Impression formation through social interaction: Effects of ethnicity in the Dutch context

Iris J. Traast, Bertjan Doosje, and David M. Amodio
University of Amsterdam

Keywords: Ethnicity, Prejudice, Learning, Interaction, Impression

Iris J. Traast (ORCID: 0000-0002-2132-4632), Bertjan Doosje (ORCID: 0000-0002-2479-5405), and David M. Amodio (ORCID: 0000-0001-7746-0150).

Preregistration of study design, hypotheses and analyses can be found at https://aspredicted.org/RSZ_XPT (Study 1 & Study 2) and https://aspredicted.org/WK2_HT6 (Study 3). Datasets and Supplemental Information are available in the OSF repository, https://osf.io/2tn54.

Correspondence regarding this article should be addressed to Iris J. Traast or David M. Amodio, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129, REC G, 1001 NK Amsterdam, NL. Email: i.j.traast@uva.nl or d.m.amodio@uva.nl.

Abstract

Research conducted in the United States shows that White Americans form more positive impressions of White than Black interaction partners through instrumental learning (Traast et al., 2024). We asked whether this pattern generalizes to the cultural context of the Netherlands, which differs in norms for expressing intergroup bias. In three pre-registered studies (*N*s= 66/83/80), White Dutch participants played a money-sharing game, based on a reward reinforcement task, with White and Moroccan partners. Although players shared at different rates, average sharing rates for White and Moroccan players were equated. Unexpectedly, and despite anti-Moroccan explicit and implicit attitudes, participants displayed a pro-Moroccan choice preference across studies. Nevertheless, computational modeling indicated the same learning effects of ethnicity as in past research: ethnicity biased initial reward expectations, and these were updated via group-specific learning rates. We discuss potential explanations for this unexpected pattern and broader implications for cross-cultural research on intergroup social cognition.

149 words

The impressions we form of others are often influenced by their race or ethnicity. Decades of research conducted in the United States finds that White Americans tend to form more positive impression of White than Black individuals, even when attributes other than race are held constant (Dovidio et al., 2010; Richeson & Sommers, 2016). This pattern of bias was found in a recent study of interaction-based impression formation (Traast et al. (2024), in which White Americans formed more positive impressions of White interaction partners than Black interaction partners despite identical feedback. In the present research, we asked whether the findings of Traast et al. (2024) would generalize to a cultural context with a different history and dynamic of intergroup relations: the Netherlands. By examining the effects of White and Moroccan ethnicity in the Dutch context, we sought to determine whether the effect of race/ethnicity on interaction-based impression formation replicated beyond the U.S. context or whether aspects of this process are culture-specific.

Race effects on social instrumental learning & impression formation

Racial biases, including prejudiced attitudes and group stereotypes, often affect the way people learn about, interact with, and form impressions of others (Allport, 1954; Fiske, 1988; Kawakami et al., 2017; Shelton & Richeson, 2006). Among White Americans, this bias may be expressed through the avoidance of interactions with Black Americans (Amodio & Devine, 2006; Dovidio et al., 1997), unfriendly nonverbal behaviors toward Black interaction partners (Dovidio et al., 2002; Fazio et al., 1995; McConnell & Leibold, 2001), and negative judgements of performance based on race (Biernat et al., 2010; Gaertner & Dovidio, 2000).

Traast et al. (2024) recently demonstrated that this effect of race extended to the formation of preferences through repeated direct interaction—that is, an impression based on a partner's responses to one's own actions over time, rooted in the process of instrumental learning (Amodio 2019). Instrumental learning is a form of reward reinforcement, in which

an agent learns the reward value of approaching an object (or person) through choice and feedback: choices that result in positive feedback are repeated, whereas those resulting in negative feedback are avoided (Sutton & Barto, 1998). Instrumental learning is an action-based form of learning, supported by dopaminergic activity in the striatum and represented in terms of reward value (O'Doherty et al., 2004), and it has been proposed to support the process of forming social preferences through direct interaction (Hackel et al., 2015). This form of learning contrasts with the kind of semantic inference previously examined in studies of trait-based impression formation, which tends to be expressed most directly in conceptual judgments and verbal behavior (Amodio, 2019).

An instrumental learning account of impression formation is useful because it provides a theoretical bases for interaction-based social learning as well as a model for how this impression is updated and expressed. According learning theory (Sutton & Barto, 1998), reward associations are updated incrementally in response to feedback as a function of a *prediction error*, the difference between expected and actual reward feedback on a choice, and a *learning rate*, the degree to which the expected value is updated in response to a prediction error. This theory permits the formalization of specific patterns of learning that can be tested using a computational modeling approach, in which the fit of alternative models to task-based behavioral data is compared (Hackel & Amodio, 2018; Lockwood & Klein-Flügge, 2021).

Traast et al. (2024) used an instrumental learning approach to investigate the effect of race on interaction-based impression formation. In their experiments, White American participants interacted with four Black and four White players in a reinforcement learning task, presented as a money-sharing game. On each trial, participants viewed two players—one Black and one White—and choose to interact with the player expected to share a point (later converted to cash). Although individual players varied in their sharing rate, the sharing

rate was identical between Black and White players on average. Nevertheless, participants formed stronger reward associations with White compared with Black players, as indicated by their choice preferences. This effect was moderated by participants' racial attitudes, such that is emerged only for participants with relatively high anti-Black explicit prejudice and low internal motivation to respond without prejudice (Plant & Devine, 1998). These results provided a first demonstration of an effect of race on instrumentally-learned impressions.

To determine the cognitive mechanisms through which this effect of race occurred, Traast et al. (2024) tested a computational model of race-based instrumental learning. Their model, adapted from a model of stereotype-biased learning (Stillerman et al., 2022), proposed that race (a) biased White participants' initial expectancies of a player's feedback behavior before an interaction, modeled as a *prior*, and then (b) led participants to update reward representations for White and Black players with separate updating rules, modeled as *learning rates*. In comparisons with alternative models, this hypothesized *prior+learning* model provided the best fit to behavioral data, revealing that race can influence the process of learning in addition to biasing initial expectancies.

Generalization beyond the US context: Ethnic prejudice in the Netherlands

Although the results of Traast et al. (2024) comport well with existing research on how race affects impression formation and intergroup behavior (e.g., Devine, 1989; Dovidio et al., 2002; Macrae & Bodenhausen, 2000), it is unclear whether it describes a general effect of race/ethnicity on interaction-biased impression formation. Indeed, expressions of racial bias in the US reflect the unique history and current dynamics of race relations. In this section, we consider similarities and differences between the US and Dutch contexts that may relate to how race/ethnicity may influence impression formation in the Netherlands.

In Dutch society, prejudice based on race is prevalent (Essed & Hoving, 2014; Verkuyten & Thijs, 2002), with White European ancestry being the racial background for native and majority Dutch individuals (Essed & Trienekens, 2008; Mok & Mok, 1999). However, the focus of both political and public discourse lies predominantly in a person's ethnicity or national identity rather than race (Essed & Hoving, 2014; Essed & Trienekens, 2008). In the Netherlands, there are multiple ethnic minority groups due to migration from (former) colonies such as Surinam, Indonesia, and the Dutch Antilles, as well as labor migration from Morocco and Turkey, and refugee migration from countries like Syria, Ethiopia, Eritrea, Bosnia, Iran, Iraq, Somalia, and Rwanda among others.

Of these minority groups, the Moroccan Dutch population is generally subjected to the most discrimination (Andriessen et al., 2012; Hagendoorn & Pepels, 2017) and may thus be most comparable to Black Americans in the United States. Similar to Black Americans (Bleich et al., 2019; Bowleg et al., 2020; Wingfield & Chavez, 2020), Moroccan Dutch individuals face discrimination on the job market (Andriessen et al., 2012), in health care (Lamkaddem et al., 2012), and during encounters with law enforcement (Bonnet & Caillault, 2015). Stereotypes portraying both Black and Moroccan individuals as aggressive and violent are prevalent in both Dutch and American societies (Bleich et al., 2019; de Jong, 2007; Hagendoorn, 2017; Kleider-Offutt et al., 2017). This negative stereotype and stigma are reflected in behavior often displayed by White Dutch individuals: they maintain more distance from Moroccan avatars compared with White Dutch avatars (Dotsch & Wigboldus, 2008), and Dutch students were faster to decode anger on a Moroccan face compared with a White Dutch face (Bijlstra et al., 2014), echoing findings for Black Americans (Bleich et al., 2019).

Despite these parallels between US White-Black and Dutch White-Moroccan relations, there are several differences. Whereas White Americans sometimes perceive a

superordinate American identity that includes all groups (Hehman et al., 2012), ethnic minorities in the Netherlands are often labeled as immigrants, even when they born in the Netherlands and speak Dutch as their native language (da Silva et al., 2022). Moroccan individuals in the Netherlands are also distinguished from White Dutch by the Islamic faith (De Graaf et al., 2011), whereas both White and Black Americans are both predominantly Christian (Kramer et al., 2022).

Finally, social norms regarding intergroup interactions may differ between the US and the Netherlands. In the US, there exists a strong norm against the expression of prejudice toward Black people and other racial minorities (Crandall et al., 2002). The degree to which a person's intergroup responses are influenced by this norm can be measured in terms of their external motivation to respond without prejudice—that is, the motivation to respond without prejudice to avoid social disapproval (Plant & Devine, 1998). External motivation is distinct from internal motivation (i.e., based on one's personal beliefs), and the influence of external motivation is particularly strong in public situations (Plant & Devine, 1998; Plant et al., 2003). By contrast, norms prohibiting the expression of prejudice may be weaker in the Netherlands, where a premium is placed on directness and uninhibited expression (Rottier et al., 2011). These cultural norm differences may also influence the effect of ethnic prejudice in interaction-based impression formation.

Research Overview

In three preregistered experiments, we investigated the effect of ethnicity on social instrumental learning. These studies were conducted in a Dutch context, where the dominant ethnic majority group is White Dutch and the primary minority group target of prejudice is Moroccan Dutch (i.e., Moroccan Dutch nationals or immigrants). Therefore, in these studies White Dutch participants completed a social probabilistic reinforcement learning task,

presented as a money-sharing game, in which they interacted with White and Moroccan players. By choosing players to interact with and receiving immediate feedback, participants were able to learn who was more likely to share money than others and thus form player-specific reward representations.

Following Traast et al. (2024), our main hypothesis was that ethnicity would modulate impression formation, such that White Dutch participants would form more positive instrumental reward representations for White than Moroccan players, despite identical reward feedback from each group. We further expected that the effect of ethnicity on learning would be moderated by participants' internal motivation to respond without prejudice and their explicit prejudice, such that this effect would be greater for participants with lower internal motivation and stronger prejudice.

In addition, we hypothesized that the effect of ethnicity on impression formation would stem from two mechanisms: (a) different initial reward expectancies for the ethnic groups (*group-based prior*), such that participants would begin the task expecting more frequent rewards from White than Moroccan players, and (b) separate updating rules for White and Moroccan players (*group-based learning rates*), such that participants would maintain separate representations of White and Moroccan players and update them at separate rates in response to reward feedback. This hypothesis was investigated using computational model fitting, following prior work (Schultner et al., 2024; Stillerman et al., 2022; Traast et al., 2024).

Study 1

Method

Participants

Participants were 74 self-identified White Dutch psychology students from the University of Amsterdam in the Netherlands who completed the study in person in the laboratory. At the end of the experiment, participants indicated their ethnicity, prompted by the question "Please select all categories that you feel apply to you," and chose among the following: *Dutch, Moroccan, Turkish, Antillean, Surinamese, A different group, namely* (an open-ended response). They next indicated whether they were born in the Netherlands. The unique selection of "Dutch" was interpreted as White Dutch, given the White European background of Dutch people and the usage of this term in the Netherlands. Participants indicated their gender as either *female*, *male*, *other*, or *choose not to respond*.

Following exclusions based on preregistered criteria (https://aspredicted.org/RSZ_XPT) for below-chance learning (under 50% choice accuracy; 2 participants) or extremely fast reaction times (median RT<500 ms; 6 participants), the final sample for analysis included 66 participants (47 female-identified, 12 male-identified, and 6 did not indicate gender; M_{age}=19.80, SD_{age}=1.95). The preregistered stopping goal was N = 100 Dutch participants, following previous studies using a similar task design (Stillerman et al., 2022). However, due to the COVID-19 outbreak and ensuing lockdown, we were forced to end in-person data collection at 74 White Dutch participants. At this point, we decided to proceed with data analysis, in conjunction with the planning of additional pre-registered online replications (Studies 2 and 3). Participants received one research credit plus a performance-based bonus ranging from €1.00 to €2.00.

Procedure

In-person data collection occurred during February and March of 2020. Upon arrival at the laboratory, participants provided informed consent and then received instructions regarding the tasks. Participants completed the main learning task, followed by a set of post-task questionnaires. The task and questionnaires were administered on a laboratory computer via the open-source framework psiTurk (v3.3.0; (Eargle et al., 2020; Gureckis et al., 2016).

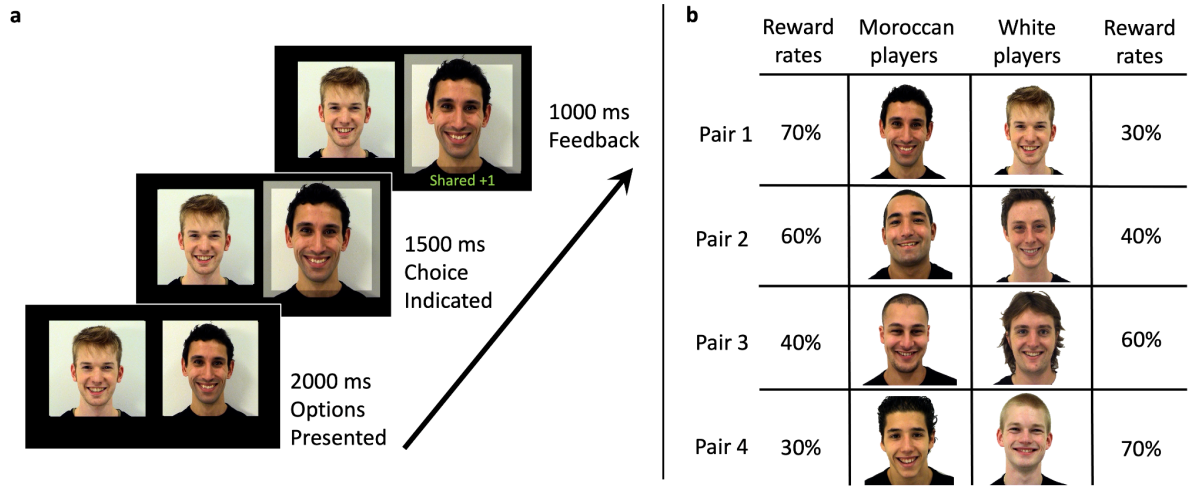
Task and measures

Social Reinforcement Learning task. Participants engaged in an interactive money-sharing task based on a probabilistic reward reinforcement paradigm (Frank et al., 2004) and adapted for the study of social instrumental learning (Hackel et al., 2015, 2022; Stillerman et al., 2022; Traast et al., 2024). Participants were informed that they would participate in a point-sharing game with eight other players, with the aim of choosing players most likely to share in order to accumulate the maximal points for themselves (converted to a cash bonus at the study's conclusion). Other players were presented as real participants who completed the task previously and whose sharing responses for each trial were taken from this prior study. In actuality, players were fictional and shared according to predetermined fixed reward rates (Figure 1b).

The eight players represented members of two groups, four with a White appearance and four with a Moroccan appearance. The gender of players was counterbalanced between participants, such that a participant interacted with either all male or all female players. Faces representing players were selected from the Amsterdam Dynamic Facial Expression Set (A.D.F.E.S.; van der Schalk et al., 2011; see *SI* for model numbers). All faces displayed smiles, consistent with the cover story that players were past participants who posed for a picture in their session.

Figure 1

Trial Sequence and Player Reward Rates



Note. Panel a shows a sample trial sequence of the training phase. Participants viewed two player faces, chose one to interact with (player on the right in the current trial), and then received feedback ('Shared: +1' or 'Shared: 0'). Panel b displays reward rates for player pairs during the training phase. Player images were randomized, and gender was counterbalanced.

The learning task included two phases: a *training phase* and a *test phase*. The training phase comprised two blocks of 80 trials. On each trial of the training phase, participants viewed a pair of faces—always one Moroccan and one White player—and chose which they would like to interact with, based on their expectation of who was more likely to share. Following each choice, participants received immediate feedback on whether the chosen player shared 1 or 0 points (Figure 1a). Participants knew that only one player would share per trial. If no response was given within 2.5 seconds, the trial ended without reward feedback, and a "too slow" message was displayed before proceeding to the next trial. During the training phase, participants chose among four fixed pairs of faces (Figure 1b). The respective reward rates of Moroccan and White players in each pair differed (70/30, 60/40,

40/60, or 30/70), such that in some pairs, the Moroccan player shared more often, whereas in other pairs, the White player shared more often. Critically, although the reward rates of players within each group varied, the average reward rate between White and Moroccan groups was equated at 50%. The face stimuli assigned to each reward rate and pair, face gender, and trial order was randomized across participants, and the presentation side for faces in each pair was randomized across trials.

Next, participants completed the test phase, which provided an assessment of what was learned. To this end, no feedback was given, and participants were presented with all possible White-Moroccan player pairings in order to assess more fine-grained preferences generalizing beyond those presented in the training phase. Participants were again instructed to choose the player who was more likely to share, and although no feedback was given, points for correct responses were added to participants' final monetary bonus. Reward learning was indicated by the degree to which participants selected players according to their reward rate during training. Critically, because reward rates were equated between groups, any group-based choice preference during the test phase would represent a group preference.

Perceived reward rates. After completing the learning task, participants were asked to explicitly estimate the reward rate of each player. These estimates provided a measure of participants' subjective perceptions of reward rates. Participants viewed the faces of each player one at a time, presented in random order, and rated each by typing in a number ranging from 0 to 100% on "What percentage of the time did this player share with you?".

Feeling thermometers. Participants' explicit prejudiced attitudes were measured using ethnicity-based feeling thermometers with the same groups and wording as in Verkuyten & Thijs (2010). Participants indicated their warmth toward five major immigrant groups in the Netherlands, as well as White Dutch people, on a scale of 0 (very cold) to 100 (very warm) degrees. These included Dutch people with no migration background (i.e., White

Dutch), a Moroccan migration background, a Turkish migration background, an Antillean migration background, a Surinamese migration background, and a western migration background (i.e., immigrants from other Western, Educated, Industrialized, Rich, and Democratic countries).

External & Internal motivation to respond without prejudice scales. Internal and external motivations to respond without prejudice toward Dutch Moroccan people were measured using an adapted and translated version of Plant and Devine's scales (see S.I.; (Plant & Devine, 1998). The internal motivation scale (IMS) assesses one's personal motivation for responding without prejudice, whereas the external motivation scale (EMS) assesses one's motivation due to real or perceived normative pressure. An example IMS item is "I am personally motivated by my beliefs to be unprejudiced toward Moroccan Dutch people." An example EMS item is "I try to hide any negative thoughts about Moroccan Dutch people in order to avoid negative reactions from others." Participants rated their agreement with each item on a scale ranging from 1 (strongly disagree) to 9 (strongly agree). Each participant was assigned a single EMS score, calculated as the average of the scores for the EMS items. The IMS score was computed in the same manner.

Additional post-task measures, which were not analyzed and are not reported in the main text, are described in the Supplemental Information.

Results

Descriptive statistics and intercorrelations for key variables in Table 1.

Table 1
Means, Standard Deviations, and Intercorrelations for Key Variables in Study 1

Variable	1	2	3	4	5
1. Ethnic difference in choice preference					
2. IMS	.07				
3. EMS	.25*	-.13			
4. Ethnic difference in perceived reward	.85**	-.02	.24		
5. Explicit prejudice	-.04	-.62**	.11	.00	
M	0.56	7.82	4.32	5.33	10.66
SD	0.13	0.90	1.27	15.00	14.69

Note. Ethnic difference in choice preference = proportion Moroccan over White player choices in test phase, from 0 (choosing only White players) to 1 (choosing only Moroccan players). IMS = internal motivation scale (range: 5.2 - 9; $\alpha = 0.68$). EMS = external motivation scale ($\alpha = 0.62$; range: 1.3 - 7.5). Ethnic difference in perceived reward = perceived reward rate for Moroccan – White players (scored -100 to 100). Explicit prejudice = feeling thermometer difference score for White Dutch -– Moroccan Dutch; higher scores represent more positive attitudes for Whites compared with Moroccans.
95% confidence intervals for correlation shown in brackets
*$p < .05$. **$p < .01$.

Explicit prejudice

Prior to examining our primary hypotheses, we tested whether participants showed explicit prejudice consistent with patterns of discrimination in the Netherlands. Indeed, participants reported more positive feeling thermometer ratings of White Dutch ($M = 74$, $SD = 14.06$) than Moroccans ($M = 63.34$, $SD = 14.57$), $t(64) = 5.85$, $p < .001$, Cohen's $d = 0.74$, 95% CI [0.46, 1.03]). Relative to other social groups (Turks: $M = 65.03$, $SD = 14.62$, Antilleans: $M = 66.92$, $SD = 13.97$, Surinamese: $M = 69.2$, $SD = 14.07$, Westerners: $M = 70.92$, $SD = 13.02$), participants' attitudes were numerically most positive toward White Dutch people and least positive toward Moroccans. Because our research question concerned anti-Moroccan prejudice, we created an explicit prejudice score computed as the difference in

ratings for White Dutch and Moroccan Dutch people, such that a higher score reflected a more pro-White/anti-Moroccan attitude.

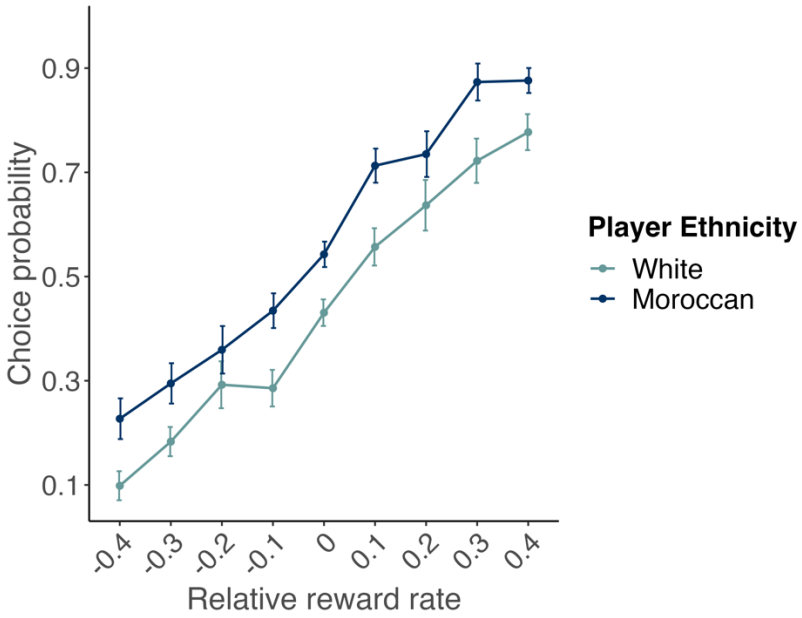
Effects of ethnicity on instrumental learning

Our primary hypothesis was that ethnicity would moderate instrumental learning. We expected White Dutch participants to form more positive reward associations with White Dutch players than with Moroccan players, in addition to forming preferences for higher-reward players. To test this prediction, we conducted a generalized linear mixed model with relative reward rate and ethnicity as predictors, using random slopes grouped within participants, and choice behavior as the outcome (R lme4 package, v1.1-26; Bates et al., 2015).

This analysis produced the expected main effect of relative reward, OR = 382.43, 95% CI = [133.89, 1092.30], $p < .001$, such that participants learned to choose high-rewarding players over low-rewarding players. This analysis also produced an effect of ethnicity on choice behavior, OR = 2.13, 95% CI = [1.44, 3.13], $p < .001$. However, the direction of this effect was *opposite* to our expectations: participants exhibited more positive reward associations with Moroccan players than with White players (Fig. 2). The direction of the ethnicity effect was especially surprising given participants' anti-Moroccan explicit attitudes, in addition to anti-minority patterns observed in previous research (Traast et al., 2024; Stillerman et al., 2022). To explore whether the ethnicity effect differed as a function of reward level, we reran the regression analysis above with the addition of the Relative Reward x Ethnicity interaction. This interaction effect was not significant, OR = 0.92, 95% CI = [0.49, 1.73], $p = .79$, nor did it change the main effects of relative reward and ethnicity. In what follows, we describe additional analyses aimed to provide insight into this unexpected result.

Figure 2

Effects of Ethnicity and Reward on Choice



Note. Effects of ethnicity and relative reward rate of player on choice during test phase in Study 1, showing a preference for choosing high-rewarding players, and for choosing Moroccan players over White players across relative reward rates. Relative reward rate (difference between training-phase reward rates of a choice pair) is displayed on the x-axis, and choice probability (probability of choosing a player) is displayed on the y-axis. Error bars represent standard errors.

Individual difference effects on choice preference

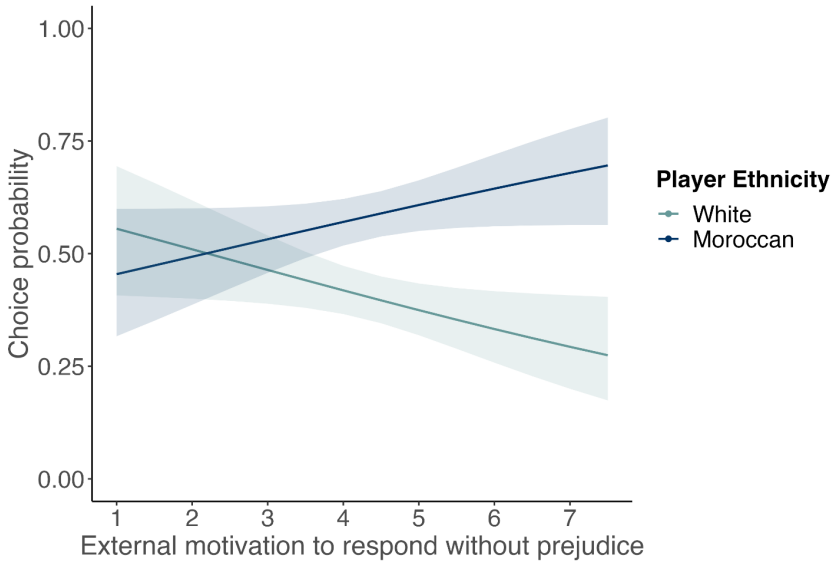
We first explored whether internal motivation moderated the effect of ethnicity on instrumental learning. Previous research by Traast et al. (2024) found that choice preference for White compared with Black players was moderated by participants' internal motivation, such that participants with lower IMS scores expressed greater anti-Black bias in their choice preferences. Based on this finding, we would expect anti-Moroccan choice preferences among low IMS participants, but pro-Moroccan preferences among high IMS participants, relative to White preferences. We tested this using a GLMM with relative reward rate, ethnicity, IMS, and an Ethnicity x IMS interaction as predictors, with choice behavior as the

outcome. The Ethnicity x IMS interaction was not significant, OR = 1.17, 95% CI = [0.76, 1.80], p = .481, indicating that the pro-Moroccan choice preference could not be explained by internal motivation.

We then tested whether explicit prejudice moderated the ethnicity effect. Traast et al. (2024) also found that White choice preference was moderated by explicit prejudice, such that participants with higher explicit prejudice displayed more anti-Black bias in their choices. To test whether explicit prejudice moderated choice preferences in the present study, we ran a GLMM with relative reward rate, ethnicity, explicit prejudice, and an Ethnicity x Explicit Prejudice interaction as predictors, and choice behavior as the outcome. The Ethnicity x Explicit Prejudice interaction was not significant, OR = 0.99, 95% CI = [0.97, 1.02], p = .602, and thus the pro-Moroccan choice effect also could not be explained by participants' explicit prejudiced attitudes.

Figure 3

EMS x Ethnicity Interaction Effect on Choice



Note. Ethnicity x EMS interaction effect on choice in Study 1, showing a stronger effect of ethnicity on choice preference among participants with higher external motivation. EMS displayed on the x-axis and choice probability (probability of player being chosen) on the y-axis. Shaded areas represent the 95% confidence interval.

Finally, we speculated that participants may have been motivated to appear nonprejudiced for external reasons—that is, to avoid negative social evaluation. A GLMM testing main effects of reward rate, ethnicity, and EMS, as well as an interaction between Ethnicity x EMS produced a significant interaction effect between Ethnicity x EMS, OR = 1.40, 95% CI = [1.04, 1.90], $p = .027$, such that people with higher EMS scores showed a preference for Moroccan players ($\beta = 1.15$, $t = 4.17$, $p < .001$) whereas people with low EMS showed no ethnic choice preference ($\beta = 0.29$, $t = 1.06$, $p = .287$; Figure 3). This interaction remains significant when IMS and the IMS x Ethnicity effects were added as covariates, OR = 1.44, 95% CI = [1.06, 1.96], $p = .018$.

If the unexpected pro-Moroccan preference was due to external motivation, then we would expect this preference to be present prior to learning, at the beginning of training phase. To examine this possibility, we tested whether a pro-Moroccan preference was already evident in the first 50 trials of training (following Traast et al., 2024, who used this early timeframe to examine priors). Indeed, we found an initial preference for Moroccan players over White players, OR = 1.37, 95% CI = [1.10, 1.72], $p = .005$, indicating that participants showed a choice preference for Moroccan players from the start of the experiment. However, this initial preference for Moroccan players was not more pronounced in participants with high EMS (EMS x Ethnicity interaction: OR = 1.05, 95% CI = [0.88, 1.26], $p = .566$). Thus, there appeared to be an initial pro-Moroccan choice preference across all participants during training that was unrelated to the effect of external motivation on test-phase choice preferences.

Ethnicity effects on perceived reward rates

Next, we examined whether participants subjectively perceived a difference in the reward rates of Moroccan and White players. Self-reported perceived reward rates were submitted to a multilevel linear regression model with ethnicity and actual player reward rate as predictors. This analysis showed that, in addition to tracking players' actual reward rates, $\beta = 0.89$, 95% CI = [0.76, 1.02], $p < .001$, participants self-reports were also influenced by ethnicity in their estimates, $\beta = 2.67$, 95% CI = [0.85, 4.48], $p = .004$, such that they estimated higher reward rates from Moroccan than White players, despite equated feedback from each group.

A final analysis tested whether the behavioral expression of preference in test phase choices simply reflected their subjective (mis)perception of more rewards from Moroccan players. This was not the case: Although an ethnic different in perceived rewards predicted behavioral choice preferences (Perceived Reward Difference x Ethnicity: OR = 1.09, 95% CI = [1.07, 1.11], $p <.001$), the ethnicity effect on choice behavior remained significant after controlling for perceived rewards (when including the Perceived Reward Difference x Ethnicity interaction, the main effect of ethnicity remained significant, OR = 1.33, 95% CI = [1.05, 1.69], $p = .018$).

Computational modeling results

Despite the surprising finding of pro-Moroccan choice preferences, we wondered whether the effect of ethnicity on learning occurred through similar mechanisms as in prior research. Traast et al (2024; see also Schultner et al., 2024, and Stillerman et al., 2022) investigated how race influenced the formation of preferences by fitting trial-by-trial behavioral data to different computational models. They found that choice behavior was best predicted by a *prior + learning* model, which included initial group-based expectancies

(prior) and separate group representations for updating (learning rates). Despite the unexpected finding of a pro-Moroccan choice preference, we tested whether group cues influenced choice behavior through the same set of processes.

Reward representations for the different groups were updated using the Rescorla-Wagner learning rule:

$$Q_{i,White}^{t+1} = Q_{i,White}^{t} + a_{White}(R^t - Q_{i,White}^{t})$$

$$Q_{i,Moroccan}^{t+1} = Q_{i,Moroccan}^{t} + a_{Moroccan}(R^t - Q_{i,Moroccan}^{t})$$

Priors were modeled as:

$$Q_{White}^{t=0} = P, \text{ and } Q_{Moroccan}^{t=0} = -P$$

To test this hypothesis, we examined the fit of our behavioral data to this hypothesized prior + learning model compared with three alternatives namely:

(a) An *unbiased model*, which contains no prior and a single learning rate applied across player ethnicity. In this model, ethnicity does not influence expectations or learning.

(b) A *group-based prior model*, which contains a group-based prior; in this model, participants begin with different initial reward representations for White and Moroccan players but updated according to a single learning rate. This model aligns with classic stereotyping frameworks in which stereotypes shape initial expectations which are replaced with individuated learning over time (Darley & Gross, n.d.; Srull & Wyer, 1988).

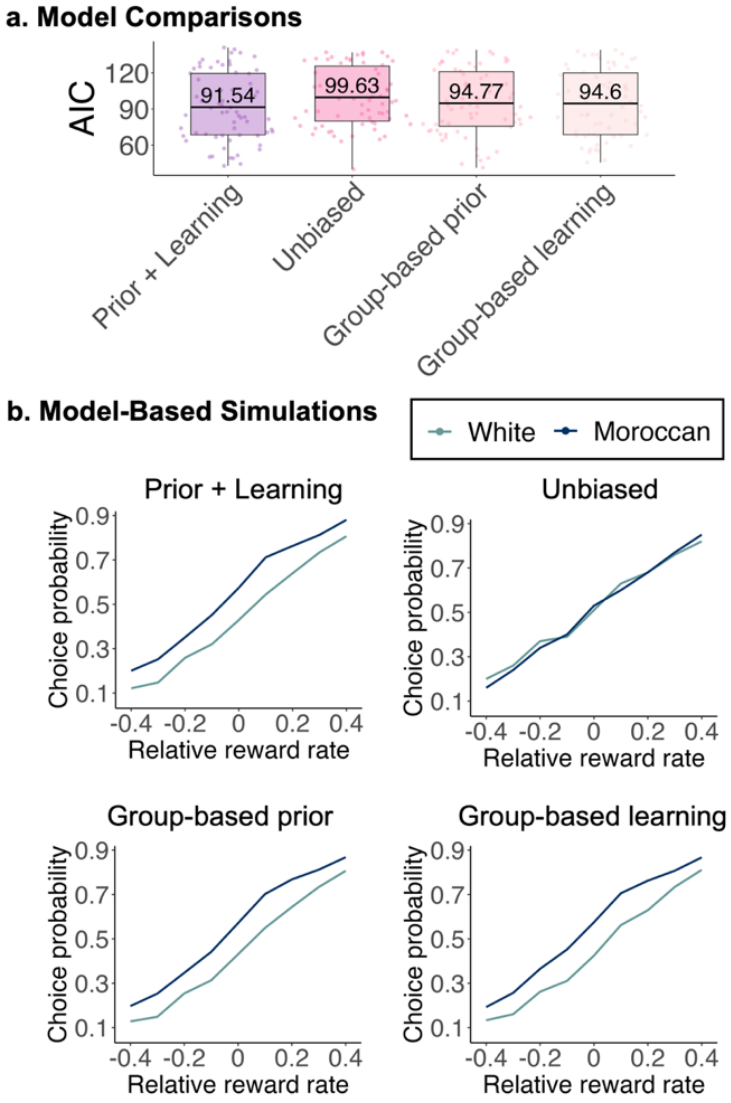
(c) A *group-based learning model*, which contains no prior but separate learning rates for White and Moroccan players; in this model, participants begin with no group-based expectancies but form group preferences according to separate updating rules.

The Akaike Information Criteria (AIC; Cavanaugh & Neath, 2019; Sakamoto et al., 1986) was used to identify the best fitting model. Model comparisons revealed that the

hypothesized prior + learning model provided the best fit the data, explaining the greatest amount of variation with the fewest possible parameters, as indicated by the lowest average AIC (Figure 4a). The difference in AIC between the prior + learning model (AIC = 91.54) and the competing models (unbiased model: AIC = 99.63, ΔAIC = 8.09; the group-based prior model: AIC = 94.77, ΔAIC = 3.23; group-based learning model: AIC = 94.60, ΔAIC = 3.06) suggests that participants did, in fact, form and sustain a group bias through the combination of adopting initial reward expectancies based on group identity, consistent with the pro-Moroccan preferences observed in early training behavior, and the updating these reward associations using separate learning rules for Moroccan and White players. This modeling result replicates prior studies of group-based effects on learning (Stillerman et al., 2022; Traast et al., 2024).

Figure 4

Computational Model Comparisons and Simulated Data



Note. (a) Model comparisons between the hypothesized prior + learning model with the unbiased model, group-based prior model, and group-based learning model in Study 1. (b) Model-based simulations for each model in Study 1. AIC = Akaike information criterion

Discussion

In Study 1, we investigated the effect of ethnicity on social impression formation in the Dutch cultural context. Although we hypothesized that ethnicity would modulate impression formation, such that White Dutch participants would form stronger choice-based preferences for White than Moroccan players, we found the opposite pattern: White Dutch

participants displayed a preference for Moroccan players, indicating that they formed stronger reward representations of the Moroccan players compared with White players. This behavioral preference was consistent with participants' reported perception that Moroccan players shared more frequently than White players (despite no actual difference). However, this effect emerged despite participants' explicit prejudice toward Moroccans as a group.

What might explain this unexpected pattern? Our exploratory analyses suggested that participants' pro-Moroccan choice responses during the task were driven by external motivation to respond without prejudice. External motivation is typically pronounced in public contexts, where one's behavior may be evaluated by others (Plant & Devine, 1998; Plant et al., 2003). It is possible that participants experienced the interactive task and the lab-based experimental session as a public context, despite their private responses and confidential identity.

Another possible cause of the unexpected effect was the use of smiling faces to depict players, which could have mitigated a prejudiced response (Raissi & Steele, 2021). However, while smiling expressions could have reduced anti-Moroccan prejudice, it would not be expected to create anti-White prejudice.

Despite observing an unexpected pro-Moroccan choice bias, computational modeling indicated that player ethnicity affected participants' behavior by inducing initial group-based reward expectancies and through separate learning rates for White and Moroccan players. This finding suggests that, despite an unexpected pattern of bias, the basic mechanisms through which group membership influenced instrumental learning were the same as in past research (e.g., Traast et al., 2024).

Study 2

Study 2 was designed to retest our original hypotheses—that participants would form choice preferences for White over Moroccan players—while controlling for factors that could have produced the unexpected result of Study 1. To this end, the facial expression of the players was changed from happy to neutral and, to reduce external motivation, the experiment was conducted online. Thus, we adhered to the same preregistration used for Study 1.

Method

Participants

Participants included 100 self-identified White Dutch Psychology students from the University of Amsterdam in the Netherlands, recruited via a test portal of the University of Amsterdam. Participants indicated their ethnicity, and gender in the same way as in Study 1. As preregistered, data collection stopped at 100 self-identified Dutch participants with the goal of obtaining valid data from at least 80 participants. Following the same preregistered exclusion criteria as in Study 1, exclusions for below-chance learning (under 50% choice accuracy; 8 participants) or extremely fast reaction times (median RT<500 ms; 9 participants), resulted in a final sample for analysis including 83 participants (58 female-identified, 24 male-identified, and 1 other-identified; M_{age} =20.46, SD_{age}=3.18). Participants received one research credit plus a performance-based bonus ranging from €1.00 to €2.00.

Procedure

Online data occurred from June to September 2020. Participants were forwarded to the informed consent and the task via a weblink. The task and questionnaires were hosted via

psiTurk (Gureckis et al., 2016). Post-task questionnaires were the same as in Study 1, except for four additional exploratory questions that are not discussed here (see *SI*).

Tasks and measures

Probabilistic Reinforcement Learning Task. The probabilistic reinforcement learning task was the same as in Study 1. We used the same models as in Study 1 but with neutral faces representing the players.

Post-task measures. As in Study 1, participants completed estimates of player reward rates, feeling thermometer ratings for major Dutch ethnic groups, and the IMS/EMS.

Results

The aim of Study 2 was to retest our original hypothesis that Dutch participants would form more positive impressions of White players compared with Moroccan players (opposite to what was found in Study 1). For this purpose, the same analysis-plan was used as in Study 1. Descriptives and intercorrelations are shown in Table 2.

Table 2.
Means, Standard Deviations, and Correlations of Key Variables in Study 2

Variable	1	2	3	4	5
1. Ethnic difference in choice preference					
2. IMS	.10				
3. EMS	.01	.08			
4. Ethnic difference in perceived reward	.82**	.08	.02		
5. Explicit prejudice	-.06	-.36**	.02	-.00	
M	0.56	7.55	4.80	3.97	10.18
SD	0.17	1.22	1.58	17.72	12.70

Note. Group player choice preference = proportion Moroccan over White player choices in test phase, from 0 (choosing only White players) to 1 (choosing only Moroccan players). IMS = internal motivation scale, range: 3.8 - 9.0, $\alpha = 0.77$. EMS = external motivation scale, range: 1 - 8.8, $\alpha = 0.68$. Ethnic difference in perceived reward = perceived reward rate for Moroccan – White players, scored -100 to 100. Explicit prejudice = feeling thermometer difference score for White Dutch -– Moroccan Dutch; higher scores represent more positive attitudes for Whites over Moroccans, range: -10 - 50.
95% confidence intervals for correlation are shown in brackets.
* indicates $p < .05$. ** indicates $p < .01$.

Explicit prejudice

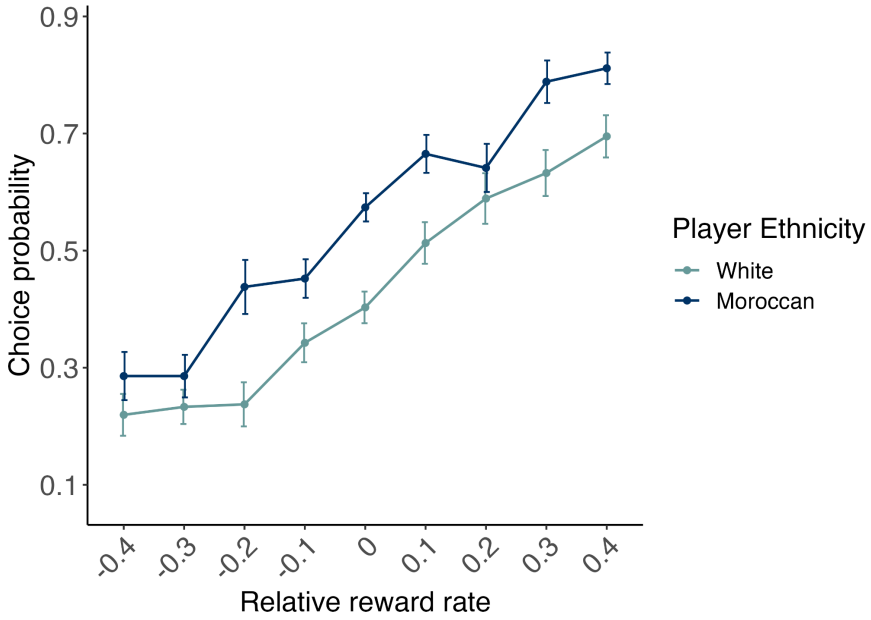
As in Study 1, participants reported more positive attitudes toward White Dutch (M = 76.02, SD = 12.41) than toward Moroccans (M = 65.84, SD = 14.81), $t(82) = 7.30$, $p < 0.001$, Cohen's $d = 0.74$, 95% CI [0.51, 0.96]. Again, attitudes were numerically most positive towards White Dutch and least positive towards Moroccans, relative to ratings of other groups (Turks: M = 67.71, SD = 14.51, Antilleans: M = 70.48, SD = 14.07, Surinamese: M = 73.25, SD = 13.51, Westerners: M = 73.86, SD = 12.96).

Effects of ethnicity on instrumental learning

Using the same regression model as in Study 1, we expected to find the originally predicted pattern of a pro-White choice preference. However, contrary to this prediction, results again showed the opposite effect: participants displayed a choice preference for Moroccan over White players, OR = 2.20, 95% CI = [1.37, 3.53], $p < .001$ (Figure 5), in addition to an effect of relative reward, OR = 88.37, 95% CI = [34.27, 227.89], $p < .001$. As in Study 1, a separate analysis showed no Ethnicity x Relative Reward interaction, OR = 1.44, 95% CI = [0.84, 2.48], $p = .187$. Thus, this analysis replicated the unexpected result of Study 1.

Figure 5

Effects of Ethnicity and Reward on Choice



Note. Effects of ethnicity of player, and relative reward on choice during test phase in Study 2, showing a preference for choosing high-rewarding players, and for choosing Moroccan players over White players across relative reward rates. Relative reward rate (difference between training-phase reward rates of a choice pair) is displayed on the x-axis, and choice probability (probability of choosing a player) is displayed on the y-axis. Error bars represent standard errors.

Individual differences in ethnicity effects

Because Study 2 was conducted online, we expected EMS effects to be reduced or eliminated. Consistent with this reasoning, EMS no longer moderated the effect of ethnicity on choice preference: in a GLMM containing ethnicity and relative reward rate as fixed effects and random effects, EMS as fixed effect, and an Ethnicity x EMS interaction, the EMS x Ethnicity interaction was not significant, OR = 0.94, 95% CI = [0.69, 1.27], p = .693. In a separate GLMM examining IMS effects, the IMS x Ethnicity interaction was also nonsignificant, OR = 1.25, 95% CI = [0.85, 1.83], p = .260.

As in Study 1, an analysis of choice preferences during the first 50 trials of training again revealed an initial preference for Moroccan over White players, OR = 1.25, 95% CI =

[1.00, 1.57], p = .055, suggesting that the preference was present from the start of the task and thus did not develop only through learning.

Ethnicity effects on perceived reward rates

As in Study 1, participants' subjectively perception of reward rates reflect players' actual reward rates, β = 0.03, 95% CI = [0.03, 0.04], p < .001, as well as ethnicity, β = 0.17, 95% CI = [0.01, 0.33], p = .034, such that participants perceived more frequent rewards from Moroccan players than White players (despite equated actual reward rates).

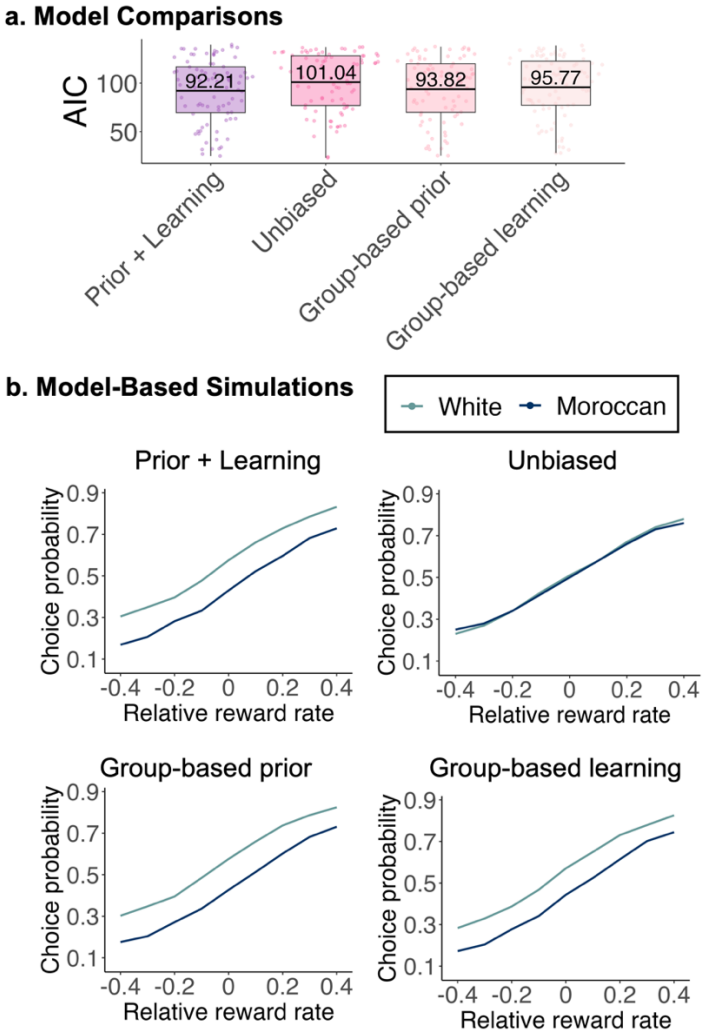
Furthermore, although perceived reward rates were associated with participants' choice preferences (Perceived Reward Difference x Ethnicity: OR = 1.10, 95% CI = [1.08, 1.12], p <.001), perceived reward did not fully explain participants' pro-Moroccan choice behavior (when the Perceived Reward Difference x Ethnicity interaction was included in this regression, the main effect of ethnicity remained significant, OR = 1.51, 95% CI = [1.13, 2.01], p = .005).

Computational modeling

As in Study 1, computational model comparison indicated that the *prior + learning* model best fit choice behavior data, suggesting that ethnicity influenced choice preferences through initial group-based expectancies and then updating according to separate learning rates (Figure 6).

Figure 6

Computational Model Comparisons and Simulated Data



Note. (a) Model comparisons between the hypothesized prior + learning model with the unbiased model, group-based prior model, and group-based learning model in Study 2. (b) Model-based simulations for each model in Study 2. AIC = Akaike information criterion.

Discussion

Study 2 replicated the unexpected main finding of Study 1: White Dutch participants showed a choice preference for Moroccan over White players, despite equal reward feedback from members of each group. As in Study 1, participants also perceived Moroccan players as sharing more frequently, although this effect of perceived reward did not fully account for the effect observed in choice behavior. These effects emerged despite participants' anti-

Moroccan explicit prejudice, as measured by feelings thermometers. In contrast to Study 1, pro-Moroccan choice preferences were no longer associated with external motivation—a pattern that may have reflected a reduction in self-presentational concerns in the online study context. Thus, while Study 2 replicated the main findings of Study 1, its results continued to leave us without an explanation for this unexpected pattern.

Study 3

In Study 3, we conducted another replication and further investigated possible reasons for the observed pro-Moroccan choice preference. First, we included a pre-task White vs. Moroccan Implicit Association Test (IAT; Greenwald et al., 1998) to determine whether participants had pro-Moroccan implicit attitudes that guided their task behavior, in contrast to their anti-Moroccan explicit prejudice. We also included a post-task IAT to test whether engaged in the interactive task changed implicit prejudice. Moreover, because ethnic group in these IATs were represented by face images of players in the task, we could test whether the pro-Moroccan choice preference in the learning task related to a preference toward the specific individuals in the task, in contrast to the measure of explicit prejudice which references abstract group representations. Second, we included a post-task questionnaire to probe possible response strategies that could lead to pro-Moroccan task behavior. Finally, in Study 3, we updated our hypothesis and preregistration to predict the pro-Moroccan effect observed in Studies 1 and 2.

Method

Participants

Participants included 100 self-identified White Dutch Psychology students from the University of Amsterdam in the Netherlands, recruited via a test portal of the University of

Amsterdam. Participants indicated their ethnicity and gender in the same manner as in Study 1 and Study 2. As preregistered (https://aspredicted.org/WK2_HT6), data collected stopped once we obtained valid data from 80 self-identified Dutch participants (62 female-identified, 18 male-identified; M_{age} = 20.5, SD_{age} = 2.29) following exclusions for below-chance learning (6 participants) or extremely fast reaction times (14 participants). Participants received one research credit plus a performance-based bonus ranging from €1.00 to €2.00.

Procedure

The study was conducted online from April to June 2021. The procedure was identical to that of Study 2 with the addition of pre-and post-task IATs and post-task questionnaires regarding response strategies.

Tasks and measures

Implicit Association Test. Participants' implicit attitudes towards Moroccans versus White Dutch were measured with an implicit association test (Greenwald et al., 1998; 2003). IATs were completed immediately before (pre-task IAT) and after (post-task IAT) the probabilistic reinforcement learning task. The pre- and post-task IATs included 80 Dutch evaluative words unrelated to ethnic stereotypes (40 pleasant, 40 unpleasant; see *SI* for exact word list; van Ravenzwaaij et al., 2011). Importantly, the eight face images represented players for participant in the sharing game (four Moroccan, four White). Both IATs consisted of 7 blocks (Greenwald et al., 2003), which half of the participants completed with blocks with Pleasant/Moroccan assigned to the same key first, and the other half of the participants completed with blocks with Pleasant/White assigned to the same key first. Although block order was counterbalanced across participants, each participant completed pre- and post-task IATs with the same block order so that their scores on each IAT would be comparable.

Using natural log transformed reaction times for correct responses, *D* scores were computed for each participant as in Amodio & Devine (2006): compatible block RTs were subtracted from incompatible block RTs and divided by the pooled *SD* separately for practice and test blocks. These resulting scores were then averaged for the final *D* score. Change in implicit attitude was scored as post-task *D* minus pre-task *D*.

Probabilistic Reinforcement Learning Task. The probabilistic reinforcement learning task was the same as in Study 2.

Post-task measures. Following task completion, participants indicate their perceived reward estimates for each player and completed feeling thermometers, as in Studies 1 and 2. Next, they completed new questionnaire items assessing possible response strategies during the task. Participants were asked "When you made a choice for one player or the other, how much was your choice influenced by the following consideration." Participants then rated each of the following: (a) "the ethnicity of the player" (followed by ""I predominantly chose Moroccan players" or "I predominantly chose White players"); (b) "I did not want to come across as prejudiced," (c) "I wanted to choose players of my own ethnicity,"; (d) " the appearance of the players, unrelated to their ethnicity,"; (e) "whether a player shared money with me in the first few trials that I chose them." Ratings were given on a 6-point scale ranging from 0 "no influence at all" to 5 "a very strong influence."

Results

Table 3

Means, standard deviations, and correlations for key variables in Study 3

Variable	1	2	3	4	5
1. Ethnic difference in choice preference					
2. pre-task IAT	-.03				
3. post-task IAT	-.02	.32**			
4. Ethnic difference in perceived reward	.77**	.09	.04		
5. Explicit prejudice	-.15	.03	-.06	-.07	
M	0.57	0.27	0.19	8.82	10.01
SD	0.14	0.37	0.30	18.12	15.09

Note. Ethnic difference in choice preference = proportion Moroccan over White player choices in test phase, from 0 (choosing only White players) to 1 (choosing only Moroccan players). Pre-task and post-task implicit association tests (IATs) = d-score from -1 (relative preference for Moroccan) to +1 (relative preference for White). Ethnic difference in perceived reward = perceived reward rate for Moroccan minus White players, scored -100 to 100. Explicit prejudice = feeling thermometer difference score for White Dutch -minus Moroccan Dutch; higher scores represent more positive attitudes for Whites over Moroccans, range: -10 - 50.
95% confidence intervals for correlation are shown in brackets.
* indicates $p < .05$. ** indicates $p < .01$.

Explicit prejudice

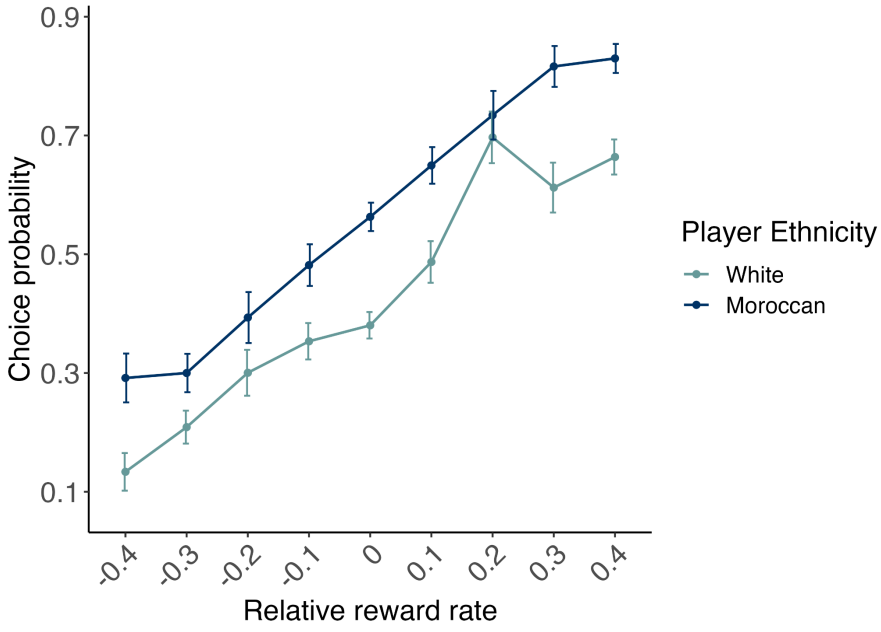
As in Studies 1 and 2, participants reported more positive attitudes toward White Dutch ($M = 78.31$, $SD = 13.17$) than Moroccans ($M = 68.30$, $SD = 16.70$), $t(79) = 5.94$, $p < 0.001$, Cohen's $d = 0.66$, 95% CI [0.42, 0.90], and attitudes were numerically most positive towards White Dutch and least positive towards Moroccans, relative to other groups (Turks: $M = 69.84$, $SD = 16.45$, Antilleans: $M = 73.28$, $SD = 16.51$, Surinamese: $M = 75.59$, $SD = 16.17$, Westerners: $M = 75.71$, $SD = 15.35$).

Effects of ethnicity on instrumental learning

As in the previous studies, we investigated the effect of ethnicity on instrumental learning with a general linear mixed model. A significant ethnicity effect indicated that participants preferred Moroccan over White players, OR = 2.40, 95% CI = [1.65, 3.50], $p <$ .001 (Figure 7), in addition to preferring high-rewarding players over low-rewarding players, OR = 76.87, 95% CI = [36.79, 160.60], $p <$ .001. The ethnicity effect was not moderated by actual reward rate (Ethnicity x Reward Rate: OR = 1.59, 95% CI = [0.94, 2.27], p = .09). These results replicated those of Studies 1 and 2.

Figure 7

Effects of Ethnicity and Reward on Choice



Note. Effects of ethnicity of player and relative reward rate on choice during test phase in Study 3, showing a preference for choosing high-rewarding players, and for choosing Moroccan players over White players across relative reward rates. Relative reward rate (difference between training-phase reward rates of a choice pair) is displayed on the x-axis, and choice probability (probability of choosing a player) is displayed on the y-axis. Error bars represent standard errors.

To investigate initial choice preference, we again examined preferences during the first 50 trials of training. Again, participants' preference for Moroccan players was already evident in the first 50 trials, OR = 1.71, 95% CI = [1.44, 2.04], $p < .001$. As in Studies 1 and 2, this finding suggests participants began the task with a preference for Moroccan players.

Implicit attitude effects

Participants exhibited an implicit preference for White over Moroccan faces on both the pre-task IAT ($M = 0.27$, $SD = 0.37$; $t(79) = 6.65$, $p < .001$) and the post-task IAT ($M = 0.19$, $SD = 0.30$; $t(79) = 5.59$, $p < .001$). This pattern was consistent with participants' average explicit preference for White over Moroccan people but contrasted with their preference for Moroccan players in choice behavior and perceived reward rates. Neither pre-task or post-task implicit attitudes were correlated with Moroccan choice preference (see intercorrelations in Table 3), and therefore participants' Moroccan choice bias did not reflect their implicit attitudes. Notably, the pro-White direction of implicit preference contrasted with participants' pro-Moroccan choice preferences during the learning task, suggesting that their task behavior was not due to their liking for the specific players.

Next, we investigated our hypothesis that White participants' implicit attitudes towards Moroccan faces would become more positive following their pro-Moroccan choices in the learning task. Given our directional prediction such that implicit attitudes would be less in the post-task IAT than in the pre-task IAT, we tested our hypothesis with a one-tailed paired t-test. As expected, post-task IAT d-scores were significantly lower (i.e., closer to zero) than pre-task IAT d-scores, $t(79) = 1.89$, $p = .031$, suggesting that participants' implicit attitudes were less anti-Moroccan after the learning task than before.

This change in IAT score could have reflected attitude change in response to task-based interactions, or it could have reflected an IAT practice effect. A general linear mixed

model testing main effects of reward rate, ethnicity, and implicit attitude difference, as well as an interaction between Ethnicity x Implicit Attitude Change, did not produce a significant Ethnicity x Implicit Attitude Change interaction, OR = 1.03, 95% CI = [0.40, 2.68], p = .94. Thus, the observed change in implicit prejudice was not associated with task behavior.

Ethnicity effects on perceived reward rates

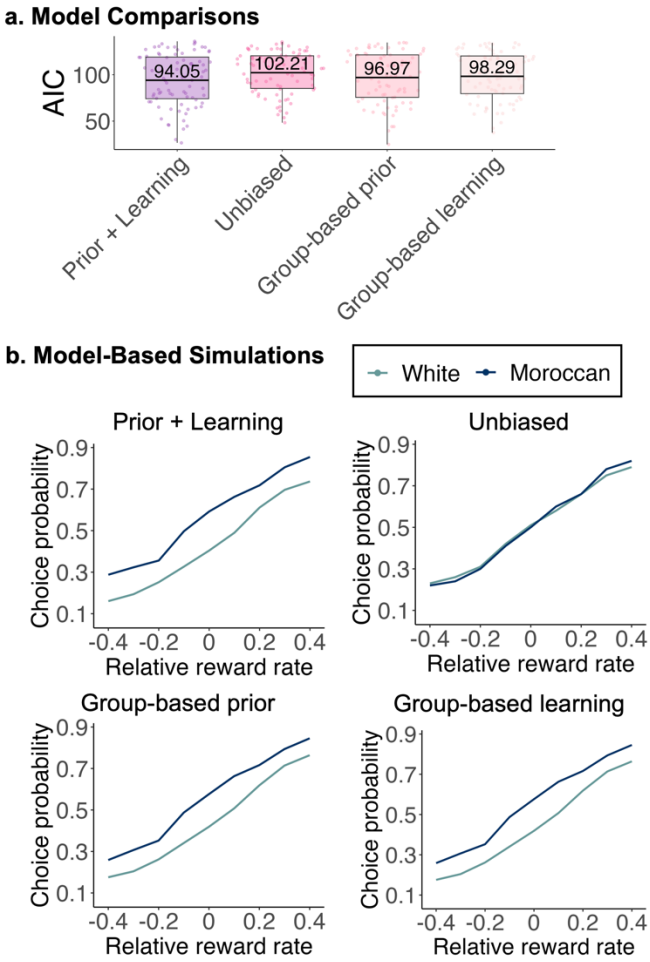
As in Studies 1 and 2, perceived reward rates of the players coincided with actual reward rates (β = 0.86, 95% CI = [0.72, 0.99], p < .001), and were influenced by ethnicity, β = 4.41, 95% CI = [2.50, 6.33], p < .001, such that they were higher for the Moroccan players compared to the White players. Participants' perception of higher rewards from Moroccan players was again associated with their choice behavior preference (Perceived Reward Difference x Ethnicity: OR = 1.08, 95% CI = [1.06, 1.09], p <.001), as in Studies 1 and 2. Unlike past studies, however, inclusion of Perceived Reward Difference x Ethnicity interaction reduced the main effect of ethnicity to nonsignificance, OR = 1.28, 95% CI = [0.97, 1.68], p = .080; that is, participants' self-reported perceptions largely explained their behavioral choice preferences in this study.

Computational modeling

As in the previous studies, computational model comparison indicated that the prior + learning model best fit choice behavior data, suggesting that ethnicity influenced choice preferences through group-based expectancies and separate learning rates (Figure 8), indicating that participants acquired and maintained a group bias through a combination of group-based initial expectancies and the updating of separate representations for Moroccan and White players. This result replicated Studies 1 and 2, and past research.

Figure 8

Computational Model Comparisons and Data Simulations



Note. (a) Model comparisons between the hypothesized prior + learning model with the unbiased model, group-based prior model, and group-based learning model in Study 3. (b) Model-based simulations for each model in Study 3. AIC = Akaike information criterion

Post-task questionnaire

Means and correlations for post-task strategy items are displayed in Table 4. Here, we describe responses to each item in turn.

Ethnicity of player. Participants indicated that, on average, ethnicity of a player had a weak influence on their decisions (M = 1.48, SD = 1.12). However, when forced to indicate whether they predominantly chose Moroccan players or White players (Table 4: variable 3),

their answer tended to reflect their task choices, OR = 0.23, 95% CI = [0.11, 0.49], $p < .001$; participants who indicated they predominantly chose Moroccan players showed a choice preference for Moroccan players during the task ($\beta = 1.61$, $t = 6.55$, $p < .001$), whereas participants who indicated they predominantly chose White players showed no ethnicity effect ($\beta = 0.16$, $t = 0.56$, $p = .574$).

Avoiding appearance of prejudice. On average, participants indicated moderate influence of wanting to avoid the appearance of prejudice in their decisions (M = 3.03, SD = 1.35). This item was not associated with an ethnicity bias in task choices, conceptually replicating the lack of an EMS effect in Study 2.

Desire to interact with players of own ethnicity. Participants generally did not report a desire to interact with players of their own ethnicity (M = 0.57, SD = 0.89). However, this item related to choice behavior, OR = 0.64, 95% CI = [0.43, 0.97], $p = .037$, such that these White participants with lower desire for own-ethnicity interaction showed a pro-Moroccan choice preference ($\beta = 1.25$, $t = 4.74$, $p < .001$), whereas this effect was nonsignificant for participants with higher desire for own-ethnicity interaction ($\beta = 0.48$, $t = 1.81$, $p = .070$). However, this variable was highly negatively skewed, suggesting that this effect was driven by a small number of participants with a strong ingroup preference.

Reciprocating player sharing. The mostly highly endorsed reason for choosing a player was that the player shared with the participant during initial trials (M = 4.05, SD = 1.09)—an explanation that did not reference ethnicity and was unrelated to ethnic preference in choice behavior.

Table 4

Correlations of Post-Task Debriefing Items with Ethnic Difference in Choice Preference

Variable	1	2	3	4	5	6	7
1. Ethnic difference in choice preference							
2. Ethnicity	-.06						
3. Perceived group choice	.45**	.09					
4. Seeming nonprejudiced	-.12	.21	-.25				
5. Own group preference	-.25*	.44**	-.21	.26*			
6. Appearance	.10	.17	-.03	.19	.20		
7. Initial reward	-.01	.11	-.06	.06	.07	.16	
M	0.57	1.48	1.41	2.73	0.57	3.03	4.05
SD	0.14	1.12	0.50	1.44	0.89	1.35	1.09

Note. Ethnic difference in choice preference = proportion Moroccan over White player choices in test phase, from 0 (choosing only White players) to 1 (choosing only Moroccan players). Ethnicity = "the ethnicity of the player," range: 1 - 4. Perceived group choice = "I predominantly chose White(1)/Moroccan(2) players." Seeming nonprejudiced = "I did not want to come across as prejudiced," range: 0 - 5. Own group preference = "I wanted to choose players of my own ethnicity," range: 0 - 3. Appearance = " the appearance of the players, unrelated to their ethnicity," range: 0 - 5. Initial reward = "whether a player shared money with me in the first few trials that I chose them," range: 0 - 5. 95% confidence intervals for correlation are shown in brackets.
* indicates $p < .05$. ** indicates $p < .01$.

Discussion

Study 3 replicated several findings from Studies 1 and 2. First, White participants again showed a choice preference for Moroccan players over White players and also perceived that Moroccan players shared more frequently, despite actually receiving equal feedback from both groups. Second, this pro-Moroccan choice bias was already observed in early training behavior, and computational modeling indicated that it reflected an existing pro-Moroccan prior combined with separate learning rates for each group. Third, despite their

pro-Moroccan task preferences, participants, on average, reported anti-Moroccan explicit prejudice.

Study 3 additionally assessed implicit prejudice and examined whether repeated interactions with Moroccan players would reduce implicit prejudice. Participants showed anti-Moroccan implicit prejudice before and after the task. They also showed a slight reduction in implicit prejudice following the task. However, this change in implicit attitudes was not associated choice behavior, and thus we could not conclude that this change was related to participants engagement with Moroccan and White players in the task. An alternative explanation—that the reduction in IAT scores reflects a practice effect (Thomas et al., 2007; Vaughn et al., 2011)—is thus more plausible.

IAT scores allowed us to address another possible explanation: that despite participants explicit prejudice toward Moroccans as a group, they might prefer individual Moroccans in direct interactions. However, scores on the IAT, which assessed responses toward the specific players in the task, showed an anti-Moroccan bias, contradicting this explanation.

Finally, Study 3 probed potential reasons for participants' task behavior. whether the pro-Moroccan choice originated from a choosing strategy in order to not come across as prejudiced. These were not particularly enlightening, as they did not suggest an explanation for the repeated finding of a pro-Moroccan choice preference.

General Discussion

We examined the effect of ethnicity on impression formation through social interaction in a Dutch context in an effort to generalize findings previously observed in the US context. In this prior research, (Traast et al., 2024), White American participants formed stronger behavioral preferences toward White than Black partners through repeated

interaction, despite equivalent reward feedback from partners. Thus, we expected White Dutch participants to form stronger preferences toward White than Moroccan interaction partners in a similar experimental task. Unexpectedly, we observed a behavioral choice preference for Moroccan players over White Dutch players in all three experiments. This unexpected pattern was robust to several design changes across studies, including a switch from using smiling to neutral faces and a switch from in-lab to online data collection, and cash incentive for accurate choices, and thus it presented a theoretical puzzle for us to solve.

An initial clue came from participants' self-reported prejudice which, as expected, showed a strong preference for Dutch people and against Moroccan people. This led us to hypothesize that pro-Moroccan behavior during the interactive learning task was due to participants' external motivation to respond without prejudice. Although this pattern was evident in Study 1, it was not observed in Study 2, based on EMS scores, or Study 3, based on a measure of task-specific desire to avoid the appearance of prejudice. If, as we contended, a move to online data collection would enhance feelings of privacy and anonymity, and thus reduce external motivation, then we would also expect to see a pro-White choice preference in Studies 2 and 3 that matched participant's explicit and implicit anti-Moroccan prejudice. But this reversal was not observed—participants continued to show a pro-Moroccan choice preference—and thus the EMS trail went cold.

It is also possible that participants intentionally chose Moroccan players for reasons other than external motivation. As in Traast et al. (2024), participants reported a higher sharing rate from Moroccan than White players, despite equated rates, and this explicit estimate correlated with participants' choice-based preferences. However, explicit sharing estimates did not fully account the pro-Moroccan effect in choice behavior, suggesting that any explicit intention to prefer Moroccan players did not completely explain this choice bias.

This suggests that participants' choice preferences were not only driven by explicit beliefs but may have reflected a degree of implicit processing.

Computational modeling results added further clues to this unexpected pattern: replicating Traast et al. (2024), they showed that the group preference was based on a combination of initial expectancies (*modeled as priors*)—in this case, an expectancy that Moroccan players were more likely to share than White players, corroborating the behavioral preference in early training trials—as well as maintaining separate representations of Moroccan and White player reward associations, as indicated by separate learning rates or each group. This pattern indicates that while the pro-Moroccan choice preference was in part due to a pre-task expectancy, it developed further through the process of learning across that task. That is, the initial expectancy may have shaped participants' perceptions of feedback during training, leading them to believe that Moroccan players were in fact sharing more often. Although this pattern does not explain why participants exhibited a pro-Moroccan preference, it may explain why participants reported higher sharing rates from Moroccan players. More broadly, these results suggest that ethnicity affects interaction-based instrumental learning via the same mechanisms as seen in prior studies of race (Traast et al., 2024) and stereotyping (Stillerman et al., 2022).

In summary, we found that in a Dutch context, ethnicity did indeed affect how participants formed impressions of White and Moroccan partners through repeated instrumental interaction—but in a pro-Moroccan outgroup direction that was unexpected. Moreover, our attempts to explain this pattern in follow-up studies were unsuccessful, and the mystery of why White Dutch participants preferred Moroccan partners, despite their anti-Moroccan implicit and explicit attitudes, remains unsolved.

Potential explanations from a cultural perspective

Further consideration of Dutch and American cultural differences may shed light on our unexpected findings. First, we considered the possibility that the nature of intergroup threat differs between contexts. Group Threat Theory states that both perceived economic threat (Quillian, 1995) and group size (Schlueter & Scheepers, 2010) contribute to perceived threat of a minority outgroup. In the Netherlands, people with a Moroccan background make up only approximately 2% of the total Dutch population (CBS, 2022) and are not typically considered an economic threat to the White Dutch majority (Andriessen et al., 2012; Hagendoorn & Pepels, 2017; Ramos et al., 2021). By contrast, in the US, Black Americans comprise approximately 14% of the U.S. population (US census, 2022), and there is widespread belief among White Americans that Black Americans and other minorities threaten their jobs (Perkins et al., 2020). Thus, it is possible that Moroccans are viewed as less threatening to White Dutch people, compared with White Americans' views toward Black Americans. However, this cultural difference cannot in itself explain our results: lower intergroup threat in the Netherlands might predict a reduction in anti-Moroccan bias but not a reversal.

Another possible difference concerns the content of Moroccan Dutch and Black American stereotypes. Moroccans in the Netherlands are often perceived as more generous and warmer compared with the White Dutch, who by contrast are stereotyped as greedy and stingy (Van Ginkel, 1996). Given that task interactions in our studies involved the sharing of money, these stereotypes could have led to a pro-Moroccan/anti-White Dutch preference in this particular context. This possibility remains plausible and could be tested in future research.

Finally, it is possible that cultural differences exist in the nature and expression of external motivation in the US and Netherlands. Whereas strong norms prohibiting the expression of prejudice toward Black people exist in the US (Plant & Devine, 1998), such

norms are relatively weaker in the Netherlands where there is a premium on directness and individual expression (Rottier et al., 2011). Although we measured external motivation using an adapted version of Plant and Devine's (1998) scale (also, e.g., Derous et al., 2009; Jargon & Thijs, 2021) and observed mean EMS scores comparable to those found in US samples, this measure might not sufficiently capture the expression of external motivation in a Dutch context. New research is needed to understand whether cultural differences in external motivation to understand their effects in non-US contexts.

Contributions to theory on intergroup impression formation

Despite our unexpected main finding, this research contributes several advances to research on social-interactive impression formation and its underlying learning mechanisms. First, it demonstrated an effect of ethnicity on the formation of individual person impressions through social interaction and replicated a computational model of race on impression formation through repeated interaction (Traast et al., 2024). Second, it provides a crucial first test of this learning process in a non-US context, raising new questions regarding cross-cultural generalization. And third, in attempting to explain unexpected findings, this research examined and ruled out multiple response strategies that may influence instrumental social learning. Ultimately, this research illuminates a previously-unidentified gap in our understanding of social-interactive impression formation processes across cultures and highlights the need for additional research on this topic.

Author contributions.

Data availability. Preregistration of study design, hypotheses and analyses can be found at https://aspredicted.org/RSZ_XPT (Study 1 & Study 2) and https://aspredicted.org/WK2_HT6 (Study 3). The datasets for the three studies, and the Supplemental Information are available in the OSF repository, https://osf.io/2tn54/?view_only=d489d37fcfc949a7bb4f19cca62d76fc. All data were analyzed using R Statistical Software (v4.3.1; R Core Team, 2023). This study complies with TOP level 2 guidelines (Nosek et al., 2015).

Acknowledgements.

Competing interests. The authors declare no competing interests.

References

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley Publishing Company. https://play.google.com/store/books/details?id=5AE7AAAAMAAJ

Amodio, D. M. (2019). Social Cognition 2.0: An Interactive Memory Systems Account. *Trends in Cognitive Sciences*, *23*(1), 21–33. https://doi.org/10.1016/j.tics.2018.10.002

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661. https://doi.org/10.1037/0022-3514.91.4.652

Andriessen, I., Nievers, E., Dagevos, J., & Faulk, L. (2012). Ethnic Discrimination in the Dutch Labor Market: Its Relationship With Job Characteristics and Multiple Group Membership. *Work and Occupations*, *39*(3), 237–269. https://doi.org/10.1177/0730888412444783

Berry, J. W. (2001). A psychology of immigration. *The Journal of Social Issues*, *57*(3), 615–631. https://doi.org/10.1111/0022-4537.00231

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: effects of formal and informal evaluation tools. *Personality & Social Psychology Bulletin*, *36*(7), 855–868. https://doi.org/10.1177/0146167210369483

Biernat, M., Sesko, A. K., & Amo, R. B. (2009). Compensatory Stereotyping in Interracial Encounters. *Group Processes & Intergroup Relations: GPIR*, *12*(5), 551–563. https://doi.org/10.1177/1368430209337469

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personality & Social Psychology Bulletin*, *40*(5), 567–577. https://doi.org/10.1177/0146167213520458

Bleich, S. N., Findling, M. G., Casey, L. S., Blendon, R. J., Benson, J. M., SteelFisher, G. K.,

Sayde, J. M., & Miller, C. (2019). Discrimination in the United States: Experiences of black Americans. *Health Services Research*, *54 Suppl 2*(Suppl 2), 1399–1408. https://doi.org/10.1111/1475-6773.13220

Bonnet, F., & Caillault, C. (2015). The invader, the enemy within and they-who-must-not-be-named: how police talk about minorities in Italy, the Netherlands and France. *Ethnic and Racial Studies*, *38*(7), 1185–1201. https://doi.org/10.1080/01419870.2014.970566

Bowleg, L., Maria del Río-González, A., Mbaba, M., Boone, C. A., & Holt, S. L. (2020). Negative Police Encounters and Police Avoidance as Pathways to Depressive Symptoms Among US Black Men, 2015–2016. *American Journal of Public Health*, *110*(S1), S160–S166. https://doi.org/10.2105/AJPH.2019.305460

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews. Computational Statistics*, *11*(3), e1460. https://doi.org/10.1002/wics.1460

Crandall, C. S., Eshleman, A., & O'brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of Personality and Social Psychology, 82*, 359-378.

da Silva, C., de Jong, J., Feddes, A. R., Doosje, B., & Gruev-Vintila, A. (2022). Where Are You Really From? Understanding Misrecognition From the Experiences of French and Dutch Muslim Women Students. *Journal of Social and Political Psychology*, *10*(1), 201–217. https://doi.org/10.5964/jspp.9395

Darley, J. M., & Gross, P. H. (n.d.). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

De Graaf, P. M., Kalmijn, M., & Kraaykamp, G. L. M. (2011). *Sociaal-culturele verschillen*

tussen Turken, Marokkanen en autochtonen: eerste resultaten van de Nederlandse LevensLoop Studie (NELLS). repository.ubn.ru.nl. https://repository.ubn.ru.nl/bitstream/handle/2066/99714/99714.pdf

de Jong, J. D. A. (2007). *Kapot moeilijk: een etnografisch onderzoek naar opvallend delinquent groepsgedrag van "Marokkaanse" jongens*. https://research.rug.nl/en/publications/kapot-moeilijk-een-etnografisch-onderzoek-naar-opvallend-delinque

Derous, E., Nguyen, H. H., & Ryan, A. M. (2009). Hiring discrimination against Arab minorities: Interactions between prejudice and job characteristics. *Human Performance, 22*, 297-320.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. https://www.ncbi.nlm.nih.gov/pubmed/12003481

Dotsch, R., & Wigboldus, D. H. J. (2008). Virtual prejudice. *Journal of Experimental Social Psychology*, *44*(4), 1194–1198. https://doi.org/10.1016/j.jesp.2008.03.003

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 3–28). SAGE Publications. https://doi.org/10.4135/9781446200919.n1

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62. https://psycnet.apa.org/journals/psp/82/1/62/

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the Nature of Prejudice: Automatic and Controlled Processes. *Journal of Experimental Social*

Psychology, *33*(5), 510–540. https://doi.org/10.1006/jesp.1997.1331

Eargle, D., Gureckis, T., Rich, A. S., McDonnell, J., & Martin, J. B. (2020). *psiTurk: An open platform for science on amazon mechanical turk (Version v2. 3.11). Zenodo*.

Essed, P., & Hoving, I. (2014). Innocence, smug ignorance, resentment: An introduction to Dutch racism. In *Dutch racism* (pp. 9–29). Brill. https://brill.com/downloadpdf/book/edcoll/9789401210096/B9789401210096-s002.pdf

Essed, P., & Trienekens, S. (2008). 'Who wants to feel white?' Race, Dutch culture and contested identities. *Ethnic and Racial Studies*, *31*(1), 52–72. https://doi.org/10.1080/01419870701538885

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fiske, S. T. (1988). Stereotyping, prejudice, and discrimination. In D. T. Gilbert (Ed.), *The handbook of social psychology, Vols* (pp. 357–411). McGraw-Hill, x. https://psycnet.apa.org/fulltext/1998-07091-025.pdf

Foerde, K., & Shohamy, D. (2011). The role of the basal ganglia in learning and memory: insight from Parkinson's disease. *Neurobiology of Learning and Memory*, *96*(4), 624–636. https://doi.org/10.1016/j.nlm.2011.08.006

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Gaertner, S. L., & Dovidio, J. F. (2000). The aversive form of racism. *Stereotypes and Prejudice: Essential Readings.*, *490*, 289–304. https://psycnet.apa.org/fulltext/2000-16592-016.pdf

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24*, 92-97.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-dependent learning in social interaction: Trait impressions support flexible social choices. *Journal of Personality and Social Psychology*, *123*(4), 655–675. https://doi.org/10.1037/pspa0000296

Hagendoorn, L. (2017). Stereotypes of ethnic minorities in the Netherlands. *Ethnic Minorities and Inter-Ethnic Relations In*. https://doi.org/10.4324/9781315209999-2/stereotypes-ethnic-minorities-netherlands-louk-hagendoorn

Hagendoorn, L., & Pepels, J. (2017). Why the Dutch maintain more social distance from some ethnic minorities than others: A model explaining the ethnic hierarchy. In W. Vollebergh, J. Veenman, & L. Hagendoorn (Eds.), *Integrating Immigrants in the Netherlands* (pp. 57–78). https://doi.org/10.4324/9781315197012-11/dutch-maintain-social-distance-ethnic-minorities-others-model-explaining-ethnic-hierarchy-louk-hagendoorn-josã©-pepels

Hehman, E., Gaertner, S. L., Dovidio, J. F., Mania, E. W., Guerra, R., Wilson, D. C., & Friel, B. M. (2012). Group status drives majority and minority integration preferences. *Psychological Science*, *23*(1), 46–52. https://doi.org/10.1177/0956797611423547

Jargon, M., & Thijs, J. (2021). Antiprejudice norms and ethnic attitudes in preadolescents: A matter of stimulating the "right reasons". *Group Processes & Intergroup Relations, 24*, 468-487.

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Chapter One - Intergroup Perception and Cognition: An Integrative Framework for Understanding the Causes and Consequences of Social Categorization. In J. M. Olson (Ed.), *Advances in Experimental Social Psychology* (Vol. 55, pp. 1–80). Academic Press. https://doi.org/10.1016/bs.aesp.2016.10.001

Kleider-Offutt, H. M., Bond, A. D., & Hegerty, S. E. A. (2017). Black Stereotypical Features: When a Face Type Can Get You in Trouble. *Current Directions in Psychological Science*, *26*(1), 28–33. https://doi.org/10.1177/0963721416667916

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402. https://doi.org/10.1126/science.273.5280.1399

Kramer, S., Hackett, C., & Beveridge, K. (2022). Modeling the future of religion in America. *Pew Research Center*. https://adam2.org/headlines/support_docs/US-Religious-

Projections_FOR-PRODUCTION-9.13.22.pdf

Lamkaddem, M., Essink-Bot, M.-L., Devillé, W., Foets, M., & Stronks, K. (2012). Perceived discrimination outside health care settings and health care utilization of Turkish and Moroccan GP patients in the Netherlands. *European Journal of Public Health*, *22*(4), 473–478. https://doi.org/10.1093/eurpub/ckr113

Lockwood, P. L., & Klein-Flügge, M. C. (2021). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience, 16*, 761-771.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*. https://www.sciencedirect.com/science/article/pii/S0022103100914707

Mok, I., & Mok, R. J. M. (1999). *In de ban van het ras: Aardrijkskunde tussen wetenschap en samenleving 1876-1992*. ASCA Press. https://play.google.com/store/books/details?id=HdjXAAAACAAJ

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*, 452-454.

Perkins, K. M., Toskos Dils, A., & Flusberg, S. J. (2022). The perceived threat of demographic shifts depends on how you think the economy works. *Group Processes &*

Intergroup Relations, 25, 227-246.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Plant, E. A., Devine, P. G., & Brazy, P. C. (2003). The Bogus Pipeline and Motivations to Respond without Prejudice: Revisiting the Fading and Faking of Racial Prejudice. *Group Processes & Intergroup Relations: GPIR*, *6*(2), 187–200. https://doi.org/10.1177/1368430203006002004

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe. *American Sociological Review*, *60*(4), 586–611. https://doi.org/10.2307/2096296

R Core Team (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raissi, A., & Steele, J. R. (2021). Does Emotional Expression Moderate Implicit Racial Bias? Examining Bias Following Smiling and Angry Primes. *Social Cognition*, *39*(5), 570–590. https://doi.org/10.1521/soco.2021.39.5.570

Ramos, M., Thijssen, L., & Coenders, M. (2021). Labour market discrimination against Moroccan minorities in the Netherlands and Spain: a cross-national and cross-regional comparison. *Journal of Ethnic and Migration Studies*, *47*(6), 1261–1284. https://doi.org/10.1080/1369183X.2019.1622824

Richeson, J. A., & Sommers, S. R. (2016). Toward a social psychology of race and race relations for the twenty-first century. *Annual Review of Psychology, 67*, 439-463.

Roggeband, C., & van der Haar, M. (2018). “Moroccan youngsters”: category politics in the Netherlands. *International Migration* , *56*(4), 79–95. https://doi.org/10.1111/imig.12419

Rottier, B., Ripmeester, N., & Bush, A. (2011). Separated by a common translation? How the

British and the Dutch communicate. *Pediatric Pulmonology, 46*, 409-411.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478680#page=6

Schlueter, E., & Scheepers, P. (2010). The relationship between outgroup size and anti-outgroup attitudes: A theoretical synthesis and empirical test of group threat- and intergroup contact theory. *Social Science Research*, *39*(2), 285–295. https://doi.org/10.1016/j.ssresearch.2009.07.006

Schultner, D.T., Lindström, B.R., Cikara, M., & Amodio, D.M. (2024). *Transmission of social bias through observational learning*. In *PsyArXiv*. https://doi.org/10.31234/osf.io/7ec3u

Shelton, J. N., & Richeson, J. A. (2006). Interracial Interactions: A Relational Approach. In *Advances in Experimental Social Psychology* (Vol. 38, pp. 121–181). Academic Press. https://doi.org/10.1016/S0065-2601(06)38003-3

Srull, T. K., & Wyer, R. S. (1988). *A Dual Process Model of Impression Formation*. L. Erlbaum Associates. https://play.google.com/store/books/details?id=JkwtAQAAMAAJ

Stephan, W. G., Boniecki, K. A., Ybarra, O., Bettencourt, A., Ervin, K. S., Jackson, L. A., McNatt, P. S., & Renfro, C. L. (2002). The Role of Threats in the Racial Attitudes of Blacks and Whites. *Personality & Social Psychology Bulletin*, *28*(9), 1242–1254. https://doi.org/10.1177/01461672022812009

Stillerman, B., Lindström, B., Schultner, D., Hackel, L. M., Hagen, D., Jostmann, N., & Amodio, D. (2022). Societal stereotypes shape learning to produce group-based preferences. In *PsyArXiv*. https://doi.org/10.31234/osf.io/mwztc

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*,

9(5), 1054–1054. https://doi.org/10.1109/tnn.1998.712192

Thomas, A., Doyle, A., & Vaughn, D. (2007). Implementation of a computer based implicit association test as a measure of attitudes toward individuals with disabilities. *Journal of Rehabilitation*, *73*, 3. https://search.proquest.com/openview/f235a2cebb3b07565a944d24a12d65d1/1?pq-origsite=gscholar&cbl=37110

Traast, I. J., Schultner, D. T., Doosje, B., & Amodio, D. M. (2024). Race effects on impression formation in social interaction: An instrumental learning account. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001523

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, *11*(4), 907–920. https://doi.org/10.1037/a0023853

U.S. Census Bureau (2023). QuickFacts. Population Estimates. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045222#PST045222

Van Ginkel, R. (1996). Foreigners' Views of the Dutch: Past and Present. *Dutch Crossing*, *20*(1), 117–131. https://doi.org/10.1080/03096564.1996.11784059

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental Psychology*, *58*(4), 271–277. https://doi.org/10.1027/1618-3169/a000093

Vaughn, E. D., Thomas, A., & Doyle, A. L. (2011). The Multiple Disability Implicit Association Test: Psychometric Analysis of a Multiple Administration IAT Measure. *Rehabilitation Counseling Bulletin*, *54*(4), 223–235. https://doi.org/10.1177/0034355211403008

Velasco González, K., Verkuyten, M., Weesie, J., & Poppe, E. (2008). Prejudice towards Muslims in The Netherlands: testing integrated threat theory. *The British Journal of*

Social Psychology / the British Psychological Society, *47*(Pt 4), 667–685. https://doi.org/10.1348/014466608X284443

Verkuyten, M. (2002). Perceptions of ethnic discrimination by minority and majority early adolescents in the Netherlands. *International Journal of Psychology: Journal International de Psychologie*, *37*(6), 321–332. https://doi.org/10.1080/00207590244000142

Verkuyten, M., & Thijs, J. (2002). Racist victimization among children in The Netherlands: the effect of ethnic group and school. *Ethnic and Racial Studies*, *25*(2), 310–331. https://doi.org/10.1080/01419870120109502

Verkuyten, M., & Thijs, J. (2010). Ethnic Minority Labeling, Multiculturalism, and the Attitude of Majority Group Members. *Journal of Language and Social Psychology*, *29*(4), 467–477. https://doi.org/10.1177/0261927X10377992

Verkuyten, M., & Zaremba, K. (2005). Interethnic Relations in a Changing Political Context. *Social Psychology Quarterly*, *68*(4), 375–386. https://doi.org/10.1177/019027250506800405

Wingfield, A. H., & Chavez, K. (2020). Getting In, Getting Hired, Getting Sideways Looks: Organizational Hierarchy and Perceptions of Racial Discrimination. *American Sociological Review*, *85*(1), 31–57. https://doi.org/10.1177/0003122419894335

Wirtz, C., van der Pligt, J., & Doosje, B. (2016). Negative attitudes toward Muslims in The Netherlands: The role of symbolic threat, stereotypes, and moral emotions. *Peace and Conflict: Journal of Peace Psychology: The Journal of the Division of Peace Psychology of the American Psychological Association*, *22*(1), 75–83. https://doi.org/10.1037/pac0000126