State and Trait Empathy are Positively Correlated, Albeit Weakly, with Prosocial Learning

Thomas G. McCauley

University of California, San Diego

Michael E. McCullough

University of California, San Diego

Author note: Data and materials associated with this project have been uploaded to the Open Science Framework, and can be accessed at https://osf.io/45mjg/?view_only=81071e337d37443b967573eaf28938a4.

This project/publication was made possible through the support of a grant from the John Templeton Foundation Grant 61539. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. This article was made possible through the support of Grant W191NF2010259 from the Army Research Institute. The views, opinions, and/or findings contained in this report (article) are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Email: tmccaule@ucsd.edu

Abstract

Empathic concern promotes prosocial behavior, but the effect of empathy is not the same for everyone. Scholars have extensively mapped the situational and personality factors that modulate the empathy-helping relationship, but only a few studies have examined the cognitive processes that link empathy with helping. Lockwood et al. (2016) found that individual differences in online simulation (the deliberate intention to imagine what another person is feeling) were associated with individual differences in the rates at which people learned how to benefit an anonymous other person in a two-armed bandit task (which we term prosocial learning). Across three experiments (N = 1,188; data collected between April 2021 and March 2024), we critically tested and extended this finding by asking participants to learn on behalf of real people in need, generalizing existing methods of prosocial learning to online research environments, conducted with participants from diverse ethnic backgrounds. Although each individual experiment provided only mixed evidence for the hypothesis that empathic concern is associated with prosocial learning, a mega-analysis that included data from all three experiments revealed that both state and trait empathic concern were weakly (but significantly) positively associated with prosocial learning. Although manipulations of empathy were unsuccessful, and had no effect upon prosocial learning, attempts to causally manipulate empathy were also ineffective, leaving the causal association between empathy and prosocial learning unclear. We also could not replicate the finding that online simulation was associated with learning.

Public Significance Statement

Across three experiments, we found limited evidence linking empathy to prosocial learning. Although previous research suggested that empathy enhances peoples' motivation to learn to earn rewards on behalf of others, our results showed only small associations, though they were statistically significant. Attempts to manipulate empathy and social evaluation did not significantly impact prosocial learning. This nuanced outcome suggests that empathy's effect on learning mechanisms in helping contexts may be weaker than previously thought.

Introduction

Recently, researchers have begun to investigate the association between empathy and learning. Here, we critically investigate and extend the empirical link between empathy and prosocial learning by focusing on how empathic concern might shape learning for a variety of social targets, including strangers in need of help.

Background

Why do people help needy strangers? Specifically, what are the cognitive mechanisms that govern people's decisions about whether to help others? The cognitive foundations of need-based helping have been investigated for generations. One of the most popular and enduring hypotheses, based on the fact that need-based helping is typically preceded by feelings of empathy for the beneficiary (Krebs, 1975) is the empathy-altruism hypothesis, which states that empathic concern for a needy target dependably elicits altruistic motivation – that is, the motivation to benefit another person as an end unto themselves (Batson, 1981). A large body of research supports the hypothesis empathy promotes help for people experiencing some kind of suffering (Batson, 2011). However, most of this research has focused on outcomes related to behaviors, but changes in prosocial motivational ought to also affect information-processing mechanisms that underwrite empathy's effect upon altruistically motivated helping behavior.

A handful of studies have investigated how cognitive processes such as attention, effort, and learning facilitate empathy and helping (for a review, see Tusche and Bas, 2021). The bulk of this research suggests that the cognitive processes underlying empathy, and the desire to help more broadly, are governed by a cost-benefit logic that calculates the value of helping others based on the benefits that the helped person would receive, and the costs the helper would need to bear in order to do so (Cameron et al., 2020; Depow et al., 2022; Evans and Rand, 2019; Hutcherson,

Bushong, & Rangel; 2015; Imas, 2015; Teoh & Hutcherson, 2022). These results dovetail with research suggesting that people empathize with and help others when the perceived benefits of helping outweigh the costs (Delton et al., 2023; Sznycer et al., 2019).

One influential study examined the association between trait empathy and people's propensity to learn how to earn rewards for other people. Lockwood and colleagues (2016) had participants complete a two-armed bandit task (Robbins, 1952) in which they learned which of two abstract symbols carries a higher probability of reward. In contrast to a typical bandit task where participants learn to earn rewards for themselves, participants in Lockwood et al.'s (2016) experiment also learned on behalf of an anonymous other person; we refer to learning in a bandit task for non-self targets as *prosocial learning*. Lockwood et al. (2016) fit participants' data to a reinforcement learning model in order to estimate the rate at which participants learned to benefit the other person, which here we call the *prosocial learning rate*. The logic of this modeling approach is that, when learning on behalf of another person, people who are more motivated to earn rewards for the other person ought to be better at tracking which symbol is more likely to return a reward, so that a higher learning rate reflects a greater desire to benefit the other person, and a lower learning rate reflects less desire to do so.

One benefit of studying prosocial motivation in this fashion is that it reduces or even eliminates many non-altruistic accounts of the individual differences that we observe. For example, the technical process by which we estimate prosocial learning rates–and probably even the idea that it is possible to use their data to estimate their prosocial learning rates at all–is surely opaque to most participants, which means that their learning rates would be less strongly influenced by non-altruistic social motivations such as the desire to avoid guilt, the desire to please the experimenters, or the desire to win the favor of the beneficiaries of their correct choices. (With

respect to the latter, the beneficiaries of their help could never even know that our participants had the ability to help them in the first place.) Furthermore, statistically removing the variation from people's prosocial learning rates that can be explained by individual learning rates allows us to control for any factor that is responsible for both, including basic cognitive processes (e.g., attention, working memory) and non-social motivations (e.g., the desire to perform well). Because these latter motivations are some of the major non-altruistic motivations for prosocial behavior (Batson, 2011), prosocial learning rates represent a promising way to measure individual differences in altruistic motivation and not just prosocial motivation in general.

Lockwood et al's (2016) key finding was that participants who scored higher on a self-report measure of online simulation–a component of trait empathy that measures the deliberate intention to imagine what other people are feeling–were more adept at learning to earn rewards on behalf of the anonymous person, relative to participants who scored lower on online simulation. Importantly, the relationship between online simulation and prosocial learning rates obtained even after controlling for subjects' general learning ability (measured as performance when learning to earn rewards for themselves), suggesting that people who are more empathic are better at flexibly learning the most effective way to help others (see also Berger, 1962; Kwak, Pearson, & Huettel, 2014; Lockwood et al., 2017; Weiss et al., 1971; Westhoff et al., 2021).

Lockwood et al's (2016) experiment, and evidence for prosocial learning in general, provides evidence for a motivational basis for learning to help others. Since the only difference between learning to earn rewards for the self versus another person, differences in learning performance can be attributed to differences in motivation. More generally, the motivation to help others shapes socio-cognitive processes such as learning, effort, and attention (Contreras-Huerta, Pisauro, & Apps, 2020). Moreover, performance on two-armed forced choice tasks predicts fluid

general intelligence, and learning about the social behaviors of others, suggesting that these tasks measure real-world learning behaviors (Vostroknutov, Polonio, & Coricelli, 2018).

Although Lockwood et al.'s (2016) results provide compelling evidence for the role of empathy in prosocial learning and motivation, they leave at least four concerns unaddressed. First, Lockwood et al. (2016) found that the online simulation component of empathy was significantly correlated with prosocial learning. Online simulation is thought to measure the deliberate intention to imagine what other people are feeling in an effort to change one's own behaviors that might affect someone else (Reniers et al., 2011). However, the term "empathy" can refer to many different phenomena involving cognition and emotion (Cuff et al., 2014), and the desire to improve another person's welfare through helping is hypothesized to be motivated by empathic concern, which captures spontaneous feelings of other-oriented concern for people in need of help (Batson, 2009), not merely online simulation. Although other emotions that are conceptually similar to empathic concern can promote prosocial behavior – like compassion (Goetz et al., 2010) or empathic joy (Depow et al., 2021) – empathic concern is thought to be the focal emotion that causes need-based helping (Batson, 2011). Since Lockwood et al. (2016) did not measure empathic concern per se, it remains unclear whether empathic concern motivates reward-learning on behalf of others. For instance, someone motivated by online simulation might engage in prosocial learning not because they wish to improve the welfare of the target, but because they think the learning for the target might be instrumental for achieving their own personal goals, such as appearing like a good cooperative partner (Barclay, 2007). A few other studies have tested the association between prosocial learning and empathy, with mixed findings (Kwak et al., 2014; Westerhoff et al., 2021). But in Kwak et al. (2014), the motivations underlying prosocial learning were confounded with motivations to earn rewards for oneself, and Westerhoff et al. (2021)

conducted their study with adolescents, rather than adults. Thus, it is unclear exactly how we should take Kwak et al.'s (2014) and Westerhoff et al.'s (2021) findings into account when considering the relationship between empathic concern and prosocial learning.

Second, Lockwood et al. (2016) tested the association between empathy and prosocial learning by having participants learn on behalf of an anonymous target. However, empathic concern is hypothesized to respond to cues that a target is in a state of need or suffering – cues that the anonymous targets in Lockwood et al. (2016) lacked. The psychological faculties responsible for processing information share a lock-and-key relationship with the features of the world that arouse those faculties: For example, olfactory receptors are sensitive to smells, and photoreceptors are sensitive to photons (Barrett, 2005). By this same logic, empathic concern should be activated in response to cues that another person is suffering. A valid test of the relationship between empathy and prosocial learning, then, should feature targets who are in a state of need, as it is targets who express suffering that arouse feelings of empathic concern, rather than targets that express other kinds of emotions (Sassenrath, Pfattheicher, & Keller, 2017). As in Lockwood et al. (2016), the suffering targets in our study were also strangers to the participants, because participants might have non-altruistic motives for helping non-anonymous targets, like earning reputational benefits, inducing reciprocity, or earning social rewards (Batson, 2011).

Third, with few exceptions (e.g., Wright, Shaw, & Jones, 1990), previous studies examining the association between empathy and prosocial learning have relied upon data from cross-sectional research designs (Kwak, Pearson, & Huettel, 2014; Lockwood et al., 2016), limiting the causal inferences that can be drawn from the data. Although causal inferences can on occasion be drawn from correlational data, such as in the context of direct acyclic graphs (Grosz, Rohrer, & Thoemmes, 2020), bivariate correlations can rarely be interpreted with causal precision

(MacKinnon, Krull, & Lockwood, 2000). For instance, the association between empathy and learning may occur because empathy causes increases in learning, because learning causes increases in empathy, or because some unmeasured third variable causes both empathy and learning.

Fourth, the situations that are hypothesized to elicit feelings of empathy are often situations in which one is incentivized to appear empathic. In such situations, observed empathy could be caused by concerns about impression management (Cialdini et al., 1987). It remains unclear, then, whether social evaluation shapes prosocial learning processes that are otherwise attributed to empathy. Although there is little evidence that feelings of empathy redound entirely to concerns about social evaluation (Fultz et al., 1986; McCauley, McAuliffe, & McCullough, 2024), there is at least some evidence that empathy is affected by social evaluation: Scores on all four facets of the IRI are moderately correlated with impression management scores, and people deflate their state empathy scores when they believe that an experimenter will know if they lie about their scores (Sassenrath, 2020). To our knowledge, though, no study has directly examined the effect of social evaluation upon prosocial learning, let alone examining the causal effect of social evaluation.

The present experiments

To address these four issues – measuring empathy via empathic concern instead of online simulation; needy targets, causally manipulating empathy; and manipulating social evaluation – we conducted three experiments to test five hypotheses concerning the relationship between empathy and prosocial learning. The basic experimental procedure in all three studies involved participants who were learning to earn rewards for a variety of targets in a two-armed bandit task.

In Experiment 1, participants learned to earn rewards for four targets: (a) themselves; (b) a non-needy anonymous person; (c) Oxfam (a charitable organization that works to alleviate global poverty and inequality through humanitarian aid); and (d) an identifiable person in need of help selected from a common philanthropy crowdsource site. Our primary focus lay in testing the prosocial learning hypothesis in the context of learning for identifiable people in need of help, but we also included Oxfam as a learning target. No past research has directly examined prosocial learning for groups of needy people (so-called statistical victims) – much less so the organizations that seek to help them – even though the psychological foundations for each of these forms of helping differ in important ways (Small & Loewenstein, 2003). For example, people appear to be motivated by social desirability concerns to a greater extent when giving to anonymous others than to charities (Livingston & Rasulmukhamedov, 2023). We also included the anonymous other as a target to more closely replicate the method described in Lockwood et al. (2016). We then tested the association between prosocial learning for the non-self targets and trait empathic concern as measured by the empathic concern subscale of the Interpersonal Reactivity Index (Davis, 1980) because scores on this measure have been universally linked with costly prosocial behavior (Chopik, O'Brien, & Konrath, 2017; Einolf, 2008; FeldmanHall et al., 2015). We made the following prediction:

***Prediction 1*: Trait empathic concern will be positively associated with prosocial learning for the anonymous other person, the identifiable needy target, and the Oxfam charitable organization.**

Since, our inclusion of Oxfam as a target of learning in Experiment 1 was exploratory, and our inclusion of the anonymous target was to directly replicate the method used by Lockwood et al. (2016), we removed the Oxfam and anonymous targets in Experiments 2 and 3 to reduce participant fatigue. We then modified the experimental design to test four additional hypotheses about the relationship between individual differences in prosocial motivation and prosocial learning.

First, in Experiments 2, and 3, we measured participants' state empathic concern prior to learning on behalf of the target, and not merely trait empathic concern. Past studies have exclusively focused on the association between trait empathic concern and prosocial learning, but trait measures of emotion assay people's semantic memories, which can be contaminated by retrospective biases that distort the accuracy of such reports (Robinson & Clore, 2002). In contrast, state measures of emotion assay episodic memories, and state measures are thought to more reliably tap people's emotional experiences (McCauley et al., 2024, preprint; Robinson & Clore, 2002). Moreover, the empathy-altruism hypothesis turns on the finding that one's empathic concern for a *specific* target predicts motives to improve the welfare of that target. Thus, a direct test of the empathy-altruism hypothesis in the context of prosocial learning necessitates measures in-the-moment feelings of empathy for a needy target, rather than the summary of past empathic experiences as measured by trait empathic concern (Eisenberg & Fabes, 1990; Eisenberg & Miller, 1987). This leads to Prediction 2.

***Prediction 2*: State empathic concern will positively predict prosocial learning for the identifiable needy target.**

Second, in Experiments 2 and 3 we causally manipulated feelings of empathy for the identifiable needy target using perspective-taking instructions (Batson et al., 1997), so that participants were assigned to either an imagine-other condition, in which they were asked to imagine the feelings and experiences described by the identifiable needy target, or a remain-objective condition, in which they were asked to objectively evaluate the facts associated with the identifiable needy target's situation. Perspective-taking instructions have been found to have a robust effect upon empathic concern (McAuliffe et al., 2020) by dampening spontaneous feelings of empathic concern via remain-objective instructions. Most studies that have tested the effect of perspective-taking instructions measured empathic concern using the same measure of state empathic concern featured in our studies (i.e., the emotion response questionnaire; see Experiment 1 for more details about this measure). If empathy causes people to become better learners on behalf of needy others, then participants who receive imagine-other instructions ought to be better at learning for the identifiable needy target, compared to those who receive remain-objective instructions. This leads to Prediction 3.

***Prediction 3*: Participants who receive imagine-other instructions will have higher prosocial learning rates than participants who receive remain-objective instructions.**

Third, in Experiments 2 and 3, we manipulated participants' beliefs about whether their prosocial learning performance would be shared with both the experimenter and the target of their reward learning in order to investigate the impact of reputational incentives for helping upon prosocial learning. People often help in order to secure instrumental benefits like social rewards, reputational benefits, or self-enhancement, and such selfish goals can cause people to engage in

substantial prosocial behaviors, and can even lead to morally exemplary behavior (Walker, 2013). Studies testing hypotheses about prosocial learning typically strive to remove reputational effects from their experiments (e.g., Lockwood et al., 2017). To our knowledge, however, no study has directly examined whether concerns about the evaluability of one's moral appearance actually influences prosocial learning. This leads to Prediction 4:

***Prediction 4*: Participants who complete the bandit task under high social evaluation will have higher prosocial learning rates, compared to participants who complete the task under low social evaluation.**

We also included the online simulation subscale from the QCAE in Experiments 2 and 3 in order to directly replicate Lockwood et al.'s (2016) primary finding with a larger sample of participants. Lockwood et al.'s (2016) test of the association between empathy and prosocial learning was based on a small sample size, so it is plausible that the relationship Lockwood et al. (2016) observed between learning and the online simulation component of empathy was a false positive because studies with small sample sizes are more likely to have unstable and inflated effect sizes (Schönbrodt & Perugini, 2013). A larger sample size is necessary to adequately test the hypothesis that empathy of any kind promotes prosocial reward-learning. If online simulation promotes prosocial learning, then we ought to be able to recover the statistical relationship in a much larger sample. This leads to Prediction 5:

***Prediction* 5*:* The online simulation component of empathy will be associated with prosocial learning.**

Finally, we combined the data from Experiments 1, 2, and 3 into a single dataset, and conducted a mega-analysis of the associations that prosocial learning shared with online simulation, trait empathic concern, state empathic concern, the perspective-taking manipulation, and the social evaluation manipulation.

Transparency and Openness

We report how we determined our sample size for each study. Study materials, data, and analysis code for all studies can be found at https://osf.io/45mjg/?view_only=81071e337d37443b967573eaf28938a4. The pre-registration for Experiment 2 can be viewed online at https://osf.io/yp2th?view_only=81071e337d37443b967573eaf28938a4, and the pre-registration for Experiment 3 can be viewed online at https://osf.io/h9npt?view_only=81071e337d37443b967573eaf28938a4. In the Supplemental Materials, we report additional, non-preregistered analyses. We also preregistered a number of hypotheses in Experiments 2 and 3 that we do not report in the current manuscript, as they are part of a broader data collection effort. More information about these hypotheses can be found in the Supplemental Materials.

Data were analyzed using R version 4.0.3 (R Core Team, 2013), and plots were made using the package ggplot2, version 3.5.1 (Wickham, 2011). Computational modeling analyses were conducted using the HDDM.hrl function from version 0.80 of the HDDM toolbox (Wiecki, Sofer, & Frank, 2013). All materials and procedures were approved by the institutional review board the

University of California, San Diego. We report how we determined our sample size, as well as all measures, manipulations, and exclusions in these experiments (additional measures not included in analyses are described in the Supplemental Materials).

Experiment 1

Method

Participants

215 UC San Diego undergraduate students participated in Study 1 (M_{age} = 20.39, SD_{age} = 2.28; five participants failed to provide age data; 0.9% = American Indian/Alaskan Native, 46% = Asian, 0.5% = Native Hawaiian or Other Pacific Islander, 1.9% = Black or African-American, 24.9% = White, 25.8% = Other). Based on Lockwood et al.'s (2016) results, which found a correlation of r = 0.44 (based on n = 31) between online simulation and prosocial learning, we estimated that we would need a minimum of n = 38 to detect effects as large or larger. We aimed to collect data from at least n = 194, which would provide us with adequate power to detect correlations as small as r = 0.20. We oversampled to ensure we would have a sufficient sample size in case of unforeseen reasons participants might need to be excluded, although no participants needed to be removed.

Participants completed the study remotely on Qualtrics via their computers or mobile devices. They received course credit for participating, and were told they could earn up to $2.40, depending on their performance in the task. Data collection took place from April to June 2021.

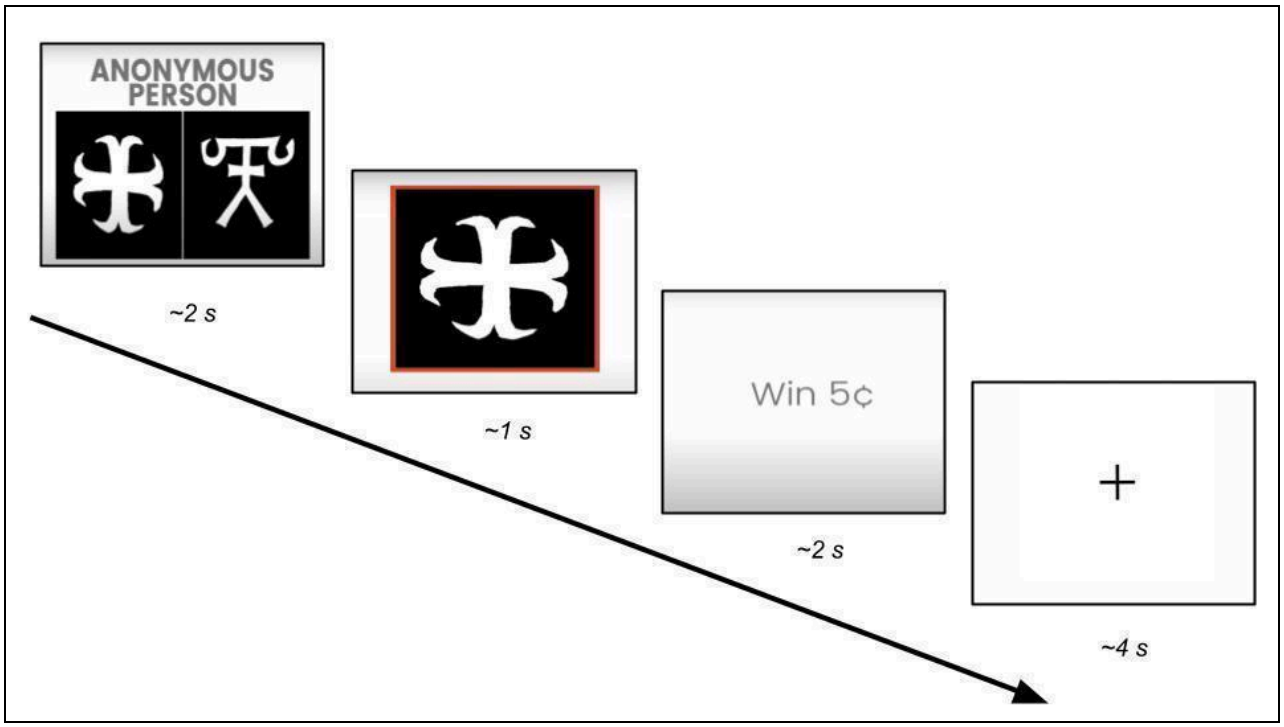
Procedure

Two-armed bandit task. After providing informed consent, participants completed a two-armed bandit task in which they were tasked with learning which of two abstract symbols carried a higher probability of reward over a series of trials (Robbins, 1952). Participants learned

about the value of each symbol by choosing between the two symbols over 16 sequential trials. One of the two symbols returned a reward 75% of the time that it was selected, and the other symbol returned a reward 25% of the time. The symbols were Agamemnon alphabet letters, identical to the symbols used in Lockwood et al. (2016). Participants earned 5¢ for trials in which the symbol returned a reward, and 0¢ otherwise. For trials in which participants failed to select either symbol (i.e., missed trials), they earned 0¢ for the trial and did not receive feedback about the rewardingness of either symbol. The location of the rewarding symbol (i.e., whether it appeared on the left or right of the screen) was randomized. See Figure 1 for a diagram of a trial from the bandit task, and a video showing the task in real-time can be viewed at youtube.com/watch?v=e4ziY_Jmr6M.

Figure 1

A Trial From the Two-Armed Bandit Task From the Participant's Point of View When Learning on Behalf of the Anonymous Target.



After completing a block of practice trials, participants learned to earn rewards on behalf of four different targets: Themselves, an anonymous other person, Oxfam, and an identifiable person seeking help via a common philanthropy website (more information about the Oxfam and identifiable targets is provided in the Supplemental Materials). For each target, participants completed three blocks of 16 trials, for a total of 48 trials per target, and 192 trials overall. The order of the targets was randomized, and participants completed all three blocks for each target in a sequential set. We randomized the order of the blocks for each target within each set. For all targets, we informed participants that any money they earned on behalf of a target would actually be given to that target.

For all three non-self targets, we told participants that the targets wouldn't know how much money the participant earned for them until the end of the experiment (at which point we would make an anonymous donation on the behalf of all participants in the experiment), and that the target was unaware that the participant was performing a task where participants could win money

on their behalf. Before learning for Oxfam, we provided participants with Oxfam's mission statement and a link to Oxfam.org, and we told them that "Donations to Oxfam helps to save lives during a disaster, get clean water running in the most remote areas, send children, especially girls to school and stand up for the rights of women." (See Supplemental Materials for the full text of the mission statement that participants were provided.) Likewise, for the blocks in which participants learned on behalf of an identifiable needy target, we provided participants with a picture of the target and details about the target's reason for seeking help, along with the target's background, goals, and anything else they wished to share. Participants were also provided with a link to the target's web page on the philanthropy site so that they could verify that the target was indeed a real person[1]. All participants saw the same identifiable needy target. When learning for the anonymous person, we told participants that they would play for the anonymous person.

After the experiment ended, we donated a total of $268.80 to Oxfam, and a total of $278.25 to the identifiable needy targets. We paid participants a flat fee of $2.40, reflecting the maximum amount of money that they could possibly have earned for themselves in the 48 trials in which they learned for themselves. To satisfy the anonymous payment, we paid each participant an additional $2.40, so that participants had effectively earned money for each other in the anonymous target condition (although they were unaware that this was the case during the study).

Interpersonal Reactivity Index. After completing the bandit task, participants completed the empathic concern subscale of the Interpersonal Reactivity Index (IRI; Davis, 1983). The IRI's empathic concern subscale includes seven items, each of which was measured on a scale from A (does not describe me well) to E (describes me very well): "Sometimes I don't feel very sorry for other people when they are having problems" (reverse-scored); "I often have tender, concerned

[1]We do not provide identifying information about the identifiable needy targets, or the philanthropy site, in order to maximally protect the privacy of those needy individuals.

feelings for people less fortunate than me"; When I see someone being taken advantage of, I feel kind of protective towards them"; "I am often quite touched by things that I see happen"; "Other people's misfortunes do not usually disturb me a great deal" (reverse-scored); "I would describe myself as a pretty soft-hearted person"; and "When I see someone being treated unfairly, I sometimes don't feel very much pity for them." (reverse-scored). We formed a composite by adding the seven items together, and dividing the sum by seven ($M = 2.73$, $SD = 0.71$; *McDonald's* $\Omega = 0.88$).

Participants also completed 23 additional self-report measures as part of a broader data collection effort. We administered the measures (including the IRI) in randomized order, and also randomized the items within each measure. See the Supplemental Materials for a list of all the measures included in the study.

Debriefing and dismissal. Finally, participants reported their demographic information, and completed a funnel questionnaire where they answered open-ended questions about the experiment to determine whether they were aware of the experimental hypothesis, and if they were suspicious about the veracity of the experiment. Although our experiment did not involve suspicion, we thought participants might still be skeptical about aspects of the experimental procedure (e.g., whether the targets were real, or if they would actually earn money for their performance), which could influence their responses.

Data Analysis

Prior to computational model fitting, we removed trials in which participants failed to provide a response. Participants missed 6.93% of all possible trials. We planned to remove participants that missed more than half of their trials, but none met this criteria, so all participants were included in model estimation. Computational models were estimated in Python using the

HDDM.hrl function from version 0.80 of the HDDM toolbox (Wiecki, Sofer, & Frank, 2013), which implements hierarchical Bayesian estimation methods to simultaneously estimation both participant and group parameters, thereby providing a more accurate estimate of model parameters than non-Bayesian approaches to model estimation (Daw, 2011). The variables included for the computational model estimation were as follows: The trial number (an integer ranging from 1-16); the participants' response (a dichotomous variable that encodes whether participants selected the symbol that rewarded 25% of selections, or the symbol that rewarded 75% of selections); the feedback for the participant's selection (a dichotomous variable that encodes whether the participant's selection was rewarded); and the target of learning (a categorical variable from 1 to 4). Output from the computational models, and their association with self-report measures, were analyzed in R (R Core Team, 2013). All data, analysis code, and research materials are available at https://osf.io/45mjg/?view_only=81071e337d37443b967573eaf28938a4.

Computational model selection

Next, we fit the data from the bandit task to competing hierarchical Bayesian reinforcement learning models (Pedersen & Frank, 2020). All three models had the same basic form, which can be represented as two equations: The first equation reflects the trial-by-trial updating (i.e., the degree to which participants learn the value of each symbol), and the second equation reflects the probability that a participant will choose an action x on trial t (i.e., a softmax link function). The first equation for trial-by-trial updating is as follows:

$$Q_{t+1}(x) = Q_t(x) + (\alpha * \delta_t)$$

In this equation, expectations of future reward for action x, represented as $Q_{t+1}(x)$, is a function of current expectations for the value of x, represented as $Q_t(x)$, and the discrepancy between the expected and actual value of x (i.e., the prediction error), which is represented as δ_t.

The extent to which the prediction error updates $Q_t(x)$ is scaled by α, the learning rate. Larger values of α indicates greater updating in response to prediction errors and thus, more learning. In this equation, α is the only free parameter that needs to be estimated.

The second equation for the softmax link function is as follows:

$$p_t(x|Q_t(x)) = \frac{e^{(Q_t(x)/\beta)}}{\Sigma_{x'} e^{(Q_t(x')/\beta)}}$$

In this equation, the softmax link function estimates the probability that a participant will select action *x* versus *x'*, weighted by the ratio of expected values for *x* and *x'*. The ratio between the expected values is influenced by the inverse temperature parameter, represented as β. Larger values of β indicate more random selections of *x* versus *x'*, while smaller values of β indicate more consistent selections. In the softmax link equation, β is the only free parameter that needs to be estimated.

We estimated and compared three competing reinforcement learning models, where the three models varied with respect to the number of α and β parameters that were estimated. First, we estimated a model in which a single α parameter and a single β parameter were estimated across conditions, so that there were a total of two parameters estimated per participant. Second, we estimated a model in which separate α parameters were estimated for each of the four conditions, so that there were unique α parameters for learning about the anonymous person, the participants themselves, Oxfam, and the identifiable needy target; and a single β parameter was estimated across the four conditions, for a total of five parameters estimated per participant. Finally, we estimated a model in which both separate α and β parameters were estimated for each of the four conditions, for a total of eight parameters estimated per participant. We considered these three particular models based on the model-fitting procedures described in past studies (Lockwood et al., 2016).

For each of these three models, we conducted 10,000 iterations of the Markov chain Monte Carlo (MCMC) slice sampler, which samples from the posterior distribution of plausible values for the parameters of interest (Neal, 2003), discarded the first 5,000 iterations as burn-ins, allowing the chain to move from its arbitrary starting values, and retained every fifth sample from the remaining 5,000 iterations, a process known as thinning. We selected large values for sampling, burn-ins, and thinning, because longer chains help to ensure model convergence (Pedersen & Frank, 2020). In total, then, the posterior distribution for each parameter value reflects samples culled from 1,000 iterations.

To determine whether the models successfully converged, we visually inspected the trace plots, autocorrelation plots, and histograms of the posterior distributions (Pedersen & Frank, 2020). The trace plots showed that the mixing of the chains for each parameter were stationary with a small amount of variance; the autocorrelations between samples varied around zero; and the histograms for the participant and group means distributions were all approximately normally distributed. Each of these results are indicative of successful model convergence (Pedersen & Frank, 2020), indicating that the parameters from our model could be interpreted. (See Figures S1-S3 in the Supplemental Materials for the model convergence plots for all three models).

Next, we adjudicated among the models by comparing their Deviance Information Criterion (DIC) fit statistics, selecting the model with the small DIC value (Spiegelhalter et al., 1998); we used DIC to adjudicate between models because DIC is the only fit statistic available in hddm, but note that DIC outperforms competing model fit statistics used to compared repeated-measures data (Du et al., 2024). The model in which separate α and β parameters were estimated for each of the four conditions had the smallest DIC statistic (DIC = 43,801.27), compared to the model in which separate α parameters and a single β were estimated (DIC =

44,209.66), and the model in which only a single α and a single β parameter were estimated (*DIC* = 45,190.66). We retained the model with separate α and β parameters for further analysis.

We also conducted a parameter recovery study and posterior predictive check for further model validation. Figure 2 shows the trial-by-trial responses for each of the four conditions, for the observed and simulated datasets. A visual inspection of the results shows that the simulated data generally tracked the observed data, with the simulated results slightly underestimating the frequency with which participants selected the more frequently rewarding symbol, compared to the observed results. Further details about parameter recovery and posterior predictive check analyses, including the results for the α and a single β parameters, are reported in the Supplemental Materials.

Figure 2

Trial-By-Trial Response Data for the Observed and Simulated Datasets in Experiment 1.

Results

Although α in reinforcement learning models is typically bounded between 0 and 1, HDDM internally transforms α so that it takes on unbounded values in order to improve sampling

during model estimation. In order to transform α to the more interpretable 0 to 1 range, we applied the inverse logit to each of the α parameters, $e^{\alpha}/(1 + e^{\alpha})$. We report results for these transformed α values.

Did participants learn the differential value of the symbols?

Prior to analyzing the computational model data, we first sought to determine whether participants learned over the course of the experiment by fitting participants' selections on each trial (0 = symbol that rewarded 25% of selections, 1 = symbol that rewarded 75% of selections) to a generalized linear mixed model (GLMM), so that trials were nested within participants. Predictors included the trial repetition number (a level-1 predictor, with values ranging from 1-16), and *k*-1 dummy-coded predictors that reflected the target whom participants learned for, with the anonymous target serving as the reference group (a level-2 predictor). We also included random intercepts for each participant. Trial repetition number significantly predicted the selection that participants made ($b = 0.042$, $SE = 0.002$, *95% CI* = [0.037, 0.047], $Z = 16.82$, $p < .001$; *Odds ratio (OR)* = 1.043, *95% CI* = [1.038, 1.048]): The odds ratio indicates that each trial increased subjects' odds of making the income-maximizing choice by 4.3%. Each of the three dummy coded variables reflecting the target of learning were also significant predictors (ps < .01). These results indicate that participants learned to select the more rewarding symbol over the course of each learning block. We also estimated a second GLMM that was identical to the first model, except that the dependent variable was the rewardingness of the selection for each trial, regardless of whether participants selected the frequently or infrequently rewarding symbol. The results were qualitatively identical (see Supplemental Materials for more information).

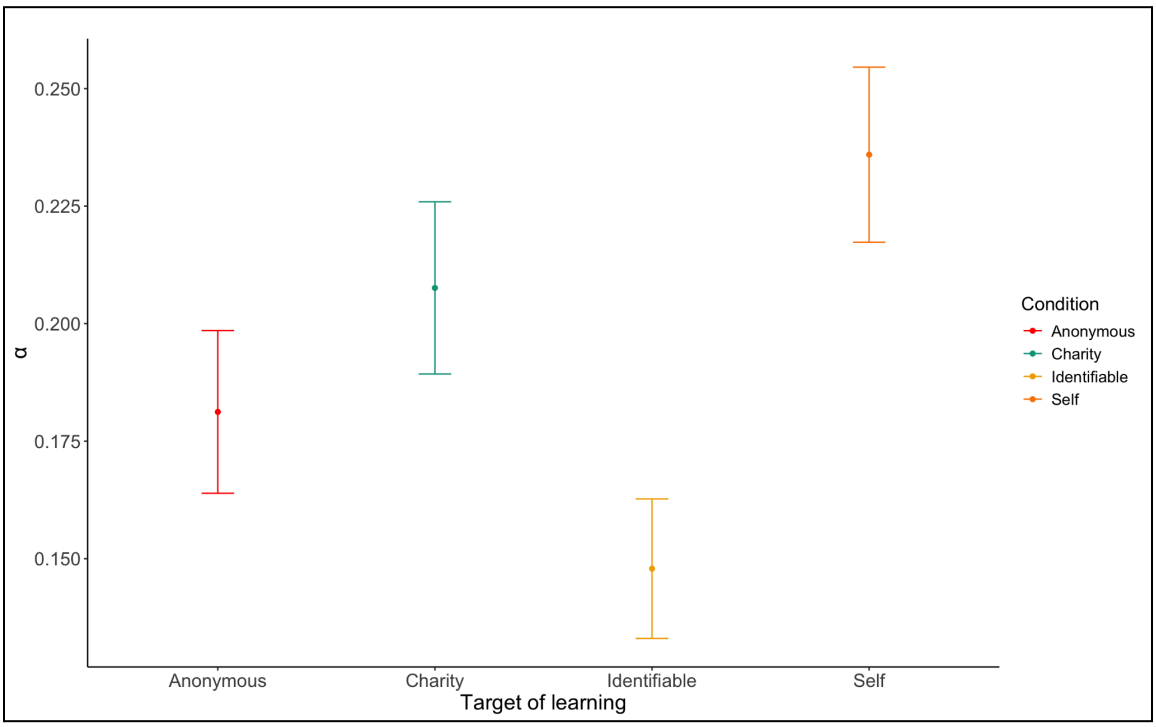
Do people differentially learn to earn reward for themselves vs. non-self targets?

We conducted a one-way within-subjects ANOVA that included α as the dependent variable, and the target of learning (i.e., anonymous, Oxfam, identifiable needy, and self) as the independent variable. There was a significant main effect for condition assignment ($F(3, 630) = 8.11$, $p < .01$, $\eta_p^2 = 0.037$, *95% CI* = [0.011, 0.067]), indicating that participants learned how to help the four different types of beneficiaries at different rates. Follow-up pairwise *t*-tests with Bonferroni-corrected *p*-values indicated that there were significant differences in α values for the self ($M = 0.24$, $SD = 0.27$) and the identifiable needy target ($M = 0.15$, $SD = 0.22$) conditions ($t(211) = 4.67$, $p < .01$, *Cohen's D* = 0.32, *95% CI* = [0.19, 0.46]); the self and anonymous ($M = 0.18$, $SD = 0.25$) conditions ($t(211) = 2.67$, $p = .05$, *Cohen's D* = 0.18, *95% CI* = [0.05, 0.34]); and the Oxfam ($M = 0.21$, $SD = 0.27$) and identifiable needy target conditions ($t(211) = 3.58$, $p < .01$, *Cohen's D* = 0.24, *95% CI* = [0.11, 0.38]; Figure 3). In general, people learned how to make money for themselves most quickly, and how to make money for the identifiable needy targets least quickly, with middling α values for Oxfam and for anonymous participants.

We also conducted additional analyses examining differences in β for self vs. non-self targets, and correlations between the α and β parameters. All parameters were significantly correlated with each other ($0.25 < r\text{s} < 0.51$; $p\text{s} < .02$), save for the correlation between the β and α parameters for the identifiable needy target ($r(212) = 0.11$, $p = .109$). (See the Supplemental Materials for more details about these analyses.)

Figure 3

Means and standard errors for the αs in each condition.



Note: The y-axis is truncated so that readers can more clearly inspect the differences (or lack thereof) between conditions.

Prediction 1: Does trait empathic concern predict prosocial learning?

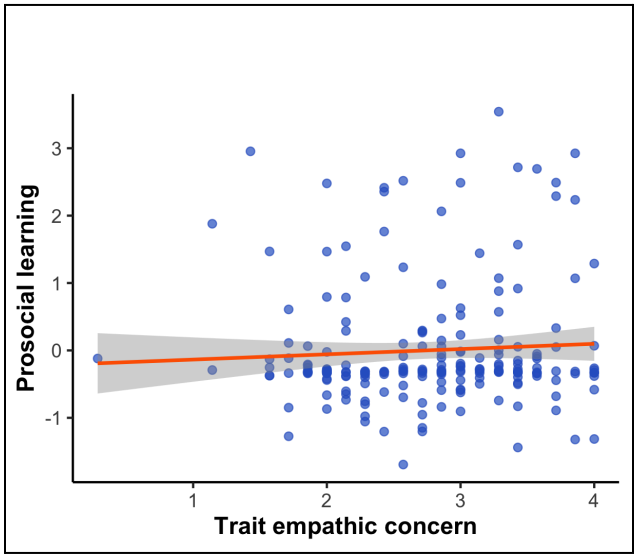
Next, we examined whether trait empathic concern, as measured by scores on the empathic concern subscale of the IRI, predicted prosocial learning. In order to control for the influence of general learning ability upon prosocial learning (as in Lockwood et al., 2016), we regressed each of the prosocial α parameters for the anonymous, Oxfam, and identifiable needy targets on α for the self in three separate regression equations, and saved the resultant residuals from each regression[2]. The residualized α values for each target (which we term $\alpha_{Anonymous}$, α_{Oxfam}, and $\alpha_{Identifiable}$) then served as the dependent variable in the remainder of analyses.

[2]The residual-saving approach described here produces results that are statistically equivalent to regressing the dependent variable on the α parameters for the self, controlling for α parameters representing the non-self target of interest.

Contrary to our predictions, we found that, controlling for α_{self}, trait empathic concern was uncorrelated with $\alpha_{Anonymous}$ ($r(210) = -0.06$, $p = .406$, *95% CI* = [-0.19, 0.08]), α_{Oxfam} ($r(211) = 0.08$, $p = .222$, *95% CI* = [-0.05, 0.22]), or $\alpha_{Identifiable}$ ($r(210) = 0.06$, $p = .381$, *95% CI* = [-0.07, 0.19]; Figure 4). Trait empathic concern was also uncorrelated with α_{self} ($r(211) = 0.04$, $p = .549$, *95% CI* = [-0.09, 0.17]). To better understand the strength of these non-significant effects, we report the Bayes Factors (Kass & Raftery, 1995) for each statistical result in the Supplemental Materials.

Figure 4

Scatterplot of the Association Between Trait Empathic Concern With Prosocial Learning ($\alpha_{Identifiable}$) in Experiment 1.



Experiment 1 Discussion

In Experiment 1, we partially replicated past research (Lockwood et al., 2016) showing that participants were more adept at learning to earn rewards for themselves, compared to other non-self targets. However, we did not find much support for the hypothesis that trait empathic

concern was associated with prosocial learning for either anonymous targets, the Oxfam charitable organization, or an identifiable needy other.

Experiment 2

In Experiment 2, we made five changes to the experimental protocol. First, we included only two learning conditions: Learning for the self, and learning for the identifiable needy target. Second, we attempted to experimentally manipulate participants' empathy for the plight of the identifiable needy target via perspective-taking instructions (Batson, Early, & Salvarani, 1997): Participants were instructed to either objectively focus on the facts about the target (low empathy condition), or imagine how the target feels (high empathy condition). Third, we manipulated participants' beliefs about the evaluability of their learning performance. We led participants to believe either that their performance would remain private, such that they alone were aware of how much money they earned for the target (low social evaluation condition), or that their performance would become public, such that the experimenter and the identifiable needy target would also learn how much money the participant had earned on their behalf (high social evaluation condition). Fourth, to increase generalizability (Yarkoni, 2022), we included four different identifiable needy targets in our study, and participants were randomly shown one of the four targets. Fifth, we examined associations between prosocial learning and a measure of the online simulation component of empathy from the QCAE in an effort to directly replicate Lockwood et al.'s (2016) finding.

We preregistered the study design and analyses for Experiment 2. The preregistration can be viewed online at https://osf.io/yp2th?view_only=81071e337d37443b967573eaf28938a4[3].

[3] We made additional preregistered predictions for Experiment 2 that we do not report in this manuscript, as they are part of a broader research effort.

Method

Participants

Participants were $N = 425$ UC San Diego undergraduate students ($M = 20.49$, $SD = 2.91$; three participants failed to provide age data; 1.2% = American Indian/Alaskan Native, 47.2% = Asian, 0.7% = Native Hawaiian or Other Pacific Islander, 2.4% = Black or African-American, 23.3% = White, 25.2% = Other) who completed the study on Qualtrics remotely via their computers or mobile devices. We selected our sample size based on a preregistered goal of sampling $n = 450$, with the expectation that we'd still reach our target sample size even in the event that 5% of our sample drops out, and 5% of subjects are removed due to missing trials, with the overall goal of $n = 400$ eligible for inclusion in analyses. Participants received course credit for participating, and were told they could earn up to $2.40 in money, depending on their performance in the task. Data collection took place from April to May 2022.

Procedure

The procedure for Experiment 2 was identical to that of Experiment 1, save for several changes detailed below.

Empathy manipulation. Before completing the two-armed bandit task, participants were informed that they would be provided with information about the other person. Participants were given mindset instructions that included the following text before they were given the information: "We'd like you to read about [person's name] before earning money for them. Research has found that a person's mindset can affect the impressions he or she has when learning about other people. Thus, it is crucial for all participants to have the same mindset when reading about [person's name]. Please adopt the mindset described on the next page before reading about [person's name] campaign."

Participants in the low-empathy condition were then given instructions to remain objective when reading about the other person: “While you are reading about [person’s name], try to pay careful attention to the information presented. Try to be as objective as possible, attending to all the facts that [person’s name] presented about their life. Try not to concern yourself with how [person’s name] feels about what they wrote. Just concentrate on trying to understand the information presented objectively.” In contrast, participants in the high-empathy condition were given instructions to imagine how the other person feels: “While you are reading about [person’s name], try to imagine how [person’s name] feels. Try to take the perspective of [person’s name], imagining how he/she feels what they wrote about and how it has affected their life. Try not to concern yourself with attending to all the information presented. Just concentrate on trying to imagine how [person’s name] feels.”

After receiving instructions, participants in both the low-empathy and high-empathy conditions viewed a picture of the identifiable needy target and read a short autobiography about the reason they were soliciting money.

Emotion Response Questionnaire. After reading about the target, participants were presented with a list of 18 emotion adjectives, and asked to indicate how much they felt each emotion while reading about the target. Five of the items mapped onto empathic concern: “Sympathetic,” “compassionate,” “tender,” “moved,” and “softhearted.” The 13 remaining items served as distractors. Response options for each item ranged from 1 (Not at all) to 7 (Extremely). We formed a composite from responses to the five empathic concern items ($M = 4.70$, $SD = 1.44$; *McDonald’s* $\Omega = 0.91$).

Social evaluation manipulation. Next, participants were assigned to either a low or high social evaluation condition. In the low-evaluation condition, participants were informed that their

performance when learning on behalf of the identifiable needy target would be private. The low-evaluation instructions included the following text: “Please keep in mind that [person’s name] will actually receive any money you win for them. The money you win for them will be given to them at the end of the study, but [person’s name] won’t know that it was you that earned it. They also do not know that you are performing a task where you could win extra money for them - we will simply make an anonymous donation to [philanthropy organization] once the study concludes.”

In contrast, participants in the high-evaluation condition were informed that their performance when learning on behalf of the identifiable needy target would be shared with both the professor in charge of the experiment and the target, so that their instructions read as follows: “Please keep in mind that [person’s name] will actually receive any money you win for them. The money you win for them will be given to them at the end of the study. After the study ends, Dr. McCullough will check to see how much money that you have personally earned for [person’s name]. Then, he will make a donation reflecting the amount of your earnings to [person’s name] and tell [person’s name] that the donation came from a student at UCSD with your age information, made on today's date.”

Two-armed bandit task. Participants then completed a two-armed bandit that was identical to the task from Experiment 1, except that they completed the task for only two targets: themselves, and an identifiable needy target. Participants were randomly assigned to learn on behalf of four different identifiable needy targets. After the experiment ended, we donated a total of $562.80 to the identifiable needy targets, earmarking the specific amount of money that had been earned for each target, and paid each participant a flat fee of $2.40 for their performance when learning for themselves.

Social evaluation manipulation check. After completing the bandit task for both themselves and the identifiable needy target, participants were asked to answer six questions about their experience in the symbol learning task. Two of these questions tapped beliefs about the privacy of participants' performance on the task: "My performance on the symbol learning task was anonymous," and "[Person's name] would know about my performance when playing for them on the symbol learning task" (reverse scored). The other four questions served as distractors. Each item was measured on a self-report scale, with response options ranging from 1 (Definitely not) to 7 (Definitely yes). We formed a composite of the two questions pertaining to privacy (M = 2.85, SD = 1.30; *Cronbach's* α = 0.33[4]).

Interpersonal Reactivity Index. We formed a composite from responses to the IRI (M = 2.76, SD = 0.69; *McDonald's* Ω = 0.87).

Online Simulation. Online simulation was measured using the online simulation subscale from the Questionnaire of Cognitive and Affective Empathy (QCAE; Reniers, 2011). The online simulation subscale includes nine items, each of which was measured on a scale from 1 (strongly disagree) to 4 (strongly agree). We formed a composite from the nine items (M = 2.96, SD = 0.45; *McDonald's* Ω = 0.87).

Data analysis

Data analysis procedures and code were identical to those we used in Experiment 1. We removed data from missing trials (6.09% of all possible trials). One participant missed more than half of their trials and was removed from model estimation, so that the final dataset included N = 424 participants.

[4] We attempted to compute McDonald's Ω, but there were too few items to estimate the underlying model needed to compute Ω.

Computational model selection

As in Experiment 1, we adjudicated among three competing hierarchical Bayesian reinforcement learning models that included α and β parameters representing the self and the identifiable needy target. A model in which separate α and β parameters were estimated for each of the self and identifiable needy targets had the smallest DIC statistic (DIC = 44,188.13), compared to the model in which separate α parameters and a single β were estimated (DIC = 44,341.98), or the model in which only a single α and a single β parameter were estimated (DIC = 44,840.93). We therefore retained the model with separate α and β parameters. We conducted a parameter recovery test, and simulated data from the winning model. See the Supplemental Materials for more information about the parameter recovery test, and other information related to model-fitting, such as trace plots, autocorrelation plots, and other model evaluation information.

Results

As in Experiment 1, participants learned the differential value of the symbols, and differentially learned to earn reward for themselves vs. non-self targets. We also conducted analyses pertaining to β, and correlations between the various α and β parameters, and found that all parameters were significantly correlated with one another ($0.24 < r\text{s} < 0.43$; $p\text{s} < .001$). See the Supplemental Materials for more details about these analyses and results.

Empathy and social evaluation manipulation checks

We ran a number of statistical tests to examine the effectiveness of our two manipulations. For the empathy manipulation, we conducted a two-samples t-test comparing state empathic concern (as measured by the ERQ) between the low-empathy and high-empathy conditions. There was no difference in state empathic concern between participants in the low-empathy condition (M

= 4.65, *SD* = 1.48) and the high-empathy condition (*M* = 4.75, *SD* = 1.40; *t*(413) = 0.76, *p* = .448, *Cohen's D* = 0.07, *95% CI* = [-0.12, 0.27].

For the social evaluation manipulation, we conducted a two-samples *t*-test comparing self-reported anonymity (as measured by the two-item composite of self-reported anonymity culled from the social evaluation manipulation check) between the low-evaluation and high-evaluation conditions. Participants in the high-evaluation condition expressed greater belief that their performance on the learning task would become known to others (*M* = 3.21, *SD* = 1.19) than did participants in the low-evaluation condition (*M* = 2.43, *SD* = 1.29; *t*(420) = 6.39, *p* < .001, *Cohen's D* = 0.62, *95% CI* = [0.43, 0.82]).

Thus, the perspective-taking instructions failed to manipulate empathic concern for the identifiable needy targets, but the social evaluation manipulation did cause participants in the high social evaluation condition to believe that their performance would become public knowledge.

Prediction 1: Does trait empathic concern predict prosocial learning?

No. We correlated participants' trait empathic concern with their scores on the residualized $\alpha_{Identifiable}$. We found that trait empathic concern was not significantly associated $\alpha_{Identifiable}$ ($r(412) = 0.07$, $p = .176$, *95% CI* = [-0.03, 0.16]), so that we failed to find support for prediction 1. A scatterplots of this correlation can be found in the Supplemental Materials. Correlations between all measures are shown in Table 1.

Table 1

Correlations Between All Measures in Experiment 2.

Variable	1	2	3	4	5	6	7
1. α_{Self}							
2. *raw* $\alpha_{Identifiable}$	.43**						
3. $\alpha_{Identifiable}$	-.00	.90**					
4. Trait empathic concern	.09	.10*	.07				
5. State empathic concern	.01	.13**	.14**	.40**			
6. Online simulation	-.09	-.05	-.01	.36**	.25**		
7. Empathy manipulation	.03	-.00	-.02	-.08	.04	.03	
8. Social evaluation manipulation	-.03	-.09	-.09	-.01	-.02	-.02	.03

Note. The variable *raw* $\alpha_{Identifiable}$ refers to the learning rate for the identifiable other before residualizing out the variance associated with α_{Self}. * indicates $p < .05$. ** indicates $p < .01$.

Prediction 2: Does state empathic concern predict prosocial learning?

Yes. We correlated participants' state empathic concern with their scores on the residualized $\alpha_{Identifiable}$. State empathic concern was positively and significantly associated with

$\alpha_{Identifiable}$ ($r(412) = 0.14$, $p = .005$, *95% CI* = [0.04, 0.23]): Participants who reported experiencing more empathic concern for the identifiable needy targets were more effective at learning for the targets, providing support for Prediction 2. A scatterplots of this correlation can be found in the Supplemental Materials.

Predictions 3 and 4: Do manipulations of empathy and social evaluation predict prosocial learning?

No. We regressed the residualized $\alpha_{Identifiable}$ on three predictor terms: A dummy-coded predictor for the empathy condition (0 = low empathy, 1 = high empathy); A dummy-coded prediction for the social evaluation condition (0 = low social evaluation, 1 = high social evaluation); and the interaction between the dummy-coded predictor terms. Neither of the main effects, nor their interaction, significantly predicted $\alpha_{Identifiable}$ (ps > .275). Thus, we did not find support for prediction 3 (i.e., that a manipulation of empathy would increase prosocial learning) or prediction 4 (that a manipulation of social evaluability would increase prosocial learning).

We conducted an additional non-preregistered analysis to test whether the association between the empathy manipulation and prosocial learning was mediated by state empathy. We fit a structural equation model in which we regressed prosocial learning on the empathy manipulation, state empathy, and learning for the self, and regressed state empathy on the manipulation. The model had excellent fit ($\chi^2(1) = 0.045$, $p = .831$, *CFI* = 1.00, *RMSEA* = 0.00, *SRMR* = 0.003). Even so, we found that the estimated indirect effect linking the empathy condition, state empathy, was non-significant ($b = 0.002$, *SE* = 0.002, *B* = 0.005, *95% CI* = [-0.007, 0.018], $Z = 0.80$, $p = .425$).

Prediction 5: Does the online simulation component of empathy predict prosocial learning?

No. We correlated participants' scores on the online simulation measure of empathy with their scores on the residualized $\alpha_{Identifiable}$. Online simulation was not significantly correlated with

$\alpha_{Identifiable}$ ($r(197) = -0.01$, $p = .875$, *95% CI* = [-0.15, 0.13]). Thus, we did not find support for prediction 5[5].

Experiment 2 Discussion

In Experiment 2, we retested a prediction 1 from Experiment 1, in addition to four new predictions about the association between prosocial motivation and prosocial learning. We found that state empathic concern for identifiable needy targets was positively associated with prosocial learning on behalf of those targets, providing support for the empathy-altruism hypothesis. We successfully manipulated the visibility of participants' performance on the task, but we did find that publicizing the amount of money that participants earned for the identifiable needy target affected their performance, suggesting that effort-based tasks provide a measure of prosocial motivation that is relatively uncontaminated by social desirability bias.

Contrary to our predictions, neither trait empathic concern nor the online simulation component of empathy were associated with prosocial learning. In addition, our manipulation of empathy via perspective-taking instructions did not predict prosocial learning nor state empathic concern, but we cannot draw strong conclusions because the perspective-taking instructions also failed to manipulate self-reported empathic concern.

Experiment 3

The study design and analyses for Experiment 3 were preregistered[6]. The preregistration can be viewed online at https://osf.io/h9npt?view_only=81071e337d37443b967573eaf28938a4. We made one change to our preregistered predictions with respect to the effect of social evaluation. In Experiment 2, we predicted that social evaluation would increase prosocial learning, but found

[5]Due to a programming error, $N = 218$ participants did not receive the online simulation subscale. The data were missing completely at random (MCAR).
[6]We made additional preregistered predictions for Experiment 3 that we do not report in this manuscript, as they are part of a broader research effort.

no effect. In light of this null finding, we preregistered the competing prediction that social evaluation would have no effect upon prosocial learning. The updated predictions for social evaluation are as follows:

***Prediction 4a*: Participants who complete the bandit task under high social evaluation will have larger prosocial learning rates, compared to participants who complete the task under low social evaluation.**

***Prediction 4b*: Participants who complete the bandit task under high social evaluation will not differ in their residualized prosocial learning rates, compared to participants who complete the task under low social evaluation.**

Method

Participants

Participants were $N = 776$ UC San Diego undergraduate students. Due to a protocol error, 210 participants saw a photo of one identifiable needy target but read about another, and were removed prior to data analysis. 18 participants were also removed from data analysis for failing to complete the entire study. Overall, data from 548 participants were included in analyses (M_{age} = 20.49, SD_{age} = 3.41; 1.1% = American Indian/Alaskan Native, 42.2% = Asian, 0.7% = Native Hawaiian or Other Pacific Islander, 3.5% = Black or African-American, 26.5% = White, 26.1% = Other). As in Experiment 2, we selected our sample size based on our preregistered goal of sampling $n = 450$, with the target of $n = 400$ participants for analyses. We oversampled participants

due to unexpectedly larger available financial resources for paying participants. Data collection took place from October 2022 to March 2024.

Procedure

The procedure was nearly identical to Experiment 2–participants learned to earn rewards for themselves and an identifiable needy target–except for three changes. First, we had participants complete the study in the laboratory, rather than remotely, to encourage more engagement.

Second, in Experiments 1 and 2, participants completed the IRI empathic concern subscale as part of a larger questionnaire battery in random order. We reasoned that participants might have engaged in satisficing so that, for instance, if they completed the subscale towards the end of the battery, the quality of their responses might be worse. We note that reliability indices for the empathic concern subscale in Experiments 1 and 2 suggested that the items had good reliability, suggesting that participants were not satisficing. Still, in Experiment 3, we had participants complete the subscale as one of the first four measures of the questionnaire battery in randomized order.

Third, midway through experiment 3 we shortened the scales from the questionnaire battery to reduce participant fatigue. Three hundred twenty-five participants from Experiment 3 completed all seven items from the empathic concern subscale, and 223 completed the four-item version of the subscale; we report results for both the full and shortened version of the subscale. We note that the questionnaire battery followed the bandit task and experimental manipulations for both the shortened and full-length version of the questionnaire battery, so inferences regarding performance on the bandit task should not have changed. We also note that we did not shorten the online simulation subscale in the interest of directly replicating Lockwood et al.'s (2016) results.

Finally, after the experiment ended, we donated a total of $797.10 to the identifiable needy target[7], earmarking the specific amount of money that had been earned for each target, and paid participants a flat fee of $2.40 for their performance when learning for themselves.

Data Analysis

Data analysis procedures and code were identical to those used in Experiments 1 and 2. We removed data from missing trials (3.15% of all possible trials), and no participants missed more than half of their trials.

Computational model selection

We followed the same model-selection and evaluation procedure as in Experiments 1 and 2. Replicating our results from Experiments 2, the model in which both separate α and β parameters were estimated for each of the self and the identifiable needy target had the smallest DIC statistic (*DIC* = 53,167.43), compared to the model in which separate α parameters and a single β were estimated (*DIC* = 53,323.56), or the model in which a single α parameter and a single β were estimated (*DIC* = 54,369.35). We therefore retained the model with separate α and β parameters. Parameter recovery results and other model-fitting information is reported in the Supplemental Materials.

Results

As in Experiments 1 and 2, participants learned the differential value of the symbols and differentially learned to earn reward for themselves vs. non-self targets. We also conducted analyses pertaining to the β parameter, and correlations between the α and β parameters. See the Supplemental Materials for more details about these analyses and results.

[7] Although we removed the data associated with participants who saw the incorrect image of the identifiable needy target described, we donated the money that all $N = 776$ participants earned for the intended identifiable needy target, regardless if they saw the correct or incorrect image, and regardless if they were removed from data analysis.

Empathy and social evaluation manipulation checks

For the empathy manipulation, we conducted a two-samples *t*-test comparing state empathic concern between the low-empathy and high-empathy conditions. Replicating results from Experiment 1, we found no difference in state empathic concern between participants in the low-empathy condition ($M = 4.44$, $SD = 1.49$) versus the high-empathy condition ($M = 4.53$, $SD = 1.50$; $t(537) = 0.69$, $p = .489$, *Cohen's D* = 0.06, *95% CI* = [-0.11, 0.23]).

For the social evaluation manipulation, we conducted a two-samples *t*-test comparing self-reported anonymity between the low and high social evaluation conditions. Participants in the high social evaluation condition expressed greater belief that other people would be aware of their performance on the learning task ($M = 3.50$, $SD = 1.36$; *Cronbach's* $\alpha = 0.23$), relative to participants in the low social evaluation condition ($M = 2.25$, $SD = 1.36$; $t(544) = 10.79$, $p < .001$, *Cohen's D* = 0.92, *95% CI* = [0.75, 1.10]).

Thus, as in Experiment 2, the perspective-taking instructions failed to manipulate empathic concern for the identifiable needy targets, but the social evaluation manipulation effectively induced concern about social evaluation.

Prediction 1: Does trait empathic concern predict prosocial learning?

No. We computed the correlation between the trait empathic concern and scores on the residualized $\alpha_{Identifiable}$ using two methods: (1) The correlation amongst participants who completed all seven items from the IRI ($M = 2.88$, $SD = 0.65$; *McDonald's* $\Omega = 0.85$), and (2) The correlation amongst all participants, using the four items that were shared from the short and long version of the IRI ($M = 2.92$, $SD = 0.71$; *McDonald's* $\Omega = 0.75$).

With respect to the correlation amongst participants who completed the seven-item version of the IRI empathic concern subscale, we found that the correlation was non-significant ($r(322) =$

0.08, p = .167, *95% CI* = [-0.03, 0.18]). Likewise, we found that the correlation amongst participants who completed the four-item version of the subscale was non-significant (r(545) = 0.04, p = .309, *95% CI* = [-0.04, 0.13]). Thus, we failed to find support for prediction 1.

Scatterplots of the correlation between both the short and long versions of trait empathic concern, and prosocial learning, can be found in the Supplemental Materials. Correlations between all measures are shown in Table 2.

Table 2

Correlations Between All Measures in Experiment 3.

Variable	1	2	3	4	5	6	7	8
1. α_{Self}								
2. *raw* $\alpha_{Identifiable}$	.17**							
3. $\alpha_{Identifiable}$	.00	.99**						
4. Seven-item IRI trait empathic concern	-.00	.07	.08					
5. Four-item IRI trait empathic concern	-.00	.04	.04	.94**				
6. State empathic concern	.03	.04	.03	.44**	.43**			
7. Online simulation	.03	.07	.06	.44**	.40**	.18**		

8. Empathy manipulation	-.07	.05	.06	-.06	-.04	.03	-.01	
9. Social evaluation manipulation	.02	.03	.02	.08	.01	-.00	-.04	-.05

Note. The variable raw $\alpha_{Identifiable}$ refers to the learning rate for the identifiable other before residualizing out the variance associated with α_{Self}. * indicates $p < .05$. ** indicates $p < .01$.

Prediction 2: Does state empathic concern predict prosocial learning?

No. We found that state empathic concern ($M = 4.48$, $SD = 1.49$; *McDonald's* $\Omega = 0.93$) was uncorrelated with $\alpha_{Identifiable}$ ($r(536) = 0.03$, $p = .445$, *95% CI* = [-0.05, 0.12]), failing to find support for prediction 2. Scatterplots of this correlation can be found in the Supplemental Materials.

Predictions 3 and 4: Do manipulations of empathy and social evaluation predict prosocial learning?

No. We regressed the residualized $\alpha_{Identifiable}$ on three predictor terms: A dummy-coded predictor for the empathy condition (0 = low empathy, 1 = high empathy); A dummy-coded prediction for the social evaluation condition (0 = low social evaluation, 1 = high social evaluation); and the interaction between the dummy-coded predictor terms. Neither of the main effects, nor the interaction effect, were significantly associated with $\alpha_{Identifiable}$ (ps > .165), failing to provide support for predictions 3 and 4.

Prediction 5: Does the online simulation component of empathy predict prosocial learning?

No. We correlated participants' scores on the online simulation component of empathy ($M = 3.05$, $SD = 0.44$; *McDonald's* $\Omega = 0.86$) with their scores on the residualized $\alpha_{Identifiable}$. Online

simulation did not significantly predict $\alpha_{Identifiable}$ ($r(540) = 0.06$, $p = .153$, *95% CI* = [-0.02, 0.14]). Thus, we failed to find support for prediction 5.

Experiment 3 discussion

Consistent with the results of Experiment 2, we did not find effects for trait empathic concern, the empathy and social evaluation manipulations, nor online simulation. In contrast to Experiment 2, we found that empathic concern failed to predict prosocial learning. Additionally, both the empathy and social evaluation manipulations failed a manipulation check. In light of the mixed findings we found in Experiments 1-3, we combined the data from all three studies together in a single dataset so that we could test whether empathy was associated with prosocial learning in a much larger sample.

Mega-analysis

To investigate our hypotheses by taking into account all of the data from all three experiments at once, we combined the data from all three studies, and analyzed them as a single dataset ($N = 1{,}188$). For each analysis, the outcome variable was the residualized $\alpha_{Identifiable}$. We also included dummy codes representing each study to control for between-study differences. For the analyses that involved trait empathic concern, we included data from Experiments 1, 2, and 3, and thus included two dummy codes encoding the study number. Experiment 1 served as the reference condition. For the analyses that involved state empathic concern, the empathy and social evaluation manipulations, and trait online simulation, we included data from Experiments 2 and 3, and included one dummy code. Experiment 2 served as the reference condition. For the data pertaining to Experiment 3, we ran analyses using data from the seven-item version of trait empathic concern, and did not include the shortened versions of those scales.

For each analysis, we regressed $\alpha_{Identifiable}$ on (1) The appropriate dummy codes encoding study number; and (2) The predictor variable(s) relevant to each prediction. Results are summarized in Tables 3-6. We found that trait empathic concern (β = 0.09, b = 0.07, SE = 0.04, p = .035, *95% CI* = [0.01, 0.18]) and state empathic concern (β = 0.08, b = 0.01, SE = 0.004, p = .019, *95% CI* = [0.001, 0.02]) were significantly associated with $\alpha_{Identifiable}$. The empathy manipulation, social evaluation manipulation, and trait online simulation did not significantly predict $\alpha_{Identifiable}$ (ps > .163). These mega-analytic results suggest that trait empathic concern and state empathic concern indeed positively predict prosocial learning, but the associations are extremely weak–so weak in fact that future research probably wouldn't be able to recover them without thousands of subjects worth of data.

Table 3

Mega-Analysis For the Model in Which Trait Empathic Concern Predicts the Residualized $\alpha_{Identifiable}$.

Predictor	β	*b (SE)*	*p*	*N for analysis*
Trait empathic concern	0.09	0.07 (0.04)	.035*	960
Study 2 dummy code	-0.001	-0.001 (0.07)	.983	
Study 3 dummy code	-0.03	-0.05 (0.08)	.524	

Note: * indicate $p < .05$.

Table 4

Mega-Analysis For the Model in Which State Empathic Concern Predicts the Residualized $\alpha_{Identifiable}$.

Predictor	*β*	*b (SE)*	*p*	*N for analysis*
State empathic concern	0.08	0.01 (0.004)	.019*	952
Study 3 dummy code	0.01	0.01 (0.06)	.818	

Note: * indicate $p < .05$.

Table 5

Mega-Analysis For the Model in Which the Empathy Manipulation, the Social Evaluation Manipulation, and their Interaction Predicts the Residualized $\alpha_{Identifiable}$.

Predictor	*β*	*b (SE)*	*p*	*N for analysis*
Empathy manipulation	0.07	0.12 (0.08)	.163	971
Social evaluation manipulation	0.01	0.02 (0.09)	.795	
Social evaluation x Empathy	-0.06	-0.13 (0.12)	.290	
Study 3 dummy code	-0.002	-0.004 (0.06)	.949	

Table 6

Mega-Analysis For the Model in Which Trait Online Simulation Predicts the Residualized $\alpha_{Identifiable}$.

Predictor	*β*	*b (SE)*	*p*	*N for analysis*
Online simulation	0.04	0.01 (0.01)	.253	741
Study 3 dummy code	-0.03	-0.06 (0.08)	.485	

General Discussion

Learning is central to helping, as people must update their beliefs about who and how to help others. Some research has suggested that learning may also be an integral component of empathy. However, empathy's effect upon learning is obscured by methodological shortcomings in past studies that (a) measured empathy, (b) featured anonymous, rather than needy, targets, and (c) measured, but did not manipulate, empathy. We addressed these problems by conducting three studies – including two pre-registered studies – that investigated how empathy influences prosocial learning, drawing upon model-free reinforcement-learning models of prosocial learning that formalize how learning occurs when earning benefits for oneself versus needy targets (Daw, 2011; Lockwood et al., 2016).

We found mixed evidence that empathic concern was associated with prosocial learning on behalf of needy targets. We did not find an effect for trait empathic concern in any study, and only in Experiment 2 did we find an effect for state empathic concern; when we analyzed data that included participants from all three studies, though, we found small but significant effects for both state and trait empathic concern. We cautiously interpret these results as evidence in favor of the hypothesis that individual differences in empathic concern, and in-the-moment feelings of empathic concern, influence cognitive processes involved in prosocial learning. However, these associations appear to be extremely small.

We did not find causal evidence for empathy's effect upon prosocial learning. However, the perspective-taking instructions failed to influence state empathy scores, so our results are inconclusive as to empathy's causal effect upon learning. We also could not replicate Lockwood et al.'s (2016) results with respect to the effect of online simulation, despite using a sample more than 25 times larger than theirs. We also found that social evaluation did not share a causal relationship

with prosocial learning, consistent with our hypothesis that prosocial learning ought to be unrelated to social evaluation. We did replicate and extend the finding from Lockwood et al., (2016) that, regardless of empathy's effect upon learning, participants were better at learning for themselves compared to anonymous targets, charitable organizations, and identifiable needy targets.

Finally, the present research advanced the generalizability of research regarding prosocial learning with respect to three methodological features. First, we actively recruited significant representation across ethnicities, with the vast majority of participants reporting as non-White. Second, our experiments generalized the targets of prosocial learning. Most notably, our study included real targets who expressed need of help, whereas past studies focused primarily on anonymous targets who were not in need of help. Moreover, the targets were of diverse ethnic backgrounds, including men, women, Black, White, older, and younger. Second, whereas past prosocial learning experiments were conducted in the laboratory, our experiments were conducted both online and in the laboratory, broadening the scope with which prosocial learning experiments have been conducted. We found that results appeared to be largely similar across the two studies – and not statistically significant, according to the dummy variables included as predictors in the mega-analysis. These results suggest that the effect of empathic concern upon prosocial learning does not depend on laboratory environments, so that future researchers might have confidence conducting their studies in online environments.

Why was empathic concern's effect upon learning so small?

In contrast to previous research, we found that the associations between empathy and learning were small, regardless of the empathy measure that we used. We designed our experiment on the premise that manipulating the target of learning ought to cause between-subjects differences

in peoples' motivations to help because people are generally less motivated to learn for others than for themselves (Contreras-Huerta, Pisauro, & Apps, 2020). Empathy is thought to be important because it might buffer the motivational decline that people typically exhibit when learning for someone else. Since we found a robust difference between learning for the self versus non-self targets, it does seem that there is an appreciable motivational gap between learning for the self versus others within-subjects. If it's true that empathy does buffer against this motivational gap, then why did we obtain such small statistical associations between empathy and prosocial learning? We can think of two reasons.

First, the existence of *within*-subjects heterogeneity in learning for the self versus others does not necessarily imply that there is substantial *between*-subjects heterogeneity in learning for others. Historically, the within-subjects experiments that predominate in cognitive psychology are explicitly designed to minimize between-subjects variance (Borsboom & Haslbeck, 2024; Fisher et al., 2018; Mattoni et al., 2025). In contrast, research on individual differences, including trait empathy, emphasizes identifying between-subjects variance (Borsboom et al., 2009). One reason we found small effects for empathy's effect upon prosocial learning, then, might be that the bandit task minimizes the amount of between-subjects variance to be explained. Although it is true that we did find that there was at least some variation in the residualized α parameters when learning on behalf of the other person, it is unclear whether those deviations reflect meaningful motivational differences, or the sum of noise that accrues over the course of experimental trials. This latter speculation is supported by the fact that very large numbers of trials are generally required to decrease the amount of measurement error in trial-by-trial tasks (Rouder and Haaf, 2019). To our knowledge, virtually all studies that have tested the association between learning and empathy have relied upon fewer than 100 trials per experimental condition; for instance, both our

experiment and Lockwood et al. (2016) included 48 trials per condition. Indeed, Rouder et al. (2023) suggested that as many as 1,600 trials per condition may be necessary to recover correlations between cognitive tasks, because cognitive tasks typically have poor reliability (Haines et al., 2023; Sullivan-Toole et al., 2022). Troublingly, hierarchical models may do little to minimize the trial-by-trial noise that they are designed to combat, so that running more trials per participant may be the only recourse for obtaining psychometrically sound trial-by-trial data (Rouder et al., 2023).

Second, and relatedly, our study featured a within-subjects manipulation and data structure, but our analysis was conducted at the between-subjects level: α is a summary variable that reflects learning across all trials, and empathy was measured at a single time point at both the state and trait level. Analyzing the data using between-subjects statistical methods may have undermined our ability to detect a meaningful effect, since the within-subjects association between learning and empathy could differ in magnitude or direction from the between-subjects association observed here (Hamaker, 2012).

More generally, the field of computational psychology prizes formal mathematical models that represent the generative process by which cognitive processes produce behaviors. Generative models are valuable for building theory (Haines et al., in press). However, parameters from these models may be poorly suited for estimating associations between cognitive task performance and individual differences because of the aforementioned psychometric problems associated with summary parameters drawn from cognitive tasks. Luckily, a solution does seem to exist: Trial-by-trial data can be analyzed using multilevel models that capture the trial-level variance, while also disaggregating within- and between-subjects effects. Although we did analyze trial-by-trial learning using a multilevel model, we measured state and trait empathy at a single

time point, so we were unable to test the association between participants' behavioral responses to each trial, and the amount of empathy they felt at each trial. Future research ought to disentangle the empathy and learning association by directly measuring the amount of emotion felt during each trial, and then testing the association between learning and empathy entirely at the within-subjects level of analysis (e.g., FeldmanHall & Heffner, 2022).

Why did our manipulation of empathic concern fail?

We could not validly test the causal effect of empathy upon learning because the perspective-taking instructions did not increase empathic concern. We struggle somewhat to explain this failure. Batson and colleagues have argued that laboratory experiments designed to test the association between empathy and helping behavior require real instances of needy others, and real opportunities to help for empathic concern to have an effect upon helping (Batson, 2010). This led us to wonder (post hoc) whether our subjects perceived that they really were in a position to help needy others. Reassuringly, we found that only 26 (2.2%) of all participants expressed suspicion either that the identifiable needy targets were a sham, or that they wouldn't receive the money that the participant had earned on their behalf. These results suggest that subjects accurately perceived the experimental situation as a genuine opportunity to help needy others.

We also note that many of the canonical studies using perspective-taking manipulations to manipulate empathy for real people involve first-person interactions between subjects and targets–for example, through the exchange of notes in what appears to be real time (e.g., Fultz et al., 1986; McCauley et al., 2024). In our experiments, subjects did not interact with the potential beneficiaries of their learning, even though they did learn facts about their needs. Perhaps perspective-taking instructions don't increase empathy reliably in the absence of direct real-time interaction. Others have failed to find that perspective-taking instructions increase subjects'

empathy for needy people with whom they don't interact (e.g., Wright et al., 1990). Surprisingly, we know of no experiments that have evaluated this question directly, although a recent meta-analysis suggests that people might experience more empathy in response to perspective-taking instructions (versus remain-objective instructions) when interaction with the needy person involves audio (McAuliffe et al., 2020).

One other reason that the perspective-taking instructions might have failed to manipulate empathic concern may be that participants were not only provided textual information about the plight of the identifiable needy targets, but were also shown their photos. It is possible that images of the identifiable needy targets interfered with participants' instructions to "focus on the facts" in the remain-objective condition, thereby increasing those subjects' empathy in a way that reduced the typical effects of the remain-objective condition (McAuliffe et al., 2020).

Social evaluation and prosocial learning

The results from our manipulation check indicate that participants thought that their prosocial learning performance was monitored. However, it's not clear that participants believed that their performance was necessarily being evaluated for moral content. In other words, it is not entirely clear that participants necessarily thought that someone else would judge them harshly on the basis of their performance on the prosocial learning task. The items included in the manipulation check could have been interpreted by participants as meaning that someone else would find out that they had performed the task, but they wouldn't necessarily have reason to believe that others could infer their desire to help from the participant's performance. For instance, participants may have interpreted the item "[Target name] would know about my performance when playing for them on the symbol learning task" to mean that their name would be shared with the identifiable needy target simply as confirmation that the participant had taken part in the

experiment regardless of whether they had performed well or not, so that this information didn't necessarily reflect on the participant's moral standing.

If the social evaluation instructions did cause participants to have an increased concern about their moral evaluation, there are three reasons why prosocial learning might yet be resistant to social evaluation concerns. First, the explanation as to why we failed to find an association between empathy and learning may also apply to the association between social evaluation and learning: There may simply be little between-subjects variance to explain. Second, since performance on social cognition tasks are based on both the ability and the motivation to help, they might be more difficult for participants to fake because task performance carries considerable cognitive costs (Contreras-Huerta, Pisauro, & Apps, 2020). It may be that the task is simply so difficult that participants don't have the cognitive resources available to consider what others think of them; many of the participants that left qualitative feedback at the end of the study wrote that they found the task to be tedious to work on, and expressed mental fatigue at the time and effort it took to complete the task. It is plausible that the cognitive costs associated with good task performance minimized the benefits associated with earning reputational benefit. Third, and relatedly, the performance-based nature of cognitive tasks increases the plausibility that one's poor performance on a learning task could be attributed to the difficulty of the task, rather than a lack of motivation on the part of the learner, providing moral cover for apathetic learners. In other words, observers could attribute poor performance on the task to either apathy (which would be morally relevant), or ineptitude (which would be morally excusable).

Constraints on Generality

The present findings must be interpreted with respect to several constraints on generality. First, although our sample was ethnically diverse, it was composed entirely of undergraduate

students, which limits the generalizability of our results to populations with differing life experiences or educational backgrounds. Second, prosocial learning was measured via an abstract task (i.e., the two-armed bandit task). Although the two-armed bandit task has canonically been used to measure learning in the laboratory, it may not fully capture the complexity of prosocial learning outside of the laboratory, particularly with respect to the nuanced ways in which empathically motivated learning operates in naturalistic contexts. For instance, empathy-induced learning in the real world is likely to involve acts such as determining the kind of medical care needed to assist those who are near and dear, searching for job opportunities for a struggling friend, or mundane tasks like the best way to change diapers for one's child. Furthermore, while the use of identifiable needy targets enhanced ecological validity, our empathy and social evaluation manipulations may not have effectively simulated the spontaneous and multifaceted nature of these constructs in real-world situations: Although empathy is typically studied in the context of strangers, people mostly empathize with close others (Depow, Francis, & Inzlicht, 2021). Similarly, our experiment might have yielded different results had we included close relationship partners as targets of learning, such as friends or family members.

Our results also raise questions about the robustness of these manipulations in influencing prosocial learning. In addition, it is important to note that our results might reflect a self-reward bias to some extent, as our study confounded prosocial targets with non-self targets. It could be, for instance, that participants viewed social targets (like Oxfam, the identifiable needy other, and the anonymous target) as indistinguishable from non-social targets (like learning for no one, or learning on behalf of a computer script), so that variation in learning simply reflects whether participants learn for themselves or a non-self target. Self-reward bias is part of the explanation as to why people earn more rewards for themselves versus others, because self-reward bias is the

behavioral output of differences in motivation when earning benefits for oneself versus others. Our results are not unique: Lockwood et al. (2016) not only found that there was no difference between learning for another person and a social target, but that there was no difference between for a social target and learning for no one at all. Self-reward bias also explains why people are better at foraging for resources for themselves versus other people because they are less attuned to the rewards that other people might earn compared to themselves (Contreras-Huerta et al., 2024). In other words, we think self-reward bias is an outcome of motivations to help, rather than a cause of motivations to help. Still, future research should more closely investigate the association between self-bias and prosocial learning. Lastly, the measures employed to assess state and trait empathy, while widely used, may not encompass all aspects of empathy relevant to helping behaviors or the motivational mechanisms underlying prosocial learning. Future research should aim to replicate and extend our findings across diverse demographic groups, incorporate alternative paradigms that better mimic everyday helping contexts, and refine the methodological tools used to measure empathy.

Conclusion

Across three studies, we found small but significant associations between empathy and learning, failed to find a causal effect for empathy upon prosocial learning, and did not find evidence that concerns about social evaluation influence prosocial learning. However, we also failed to manipulate state empathic concern. We cautiously interpret these results as positive evidence for a link between empathy and learning, but the association is complicated because the task used to measure learning might lack the psychometric characteristics that are necessary for detecting correlations between task performance and individual differences. Future research ought to test the learning-empathy hypothesis using more psychometrically rigorous tasks.

References

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150-166.

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind & Language, 20*(3), 259-287.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Batson, C. D. (2009). Two forms of perspective taking: Imagining how another feels and imagining how you would feel. In K. D. Markman, W. M. Klein, & J. A. Suhr (Eds.), *Handbook of Imagination and Mental Simulation* (pp. 267–279). New York, NY: Psychology Press.

Batson, C. D. (2010). Empathy-induced altruistic motivation. In M. Mikulincer & P. R. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 15–34). American Psychological Association. https://doi.org/10.1037/12061-001.

Berger, S. M. (1962). Conditioning through vicarious instigation. *Psychological Review, 69*(5), 450-466.

Chopik, W. J., O'Brien, E., & Konrath, S. H. (2017). Differences in empathic concern and perspective taking across 63 countries. *Journal of Cross-Cultural Psychology, 48*(1), 23-38.

Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A. L. (1987). Empathy-based helping: Is it selflessly or selfishly motivated?. *Journal of Personality and Social Psychology*, *52*(4), 749.

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J., & Miene, P. (1998). Understanding and assessing the motivations of volunteers: a functional approach. *Journal of Personality and Social Psychology*, *74*(6), 1516.

Cohen, T. R., Kim, Y., & Panter, A. T. (2014). The five-item guilt proneness scale (GP-5). *Differences*, *92*, 109-112.

Contreras-Huerta, L. S., Pisauro, M. A., & Apps, M. A. (2020). Effort shapes social cognition and behaviour: A neuro-cognitive framework. *Neuroscience & Biobehavioral Reviews*, *118*, 426-439.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Daw, N.D. (2011) Trial-by-trial data analysis using computational models. In Phelps, E.A., Robbins, T.W. & Delgado, M. (Eds.), *Affect, Learning and Decision Making, Attention and Performance XXIII*. Oxford University Press, New York, pp. 3–38.

Delton, A. W., Jaeggi, A. V., Lim, J., Sznycer, D., Gurven, M., Robertson, T. E., ... & Tooby, J. (2023). Cognitive foundations for helping and harming others: Making welfare tradeoffs in industrialized and small-scale societies. *Evolution and Human Behavior*, *44*(5), 485-501.

Depow, G. J., Francis, Z., & Inzlicht, M. (2021). The experience of empathy in everyday life. *Psychological Science*, *32*(8), 1198-1213.

Depow, G. J., Lin, H., & Inzlicht, M. (2022). Cognitive effort for self, strangers, and charities. *Scientific Reports*, *12*(1), 15009.

Du, H., Keller, B., Alacam, E., & Enders, C. (2024). Comparing DIC and WAIC for multilevel models with missing data. *Behavior Research Methods*, *56*(4), 2731-2750.

Einolf, C. J. (2008). Empathic concern and prosocial behaviors: A test of experimental results using survey data. *Social Science Research, 37*(4), 1267-1279.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion, 14*(2), 131-149.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage, 105,* 347-356.

FeldmanHall, O., & Heffner, J. (2022). A generalizable framework for assessing the role of emotion during choice. *American Psychologist*, *77*(9), 1017.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106-E6115.

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: an evolutionary analysis and empirical review. *Psychological Bulletin*, *136*(3), 351.

Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. *Journal of Applied Psychology*, *93*(1), 48.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243-1255.

Haines, N., Kvam, P. D., Irving, L., Smith, C. T., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., and Turner, B. M. (in press). A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. In press at *Psychological Methods*.

Haines, N., Sullivan-Toole, H., & Olino, T. (2023). From classical methods to generative models: Tackling the unreliability of neuroscientific measures in mental health research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 822-831.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Hu, J., & Liden, R. C. (2015). Making a difference in the teamwork: Linking team prosocial motivation to team processes and effectiveness. *Academy of Management Journal*, *58*(4), 1102-1127.

Imas, A. (2014). Working for the “warm glow”: On the benefits and limits of prosocial incentives. *Journal of Public Economics*, *114*, 14-18.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/ 01621459.1995.10476572

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*(3), 371-375.

Livingston, J. A., & Rasulmukhamedov, R. (2023). On the interpretation of giving in dictator games when the recipient is a charity. *Journal of Economic Behavior & Organization*, *208*, 275-285.

Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763-9768.

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. (2017). Prosocial apathy for helping others when effort is required. *Nature Human Behaviour*, *1*(7), 0131.

Lucas, T., Alexander, S., Firestone, I., & LeBreton, J. M. (2007). Development and initial validation of a procedural and distributive just world measure. *Personality and Individual Differences, 43*(1), 71-82.

Mattoni, M., Fisher, A. J., Gates, K. M., Chein, J., & Olino, T. M. (2025). Group-to-individual generalizability and individual-level inferences in cognitive neuroscience. *Neuroscience & Biobehavioral Reviews*, 106024.

McAdams, D. P., & de St Aubin, E. D. (1992). A theory of generativity and its assessment through self-report, behavioral acts, and narrative themes in autobiography. *Journal of Personality and Social Psychology*, *62*(6), 1003.

McCauley, T. G., McAuliffe, W., & McCullough, M. *Does Measurement Bias Explain Sex Differences in Self-Reported Empathy?*. doi: 10.31234/osf.io/6ps2r.

McCauley, T. G., McAuliffe, W., and McCullough, M. (2024). Does empathy promote helping by activating altruistic motivation or concern about social evaluation? A direct replication of Fultz et al. (1986). *Emotion*, *24*, 1868–1884. doi: 10.1037/emo0001339

Mell, H., Safra, L., Algan, Y., Baumard, N., & Chevallier, C. (2018). Childhood environmental harshness predicts coordinated health and reproductive strategies: A cross-sectional study of a nationally representative sample from France. *Evolution and Human Behavior*, *39*(1), 1-8.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision making*, *6*(8), 771-781.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics, 31*(3), 705-767.

Parker, G., Tupling, H., & Brown, L. B. (1979). A parental bonding instrument. *British Journal of Medical Psychology, 52*, 1–10.

Paulhus, D. L. (1998). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding* [Unpublished manuscript]. University of British Columbia.

Perugini, M., Gallucci, M., Presaghi, F., & Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, *17*(4), 251-283.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67,* 741-763.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452-467.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*, *30*(6), 2049-2066.

Ryan, R. M., Rigby, S., & King, K. (1993). Two types of religious internalization and their relations to religious orientations and mental health. *Journal of Personality and Social Psychology*, *65*(3), 586.

Sassenrath, C., Pfattheicher, S., & Keller, J. (2017). I might ease your pain, but only if you're sad: The impact of the empathized emotion in the empathy-helping association. *Motivation and Emotion*, *41*, 96-106.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health Psychology*, *4*(3), 219.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality*, *47*(5), 609-612.

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research*, *23*, 211-238.

Schug, J., Yuki, M., & Maddux, W. (2010). Relational mobility explains between-and within-culture differences in self-disclosure to close friends. *Psychological Science*, *21*(10), 1471-1478.

Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*, 5-16.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* (pp. 98-009). Research Report, 98-009.

Sullivan-Toole, H., Haines, N., Dale, K., & Olino, T. M. (2022). Enhancing the psychometric properties of the iowa gambling task using full generative modeling. *Computational Psychiatry*, *6*(1), 189.

Sznycer, D., Delton, A. W., Robertson, T. E., Cosmides, L., & Tooby, J. (2019). The ecological rationality of helping others: Potential helpers integrate cues of recipients' need and willingness to sacrifice. *Evolution and Human Behavior*, *40*(1), 34-45.

Teoh, Y. Y., & Hutcherson, C. A. (2022). The Games We Play: Prosocial Choices Under Time Pressure Reflect Context-Sensitive Information Priorities. *Psychological Science*, 09567976221094782.

Van Dyne, L., Graham, J. W., & Dienesch, R. M. (1994). Organizational citizenship behavior: Construct redefinition, measurement, and validation. *Academy of Management Journal*, *37*(4), 765-802.

Vostroknutov, A., Polonio, L., & Coricelli, G. (2018). The role of intelligence in social learning. *Scientific Reports*, *8*(1), 6896.

Walker, L. J. (2013). Exemplars' Moral Behavior Is Self‐Regarding. *New Directions for Child and Adolescent Development,* (142), 27-40.

Weinstein, N., & Ryan, R. M. (2010). When helping helps: autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology*, *98*(2), 222.

Weiss, R. F., Buchanan, W., Altstatt, L., & Lombardo, J. P. (1971). Altruism is rewarding. *Science, 171*(3977), 1262-1263

Westhoff, B., Blankenstein, N. E., Schreuders, E., Crone, E. A., & van Duijvenvoorde, A. C. (2021). Increased ventromedial prefrontal cortex activity in adolescence benefits prosocial reinforcement learning. *Developmental Cognitive Neuroscience*, *52*, 101018.

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, *3*(2), 180-185.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics, 14*.

Wilhelm, M. O., & Bekkers, R. (2010). Helping behavior, dispositional empathic concern, and the principle of care. *Social Psychology Quarterly*, *73*(1), 11-32.

Worthington Jr, E. L., Wade, N. G., Hight, T. L., Ripley, J. S., McCullough, M. E., Berry, J. W., Schmitt, M. M., Berry, J. T., Bursley, K. H. & O'Connor, L. (2003). The Religious Commitment Inventory--10: Development, refinement, and validation of a brief scale for research and counseling. *Journal of Counseling Psychology*, *50*(1), 84.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129-166.

Wright, R. A., Shaw, L. L., & Jones, C. R. (1990). Task demand and cardiovascular response magnitude: Further evidence of the mediating role of success importance. *Journal of Personality and Social Psychology*, *59*(6), 1250-1260.

Zhou, Y., Han, S., Kang, P., Tobler, P. N., & Hein, G. (2024). The social transmission of empathy relies on observational reinforcement learning. *Proceedings of the National Academy of Sciences*, *121*(9), e2313073121.

Supplemental Materials

List of all measures included in Experiments 1, 2, and 3, but not reported in analyses

For each Experiment, we included 23 additional individual differences measures. We did not analyze data from those additional measures in the present article, because they are part of a broader project, and thus outside the scope of the current research. They are as follows:

Moral Identity Scale (MIS). The MIS includes both the five-item symbolization and five-item internalization subscales (Aquino and Reed, 2002).

Principle of Care scale (PCS). The PCS consists of eight items measuring predispositions towards care for those in need (Wilhelm and Bekkers, 2010)

General Trust scale (GTS). The GTS consists of five items measuring beliefs about one's trust of others (Yamagishi & Yamagishi, 1994).

HEXACO Personality Inventory. HEXACO includes the 16-item Honesty-Humility, 16-item Emotionality, 16-item Extraversion, 16-item Agreeableness, 16-item Conscientiousness, and 16-item Openness subscales (Ashton and Lee, 2007).

Social Dominance Orientation (SDO). The SDO includes 16 items measuring an individual's preference for social hierarchy and the extent to which they desire their own group to be superior to other groups (Pratto et al., 1994).

Balanced Inventory of Desirable Responding (BIDR). BIDR includes both the eight-item Self-Deceptive Enhancement and eight-item Impression Management subscales (Paulhus, 1998).

Religious Behaviors scale. The Religious Behaviors scale includes both the eight-item Internalization and eight-item Introjection subscales (Ryan, Rigby, & King, 1993).

Religious Commitment Inventory (RCI). The RCI consists of ten items measuring commitment behaviors related to religious thought and cognition (Worthington et al., 2003).

Volunteer Functions Inventory (VFI). The VFI includes the five-item Protective, five-item Values, five-item Career, five-item Social, five-item Understanding, and five-item Enhancement subscales (Clary et al., 1998).

Belief in a Just World Scale (BJW). The BJW includes both the seven-item Personal Belief in a Just World and six-item General Belief in a Just World subscales (Lucas et al., 2007).

Adapted Versions of the Team Prosocial Motivation (TPM) and Team Intrinsic Motivation (TIM) scales. We used adapted versions of the TPM and TIM scale using items that were selected from Hu & Liden (2015), which were themselves originally adapted from Grant (2008). Each scale consists of four items..

Organizational Citizenship Scale (OCS). The OCS includes the five-item Altruism, five-item Courtesy, five-item Conscientiousness, and five-item Sportsmanship subscales (Van Dyne, Graham, & Dienesch, 1994).

The Five-Item Guilt Proneness Scale (GP-5). The GP-5 includes five items measuring guilt proneness (Cohen, Kim, & Panter, 2014).

Personal Norm of Reciprocity Scale (PNR). The PNR includes both the nine-item Belief in Reciprocity and nine-item Positive Reciprocity subscales (Perugini et al., 2003).

Generative Behavior Checklist (GBC). The GBC includes 11 items measuring how often an individual engages in generative behaviors (McAdams & de St Aubin,1992).

Justice Sensitivity Inventory (JSI). The JSI consists of 16 items, and includes the four-item victim sensitivity subscale, the four-item observer sensitivity subscale, the four-item beneficiary sensitivity subscale, and the four-item perpetrator sensitivity subscale (Schmitt et al., 2010).

Parental Bonding Instrument (PBI). The PBI includes both the 12-item Care and 13-item Overprotection subscales (Parker et al., 1979).

Life Orientation Test (LOT). The LOT includes six items measuring expectations about future outcomes (Scheier & Carver, 1985).

Motivations to Help Scale (MHS). The MHS consists of ten items on two subscales: The Autonomous Motivations subscale, which includes five items, and the Controlled Motivations subscale, which includes five items (Weinstein & Ryan, 2010).

Social Value Orientation (SVO) Slider Measure. The SVO slider measure consists of six items measuring individual differences in prosocial orientation. The items are scored according to the criteria described by Murphy, Ackermann, and Handgraaf (2011) where subjects are classified as either broadly prosocial (altruists, cooperators, martyrs) or non-prosocial (individualists, competitors, sadists).

Brief Fear of Negative Evaluation (BFONE). The BFONE includes 12 items measuring people's concerns about negative evaluation (Leary, 1983).

Life History Questionnaire (LHQ). The LHQ consists of 32 questions measuring positive and negative childhood experiences (Mell et al., 2018).

Relational Mobility Scale (RMS). The RMS includes 12 items measuring beliefs about the ease with which a person in their culture can form new relationships (Schug, Yuki, & Maddux, 2010).

Experiment 1

Stimuli used for the Oxfam target in Experiment 1

The following is the text participants encountered with respect to Oxfam:

"You will now play for the Oxfam Charity.

Here is some more information about Oxfam from their website:

Oxfam is a global movement of people, working together to end the injustice of poverty. Working with partners, we use a combination of practical tactics and innovation to deliver development programs, public education, campaigns, advocacy and humanitarian assistance in disasters and conflicts. From sanitation and clean water to getting more girls into school, we won't stop until every person on the planet can enjoy life free from poverty.

Donations to Oxfam helps [sic] to save lives during a disaster, get clean water running in the most remote areas, send children, especially girls to school and stand up for the rights of women.

Visit this link if you're interested in finding out more information about Oxfam (but please wait until after the study is over so as not to lose your progress): www.oxfam.org/en"

Stimuli used for the identifiable target in Experiment 1

Out of respect for the privacy of the identifiable target, we do not include the actual stimuli used. Instead, we describe the identifiable target: A Black woman in her early 40's seeking financial stability and shelter for her family, and medical help.

Block position information

To ensure that the order in which participants encountered each of the four conditions (i.e., self, identifiable other, Oxfam, and anonymous other), we conducted four χ^2 goodness-of-fit tests, one for each block position, to determine if there were significant differences in the order in which each block was presented for each position (e.g., if the anonymous, identifiable, self, and charity blocks were presented frequently as often for the first block). All χ^2 goodness-of-fit tests were non-significant ($ps > .435$), indicating that there was no difference in the frequency with which each block position was presented.

We also computed the percentage of participants that viewed each condition with respect to each possible condition position (i.e., whether the condition appeared first, second, third, or fourth). Results are shown in Table S1.

Table S1

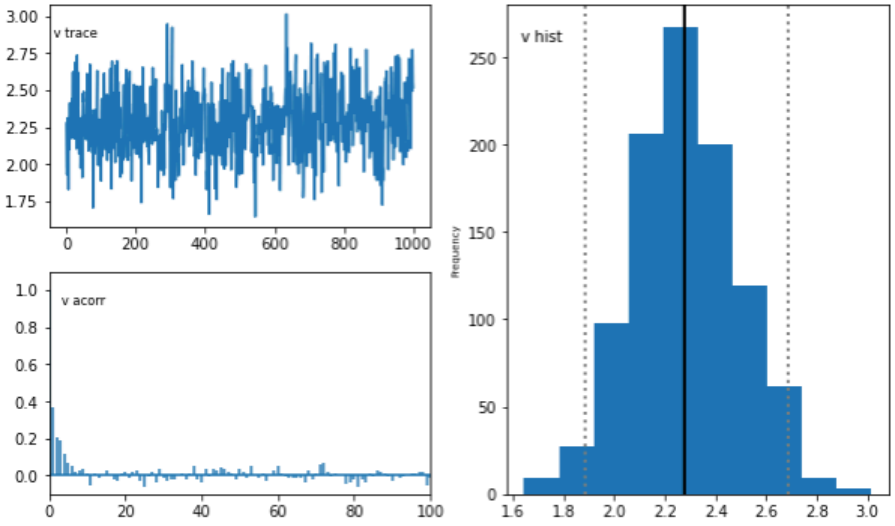
Percentage of Participants That Viewed Each Block Position In Experiment 1.

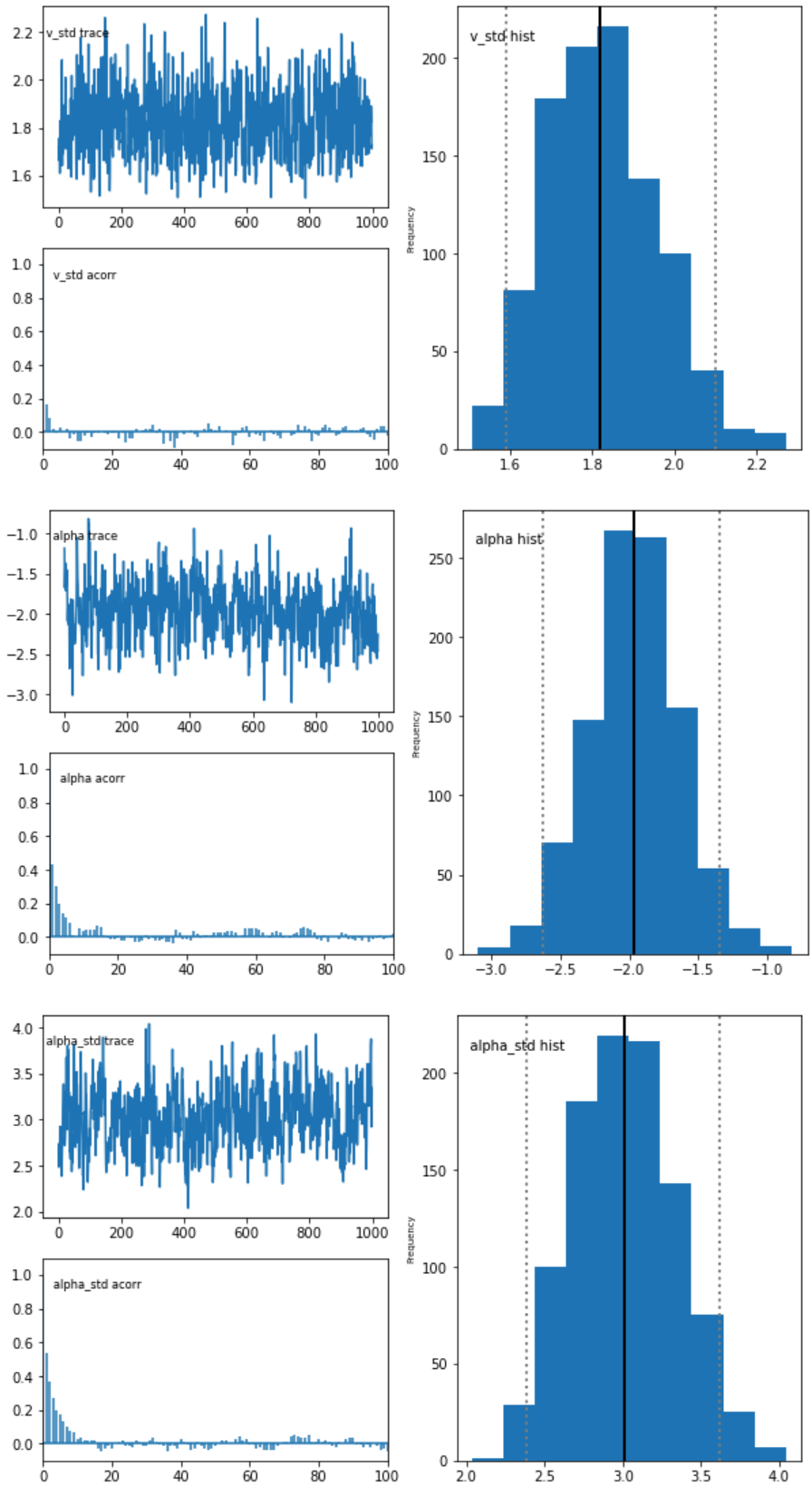
Block	Viewed first	Viewed second	Viewed third	Viewed fourth
Self	23.6%	24.7%	26.5%	25.1%
Identifiable	28.2%	20.9%	24.2%	26.5%
Oxfam	24.5%	25.6%	27%	22.8%
Anonymous	23.1%	28.8%	22.3%	25.6%

Convergence plots for model estimation in Experiment 1

Figure S1

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 1 In Which a Single α Parameter and a Single β Parameters Were Estimated Across the Four Conditions.

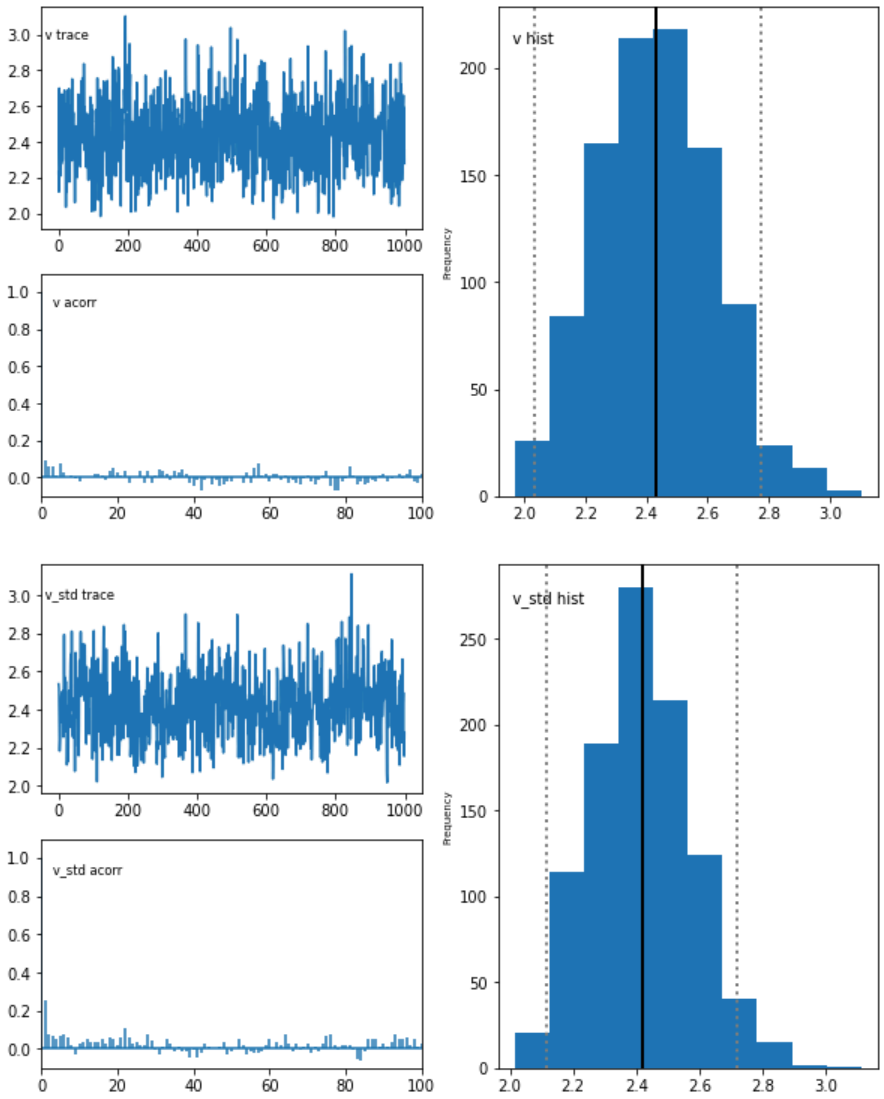


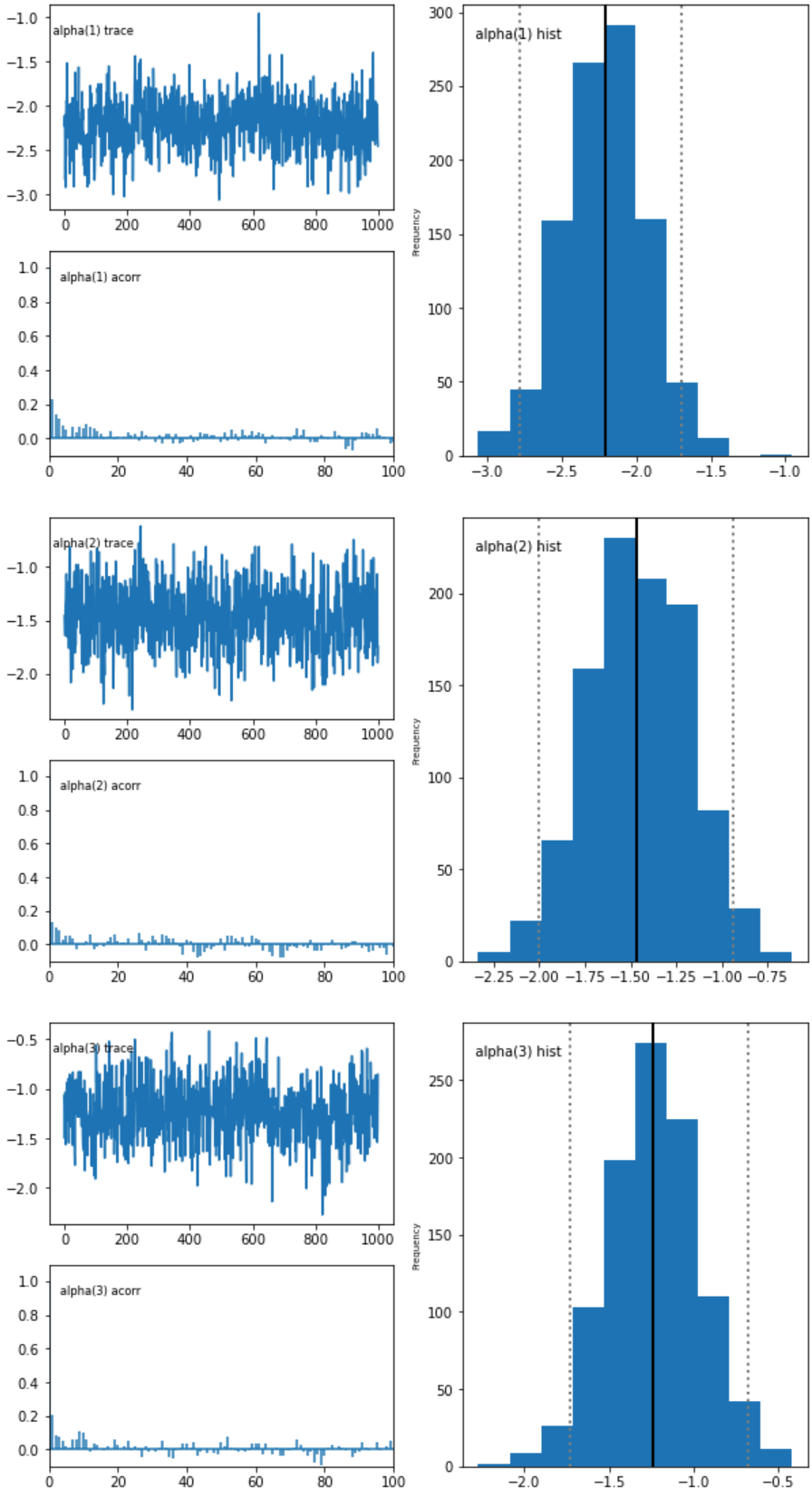

v_std trace
v_std acorr
v_std hist
alpha trace
alpha acorr
alpha hist
alpha_std trace
alpha_std acorr
alpha_std hist
Frequency

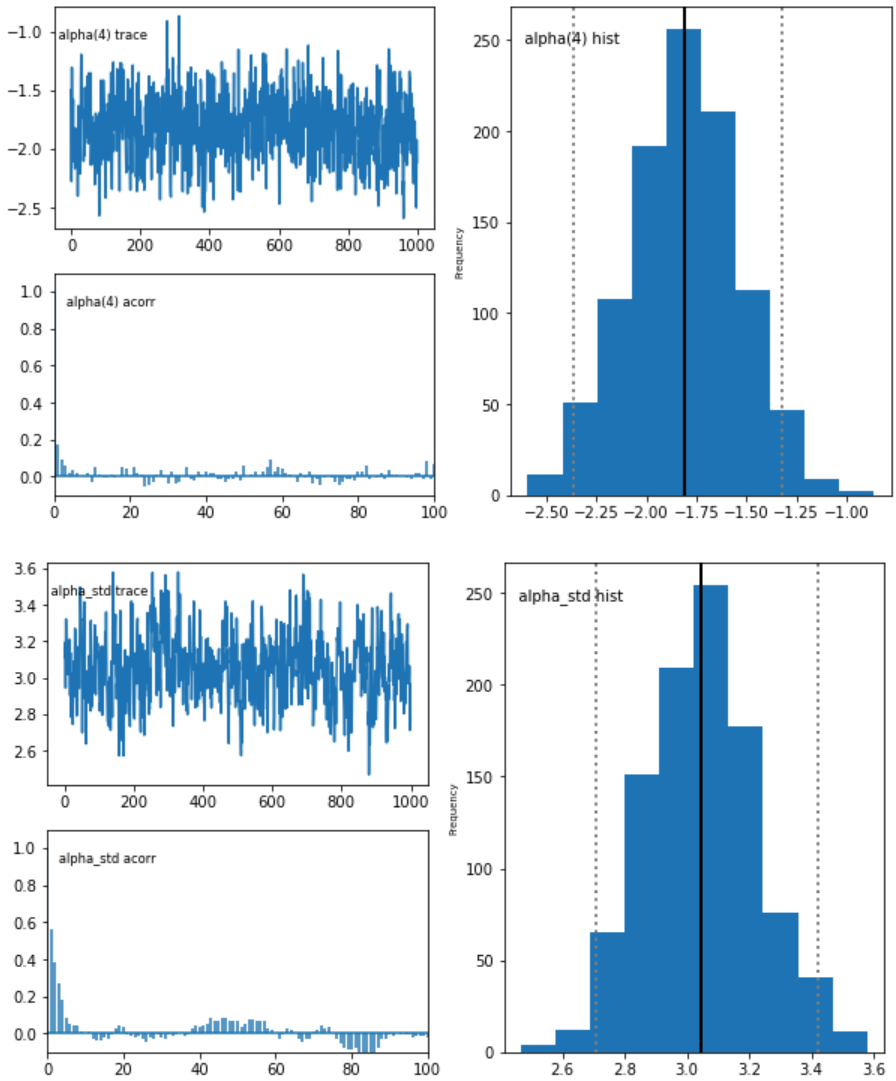
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha" refers to the α parameter, and "alpha_std" refers to the group variability for the α parameter.

Figure S2

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 1 In Which Separate α Parameters Were Estimated For Each of the Four Conditions, and a Single β Parameter Was Estimated Across the Four Conditions.



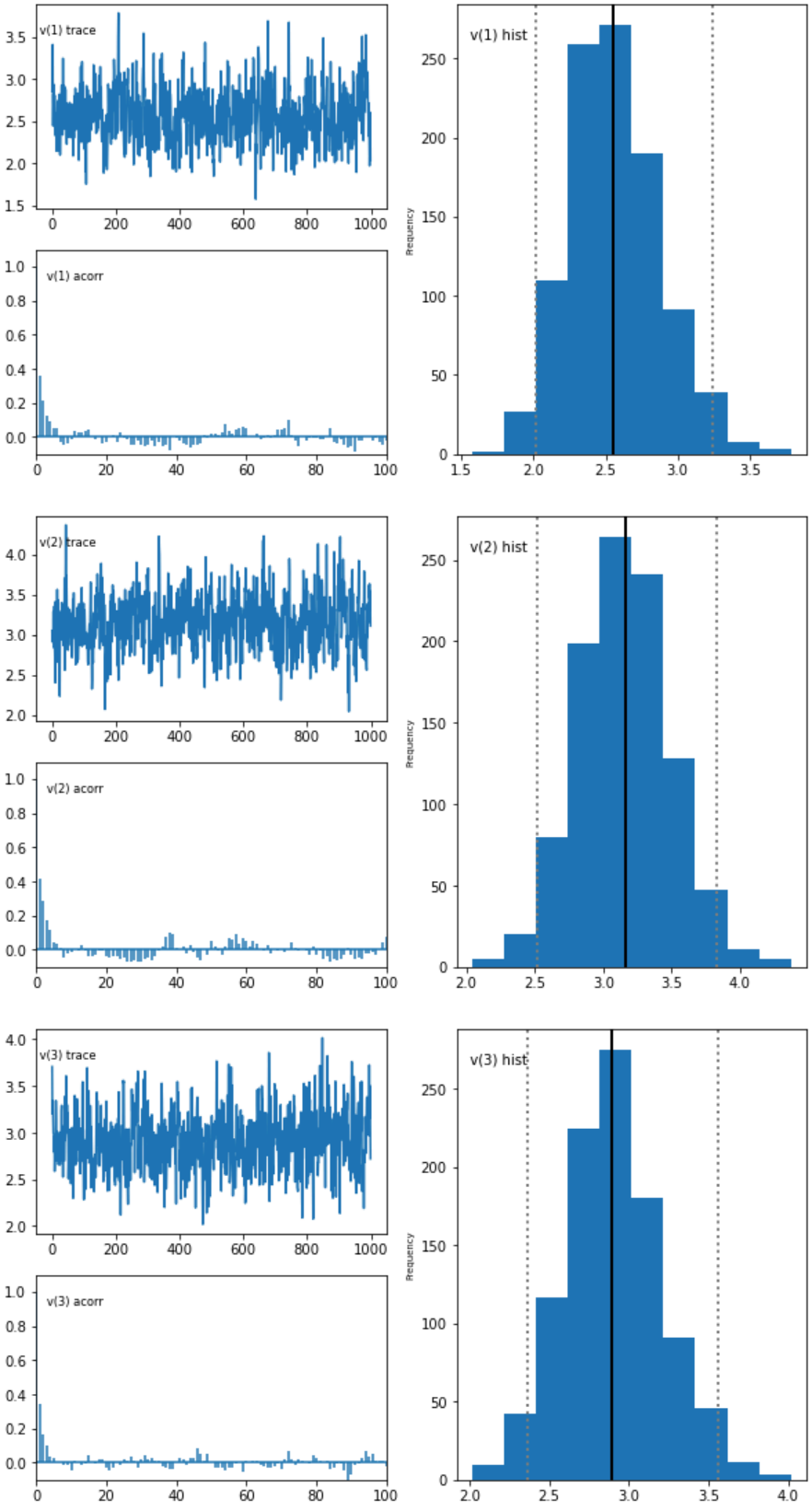

alpha(1) trace
alpha(1) acorr
alpha(1) hist
alpha(2) trace
alpha(2) acorr
alpha(2) hist
alpha(3) trace
alpha(3) acorr
alpha(3) hist
Frequency

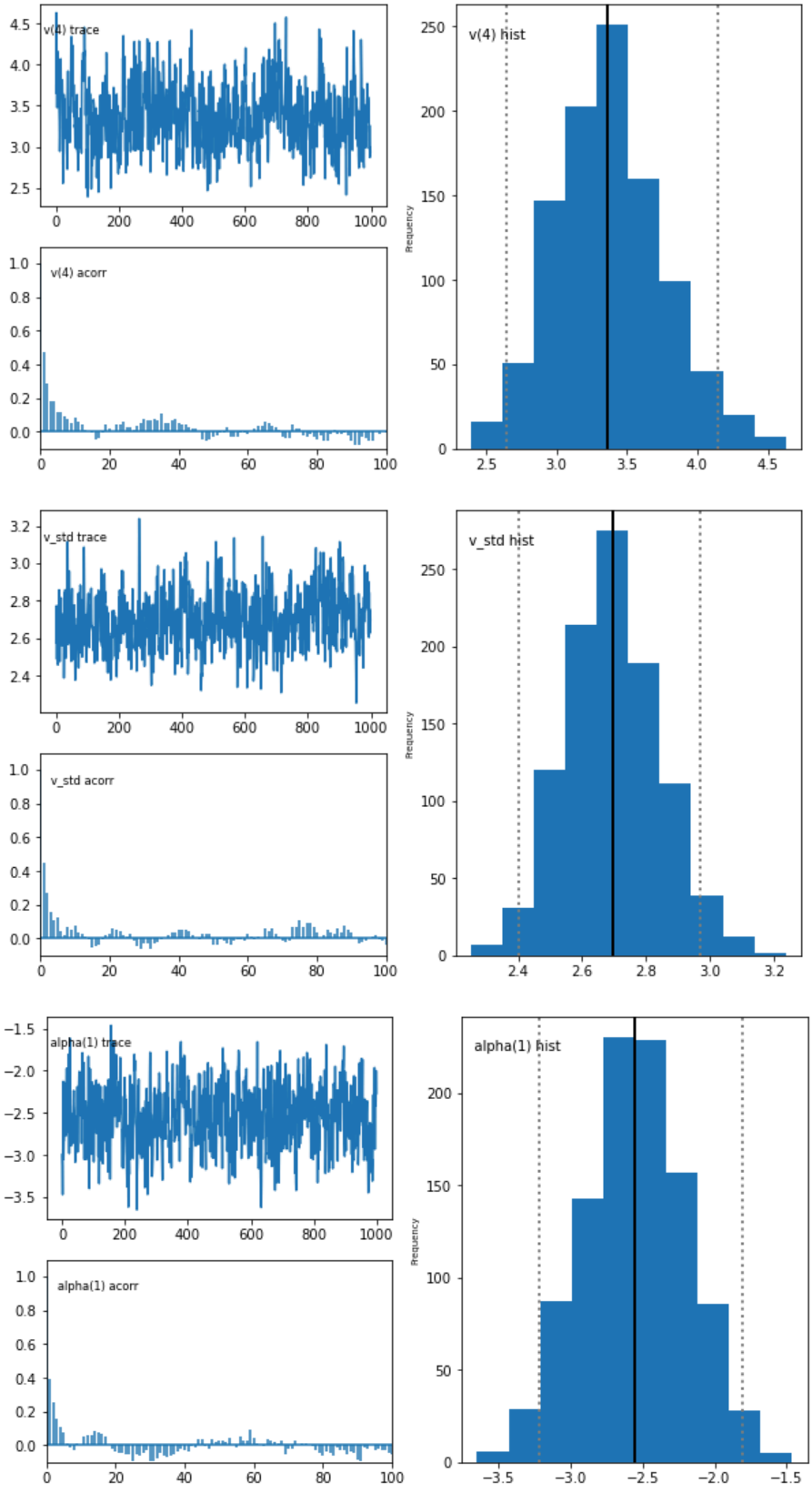


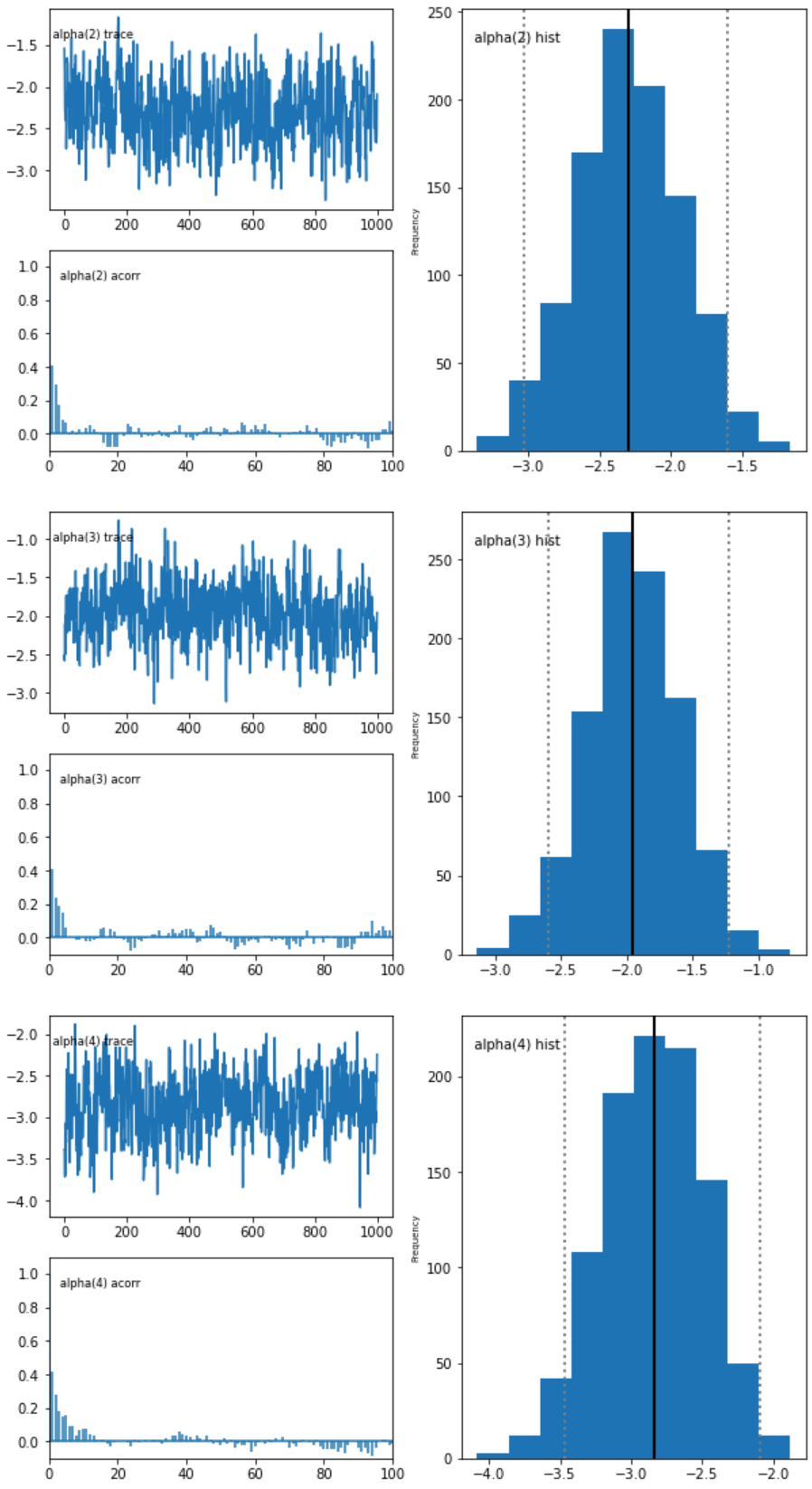
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the anonymous target, "alpha (2)" refers to the α parameter for the Oxfam target, "alpha (3)" refers to the α parameter for the self, "alpha (4)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

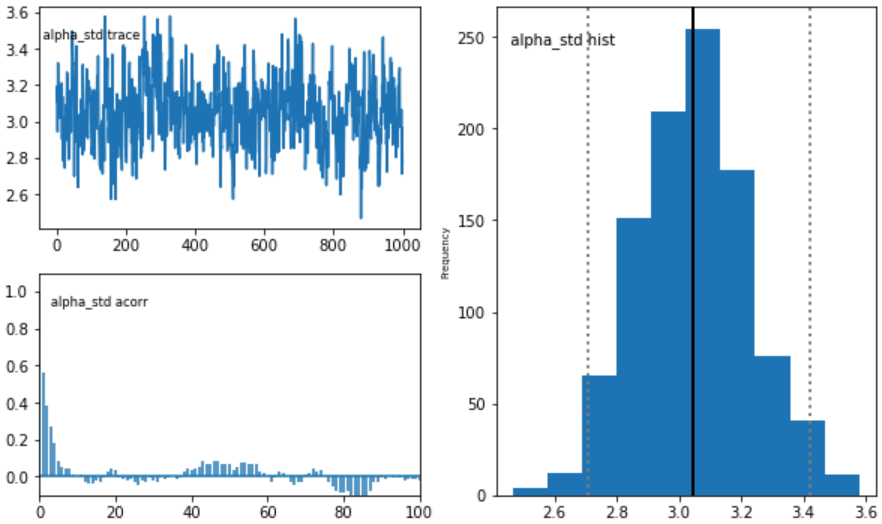
Figure S3

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 1 In Separate α Parameters and β Parameters Were Estimated For Each of the Four Conditions.


v(1) trace
v(1) acorr
v(1) hist
v(2) trace
v(2) acorr
v(2) hist
v(3) trace
v(3) acorr
v(3) hist
Frequency


v(4) trace
v(4) acorr
v(4) hist
Frequency
v_std trace
v_std acorr
v_std hist
Frequency
alpha(1) trace
alpha(1) acorr
alpha(1) hist
Frequency


alpha(2) trace
alpha(2) acorr
alpha(2) hist
alpha(3) trace
alpha(3) acorr
alpha(3) hist
alpha(4) trace
alpha(4) acorr
alpha(4) hist
Frequency



Note: "v (1)" refers to the β parameter for the anonymous target, "v (2)" refers to the β parameter for the Oxfam target, "v (3)" refers to the β parameter for the self, "v (4)" refers to the β parameter for the identifiable needy target, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the anonymous target, "alpha (2)" refers to the α parameter for the Oxfam target, "alpha (3)" refers to the α parameter for the self, "alpha (4)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

Parameter recovery analyses

We conducted a parameter recovery study to determine if we could recover the computational parameters from simulated data based on the observed data. Using the parameters from the observed data, we simulated n = 10 datasets. We then fit data from each of the simulated datasets to the winning model found for our observed data: A computational model in which separate α and β parameters were estimated for each of the four conditions, for a total of eight parameters estimated per participant for each of the models.

Figure S4 shows the αs for the observed data, compared to the αs for the simulated data. A visual inspection of the results shows that, when the simulated data were fit to the winning model, we were generally able to recover the parameters found in the observed data, although the αs in our data were slightly lower than the αs in the observed data. In other words, in the observed data, participants were faster at learning which symbol to select across all trials, which slightly depressed α since there was less for participants to learn on later trials. Symmetrically, we find that the βs slightly overestimated the degree to which participants explored the rewardinginess of different response options.

Figure S4

α Parameters for the Observed and Simulated Datasets in Experiment 1.

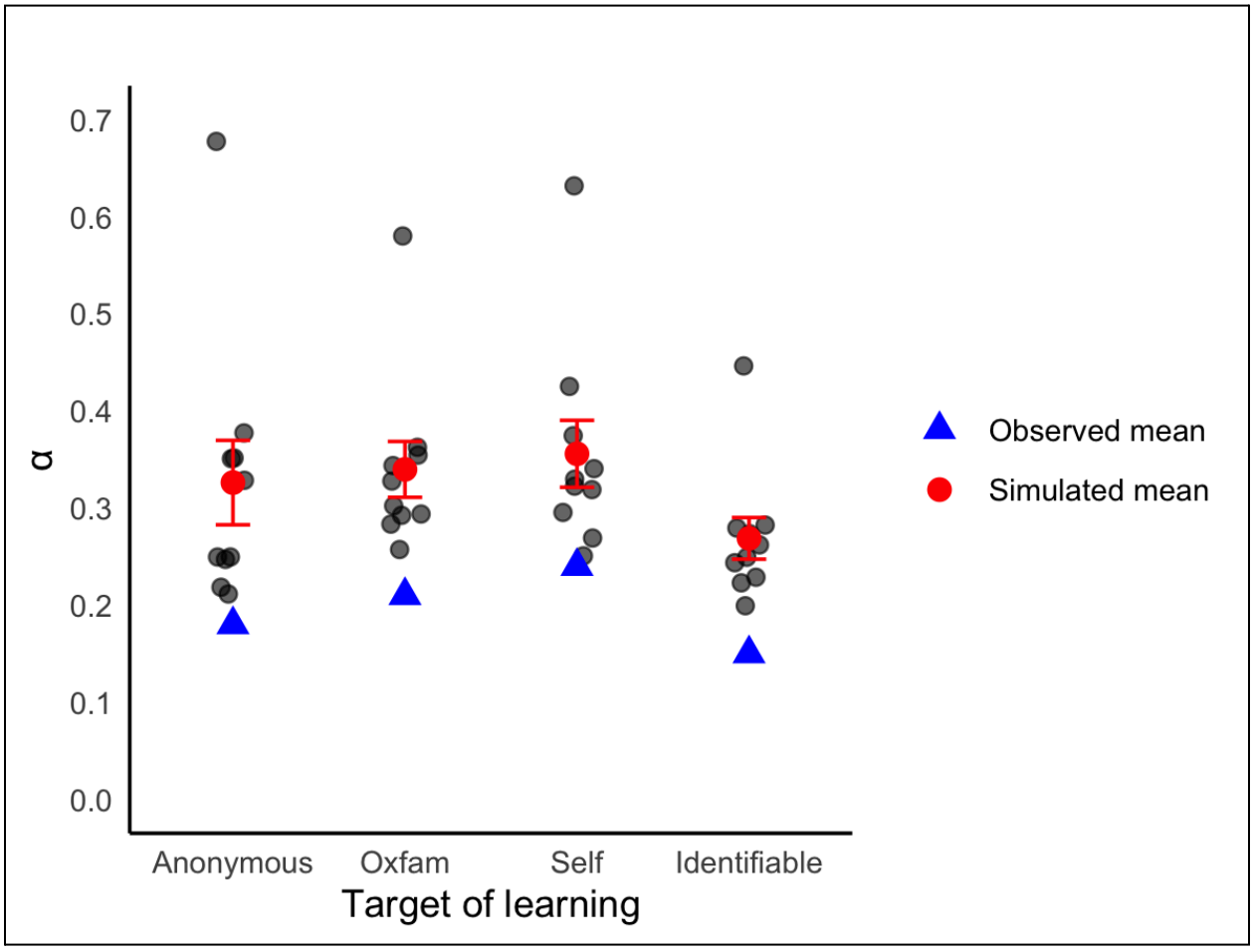
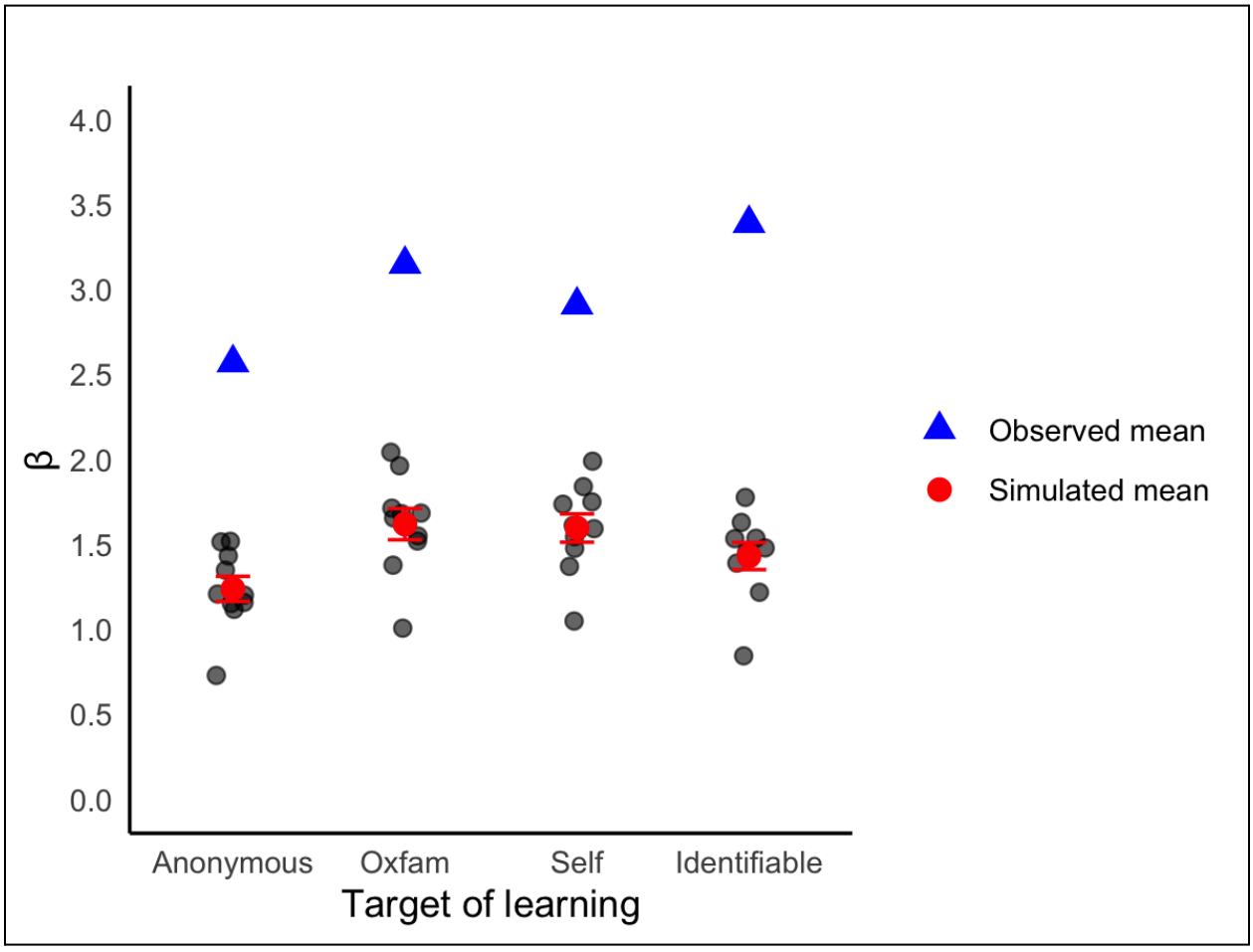


Figure S5

β Parameters for the Observed and Simulated Datasets in Experiment 1.



Posterior predictive check

We conducted a posterior predictive check as a further test of computational model validity. Using the $n = 10$ simulated datasets generated during the parameter recovery study, we computed summary statistics for two variables: (1) The choice participants made in each condition (0 = infrequently rewarding symbol, 1 = frequently rewarding symbol), and (2) Whether participants earned rewards on each trial (0 = no reward, 1 = reward). Results are shown in Tables S2 and S3. The results suggest that participants in the observed data marginally outperformed the simulated cases with respect to the choices participants made to select the frequently rewarding symbol, and the frequency with which participants earned rewards on trials.. These results were also consistent across the different targets that participants learned for.

Table S2
Summary Statistics for Frequency of Selecting Rewarding Symbol in the Observed and Replicated Datasets in Experiment 1.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.64	0.48	[0.63, 0.65]	0.005
	Identifiable	0.63	0.48	[0.63, 0.64]	0.004
	Oxfam	0.65	0.48	[0.64, 0.65]	0.005
	Anonymous	0.61	0.49	[0.60, 0.62]	0.005
Simulated	Self	0.60	0.005	[60, 0.61]	0.001
	Identifiable	0.59	0.005	[0.59, 0.59]	0.001
	Oxfam	0.60	0.004	[0.60,0.61]	0.001
	Anonymous	0.58	0.004	[0.58, 0.58]	0.001

Note: SD = standard deviation. SEM = standard error of the mean.

Table S3
Summary Statistics for Frequency of Earning Reward in the Observed and Replicated Datasets in Experiment 1.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.56	0.50	[0.55, 0.57]	0.005
	Identifiable	0.56	0.50	[0.57, 0.59]	0.005
	Oxfam	0.56	0.50	[0.55, 0.57]	0.005
	Anonymous	0.55	0.50	[0.54, 0.56]	0.005
Simulated	Self	0.55	0.007	[0.54, 0.55]	0.002
	Identifiable	0.54	0.006	[0.54,0.55]	0.002
	Oxfam	0.55	0.006	[0.54, 0.55]	0.002
	Anonymous	0.54	0.004	[0.54, 0.54]	0.001

Note: SD = standard deviation. SEM = standard error of the mean.

GLMM with rewardingness of the selection as the dependent variable

We estimated a GLMM that was identical to the first model, except that the dependent variable was the rewardingness of the selection for each trial, regardless of whether participants

selected the frequently or infrequently rewarding symbol (0 = selection was not rewarded, 1 = selection was rewarded). Predictors included the trial repetition number (a level-1 predictor, with values ranging from 1-16), and *k*-1 dummy-coded predictors that reflected the target whom participants learned for, with the anonymous target serving as the reference group (a level-2 predictor). We also included random intercepts for each participant. The results were qualitatively identical to the model in which the frequency/infrequency of the symbol served as the dependent variable ($b = 0.017$, $SE = 0.002$, *95% CI* = [0.012, 0.021], $Z = 7.11$, $p < .001$; *Odds ratio (OR)* = 1.017, *95% CI* = [1.012, 1.021]), indicating that participants not only learned to select the more rewarding symbol over the course of each block, but also became more competent at earning rewards.

Analyses involving the β parameters

We conducted a second one-way within-subjects ANOVA that featured the β parameters as the dependent variable. There was a significant main effect for condition assignment ($F(3, 630) = 12.97$, $p < .01$, $\eta_p^2 = 0.058$, *95% CI* = [0.025, 0.093]), indicating that participants exhibited different exploration rates across conditions. Follow-up pairwise *t*-tests with Bonferroni-corrected *p*-values indicated that there were significant differences in the β parameters between the Oxfam ($M = 3.15$, $SD = 1.68$) and anonymous ($M = 2.57$, $SD = 1.80$) conditions ($t(213) = 4.10$, $p < .01$, *Cohen's D* = 0.28, *95% CI* = [0.14, 0.40]), the identifiable needy target ($M = 3.39$, $SD = 1.91$) and anonymous conditions ($t(212) = 5.61$, $p < .01$, *Cohen's D* = 0.39, *95% CI* = [0.25, 0.52]), and the identifiable needy target and self ($M = 2.91$, $SD = 1.74$) conditions ($t(211) = 3.55$, $p < .01$, *Cohen's D* = 0.24, *95% CI* = [0.11, 0.40]) (Figure S7).

Finally, we examined the correlations among the α and the β parameters in each condition, as well as the correlations that the α and the β parameters shared with each other. A correlation table of the results is shown in Table S4. All correlations were significant (ps < .01), save for the correlation between the α for the identifiable needy target and the β parameter for the identifiable needy target ($p = .109$). The α and β parameters were correlated because they are intrinsically related. Increases in α will necessarily reduce Beta, because α reflects the learner homing in on the correct symbol, so that learners consistently select the same symbol, while β reflects shifts away from selecting the same symbol. In other words, participants who randomly explore the rewardingness of the two symbols will less consistently select the most rewarding symbol.

Figure S7

Means and standard errors for the β parameters in each condition.

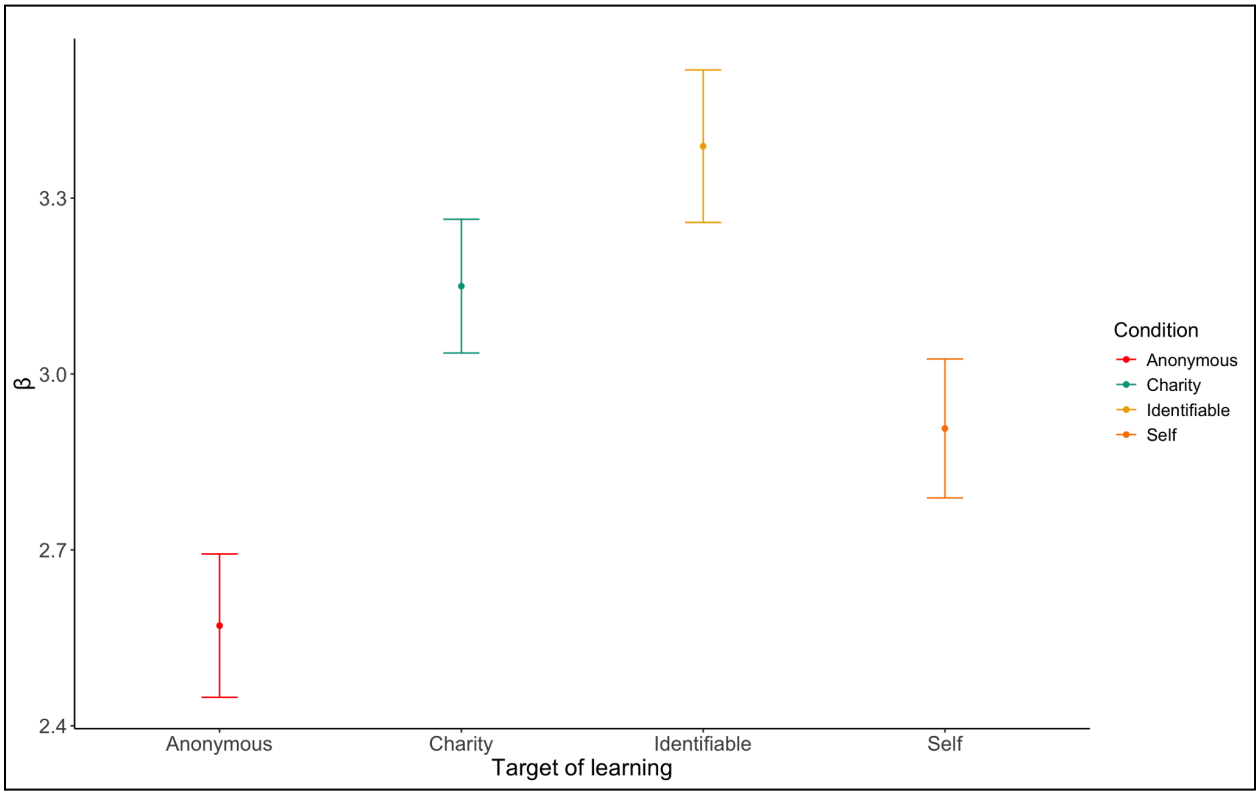


Table S4

Correlations and 95% CIs Among the α and β Parameters in Experiment 1.

Variable	1.	2.	3.	4.	5.	6.	7.
1. $\alpha_{anonymous}$							
3. $\alpha_{Identifiable}$	.44**	.50**					
4. α_{Self}	.38**	.51**	.42**				
5. $\beta_{Anonymous}$	.31**	.25**	.27**	.33**			
6. β_{Oxfam}	.30**	.31**	.34**	.31**	.31**		
7. $\beta_{Identifiable}$	.29**	.26**	.11	.28**	.34**	.44**	

8. β_{Self}	.25**	.30**	.16*	.34**	.27**	.43**	.45**

Note. * indicates $p < .05$. ** indicates $p < .01$.

Bayes factors for the Experiment 1 results

We conducted a Bayesian analysis of the hypotheses reported in Experiment 1. A Bayesian analysis can reveal which of two hypotheses (in this case, the null hypothesis that the correlation between trait empathic concern and learning is equal to zero, versus the alternative hypothesis that the correlation is larger than zero.

The association between trait empathic concern and $\alpha_{Identifiable,}$ the analysis yielded a Bayes factor of 0.23. According to conventional interpretations of Bayes factors (Jeffreys, 1961), this result suggests that the data provide more support for the null hypothesis than for the alternative hypothesis. Specifically, the observed data are approximately 4.31 times more likely under the null hypothesis than under the alternative hypothesis. Results for all analyses, along with their interpretation, shown in Table S5.

Table S5
Bayesian Analyses for the Experiment 1 Hypotheses.

Analysis	Bayes factor	Interpretation
Trait empathic concern correlation with $\alpha_{Identifiable}$	0.23	Data are approximately 4.31 times more likely under the null hypothesis.
Trait empathic concern correlation with α_{Oxfam}	0.33	Data are approximately 3.03 times more likely under the null hypothesis.
Trait empathic concern correlation with $\alpha_{Anonymous}$	0.22	Data are approximately 4.47 times more likely under the null hypothesis.
Trait empathic concern correlation with α_{Self}	0.19	Data are approximately 5.26 times more likely under the null hypothesis.

Experiment 2

Stimuli used for the identifiable target in Experiment 2

Participants learned to earn rewards for four different identifiable targets. Here is information about each of those targets. Target 1: A Black woman in her 30's seeking money to help find permanent housing for her and her children. Target 2: A Black woman in her 50's whose apartment recently flooded, damaging her laptop, and who is seeking money for a laptop and other necessities so that she can finish her schooling program. Target 3: A Black man in his 50's seeking money for medical services related to debilitating arthritis. Target 4: A White woman in her 30's who is currently homeless, and seeking money for rent.

Empathy and prosocial learning towards targets

To test whether participants expressed different amounts of state empathic concern for the four identifiable targets, we regressed the amount of state empathy participants reported on dummy-coded variables for each of the targets, with Target 3 serving as the reference condition. Participants who learned about Target 1 (M = 5.08, SD = 1.34) reported significantly more feelings of empathy, compared to Target 3 (M = 4.33, SD = 1.53; B = 0.23, b = 0.74, SE = 0.19, p < .001). There was no significant effect difference between the empathy reported for Target 2 (M = 4.68, SD = 1.40; p = .081) or Target 4 (M = 4.69, SD = 1.41; p = .071).

We then re-ran the analysis and included the prosocial learning rate as the dependent variable. There was no significant difference among any of the targets in terms of the prosocial learning rates (ps > .071).

Block position information

To ensure that the order in which participants encountered each of the two conditions (i.e., self and identifiable other), we conducted a χ^2 goodness-of-fit tests to determine if there were significant differences in the order in which the self vs. other blocks were presented for the first position (we did not conduct a second χ^2 test, since it would be redundant with the first test). The χ^2 goodness-of-fit test was non-significant (p = .265), indicating that there was no difference in the frequency with which the self vs. identifiable block position was presented. Results for the percentage of participants that viewed each condition with respect to each possible condition position shown in Table S4.

Table S6

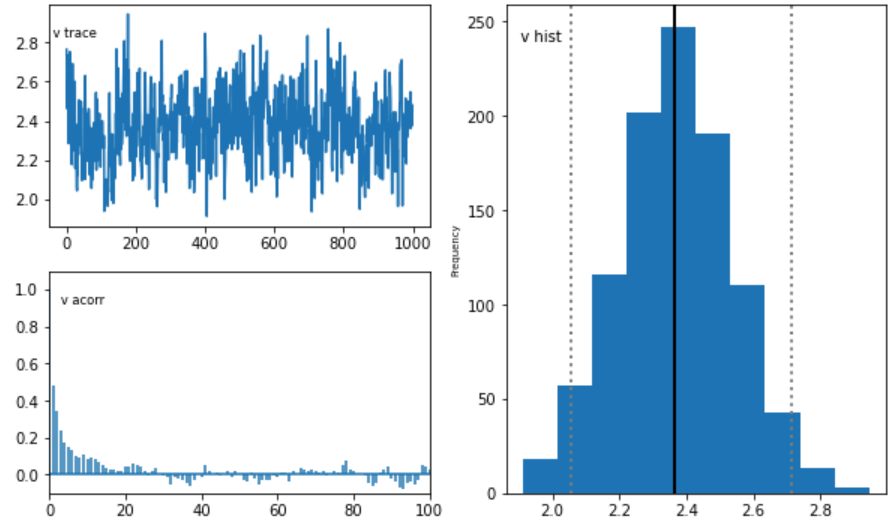
Percentage of Participants That Viewed Each Block Position In Experiment 2.

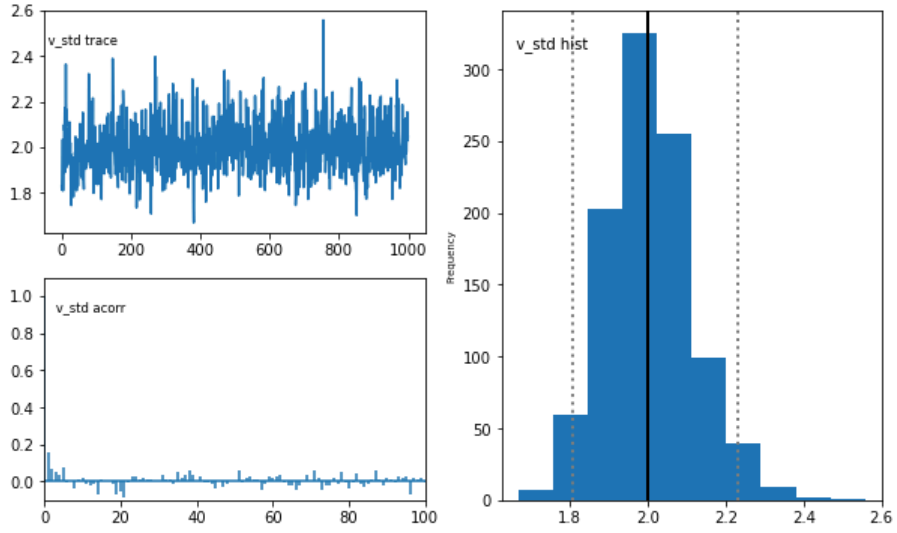
Block	Viewed first	Viewed second
Self	52.7%	47.3%
Identifiable	47.3%	52.7%

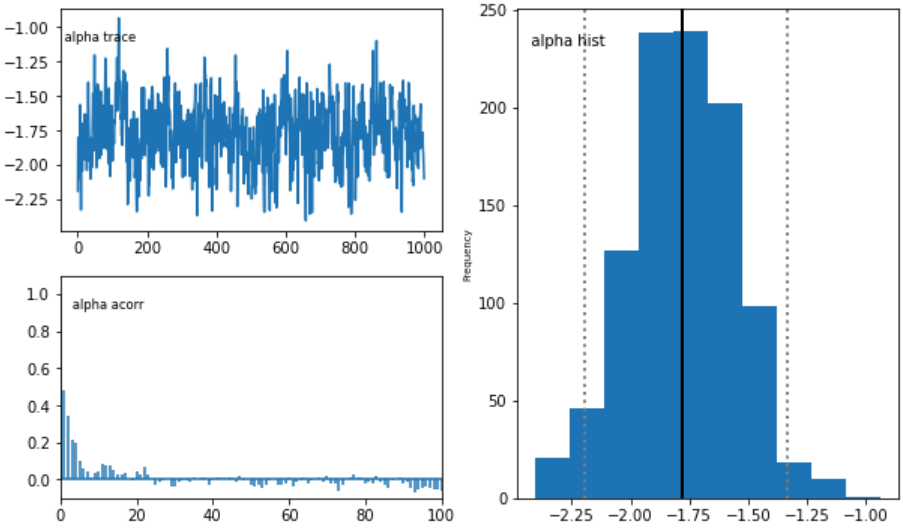
Convergence plots for model estimation in Experiment 2

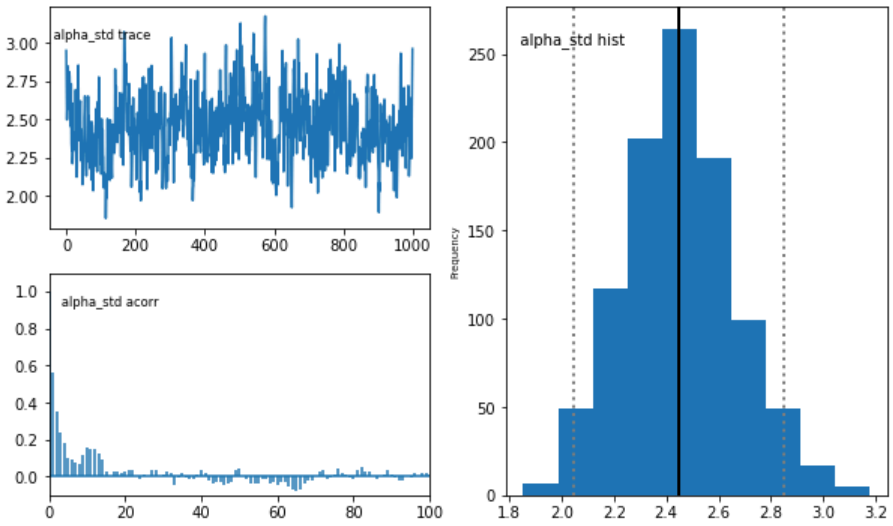
Figure S8

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 2 In Which a Single α Parameter and a Single β Parameter Were Estimated Across the Self and Identifiable Needy Target Conditions.


v trace
v acorr
v hist
Frequency


v_std trace
v_std acorr
v_std hist
Frequency

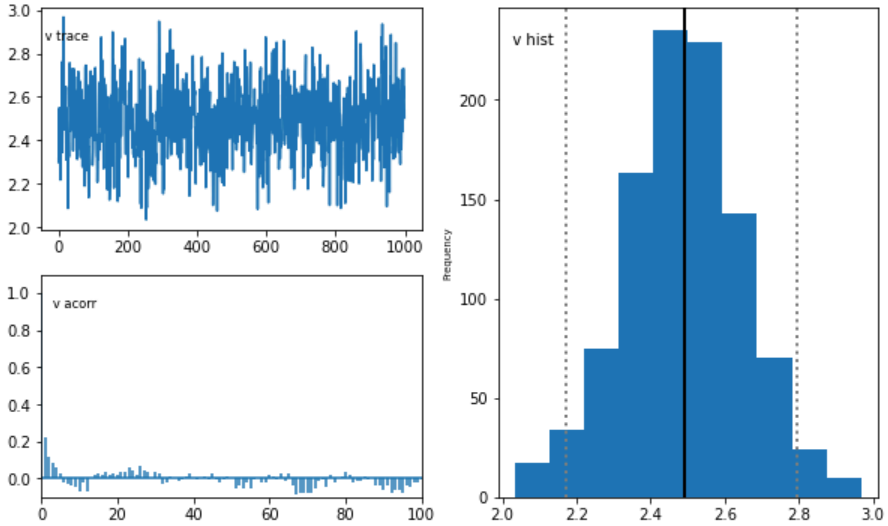

alpha trace
alpha acorr
alpha hist
Frequency

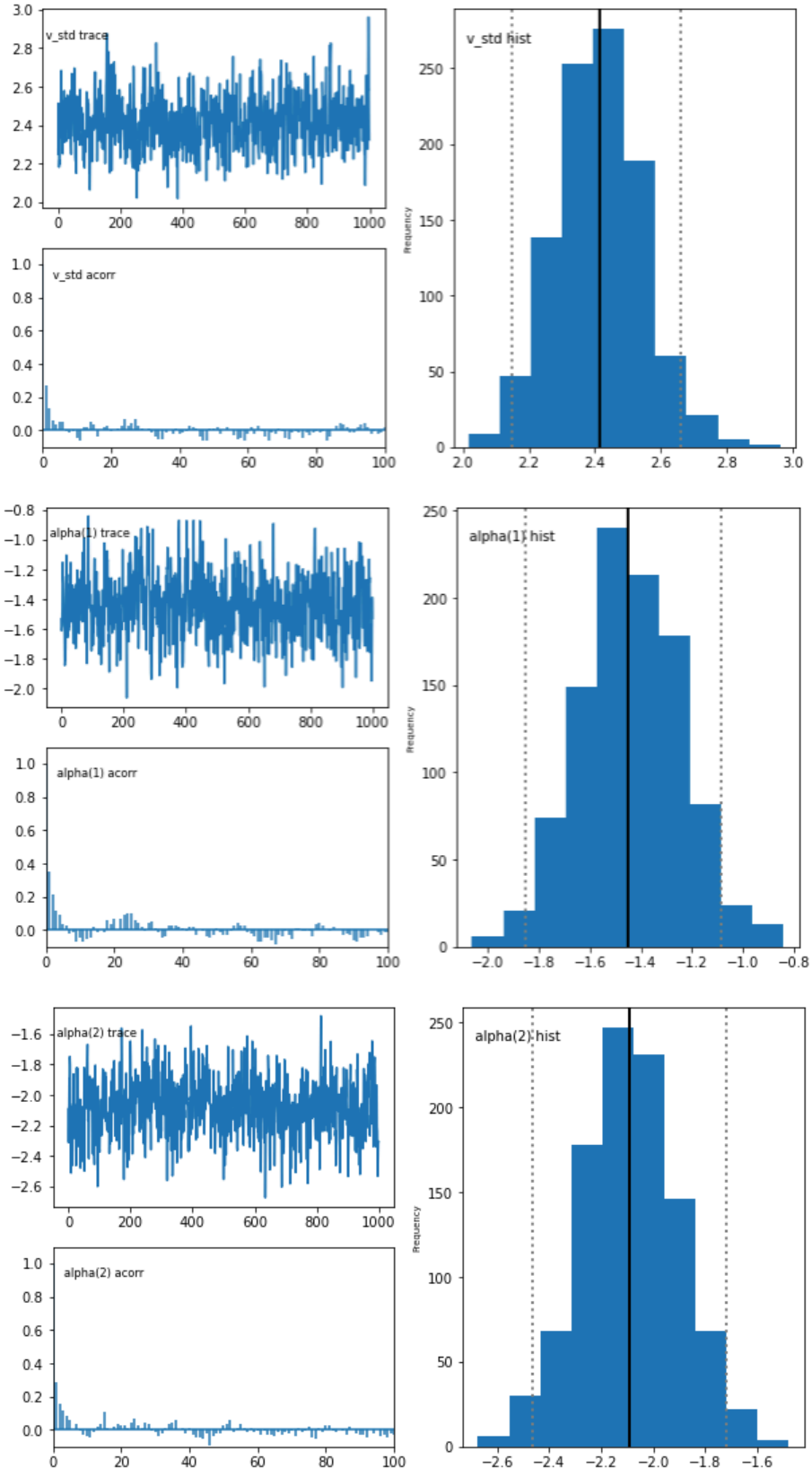


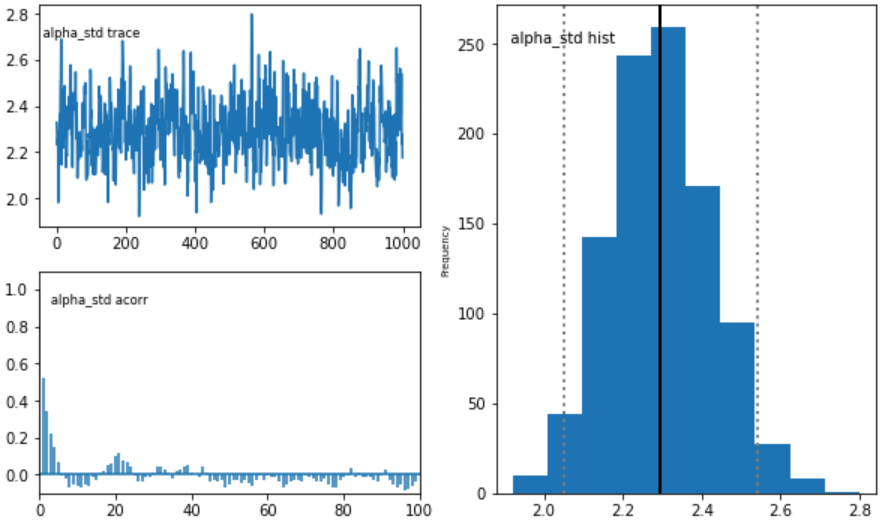
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha" refers to the α parameter, and "alpha_std" refers to the group variability for the α parameter.

Figure S9

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 2 In Which Separate α Parameters Were Estimated For Each of the Self and Identifiable Needy Target Conditions, and a Single β Parameter Was Estimated For Each of the Self and Identifiable Needy Target Conditions.



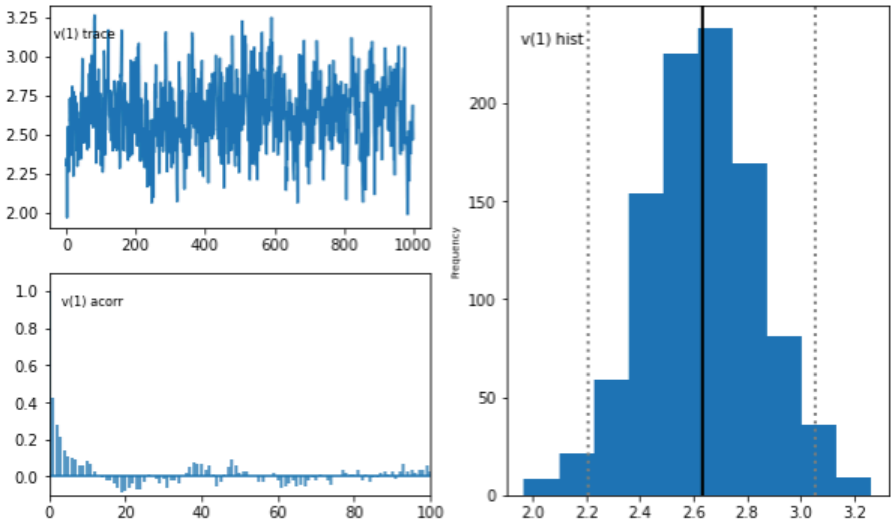

v_std trace
v_std acorr
v_std hist
alpha(1) trace
alpha(1) acorr
alpha(1) hist
alpha(2) trace
alpha(2) acorr
alpha(2) hist
Frequency

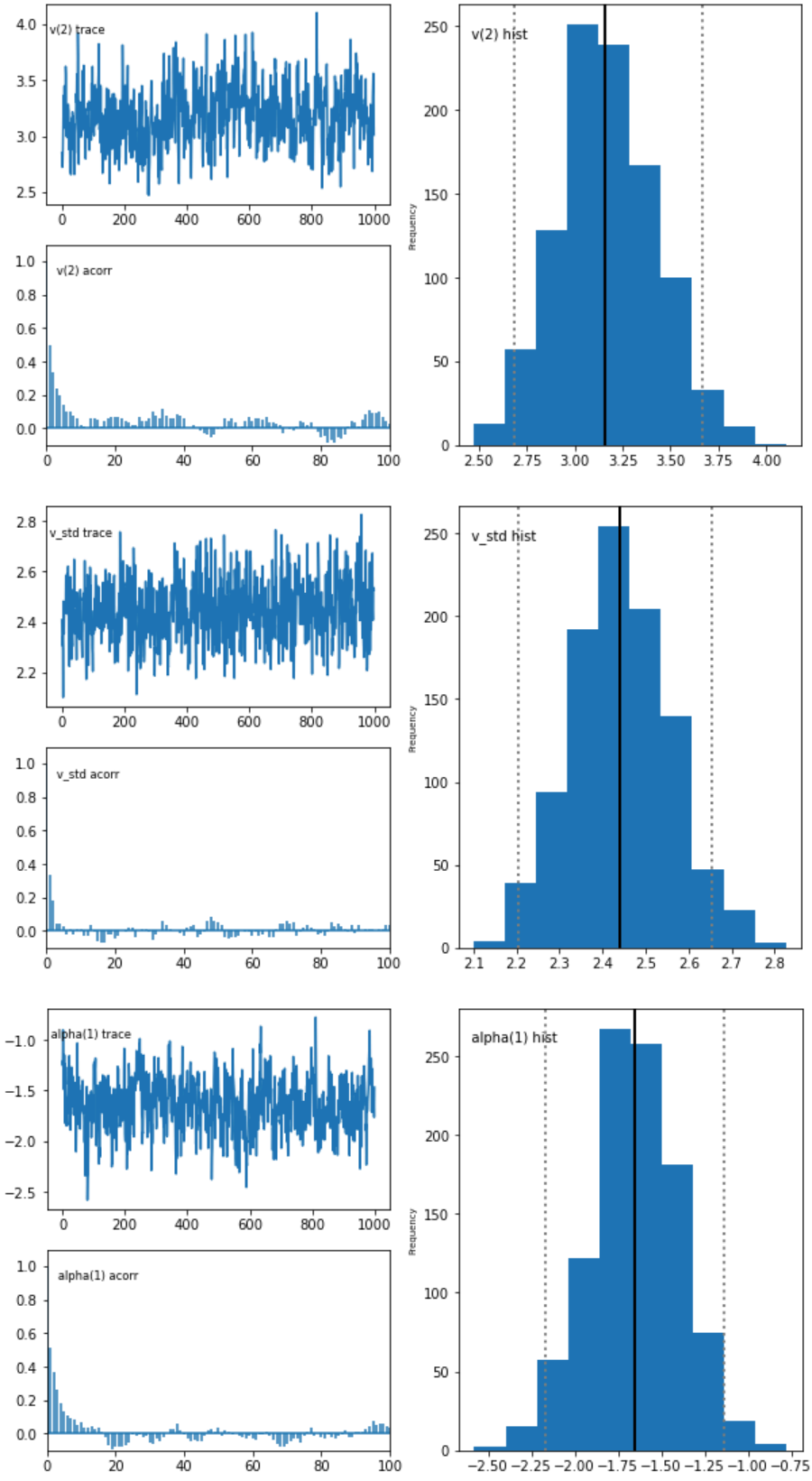


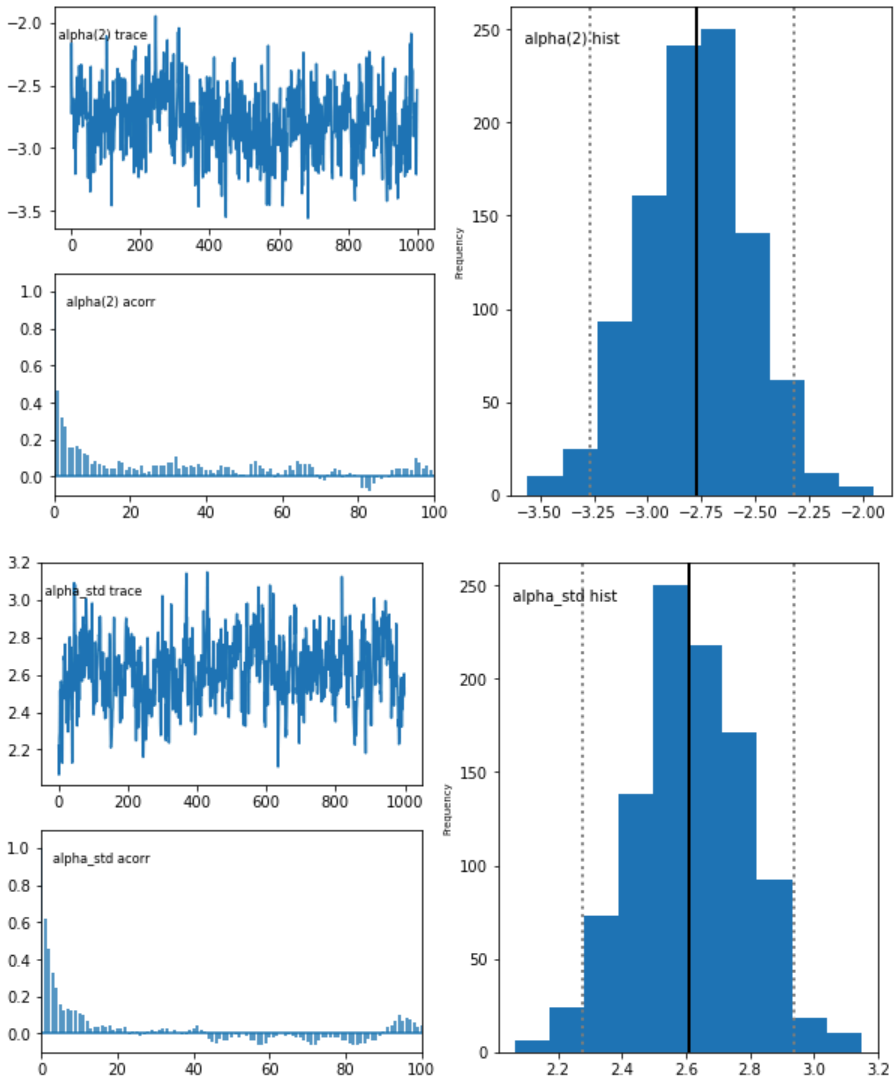
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the self, "alpha (2)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

Figure S10

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 2 In Separate α and β Parameters Were Estimated For Each of the Self and Identifiable Needy Target Conditions.




v(2) trace
v(2) acorr
v(2) hist
v_std trace
v_std acorr
v_std hist
alpha(1) trace
alpha(1) acorr
alpha(1) hist
Frequency



Note: "v (1)" refers to the β parameter for the self, "v (2)" refers to the β parameter for the identifiable needy target, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the self, "alpha (2)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

Parameter recovery analyses

We simulated $n = 10$ datasets based on the observed data from Experiment 2. Figure S12 shows the trial-by-trial responses for each of the two experimental conditions for the observed and simulated datasets. As in Experiment 1, the simulated data generally tracked the observed data, with the simulated results again slightly underestimating the frequency with which participants selected the more frequently rewarding symbol, compared to the observed results.

We then fit data from each of the simulated datasets to the winning model found for our observed data: A computational model in which separate α and β parameters were estimated for each of the two conditions, for a total of four parameters estimated per participant, per model. Figures S13 and S14 show the αs and βs for the observed data, compared to the αs and βs for the

simulated data. A visual inspection of the results shows that, when the simulated data were fit to the winning model, we were generally able to recover the parameters found in the observed data, with the αs slightly overestimated, and the βs slightly underestimated.

Figure S11

α Parameters for the Observed and Simulated Datasets in Experiment 2.

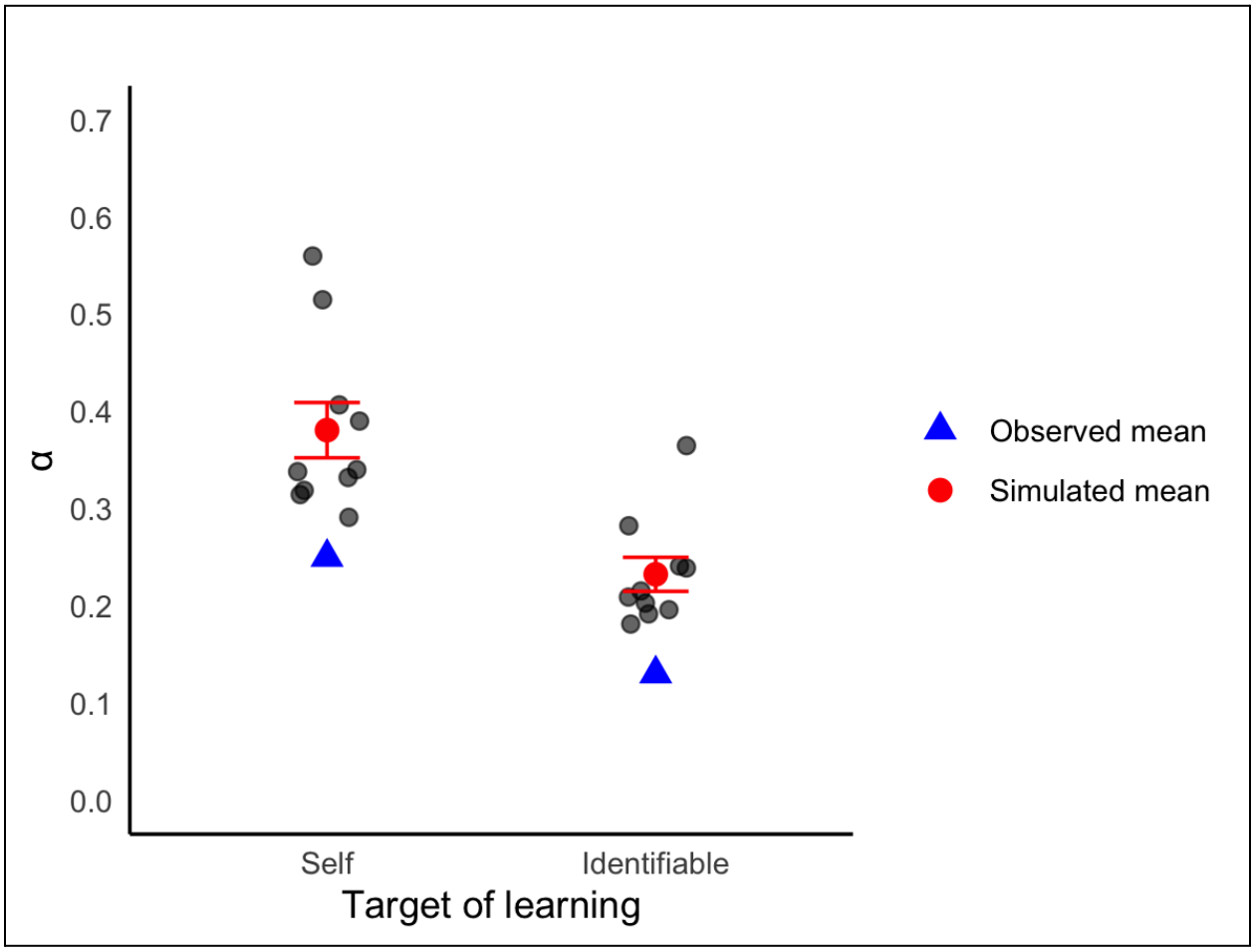
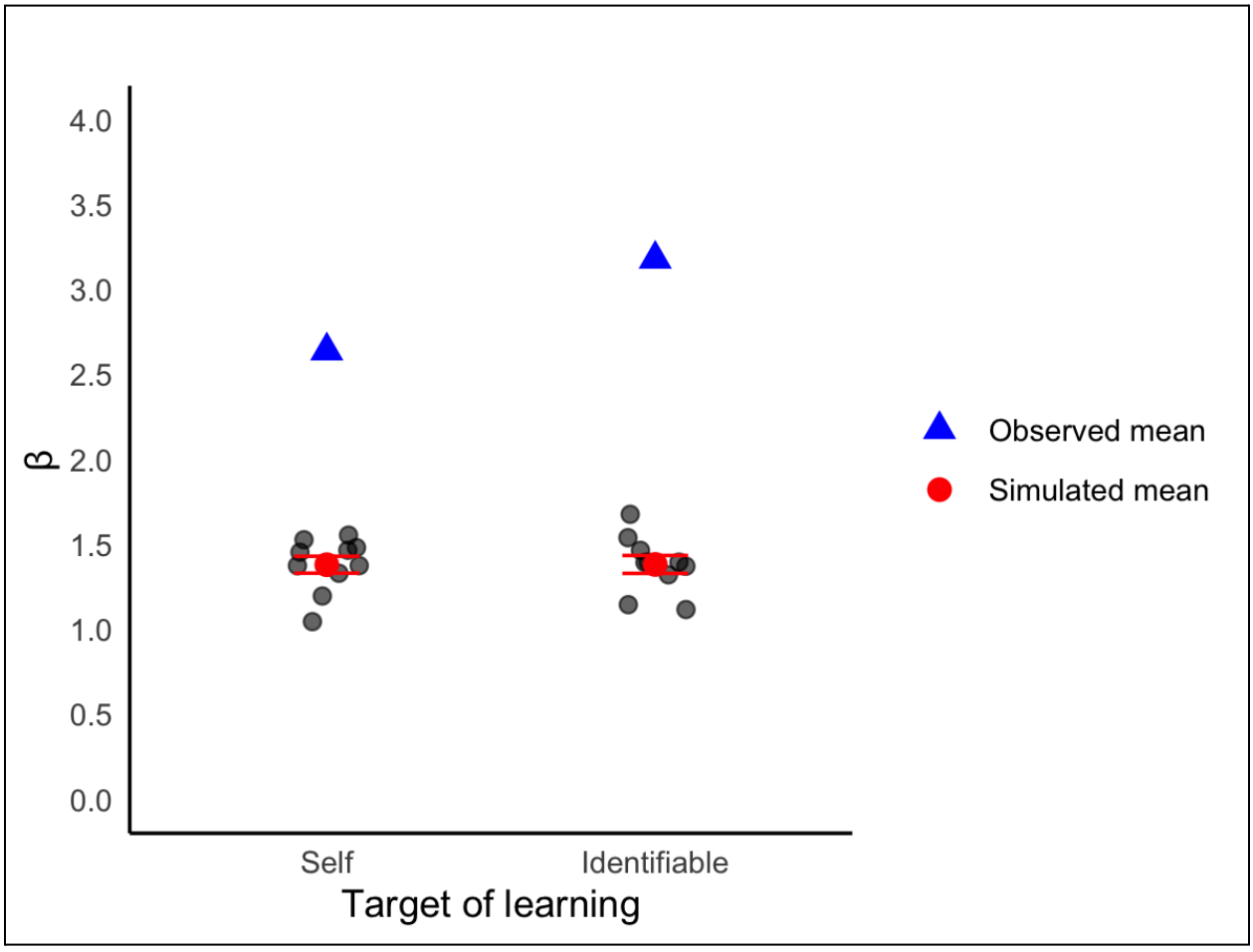


Figure S12

β Parameters for the Observed and Simulated Datasets in Experiment 2.



Posterior predictive check

We conducted a posterior predictive check using identical procedures as we did in Experiment 1. Results are shown in Figure S13 and Tables S6-S7. As in Experiment 1, the results were mostly similar for the observed and simulated data, with the observed participants selecting the rewarding symbol, and earning rewards, more frequently than cases in the simulated data.

Figure S13

Trial-By-Trial Response Data for the Observed and Simulated Datasets in Experiment 2.

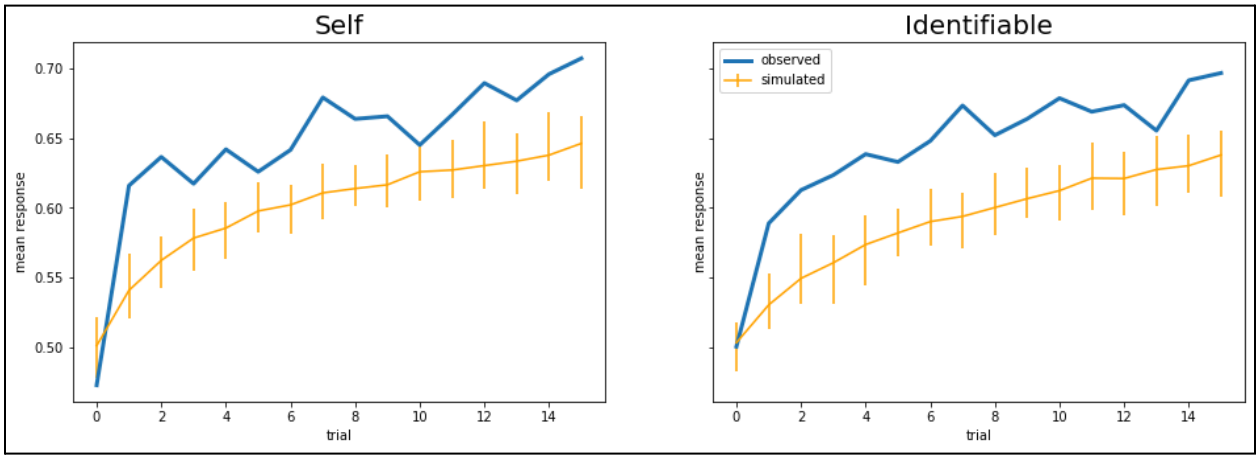


Table S6

Summary Statistics for Frequency of Selecting Rewarding Symbol in the Observed and Replicated Datasets in Experiment 2.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.64	0.48	[0.64, 0.65]	0.003
	Identifiable	0.64	0.48	[0.64, 0.65]	0.003
Simulated	Self	0.59	0.004	[0.60, 0.60]	0.0005
	Identifiable	0.59	0.003	[0.59, 0.59]	0.0005

Note: SD = standard deviation. SEM = standard error of the mean.

Table S7

Summary Statistics for Frequency of Earning Reward in the Observed and Replicated Datasets in Experiment 2.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.56	0.50	[0.56, 0.57]	0.004
	Identifiable	0.58	0.49	[0.58, 0.59]	0.004
Simulated	Self	0.55	0.004	[0.55, 0.55]	0.0007
	Identifiable	0.55	0.004	[0.54, 0.55]	0.0006

Note: SD = standard deviation. SEM = standard error of the mean.

Did participants learn the differential value of the symbols?

We fit participants' selections on each trial (0 = symbol that rewarded 25% of selections, 1 = symbol that rewarded 75% of selections) to a generalized linear mixed model (GLMM), so that trials were nested within participants. Predictors included the trial repetition number (a level-1 predictor, with values ranging from 1-16), and a dummy-coded predictor that reflected the target whom participants learned for (0 = self as target, 1 = identifiable needy person as target). We also included random intercepts for each participant.

Trial repetition number significantly predicted the selection that participants made (b = 0.04, SE = 0.002, *95% CI* = [0.03, 0.0.04], Z = 15.42, $p < .001$; *Odds ratio (OR)* = 1.04, *95% CI* = [1.03, 1.04]), which was nearly identical to the odds ratio of 1.043 that Experiment 1 revealed. The dummy coded variable encoding the target of learning was also non-significant (p = .647). We found qualitatively identical results when we re-ran the model including rewardingness of the selection as the dependent variable. See the Supplemental Materials for more details about these analyses.

GLMM with rewardingness of the selection as the dependent variable

We estimated a GLMM which featured the rewardingness of the selection for each trial as the dependent variable (0 = selection was not rewarded, 1 = selection was rewarded) produced qualitatively identical results, such that trial repetition number significantly predicted whether participants' selection was rewarded (b = 0.017, SE = 0.002, *95% CI* = [0.12, 0.21], Z = 7.29, $p <$ .001; *Odds ratio (OR)* = 1.017, *95% CI* = [1.012, 1.022]). There was also an effect for the target of learning: Participants were less likely to select the rewarding symbol on trials where they learned for themselves, relative to trials where they learned for the identifiable needy target ($p <$ .001). Overall, as in Experiment 1, participants learned to select the more rewarding symbol over the course of each learning block, and made more rewarding selections more generally.

Do people differentially learn to earn reward for themselves vs. non-self targets?

After transforming the αs to the 0 to 1 range by applying the inverse logit, we conducted a within-subjects t-test that included α as the dependent variable, and the target of learning as the independent variable. There was a significant main effect for condition assignment (t(423) = 9.62, $p < .001$, *Cohen's D* = 0.47, *95% CI* = [0.37, 0.57]): Participants were more adept at learning for themselves (M = 0.25, SD = 0. 26) than at learning on behalf of the identifiable needy targets (M = 0.13, SD = 0.18). This finding replicates the results from Experiment 1.

Analyses involving the β parameters

We conducted a within-subjects t-test that featured the β parameters as the dependent variable. There was a significant main effect for condition assignment (t(423) = -6.39, $p < .001$, *Cohen's D* = 0.31, *95% CI* = [0.21, 0.41]), such that participants exhibited higher exploration rates for the identifiable needy target (M = 3.18, SD = 1.64) than themselves (M = 2.64, SD = 1.70). Finally, we examined the correlations among the α and β parameters in each condition, as well as the correlations that the α and β parameters shared with each other. A correlation table of the results is shown in Table S8. Correlations between all β and α parameters were significant (ps $< .01$).

Table S8

Correlations and 95% CIs Among the α and β Parameters in Experiment 2.

Variable	1.	2.	3.
1. $\alpha_{Identifiable}$			
2. α_{Self}	.43**		
3. $\beta_{Identifiable}$	.24**	.26**	
4. β_{Self}	.24**	.35**	.45**

Note. ** indicates $p < .01$.

Scatterplots of correlational results

Figure S14

Scatterplot of the Association Between State Empathic Concern With Prosocial Learning in Experiment 2.

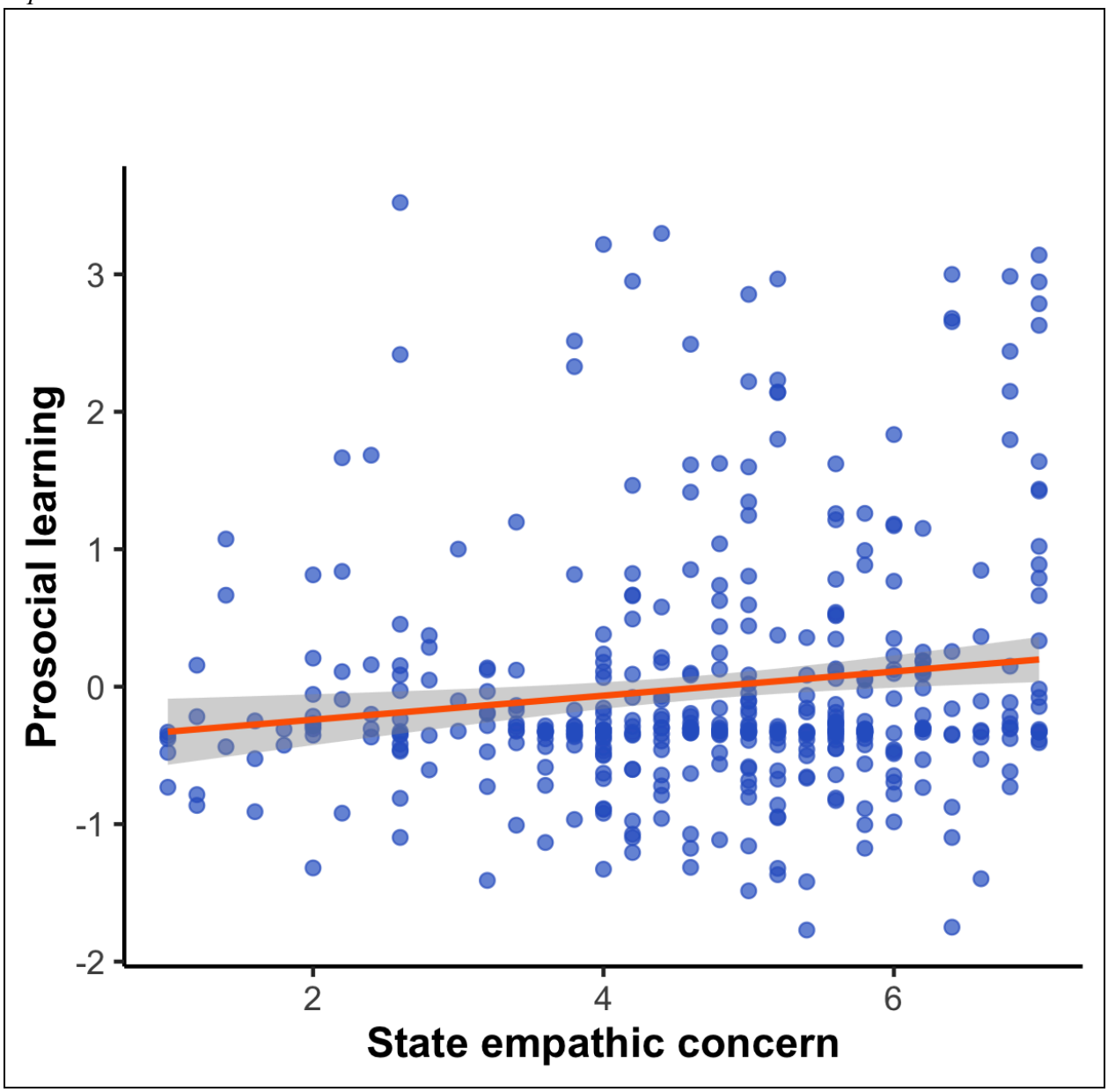
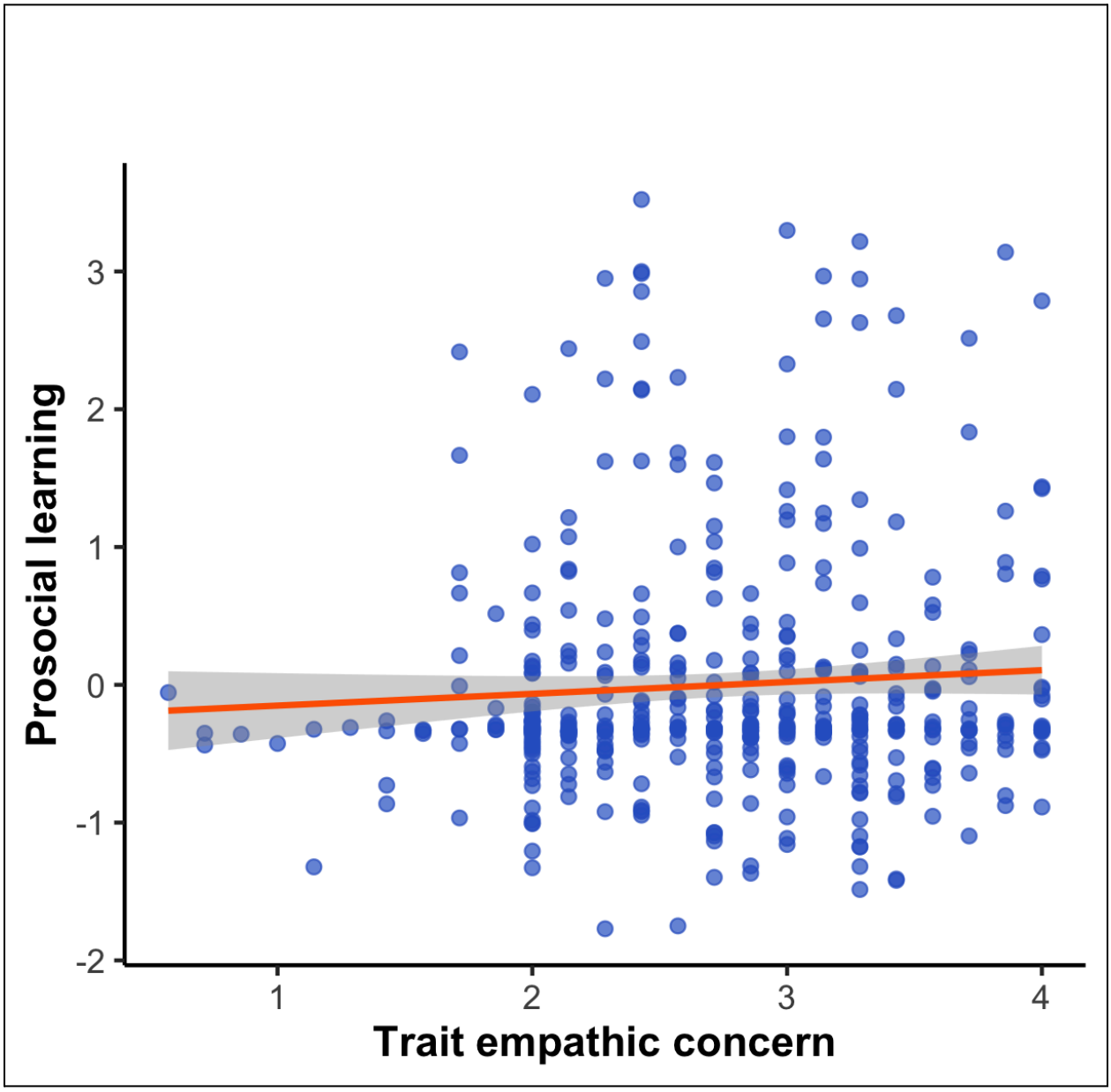


Figure S15

Scatterplot of the Association Between Trait Empathic Concern With Prosocial Learning in Experiment 2.



Bayes factors for the Experiment 2 results

We conducted a Bayesian analysis of the hypotheses reported in Experiment 2. Results for all analyses are shown in Table S6.

Table S9

Bayesian Analyses for the Experiment 2 Hypotheses.

Analysis	Bayes factor	Interpretation
Trait empathic concern correlation with $\alpha_{Identifiable}$	0.28	Data are approximately 3.57 times more likely under the null hypothesis.
State empathic concern correlation with $\alpha_{Identifiable}$	6.05	Data are approximately 6.05 times more likely under the alternative hypothesis.
Empathy manipulation predicting $\alpha_{Identifiable}$	0.11	Data are approximately 9.28 times more likely under the null hypothesis.
Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.57	Data are approximately 1.75 times more likely under the null hypothesis.
Empathy x Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.41	Data are approximately 2.44 times more likely under the null hypothesis.
Online simulation correlation with $\alpha_{Identifiable}$	0.17	Data are approximately 6.0 times more likely under the null hypothesis.

Additional preregistered analyses

In addition to the preregistered predictions tested in the current article, we also made additional preregistered predictions regarding data from Experiment 2 as part of a broader project. Below, we detail the preregistered research questions and predictions from Experiment 2 that were not reported in this article. More information about these research questions can be found at https://osf.io/yp2th?view_only=81071e337d37443b967573eaf28938a4. The numbering listed below corresponds to the numbering found in the preregistration.

Research question 9: Does moral identity internalization more strongly influence prosocial reward learning when social rewards are absent?

Prediction 9: The association between moral identity internalization and prosocial reward learning will be stronger for subjects in the low social evaluation condition, relative to the high social evaluation condition.

Research question 10: Does moral identity symbolization more strongly influence prosocial reward learning when social rewards are present?

Prediction 10: The association between moral identity symbolization and prosocial reward learning will be stronger for subjects in the high social evaluation condition, relative to the low social evaluation condition.

Research question 11: Does endorsement of the principle of care more strongly influence prosocial reward learning when social rewards are absent?

Prediction 11: The association between endorsement of the principle of care and prosocial reward learning will be stronger for subjects in the low social evaluation condition, relative to the high social evaluation condition.

Research question 12: Are there other traits, values, or emotions that are associated with learning for an identifiable needy target?

Prediction 12: A variety of individual differences predictors will be significantly associated with prosocial reward learning (see Table 1 for a list of all the individual differences variables for which we make predictions).

Experiment 3

Stimuli used for the identifiable target in Experiment 3

Participants learned to earn rewards for four different identifiable targets. Here is information about each of those targets. Target 1: A Black woman in her 30's who is pregnant and was recently laid off from her job, seeking money to help pay for rent. Target 2: A Black man in his 60's who is legally blind, and seeking money for medical services. Target 3: A Black man in his 50's seeking money for medical services related to debilitating arthritis (this is the same target from Experiment 2). Target 4: A Latina woman in her 60's who is seeking money for medical services.

Empathy and prosocial learning towards targets

We again tested whether participants expressed different amounts of empathic concern on targets, using the same coding scheme and approached as in Experiment 2. We regressed the amount of state empathy participants reported on dummy-coded variables for each of the targets, with Target 3 serving as the reference condition. Participants who learned about Target 1 (M = 4.65, SD = 1.49; B = 0.21, b = 0.53, SE = 0.20, p = .001), Target 2 (M = 4.70, SD = 1.51; B = 0.58, b = 0.13, SE = 0.20, p = .004), and Target 4 (M = 4.71, SD = 1.45; B = 0.19, b = 0.59, SE = 0.15, $p < .001$) reported significantly more feelings of empathy, compared to Target 3 (M = 4.11, SD = 1.47). We re-ran the analysis and included the prosocial learning rate as the dependent variable. There was no significant difference among any of the targets in terms of the prosocial learning rates (ps > .617).

Block position information

We conducted a χ^2 goodness-of-fit tests to determine if there were significant differences in the order in which the self vs. other blocks were presented for each position. The χ^2 goodness-of-fit tests were non-significant (p = .232), indicating that there was no difference in the frequency with which each block position was presented. Results for the percentage of participants that viewed each condition with respect to each possible condition position shown in Table S10.

Table S10

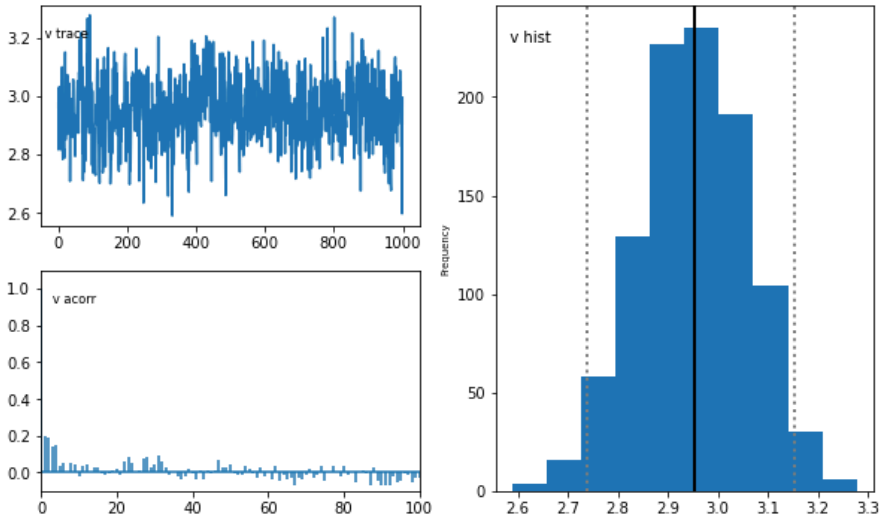
Percentage of Participants That Viewed Each Block Position In Experiment 3.

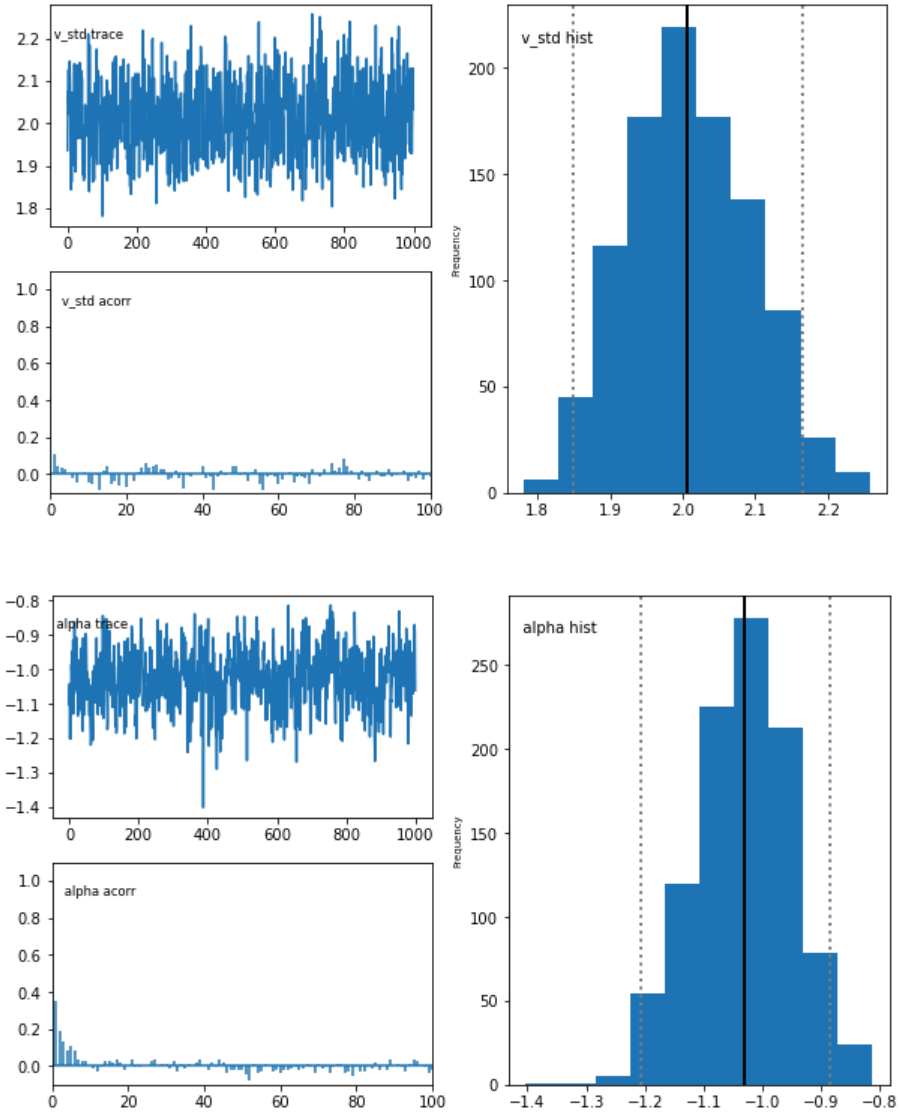
Block	Viewed first	Viewed second
Self	52.6%	47.4%
Identifiable	47.4%	52.6%

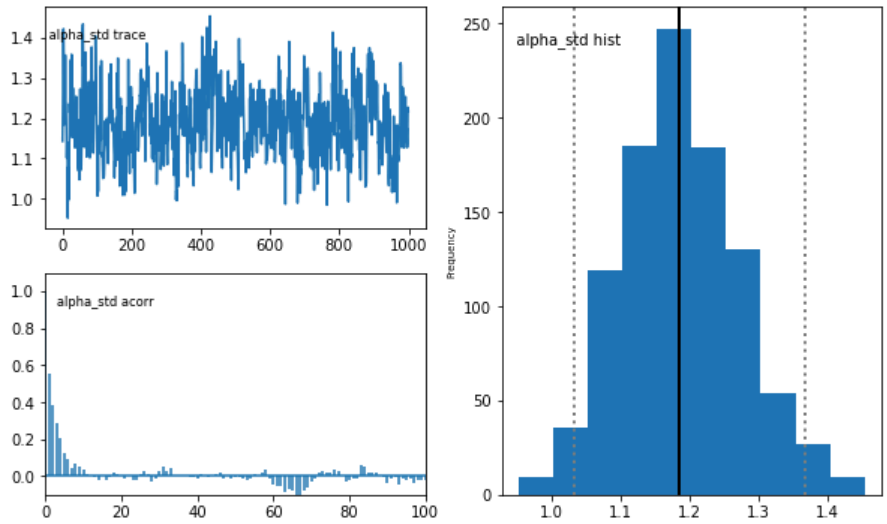
Convergence plots for model estimation in Experiment 3

Figure S16

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 3 In Which a Single α Parameter and a Single β Parameters Were Estimated Across the Self and Identifiable Needy Target Conditions.



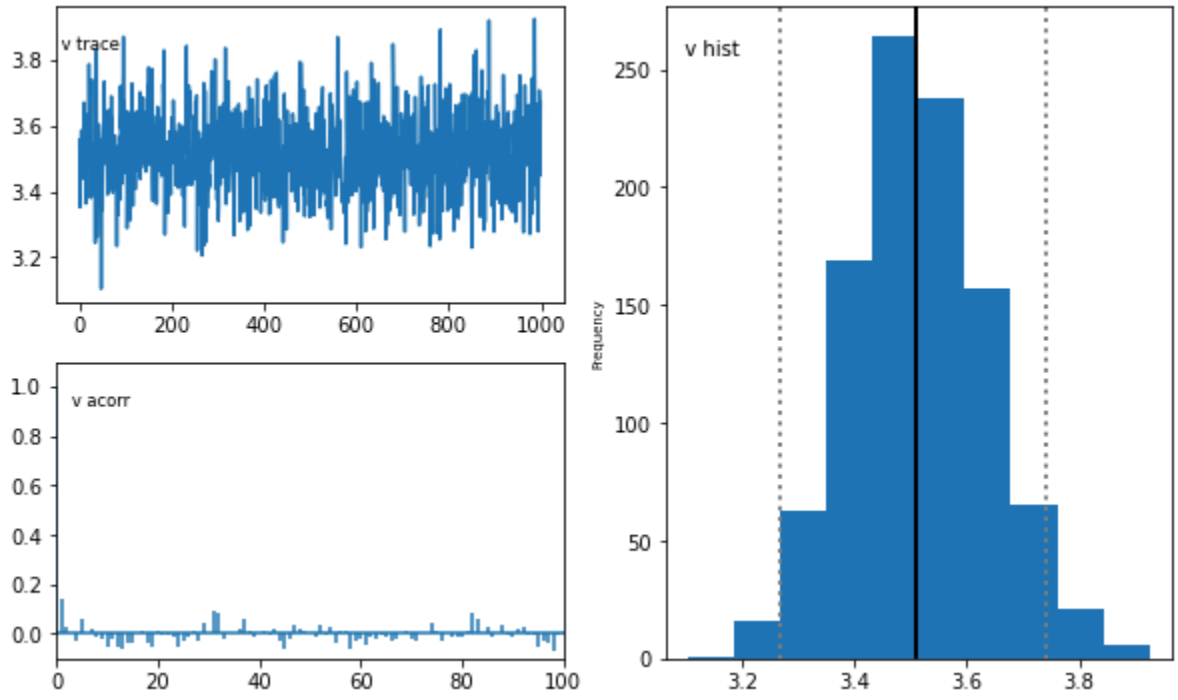

v_std trace
v_std acorr
v_std hist
Frequency
alpha trace
alpha acorr
alpha hist
Frequency

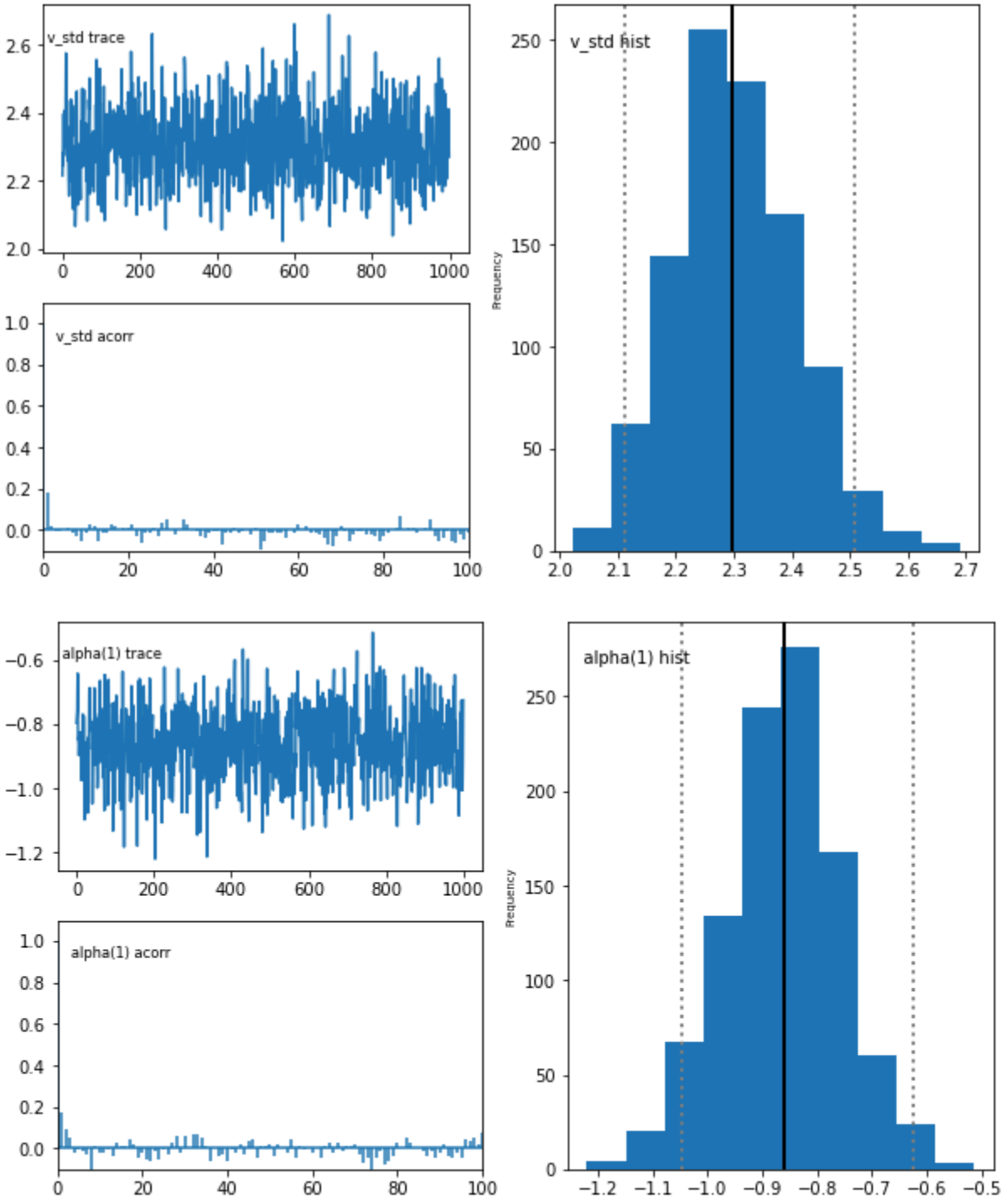


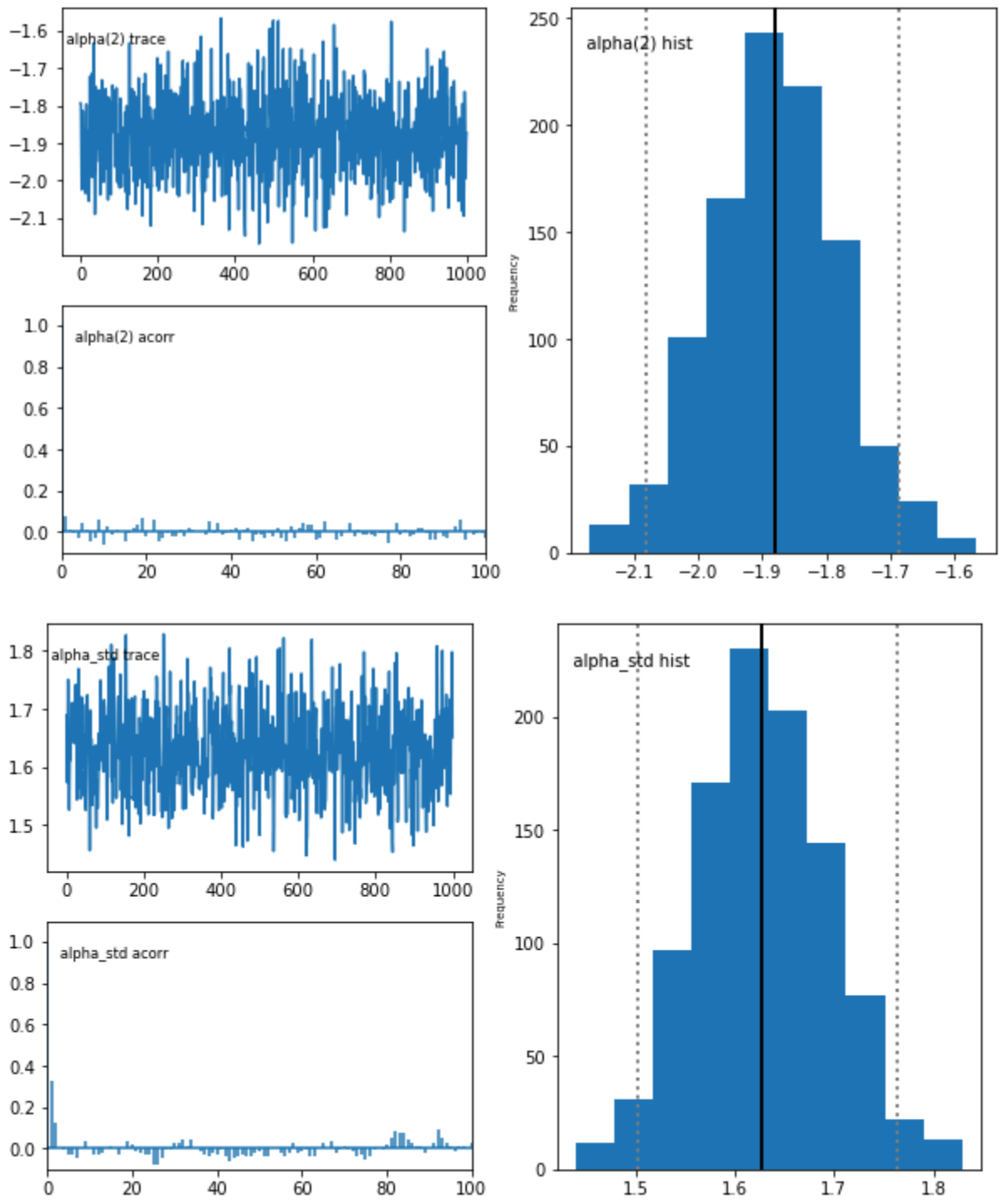
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha" refers to the α parameter, and "alpha_std" refers to the group variability for the α parameter.

Figure S17

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 3 In Which Separate α Parameters Were Estimated For Each of the Self and Identifiable Needy Target Conditions, and a Single β Parameter Was Estimated For Each of the Self and Identifiable Needy Target Conditions.



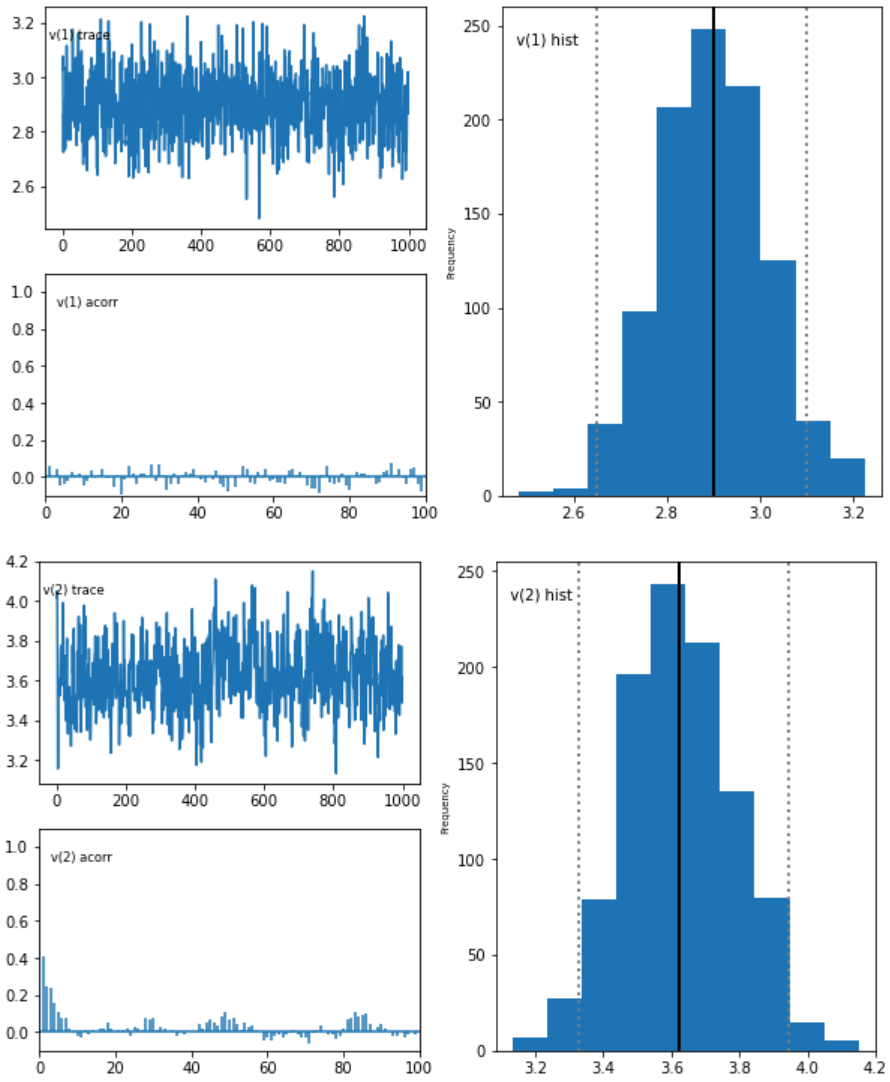

v_std trace
v_std acorr
v_std hist
Frequency
alpha(1) trace
alpha(1) acorr
alpha(1) hist
Frequency

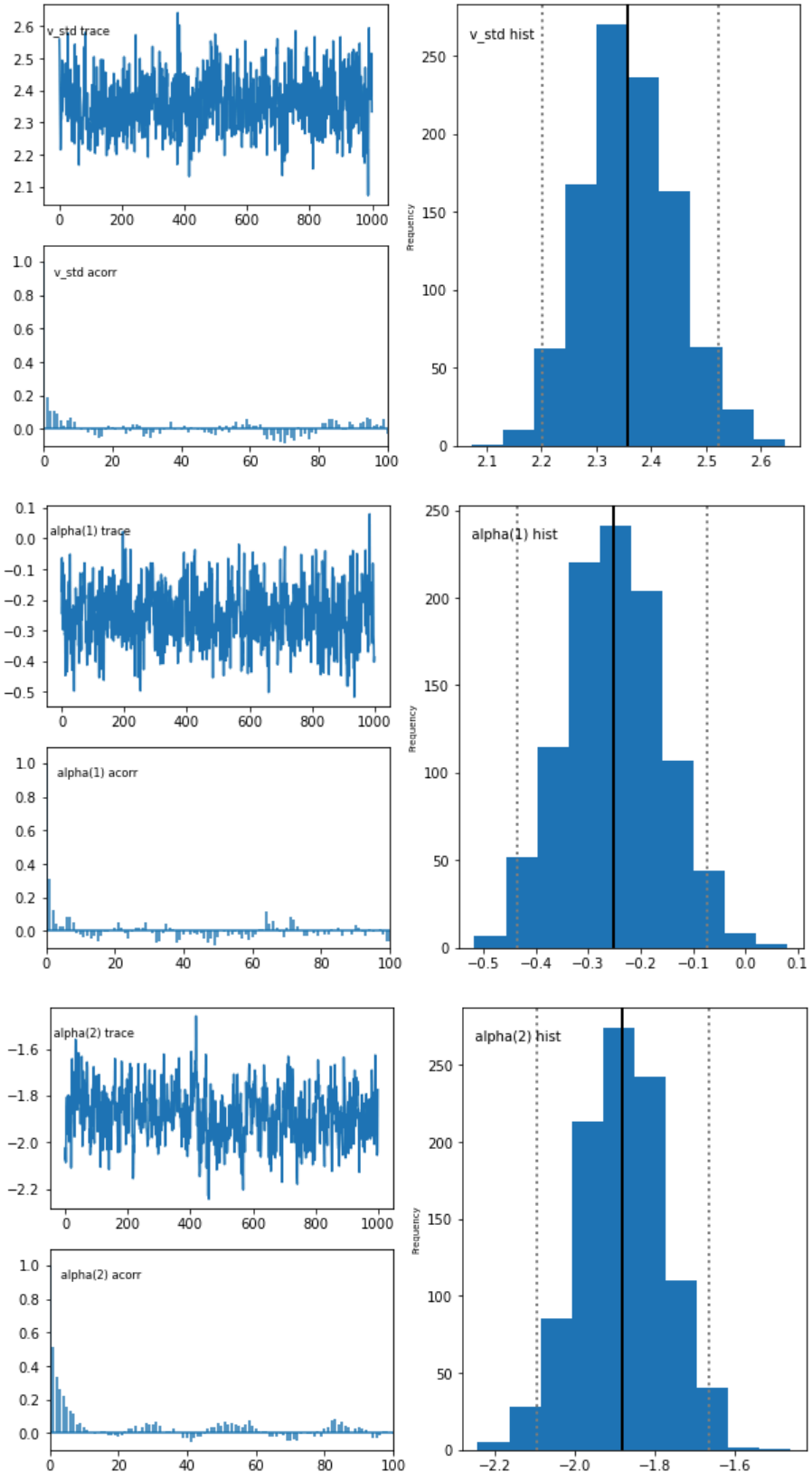


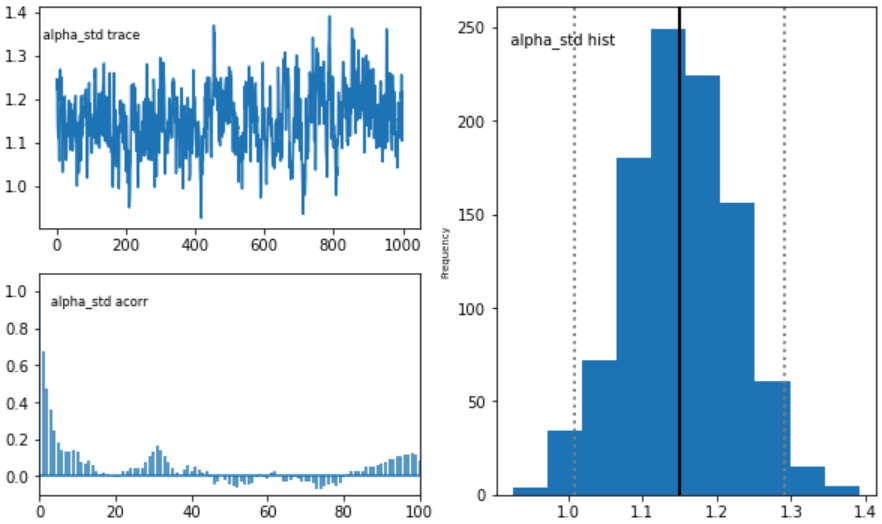
Note: "v" refers to the β parameter, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the self, "alpha (2)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

Figure S18

Trace Plots, Autocorrelations, and Histogram of the Group Mean Distributions for the Reinforcement Learning Model In Experiment 3 In Separate α Parameters and β Parameters Were Estimated For Each of the Self and Identifiable Needy Target Conditions.




v_std trace
v_std hist
v_std acorr
alpha(1) trace
alpha(1) hist
alpha(1) acorr
alpha(2) trace
alpha(2) hist
alpha(2) acorr
Frequency



Note: "v (1)" refers to the β parameter for the self, "v (2)" refers to the β parameter for the identifiable needy target, "v_std" refers to the group variability for the β parameter, "alpha (1)" refers to the α parameter for the self, "alpha (2)" refers to the α parameter for the identifiable needy target, and "alpha_std" refers to the group variability for the α parameter.

Parameter recovery analyses

We simulated n = 10 datasets based on the observed data from Experiment 3. Figure S20 shows the trial-by-trial responses. Once again, the simulated data generally tracked the observed data, following the same pattern in Experiments 1 and 2 where the simulated results again slightly underestimate the frequency with which participants selected the more frequently rewarding symbol, compared to the observed results.

We then fit data from each of the simulated datasets to the same winning model from Experiment 2 (i.e., separate α and β parameters were estimated for each of the two conditions). A visual inspection of the results shows that, when the simulated data were fit to the winning model, we were generally able to recover the same pattern of parameters found in the observed data, although the gap between the simulated and observed αs and βs was slightly larger compared to Experiments 1 and 2.

Figure S19

α Parameters for the Observed and Simulated Datasets in Experiment 3.

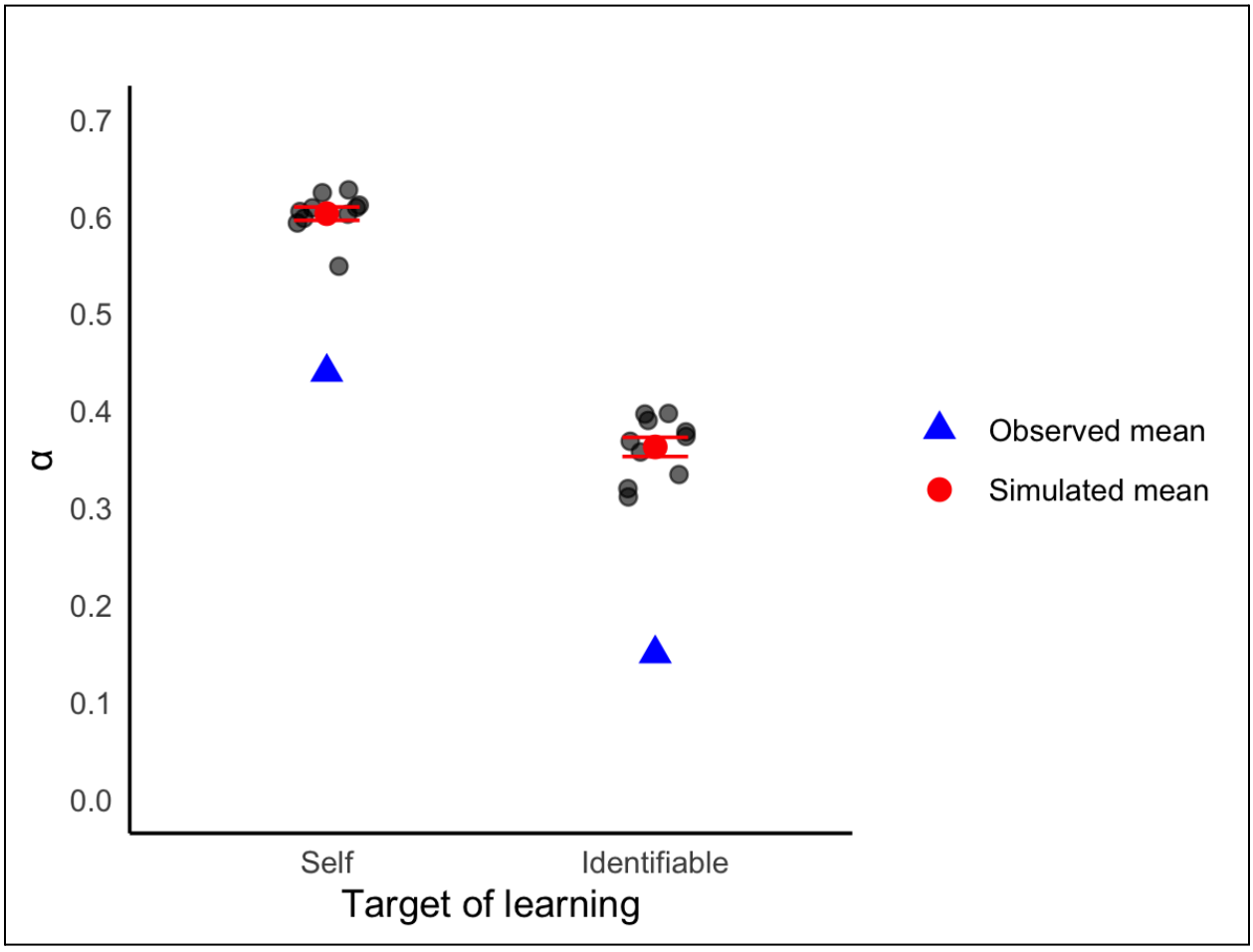
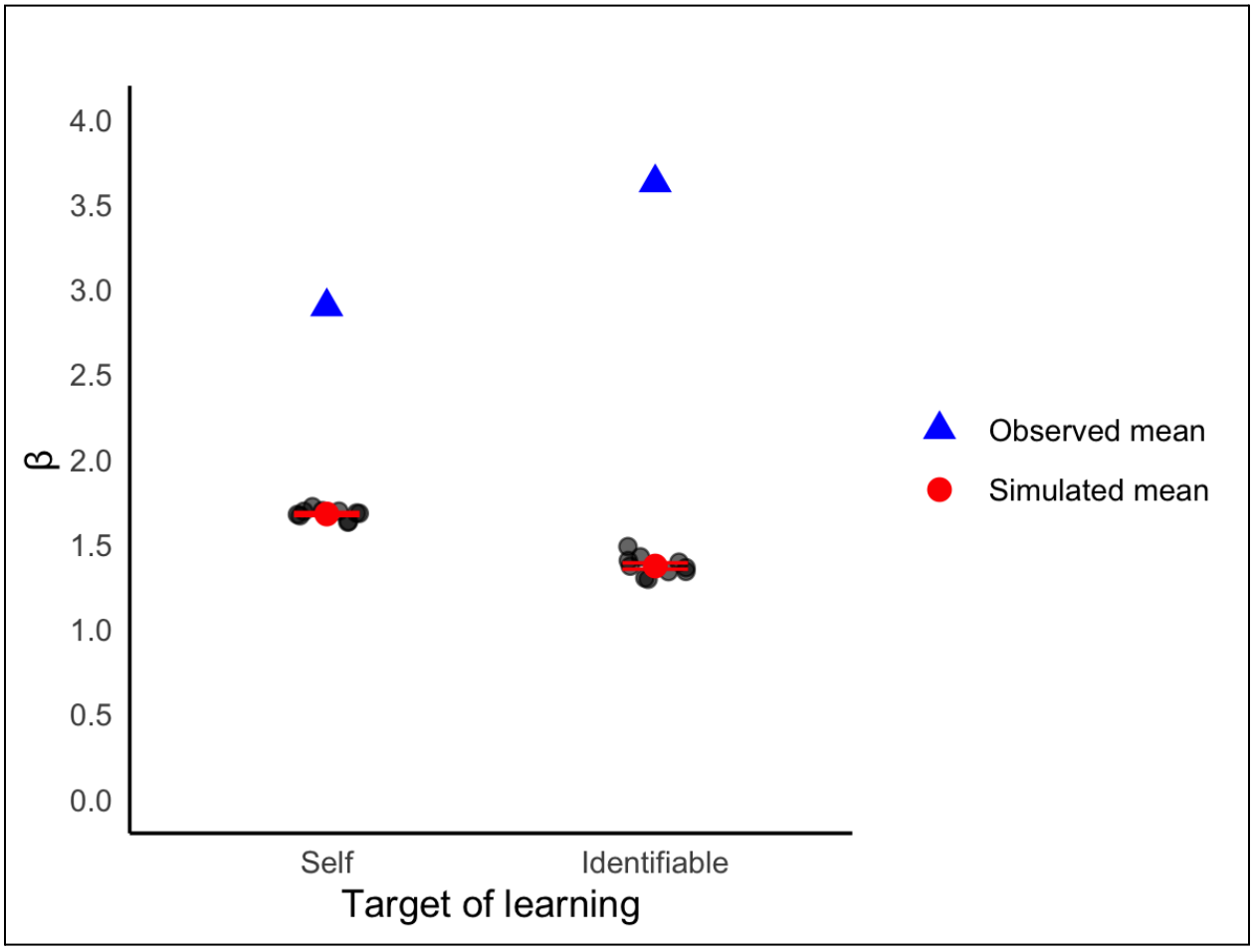


Figure S20

β Parameters for the Observed and Simulated Datasets in Experiment 3.



Posterior predictive check

We conducted a posterior predictive check using identical procedures to those in Experiment 1. Results are shown in Tables S11-S12. Figure S21 shows the learning rate for the observed data, compared to the learning rate for the simulated data. As in Experiments 1 and 2, the results were highly similar for the observed and simulated data.

Figure S21

Trial-By-Trial Response Data for the Observed and Simulated Datasets in Experiment 3.

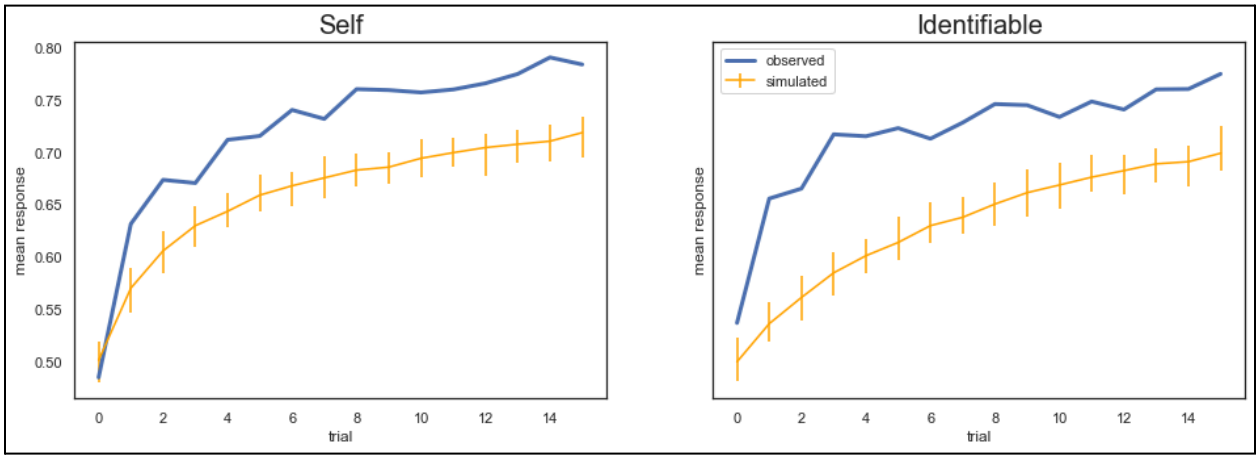


Table S11

Summary Statistics for Frequency of Selecting Rewarding Symbol in the Observed and Replicated Datasets in Experiment 3.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.72	0.45	[0.71, 0.72]	0.003
	Identifiable	0.71	0.45	[0.71, 0.72]	0.003
Simulated	Self	0.66	0.003	[0.66, 0.66]	0.0004
	Identifiable	0.63	0.004	[0.63, 0.63]	0.0006

Note: SD = standard deviation. SEM = standard error of the mean.

Table S12

Summary Statistics for Frequency of Earning Reward in the Observed and Replicated Datasets in Experiment 3.

Data	*Condition*	*Mean*	*SD*	*95% CI*	*SEM*
Observed	Self	0.59	0.49	[0.58, 0.59]	0.003
	Identifiable	0.63	0.48	[0.62, 0.63]	0.003
Simulated	Self	0.58	0.004	[0.58, 0.58]	0.0005
	Identifiable	0.56	0.003	[0.56, 0.57]	0.0005

Note: SD = standard deviation. SEM = standard error of the mean.

Did participants learn the differential value of the symbols?

We fit a GLMM with random intercepts for each participant, with participants' selections on each trial (0 = symbol that rewarded 25% of selections, 1 = symbol that rewarded 75% of

selections) predicted by the trial repetition number (a level-1 predictor, with values ranging from 1-16) and a dummy-coded predictor encoding the target of learning (0 = self as target, 1 = identifiable needy person as target). Trial repetition number significantly predicted the selection that participants made (b = 0.06, SE = 0.002, *95% CI* = [0.06, 0.07], Z = 26.83, $p < .001$; *Odds ratio (OR)* = 1.06, *95% CI* = [1.06, 1.07]). As in Experiments 1 and 2, subjects made more income-maximizing choices as the trials proceeded. The dummy coded variable encoding the target of learning was non-significant (p = .478).

GLMM with rewardingness of the selection as the dependent variable

We conducted a second GLMM, which featured the rewardingness of the selection for each trial as the dependent variable (0 = selection was not rewarded, 1 = selection was rewarded), produced qualitatively identical results, such that trial repetition number significantly predicted whether participants' selection was rewarded (b = 0.03, SE = 0.002, *95% CI* = [0.02, 0.03], Z = 14.14, $p < .001$; *Odds ratio (OR)* = 1.03, *95% CI* = [1.03, 1.03]). Replicating our results from Experiment 2, the target of learning was negatively associated with learning, such that people selected the rewarding symbol more often when learning for the identifiable needy target, relative to when they learned to themselves (b = -0.16, SE = 0.02, *95% CI* = [-0.19, -0.12], Z = -8.64, $p < .001$; *Odds ratio (OR)* = , *95% CI* = [0.82, 0.88]). However, we cautiously interpret the results from this second model, as the model was nearly unidentifiable due to a large Eigenvalue (Bates et al., 2015). Still, the results from both GLMMs suggest that, as in Experiments 1 and 2, participants learned to select the more rewarding symbol over the course of each learning block, and made more rewarding selections more generally.

Do people differentially learn to earn reward for themselves vs. non-self targets?

We transformed the αs to the 0 to 1 range by applying the inverse logit, and conducted a within-subjects t-test that included α as the dependent variable, and the target of learning as the independent variable. There was a significant main effect for condition assignment, such that participants were more adept at learning for themselves (M = 0.44, SD = 0.16) than learning on behalf of the identifiable needy targets (M = 0.15, SD = 0.09; $t(546) = 40.75$, $p < .001$, *Cohen's D* = 1.74, *95% CI* = [1.61, 1.88]).

Analyses involving the β parameters

We conducted a within-subjects t-test that featured the β parameters as the dependent variable. There was a significant main effect for condition assignment ($t(546) = 7.96$, $p < .001$, *Cohen's D* = 0.34, *95% CI* = [0.25, 0.43]), such that participants exhibited higher exploration rates for the identifiable needy target (M = 3.63, SD = 2.09) than themselves (M = 2.90, SD = 1.94). Finally, we examined the correlations among the αs and the β parameters in each condition, as well as the correlations that the α and β parameters shared with each other. A correlation table of the results is shown in Table S13. Correlations between all β and α parameters were significant (ps < .01), except for the correlations that α_{Self} shared with $\beta_{Identifiable}$ and β_{Self} (ps > .05).

Table S13

Correlations and 95% CIs Among the α and β Parameters in Experiment 3.

Variable	1.	2.	3.
1. $\alpha_{Identifiable}$			
2. α_{Self}	.17**		
3. $\beta_{Identifiable}$	.16**	-.04	
4. β_{Self}	.22**	.04	.44**

Note: ** indicates $p < .01$.

Scatterplots of correlational results

Figure S22

Scatterplot of the Association Between the Seven-Item Measure of Trait Empathic Concern With Prosocial Learning in Experiment 3.

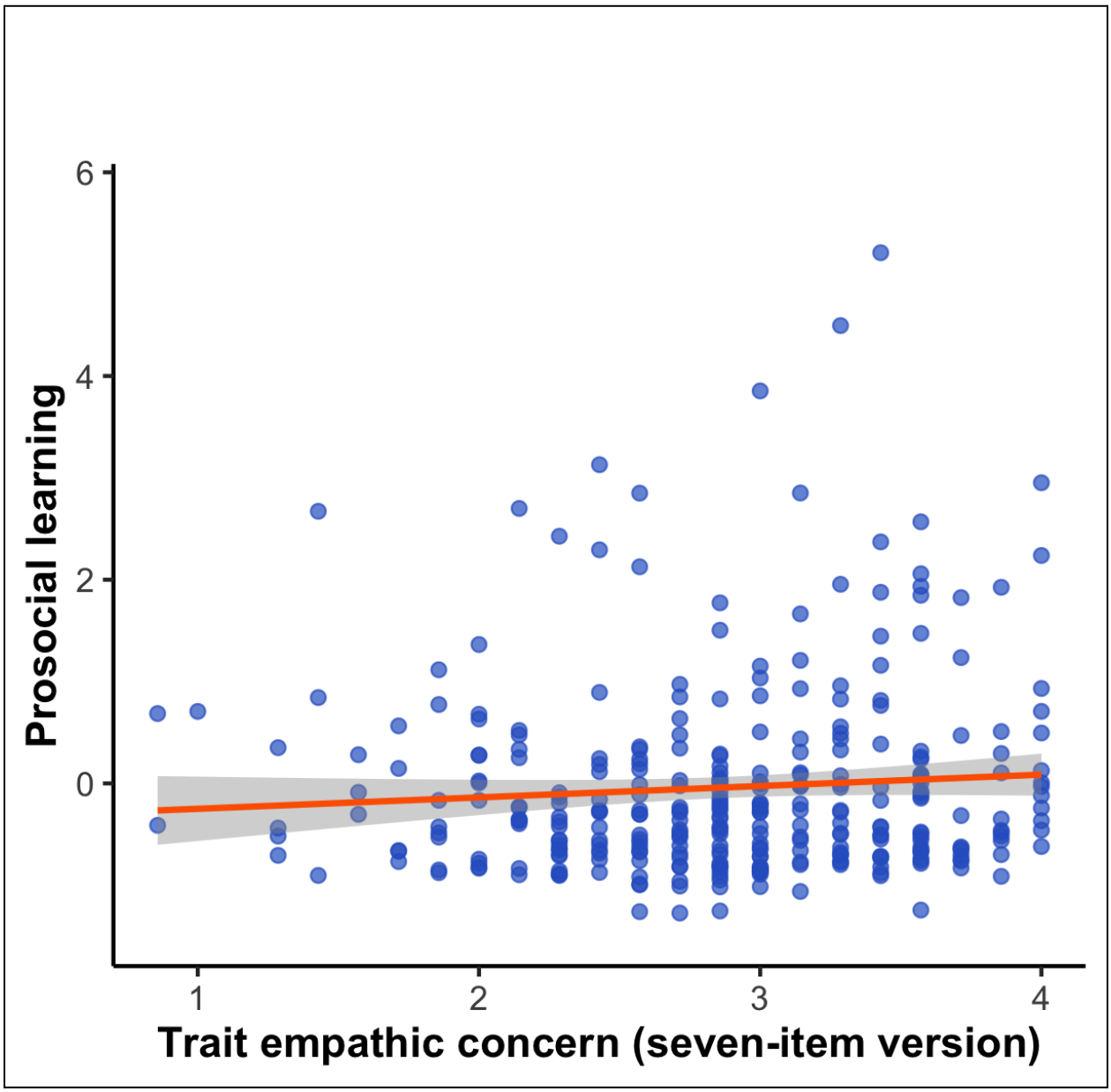


Figure S23

Scatterplot of the Association Between the Four-Item Measure of Trait Empathic Concern With Prosocial Learning in Experiment 3.

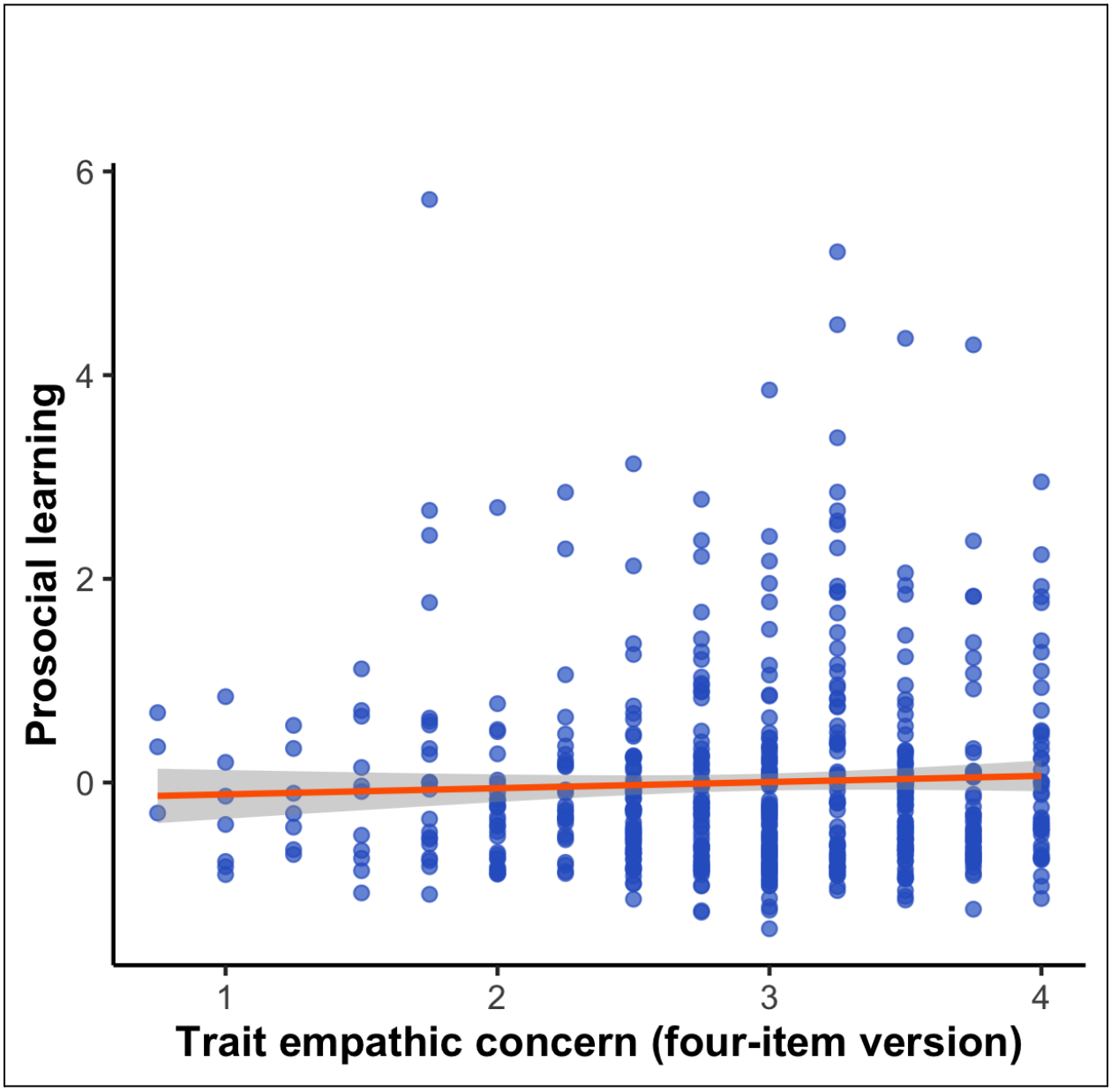
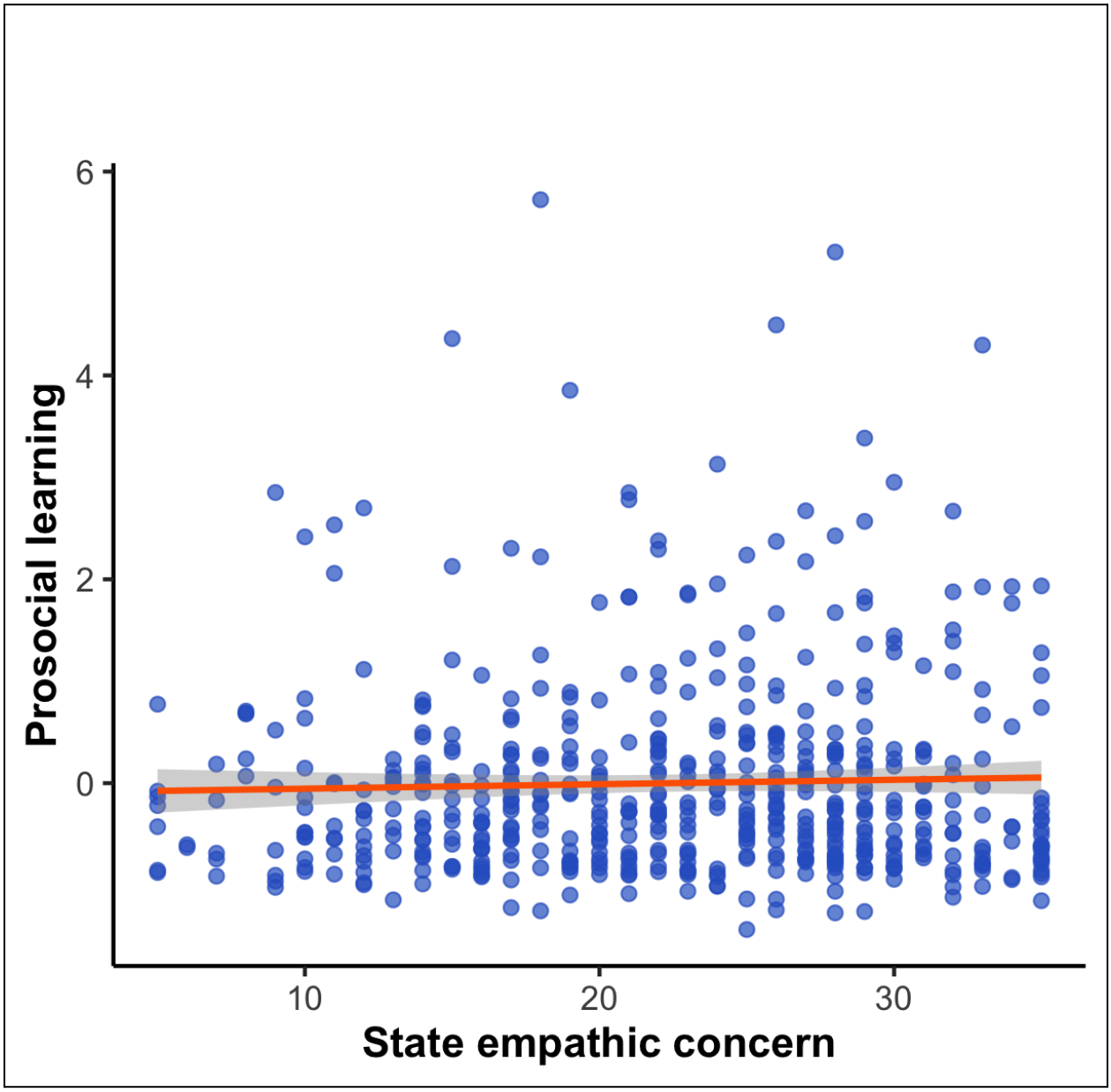


Figure S24

Scatterplot of the Association Between State Empathic Concern With Prosocial Learning in Experiment 3.



Bayes factors for the Experiment 3 results

We conducted a Bayesian analysis of the hypotheses reported in Experiment 3. Results for all analyses are shown in Table S14.

Table S14
Bayesian Analyses for the Experiment 3 Hypotheses.

Analysis	Bayes factor	Interpretation
Four-item version of trait empathic concern correlation with $\alpha_{Identifiable}$	0.17	Data are approximately 5.99 times more likely under the null hypothesis.
Seven-item version of trait empathic concern correlation with $\alpha_{Identifiable}$	0.33	Data are approximately 3.02 times more likely under the null hypothesis.
State empathic concern correlation with $\alpha_{Identifiable}$	0.13	Data are approximately 7.43 times more likely under the null hypothesis.
Empathy manipulation predicting $\alpha_{Identifiable}$	0.19	Data are approximately 5.26 times more likely under the null hypothesis.
Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.11	Data are approximately 9.09 times more likely under the null hypothesis.
Empathy x Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.15	Data are approximately 6.67 times more likely under the null hypothesis.
Online simulation correlation with $\alpha_{Identifiable}$	0.28	Data are approximately 3.64 times more likely under the null hypothesis.

Additional preregistered analyses

Below, we detail the preregistered research questions and predictions from Experiment 2 that were not reported in this article. More information about these research questions can be found at https://osf.io/h9npt?view_only=81071e337d37443b967573eaf28938a4. The numbering listed below corresponds to the numbering found in the preregistration.

Research question 8: Does internalized moral identity more strongly influence prosocial reward learning in the absence of social evaluation?

Prediction 8: Subjects who score high on a measure of internalized moral identity will be more adept at prosocial reward learning in the low social evaluation condition, relative to the high social evaluation condition. There will also be a main effect for internalized moral identity.

Research question 9: Does symbolized moral identity more strongly influence prosocial reward learning in the presence of social evaluation?

Prediction 9: Subjects who score high on a measure of symbolized moral identity will be more adept at prosocial reward learning in the high social evaluation condition, relative to the low social evaluation condition. There will also be a main effect for symbolized moral identity.

Research question 10: Does the principle of care more strongly influence prosocial reward learning in the absence of social evaluation?

Prediction 10: Subjects who score high on a measure of the principle of care will be more adept at prosocial reward learning in the low social evaluation condition, relative to the high social evaluation condition. There will also be a main effect for the principle of care.

Research question 11: Does honesty-humility more strongly influence prosocial reward learning in the absence of social evaluation?

Prediction 11: Subjects who score high on a measure of honesty-humility will be more adept at prosocial reward learning in the low social evaluation condition, relative to the high social evaluation condition. There will also be a main effect for honesty-humility.

Mega-analysis

Bayes factors for the mega-analysis results

We conducted a Bayesian analysis of the hypotheses reported in the mega-analysis. Results for all analyses are shown in Table S15.

Table S15
Bayesian Analyses for the Mega-analysis Hypotheses.

Analysis	Bayes factor	Interpretation
Trait empathic concern correlation with $\alpha_{Identifiable}$	0.57	Data are approximately 1.75 times more likely under the null hypothesis.
State empathic concern correlation with $\alpha_{Identifiable}$	1.05	Data are approximately 1.05 times more likely under the alternative hypothesis.
Empathy manipulation predicting $\alpha_{Identifiable}$	0.11	Data are approximately 9.09 times more likely under the null hypothesis.

Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.09	Data are approximately 11.11 times more likely under the null hypothesis.
Empathy x Social evaluation manipulation predicting $\alpha_{Identifiable}$	0.08	Data are approximately 12.50 times more likely under the null hypothesis.
Online simulation correlation with $\alpha_{Identifiable}$	0.15	Data are approximately 6.67 times more likely under the null hypothesis.

Instrumental variable analysis

We attempted to further probe the manipulations' effect upon self-reported empathic concern using an instrumental variable analysis (Bollen, 2012). The instrumental variable analysis allows us to test whether the empathy manipulations may have exerted a causal effect upon $\alpha_{Identifiable}$, aside from measured state empathy. For this analysis, we allowed the error term in self-reported empathy to correlate with the error term in prosocial learning to statistically control for all other routes, aside from empathy, through which the perspective-taking instructions might have exerted a causal influence upon prosocial learning.

We estimated a mediation model in which we regressed state empathy upon the empathy manipulation; regressed $\alpha_{Identifiable}$ on state empathy and the empathy manipulation; and correlated the error for $\alpha_{Identifiable}$ with the error for state empathy. The covariance between the errors for state empathy and $\alpha_{Identifiable}$ was significant ($b = 0.52$, $SE = 0.23$, $Z = 2.30$, $p = .022$). However, even after accounting for the covariance, the instrumental variable analysis also did not reveal a causal link between the empathy manipulation and $\alpha_{Identifiable}$ ($b = 0.004$, $SE = 0.009$, $Z = 0.442$, $p = .659$).