounds, in the same manner as in Study 2.

The design of Study 3 was similar to the first two studies, with additional measures added at different time points (pre-game, during-game, post-game). These measures are treated as both dependent and independent variables for different analyses (see data analysis section).

Materials and procedure. Only the differences with Study 1 are described here. First, the nationality of the outgroup was changed to only represent one country instead of four different foreign nationalities. Based on a pilot study run on a sample from the same pool of students as used in the main experiment, Austria was chosen as the country that was

[7] The sample size was based on power calculations using the coefficient and standard error of the three-way interaction of interest from Study 2. Formulas were derived from Snijders (2005), but ignored the design effect. See preregistration materials on osf.io/3vney/ for the MatLab script.

most similar in people's stereotype perceptions of trustworthiness and generosity as Britain (see supplementary materials for a description of the pilot and results).

Second, we adapted the questionnaires about expectations of the partners and ratings of trustworthiness of the partners. Instead of only asking for expectations before the game and judgments after the game, we implemented three time points for measures throughout the experiment: pre-game, during-game, and post-game. Before the game started (pre-game), participants were presented with a short 7-item in-group identification questionnaire (Cinnirella, 1997) a feeling thermometer towards Austrian people, and a 6-item semantic differential scale adapted from Terracciano et al. (2005) to measure stereotype perceptions of trustworthiness and generosity of both British and Austrian people ($\alpha = 0.73$).

Next, participants were asked to rate the individual partners in the game on trustworthiness (very untrustworthy to very trustworthy), expectations of return in the game (very unlikely to reciprocate to very likely to reciprocate), and general affect towards the partner (very cold to very warm). Participants rated each of the partners on these three traits on a 100-point slider scale. After providing these ratings, participants started with the game.

The during-game measurement took place after completing seven out of the 15 rounds with an individual partner, in which the above ratings of trustworthiness, expectations, and affect were repeated[8]. The post-game ratings of each partner occurred immediately after completing the game and receiving information about the average earnings in the game. Finally, participants were again presented with the feeling thermometer towards Austrian people and the 7-item semantic differential scale to measure changes in stereotype perceptions of British and Austrian people.

[8] As the order of the game rounds was fully randomised, the exact moment of the during-game ratings came at different points for each partner and for each participant.

Data analysis. The data analysis consisted of three consecutive models. First, the same analysis was performed as in the first two studies, where we examined the effect of group membership of the partner (ingroup vs outgroup), trustworthiness of behaviour of the partner (high vs low), and trial number (game round 1 to 15) on investments in the game using linear multilevel models. Second, the effect of changes in trustworthiness judgments, expectations of return, and affective responses (i.e. post-game minus pre-game difference scores) on changes in investments in the game (i.e. round 15 minus round one investment difference scores) were examined. Third, the effect of group membership, trustworthiness, and time (pre-game, during-game, post-game) on the ratings of the partners was examined. We created separate linear multilevel models for trustworthiness judgments, expectations of return, and affective responses as outcome variables. Differences in stereotype perceptions and outgroup feelings before and after the game were also examined, these results can be found in the supplementary materials.

Results

The Trust Game behaviour was analysed on two levels. The first level examined the effect of group, behaviour, and time, on the investments in the game, and the second level explored how changes in trustworthiness, expectations, and affect predicted changes in investments during the game.

Level one analysis. The square root model of group, reciprocity behaviour, and trial number resulted in a better fit to the data (AIC = 56057) than the linear model (AIC = 56166), $\chi^2(0) = 109$, $p < .001$. The fixed effects of this model alone explained 25% of the variance in the data. The random effects per subject increased the explained variance to 59%. We will report the results of the model including a square root effect of trial.

In the full model, significant main effects of reciprocity behaviour, $F(1, 282) = 32.05$, $p < .001$, $B = 1.56$, 95% CI [1.01, 2.09], and trial number, $F(1, 269) = 101.65$, $p < .001$, $B = 0.62$, 95% CI [0.50, 0.74], were found. Participants invested more in partners that returned high amounts than in partners that returned low amounts (see Table 3 for descriptive statistics). Furthermore, there was a linear, positive trend in investments over trials. As part of the full model, the main effect of group was not significant, $F(1, 960) = 0.17$, $p = .679$, $B = -0.01$, 95% CI [-0.50, 0.31]. No initial (first round) bias was observed towards either the ingroup ($M = 4.90$ tokens, $SD = 2.02$) or the outgroup ($M = 4.99$, $SD = 2.20$), $t(106) = -0.60$, $p = .553$, $d = 0.04$.

Table 3

Mean (Standard Deviations) Investments in the Trust Game in Study 3, for Group, Reciprocity Behaviour and Condition (Group x Reciprocity Behaviour Interaction)

Group	**Investment**
Ingroup	4.32 tokens (1.61)
Outgroup	4.70 tokens (1.76)
difference	$t(106) = -3.20$, $p = .002$, $d = 0.25$
Reciprocity behaviour	
High reciprocity	5.95 tokens (2.21)
Low reciprocity	3.06 tokens (1.81)
difference	$t(106) = -13.24$, $p < .001$, $d = 1.63$

Condition (Group x Reciprocity)

Ingroup-High	5.72 tokens (2.52)
Outgroup-High	6.19 tokens (2.52)
difference	$t(106) = -2.53$, $p = .013$, $d = 0.21$
Ingroup-Low	2.92 tokens (1.82)
Outgroup-Low	3.21 tokens (1.91)
difference	$t(106) = -3.77$, $p < .001$, $d = 0.18$

The 3-way interaction Group x Reciprocity x Trial number was not significant, $F(1, 12408) = 0.46$, $p = .498$, $B = -0.06$, 95% CI [-0.23, 0.11]. However, the Group x Trial interaction, $F(1, 12408) = 11.21$, $p < .001$, $B = 0.20$, 95% CI [0.08, 0.32], and Reciprocity x Trial interaction, $F(1, 12408) = 708.43$, $p < .001$, $B = -1.62$, 95% CI [-1.73, -1.50], were found to be significant. Investments in high reciprocating partners increased over time ($B = 0.22$, $SE = 0.01$, $p < .001$), while investments in low reciprocating partners decreased over time ($B = -0.29$, $SE = 0.01$, $p < .001$). The Group x Trial interaction showed that, over trials, investments in the ingroup decreased, ($B = -0.06$, $SE = 0.01$, $p < .001$), while investments in the outgroup did not change, ($B = 0.00$, $SE = 0.01$, $p = .759$).

A comparison of the slopes of all the conditions showed that, for highly reciprocating partners, the outgroup slope ($B = 0.26$, $SE = 0.01$) was steeper than the ingroup slope ($B = 0.19$, $SE = 0.01$), $\chi^2(1) = 11.54$, $p = .001$. However, for low reciprocating partners, the outgroup slope was less steep ($B = -0.26$, $SE = 0.01$) than the ingroup slope ($B = -0.31$, $SE =$

0.01), $\chi^2(1) = 5.88$, $p = .015$, see Figure 7. The 3-way interaction is not significant because for both high and low reciprocating partners, outgroup members receive higher investments over time than do ingroup members.

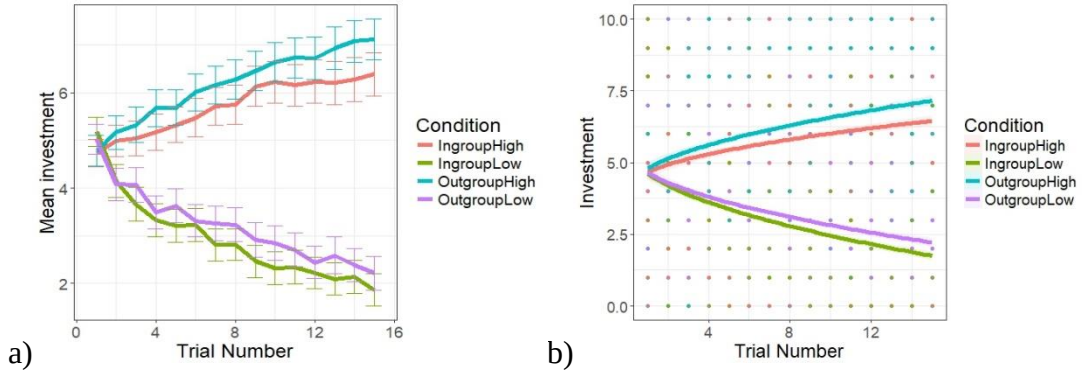


Figure 7. Investments in the Trust Game for Study 3 over sequential trials with a partner (1 to 15), for the different conditions: ingroup-high (red), outgroup-high (blue), ingroup-low (green), outgroup-low (purple). Figure 7a indicates the mean investment scores for each round and each condition. Error bars represent 95% confidence intervals. Figure 7b shows the data points with regression lines for the square root effect of time, with separate lines for each condition.

We also compared last round investments for high and low reciprocating ingroup and outgroup partners. As in Study 2, the difference in last round investment between highly reciprocating ingroup ($M = 6.21$, $SD = 3.36$) and outgroup partners ($M = 6.82$, $SD = 3.29$) was marginally significant, $t(106) = -1.77$, $p = .080$, $d = 0.17$. There was no significant difference between investments in ingroup ($M = 1.92$, $SD = 2.50$) and outgroup partners ($M = 2.10$, $SD = 2.43$) that did not reciprocate trust, $t(106) = -0.88$, $p = .380$, $d = 0.09$.

Level two analysis. For the second level of analysis, difference scores were computed. The outcome variable was the difference score between the investment in the last round with a partner (round 15) subtracting the investment in the first round with a partner (round 1). The main predictors were the difference scores between the partner ratings after

the game (trust – postgame, expectation-postgame, affect-postgame) and the partner ratings before the game (trust –pregame, expectation-pregame, affect-pregame).

A linear multilevel model was created with the difference scores for trustworthiness, expectations, and affect, predicting the investment difference score. A per subject random slope was added to account for the repeated measures design. The fixed effects of this model alone explained 58% of the variance in the data. The random effects per subject increased the explained variance to 68%. This model showed significant effects of expectations, $F(1, 827) = 58.46$, $p < .001$, $B = 0.47$, 95% CI [0.35, 0.59], affect, $F(1, 850) = 13.07$, $p < .001$, $B = 0.21$, 95% CI [0.10, 0.32], and trustworthiness, $F(1, 845) = 4.06$, $p = .044$, $B = 0.13$, 95% CI [0.00, 0.25] on the investment difference score. The coefficients of this regression model show that all difference scores of partner ratings have a positive relation with the difference scores in investments. However, the effect of expectations and affect are stronger than the effect of trustworthiness judgments.

Additionally, correlations were examined between the difference scores of trustworthiness, expectations, and affective responses. It was found that all three of the difference scores of partner ratings were highly correlated (see Table 4).

Table 4

Means, Standard Deviations, and Correlations with Confidence Intervals of the Partner Rating Difference Scores

Variable	*M*	*SD*	1	2
1. Trustworthiness post-pre	-8.09	33.23		
2. Expectation post-pre	-4.30	35.48	.91**	
3. Affect post-pre	-5.66	31.47	.91**	.88**

Note. * indicates $p < .05$; ** indicates $p < .01$. *M* and *SD* are used to represent mean and standard deviation, respectively.

Lastly, we also examined to what extend the ratings of trustworthiness, expectations, and affect were predictive of first round investments in the game. The pre-game ratings were regressed onto the first round investment with each partner, using a multilevel model including a per-subject random effect. This model showed that only expectations of return significantly predicted the first-round investment, $F(1, 847) = 12.15$, $p < .001$, $B = 0.12$, 95% CI [0.05, 0.19]. Ratings of trustworthiness, $F(1, 836) = 0.00$, $p = .984$, $B = 0.00$, 95% CI [-0.06, 0.06], and ratings of affect towards the partners, $F(1, 824) = 0.24$, $p = .625$, $B = -0.02$, 95% CI [0.09, 0.05], both did not significantly predict first round investments.

Partner ratings. To examine changes in ratings of trustworthiness, expectations, and affect over time, the effects of group, reciprocity behaviour, and time (pre-game, during-game, post-game) were examined on each of the partner ratings in separate linear multilevel models[9]. All models contained per subject random intercepts and random slopes for group, reciprocity behaviour, time, and the interaction between group and reciprocity behaviour. As the results of the three models are very similar, they are described together in this section (see Figure 8 and supplementary materials).

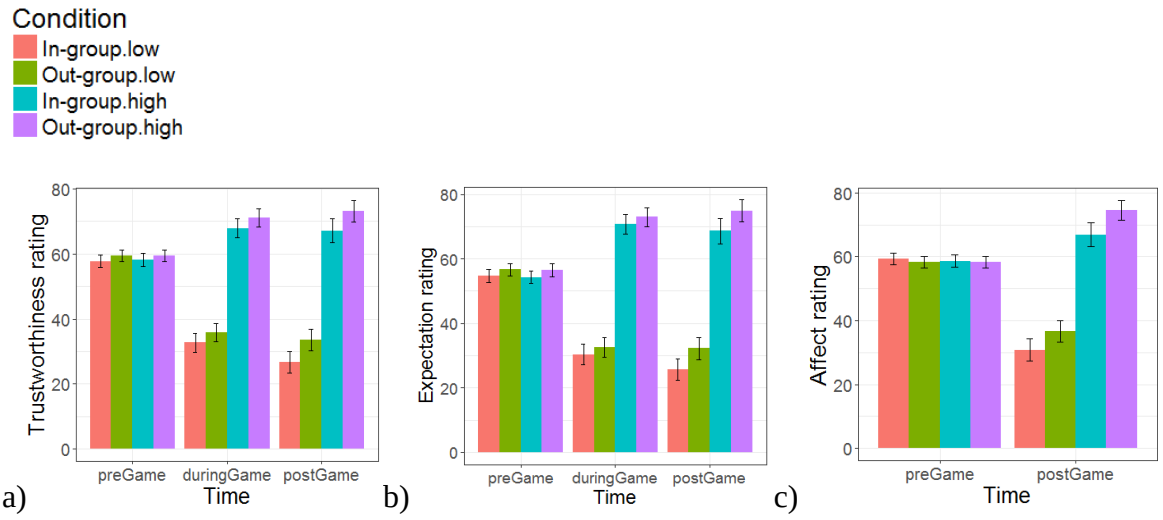
For each of the outcome ratings (trustworthiness, expectations, and affect), the Reciprocity x Time interaction was significant. Trustworthiness, $F(2, 1920) = 175.55$, $p < .001$, $B_{pre\text{-}during} = 1.34$, $B_{pre\ -\ post} = 1.54$, 95% CIs [1.16, 1.51], [1.37, 1.71][10]; expectation,

[9] Upon inspection of the data, it was discovered that a coding error had prevented the data of the affect ratings during the game to be recorded correctly. Therefore, any comparisons with the affect during-game ratings were removed from analysis.

[10] As the variable time has three levels (pregame, during game, postgame), the model creates two dummy variables. The beta coefficients and 95% confidence intervals of both these dummy variables (pre-during, pre-post) are reported. However, as the during-game ratings of affect were not accidently not recorded correctly and cannot be used, only one coefficient is reported for affect (pre-post).

$F(2,1920) = 207.60$, $p < .001$, $B_{pre\text{-}during} = 1.47$, $B_{pre\text{-}post} = 1.57$, 95% CIs [1.30, 1.64], [1.41, 1.74]; affect, $F(1,1279) = 257.08$, $p < .001$, $B_{pre\text{-}post} = 1.51$, 95% CI [1.32, 1.70]. Post-hoc multilevel models of pregame, during, and postgame subsets revealed that partner ratings of high and low reciprocating partners did not differ before the game, but did differ significantly during and after the game (see Figure 8 and supplementary materials for the outcomes of post-hoc tests). Highly reciprocating partners were always rated more positively than low reciprocating partners were.

The Group x Time interaction was not significant for trustworthiness ratings, $F(2, 1920) = 2.06$, $p = .128$, $B_{pre\text{-}during} = 0.06$, $B_{pre-post} = 0.20$, 95% CIs [-0.12, 0.22], [0.02, 0.38], and expectations ratings, $F(2, 1920) = 2.00$. $p = .135$, $B_{pre\text{-}during} = 0.01$, $B_{pre\text{-}post} = 0.17$, 95% CIs [-0.15, 0.18], [0.01, 0.34]. However, the interaction was significant for affect ratings, $F(1, 1279) = 8.86$, $p = .003$, $B_{pre\text{-}post} = 0.28$, 95% CI [0.10, 0.46]. Ratings of ingroup and outgroup partners were not significantly different before and during the game, but the outgroup partners were rated more positively than ingroup partners after the game (see Figure 8 and supplementary materials). The Group x Reciprocity interaction behaviour was not significant, the outgroup was rated more positively than the ingroup after the game for both high and low reciprocating partners.



a) b) c)

Figure 8. Mean partner ratings before, during, and after the game. Separate bars are presented for the four conditions: ingroup-low (red), outgroup-low (green), ingroup-high (blue), outgroup-high (purple). Figure 8a shows trustworthiness ratings; figure 8b shows expectations; figure 8c shows affect ratings. Error bars represent 95% confidence intervals.

Moreover, correlations were examined between the ratings of trustworthiness, expectations, and affective responses that were provided before, during, and after the Trust Game. While trustworthiness ratings, expectations, and affective responses towards the partners before the game were correlated between $r(854) = .60$, $p < .001$, and $r(854) = .66$, $p < .001$, correlations increased to $r(854) = .91$, $p < .001$ during the game. The correlation between expectations and trustworthiness ratings of partners after the game remained very high at $r(854) = .95$, $p < .001$. Meanwhile, the correlations between expectations and affective responses, $r(854) = .64$, $p < .001$, and between affective responses and trustworthiness ratings, $r(854) = .64$, $p < .001$, reduced again to be similar to the measures before the game (see supplementary materials for full correlation tables).

Ingroup identification. No significant effects of ingroup identification were observed in any of the models described above. Inferential statistics of the analyses can be found in the supplementary materials.

Discussion

In this third study, we aimed to replicate the findings from the first two studies while considering several additional factors. First, to rule out the influence of stereotype perceptions of particular foreign nationalities, the outgroup here consisted of Austrian people, a national group that was found to be stereotypically perceived as very similar to British people in a pilot study. Second, a number of partner ratings were included before, during, and

after the Trust Game to measure how trustworthiness perceptions, expectations of return, and affective responses towards the different partners changed throughout the game.

In contrast to the first two studies, here we did not observe any initial bias towards the outgroup. There was no difference between first round investments in ingroup or outgroup partners, and the ratings before the game were similar for ingroup and outgroup partners. We predicted this from the change of outgroup to Austrian people, as Austrians are a relatively unknown group that was stereotyped as having similar trustworthiness levels to the British, as observed in the pilot study (see supplementary materials). However, it should be noted that we cannot conclusively state that the lack of initial outgroup bias is caused by the similarity in stereotype perceptions of trustworthiness between Austrians and British people. It is also possible that people did not have any stereotypes about Austrians beforehand and therefore projected their views of British people onto this relatively unknown European group. Perhaps when groups are seen as stereotypically similar, learning is less divergent for ingroups and outgroups compared to situations in which the outgroup is unknown and therefore people have no stable mental representations of them.

Regardless, an outgroup preference did develop throughout the game. This was reflected in the investment slopes over time, where the slope for high reciprocating outgroup partners was steeper than the slope for high reciprocating ingroup partners, but less steep for low reciprocating outgroup partners than low reciprocating ingroup partners. By the last round of the game, this outgroup bias was still present for high trustworthy partners, but investments in low trustworthy partners again did not differ between ingroup and outgroup.

The developing outgroup bias over time was also visible in the partner ratings. Before the game, people showed similar expectations, trustworthiness ratings, and affective responses towards both ingroup and outgroup members. However, after the game, outgroup

partners were rated more positively than ingroup partners on all three factors. This outgroup bias was visible in both high and low reciprocating partners.

We examined to what extent expectations, trustworthiness ratings, and affective responses could explain the changes in investments over time. All three factors were found to relate strongly to investments in the game, where a stronger positive change in the ratings predicted a stronger positive change in investments throughout the game. Moreover, the ratings of trustworthiness, expectations, and affect were all highly correlated with each other. When all three factors were entered into the model simultaneously, the effect of trustworthiness was strongly reduced. This could mean that expectations of return and affective responses taken together explain the trustworthiness ratings. The high correlations between the rating scores indicate that perhaps the impression of trustworthiness of the person and the expectations of trustworthy behaviour are not separate constructs, but are indicators of the same construct. We will elaborate more on this in the general discussion.

In summary, Study 3 showed that even when no initial bias is present, people are still affected by group membership in how they learn about the trustworthiness of others. For highly trustworthy partners, outgroup members were trusted more than similarly behaving ingroup members over time. This pattern of behaviour was also observed in Study 1 and Study 2, and supports complexity-extremity theory (CET). However, here investments decreased at a slower rate for untrustworthy outgroup partners than for ingroup partners, in contrast to Study 1 and Study 2. At the end of the game, people showed equally low amounts of trust towards untrustworthy outgroup and ingroup partners, which was also observed in Studies 1 and 2. In this study, outgroup partners were rated somewhat higher on perceived trustworthiness, expectations of return, and affective responses after the game. This could be due to a general motivation to treat the outgroup positively (Henry, 2008).

The changes in investments over time for untrustworthy outgroup and ingroup

partners indicate that the findings are more nuanced than can be predicted solely by CET. Investments in untrustworthy partners in Study 3 were similar for ingroup and outgroup partners without an initial outgroup favouritism. This suggests that group membership becomes less influential over time when experiencing negative interactions. Both untrustworthy ingroup and outgroup partners are distrusted equally, while outgroup partners receive higher degrees of trust when interactions are positive.

General discussion

Three studies examined the effect of group biases, based on nationality, on learning about individuals' trustworthiness in a repeated Trust Game. In the first two studies, we observed that participants initially invested more in outgroup members than ingroup members. Over time, they continued to do so when the partners were reciprocating trustworthy behaviour, and they generally rated outgroup partners more positively. However, when the partners did not reciprocate trust, participants decreased their investments in outgroup members to a greater degree than for ingroup members. In the third study, in which initial outgroup bias was not present, participants showed outgroup favourability over time. This was evidenced by investments and partner ratings, particularly for high trustworthy partners. Investments in untrustworthy outgroup partners decreased less strongly, but by the end of the game ingroup and outgroup members did not differ.

Differences Between High and Low Trustworthy Partners

We saw clear effects for interactions with high trustworthy partners. Studies 1 and 2 began with positive outgroup bias, and this bias remained as overall trusting behaviour increased throughout the study. In Study 3, there was no initial bias, but overall trusting behaviour increased quicker for partners belonging to the outgroup, compared to those in the ingroup. The same pattern was also observed in partner ratings; ratings of trustworthy

outgroup members were higher than ratings of trustworthy ingroup partners after the game. The positive bias towards trustworthy outgroup members offers support for Linville's complexity-extremity theory (Linville, 1982), as applied to actual decision-making behaviour. Whilst in Studies 1 and 2, participants seemed to carry on an initial outgroup bias, the lack of initial outgroup favouritism in Study 3 particularly supports this complexity-extremity perspective as applied to trustworthy behaviour.

The effects for low trustworthy partners were slightly more nuanced. In Studies 1 and 2 the slope in investments for untrustworthy partners was steeper for the outgroup than the ingroup. In Study 3, there was a slightly steeper decrease in investments for untrustworthy ingroup partners compared to outgroup partners. However, in all three studies there was no difference between the amount invested in the ingroup and the outgroup in the last round. This suggests that, although group membership had an initial influence on investments, its importance decreased once experience became available. Consistent with the well-known negativity bias in memory and person perception (Rozin & Royzman, 2001; Skowronski & Carlston, 1989), perhaps participants individuated the untrustworthy partners more than they did with highly trustworthy partners.

Interestingly, Study 3 showed that partner ratings were more positive for the outgroup for both high and low trustworthy partners after the game, which suggests a general bias towards the outgroup. In a situation in which the outgroup and the ingroup behaved similarly overall, participants rated the outgroup more favourably. Study 3 further tested the effect of expectations, affect, and perceived trustworthiness on investment behaviour. We found that all three had a significant effect. Looking at the regression coefficients, the importance of trustworthiness is reduced when included with expectation and affect. This in not necessarily surprising, as positive feelings and expectation of behaviour are theoretical components of perceived trustworthiness (Lewicki et al., 2006; Mayer et al., 1995; McAllister, 1995).

Trustworthiness ratings, expectations of return, and affective responses were highly correlated with each other and were all highly predictive of changes in investment behaviour. From these findings, we can infer that the impression of trustworthiness of the person and the characterisation of the trustworthiness behaviour are integrated and perceived as the same construct predicting trust behaviour. However, the ratings of partners before experience became available in the Trust Game were less highly correlated then ratings during and after the game, particularly the correlation between the trustworthiness rating and the expectations of return. Additionally, we found that at the start of the game only expectations of return predicted investment behaviour, while all three ratings were significant predictors of investments over the whole game. This might indicate that people's impressions of the person and their expectations of the behaviour of the person were still perceived by people as separate constructs when no information about behaviour was available. However, while interacting with the different partners and information about behaviour became available, the impression of the person and expectation of behaviour were integrated and perceived as the same construct.

Relation to Complexity-Extremity Theory

These results are more nuanced than would be predicted by CET. All three studies show support for CET for highly trustworthy partners. Regardless of initial outgroup favouritism, people over time showed higher trust in outgroup partners than ingroup partners. However, the findings for untrustworthy partners show that initial levels of bias do influence the changes in investments as well. When initial outgroup favouritism was present, people decreased their investments in outgroup partners more quickly than investments in equally untrustworthy ingroup partners, and ended with similar low levels of trust towards both groups. When there was no initial bias, the decrease in investments was less steep for outgroup than ingroup partners, but again people ended up similarly distrusting both groups.

This suggests that, in addition to complexity effects leading to extreme investment behaviour, we also observe that negative experiences lead to reduced importance of group membership in making decisions to trust.

These findings do not support the Black Sheep Effect (BSE), as all three studies do not show ingroup favouritism, and no differences were observed between investments in untrustworthy ingroup and outgroup partners towards the end of the game. Moreover, the BSE indicates a moderating effect of identification with the ingroup (Marques et al., 1988), and effects of identification were limited in our studies.

In summary, this research shows that group-based biases influence how people update impressions and learn about individual trustworthiness. Interestingly, the influence of the group bias differs depending on the valence of the experiences. When interacting with trustworthy group members, people showed a tendency to trust outgroup members more over time. Depending on initial levels of bias, they also increased their trust more quickly. However, when group members did not reciprocate trust, initial biases disappeared over time and ingroup and outgroup members were treated similarly by the end of the game. Thus, we found that, within our student sample and using Western European national outgroups, people display a tendency to trust the outgroup more, but this tendency disappears when interacting with untrustworthy people.

On the Initial Outgroup Bias

The outgroup bias that was demonstrated and replicated in Studies 1 and 2 was surprising, as the literature mostly suggests that people favour ingroup members over outgroup members in cooperative game settings (Balliet et al., 2014). We believe that the general outgroup bias observed here is mostly due to the samples of university students, as they are often more politically liberal (Bailey & Williams, 2016), including being more egalitarian in their views about groups and having greater sensitivity about what responses

are appropriate (Henry, 2008; Peterson, 2001). An online study conducted in 2016, with undergraduate students at the university where Studies 2 and 3 were conducted, demonstrated that students express very positive attitudes towards many outgroups, including Western European immigrants (see supplementary materials for description and results of the study). Thus, we suggest the possibility that this general outgroup bias would be most likely to occur in situations in which people have egalitarian attitudes or are motivated to view the outgroup positively (e.g. Moskowitz, Salomon, & Taylor, 2000; Shepherd, Spears, & Manstead, 2013). These situations are common, such as when a liberal-minded person meets a member of a novel ethnic group, when one is a tourist in a foreign country, or when a psychologist meets an economist at an interdisciplinary conference.

Limitations and Future Directions

Nationality was chosen as group manipulation because it was our aim to use naturally occurring groups to increase the ecological validity of this research. We wanted to specifically explore learning when people already have some pre-existing knowledge or expectation about the outgroup, because that is often the case in the real world. Future research could examine several characteristics regarding the relation between ingroup and outgroup. First, it would be interesting to use more antagonistic groups to examine situations in which people have a clear motivation to favour the ingroup over the outgroup. Second, manipulating or measuring complexity representation of outgroups as well as group entitativity perceptions (i.e. the extent to which a group is considered as one entity instead of a collection of individual entities), would further the research. This variation in groups should show us whether these results would also hold towards antagonistic outgroups, due to the complexity of the group representation. Of course, situations in which there is not explicit ingroup bias, or even outgroup favourability, can and do occur (Cuddy et al., 2008; Jussim et al., 1987; Trifiletti & Capozza, 2011), and that is indeed what we found here.

With regard to entitativity perceptions, we would argue that the change to a single nationality in Study 3 increases perceptions of entitativity, although this was not explicitly measured. We would predict that higher entitative groups lead to more generalisation between individual group members, and therefore to less steep learning curves about individual trustworthiness (Crawford, Sherman, & Hamilton, 2002). Finally, it would be very interesting to examine whether our findings might generalise to other types of groups as well, such as sport affiliation, university membership, or political orientation. The effect of the type of group used might relate to the factors described above, with different effects for more or less antagonistic, complex, and entitative groups.

Conclusion

This research studied the influence of group biases on learning about trustworthiness. It was found, and replicated in two different European countries, that people demonstrate a positive bias towards outgroup members who reciprocated generous behaviour in an economic exchange. When integrating group membership information with individual experiences, the effect of group remains important throughout repeated interactions, but only with positive experiences. When individuals behave in an untrustworthy manner, group membership ceased to have an important effect on decisions to trust. This is the first research to examine how initial group-level biases interact with repeated encounters with different individuals in a laboratory environment. Our work has implications for real-world behaviour, such as when an outgroup member becomes a co-worker, neighbour, or part of a friendship group. It suggests that any favouritism earned by an outgroup member will be robust through future positive interactions. However, this favouritism can quickly turn when he or she behaves in a negative manner, which elicits behaviour similar to that induced by a negative ingroup member.

Author Contributions

The original study concept and design was developed by M. Vermue and A. Sanfey. Data collection and analysis of Study 1 was conducted by M. Vermue under supervision of A. Sanfey. Data collection and analysis of Study 2 and Study 3 was conducted by M. Vermue under supervision of C. Seger. M. Vermue and C. Seger drafted the manuscript, A. Sanfey provided critical revisions. All authors approved the final version of the manuscript for submission.

Research Disclosure Statement

The authors declare that a) they have reported the total number of excluded observations and the reasons for making these exclusions in the method sections of Studies 1 2, and 3, b) all independent variables or manipulations, whether successful or failed, have been reported in the method sections, and c) all dependent variables or measures that were analysed for this article's target research questions have been reported in the method sections. The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

Open Practice Statement

All experiment materials, analysis scripts, and datasets are publicly available through https://osf.io/78ukq/. Study 3 was preregistered through the OSF, the preregistration can be found on https://osf.io/3vney/. The authors declare that the preregistration was made before data collection commenced.

References

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, *9*, 193–208. https://doi.org/10.1007/s10683-006-9122-4

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617. https://doi.org/10.1348/000712608X377117

Bailey, M., & Williams, L. R. (2016). Are college students really liberal? An exploration of student political ideology and attitudes toward policies impacting minorities. *The Social Science Journal*, *53*, 309–317. https://doi.org/10.1016/j.soscij.2016.04.002

Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*, 1090–1112. https://doi.org/10.1037/a0030939

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological Bulletin*, *140*, 1556–1581. https://doi.org/10.1037/a0037737

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. C. Y. Trope (Ed.), *Dual-process theories in social psychology* (pp. 361–382). New York, NY, US: Guilford Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142. https://doi.org/10.1006/game.1995.1027

Buchan, N. R., & Croson, R. T. A. (2004). The boundaries of trust: Own and others' actions in the US and China. *Journal of Economic Behavior & Organization*, *55*, 485–504. https://doi.org/10.1016/j.jebo.2003.11.005

Buchan, N. R., Johnson, E. J., & Croson, R. T. A. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other

regarding preferences. *Journal of Economic Behavior & Organization*, *60*, 373–398. https://doi.org/10.1016/j.jebo.2004.03.017

Burns, J. (2006). Racial stereotypes, stigma and trust in post-apartheid South Africa. *Economic Modelling*, *23*, 805–821. https://doi.org/10.1016/j.econmod.2005.10.008

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, *61*, 87–105. https://doi.org/10.1016/j.cogpsych.2010.03.001

Cinnirella, M. (1997). Towards a European identity? Interactions between the national and European social identities manifested by university students in Britain and Italy. *British Journal of Social Psychology*, *36*, 19–31. https://doi.org/10.1111/j.2044-8309.1997.tb01116.x

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, *83*, 1076–1094. https://doi.org/10.1037/0022-3514.83.5.1076

Crisp, R. J., & Hewstone, M. (2007). Multiple social categorization. *Advances in Experimental Social Psychology*, *39*, 163–254. https://doi.org/10.1016/S0065-2601(06)39004-1

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the Trust Game. *Nature Neuroscience*, *8*, 1611–1618. https://doi.org/10.1038/nn1575

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on

trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*, 148. https://doi.org/10.3389/fnins.2012.00148

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *The Journal of Neuroscience*, *33*, 3602–3611. https://doi.org/10.1523/jneurosci.3086-12.2013

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*, 49–71. https://doi.org/10.1080/10478400802049936

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups. *Journal of Experimental Social Psychology*, *38*, 586–602. https://doi.org/10.1016/S0022-1031(02)00505-X

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology*, *52*, 536–546.

Kenworthy, J. B., & Jones, J. (2009). The roles of group importance and anxiety in predicting depersonalized ingroup trust. *Group Processes & Intergroup Relations*, *12*, 227–239. https://doi.org/10.1177/1368430208101058

Leach, C. W., van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., … Spears, R. (2008). Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of Personality and Social Psychology*, *95*, 144–165. https://doi.org/10.1037/0022-3514.95.1.144

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, *32*, 991–1022. https://doi.org/10.1177/0149206306294405

Linville, P. W. (1982). The complexity–extremity effect and age-based stereotyping. *Journal of Personality and Social Psychology*, *42*, 193–211. https://doi.org/10.1037/0022-3514.42.2.193

Marques, J. M., Yzerbyt, V. Y., & Leyens, J.-P. (1988). The "Black Sheep Effect": Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology*, *18*, 1–16. https://doi.org/10.1002/ejsp.2420180102

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*, 709–734. https://doi.org/10.5465/AMR.1995.9508080335

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*, 24–59. https://doi.org/10.2307/256727

Mlicki, P. P., & Ellemers, N. (1996). Being different or being better? National stereotypes and identifications of Polish and Dutch students. *European Journal of Social Psychology*, *26*, 97–114. https://doi.org/10.1002/(SICI)1099-0992(199601)26:1<97::AID-EJSP739>3.0.CO;2-F

Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Preconsciously controlling stereotyping: implicitly activated egalitarian goals prevent the activation of stereotypes. *Social Cognition*, *18*, 151–177. https://doi.org/10.1521/soco.2000.18.2.151

Newman, L. S., & Uleman, J. S. (1993). When are you what you did? Behavior identification and dispositional inference in person memory, attribution, and social judgment. *Personality and Social Psychology Bulletin*, *19*, 513–525.

https://doi.org/10.1177/0146167293195004

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461. https://doi.org/10.1086/323732

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Rotella, K. N., Richeson, J. A., Chiao, J. Y., & Bean, M. G. (2013). Blinding trust: The effect of perceived group victimhood on intergroup trust. *Personality and Social Psychology Bulletin*, *39*, 115–127. https://doi.org/10.1177/0146167212466114

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Shepherd, L., Spears, R., & Manstead, A. S. R. (2013). When does anticipating group-based shame lead to lower ingroup favoritism? The role of status and status stability. *Journal of Experimental Social Psychology*, *49*, 334–343. https://doi.org/10.1016/J.JESP.2012.10.012

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*, 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Sloan, L. R., & Ostrom, T. M. (1974). Amount of information and interpersonal judgment. *Journal of Personality and Social Psychology*, *29*, 23–29. https://doi.org/10.1037/h0035728

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science, 3,* 1570–1573. https://doi.org/10.1002/0470013192.bsa492

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment, 17,* 222–232. https://doi.org/10.1027//1015-5759.17.3.222

Stoddard, O., & Leibbrandt, A. (2014). An experimental study on the relevance and scope of nationality as a coordination device. *Economic Inquiry, 52,* 1392–1407. https://doi.org/10.1111/ecin.12097

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–37). Monterey, CA: Brooks-Cole.

Terracciano, A., Abdel-Khalek, A. M., Adám, N., Adamovová, L., Ahn, C., Ahn, H., … McCrae, R. R. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310,* 96–100. https://doi.org/10.1126/science.1117199

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27,* 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Trifiletti, E., & Capozza, D. (2011). Examining group-based trust with the investment game. *Social Behavior and Personality: An International Journal, 39,* 405–409. https://doi.org/10.2224/sbp.2011.39.3.405

van Leeuwen, F., & Park, J. H. (2009). Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences, 47,* 169–173. https://doi.org/10.1016/j.paid.2009.02.017

Voci, A. (2006). The link between identification and in-group favouritism: Effects of threat to

social identity and trust-related emotions. *British Journal of Social Psychology*, *45*, 265–284. https://doi.org/10.1348/014466605X52245

Williams, M. (2001). In whom we trust: Group membership as an affective context for trust development. *Academy of Management Review*, *26*, 377–396.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x