communications psychology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44271-025-00286-7

Rewards bias self-evaluations of ability

Jean Luo [1] ✉, Peter Mende-Siedlecki [2] & Leor M. Hackel[1]

How do people learn about their own abilities? Often, people receive rewards that offer information about their performance level. Yet, even when two people perform equivalently on a task, they may receive disparate rewards. In these cases, could rewards still influence self-evaluations of ability? In two behavioral experiments, we asked whether people feel more capable and confident when they receive more rewards, even when their performance is held constant, and they know how they objectively performed. Participants played a perceptual game in which they received trial-by-trial accuracy feedback; a staircase procedure held their objective performance constant. However, participants were assigned to either a high or low-reward condition, which varied the probability of receiving a reward for a correct answer. In Experiment 1 ($N = 340$), we found evidence that rewards bias overall self-evaluations of ability after the task—particularly estimations of objective accuracy. Next, in Experiment 2 ($N = 342$), we examined whether reward feedback would inflate participants' trial-by-trial expectations of their own accuracy before each round of the game. Results indicated that participants updated their expectations to a greater extent when a correct response was accompanied with a reward. These findings suggest that rewards enhance how much people integrate accuracy feedback into their dynamic self-beliefs.

When learning new skills, people track their own performance, forming an impression of their own abilities. Once developed, these self-beliefs can influence the careers people pursue and the challenges they take on, shaping future performance, outcomes, and earnings[1,2]. One way people learn how they are doing on a given task is through positive feedback, such as good grades or performance bonuses, which can offer information about one's performance level. However, the same level of performance may also garner different rewards. For instance, equally talented children may receive different amounts of praise, and equally skilled employees may receive different salaries. In these cases, rewards may still have an affective impact, leading people to feel more positively about their successes. By extension, would such variations in reward lead to biased perceptions of one's own ability?

Beliefs about one's own ability reflect one component of the self-concept, which encapsulates a broad collection of beliefs and attitudes about the self in various domains. The self-concept includes both stable components and temporary components that can be malleable in response to environmental influences[3,4]. As people form impressions of their ability in a task domain, these beliefs may serve as a momentary and temporary aspect of self-concept and, over many instances of learning, may eventually become integrated into more stable aspects of the self[5].

Past work suggests that rewards can *indirectly* influence these self-beliefs by shaping behavior. For instance, greater financial incentives can motivate people to work harder and perform better in the workplace[6,7]. This behavior change can fuel a positive feedback loop: By motivating employees to work harder, rewards may lead them to develop an identity as a *hard worker*[8]. Yet, beyond reinforcing behavior, receiving rewards typically *feels* good. Affective experiences may, in turn, drive belief updating[9–14]. In particular, people are often unaware of the source of their feelings[15], leading feelings to seep into judgments. For instance, individuals in a good mood tend to overestimate their overall life satisfaction[16], and individuals who find instructions for a task easy to read tend to underestimate the difficulty of completing the task[17]. People may similarly attribute the positive feelings induced by rewards to their own ability when performing a task, leading them to form an inflated sense of their own performance. Indeed, in previous work, monetary incentives for accurate confidence judgments biased those judgments, such that the prospect of gains increased confidence and the prospect of losses decreased confidence[18]. This work underscores that rewards can elicit positive feelings that bias confidence in task-related performance estimates. Local decision confidence then shapes the formation of general self-beliefs about ability in a given task[19]. Yet, in addition to shaping immediate confidence in a particular response, rewards during learning about one's actual performance in a task may lead to biased self-evaluations of one's performance on the whole. Altogether, beyond indirectly influencing identity by reinforcing behavior, rewards may also directly influence self-beliefs through changing affect.

When forming impressions of *others'* ability, rewards indeed influence evaluations of others' competence. In one study, participants learned about

[1]Department of Psychology, University of Southern California, Los Angeles, CA, USA. [2]Department of Psychological and Brain Sciences, University of Delaware, Newark, DE, USA. ✉e-mail: jeanluo@usc.edu

others who varied both a) in their competence in answering trivia questions and b) in the amount of rewards they earned for participants through the game, which was determined randomly by lottery scratch-off tickets. Participants subsequently rated individuals who had earned them larger material rewards as more competent, even when these individuals had identical trivia performance to those who earned them lower rewards[20]. A similar mechanism may govern self-perception, whereby rewards might bias a person's impression of their own competence even when they are aware of their objective performance.

Prior neuroimaging work similarly raises the hypothesis that rewards could become intertwined with expectations of success. Brain regions involved in reward processing often overlap with those that process accuracy feedback[21,22], track perceptions of one's own ability[23], and form confidence judgments[24,25]. Thus, reward feedback may become conflated with ability as part of an expectation of success as people evaluate their task performance—a hypothesis we tested behaviorally in the present work.

Altogether, beyond indirectly shaping self-beliefs by reinforcing specific behaviors, reward may also directly mold self-beliefs by leading people to feel more positively about their performance. Although reward feedback can sometimes hold informative value about performance, we asked whether it has an influence even when a more objective metric of performance is readily available. In two preregistered experiments, we examined the impact of rewards on self-evaluations of ability by decoupling reward feedback from objective performance feedback. Using a between-subjects design, we manipulated the frequency of monetary rewards participants received for correct answers in a perceptual game. This game allowed us to hold objective performance (i.e., accuracy) constant and give explicit accuracy feedback after each round. In this way, participants could learn based on feedback about their performance in a task that did not clearly map onto pre-existing abilities, similar to prior paradigms used to study self-beliefs[11,26–28]. We measured the effect of rewards on participants' (1) estimations of previous accuracy, (2) predictions of future accuracy, and (3) overall judgments of game competence. We refer to these measures as *task-specific self-beliefs*—measures reflecting impressions of performance, expectations of success, and generalized beliefs about ability in a task domain. Next, we examined the trial-by-trial mechanisms driving biased self-belief formation, testing whether rewards inflate immediate predictions of accuracy after the same accuracy feedback.

Given that participants received explicit feedback about accuracy, they could form self-beliefs by integrating across objective performance feedback. However, at times, people must assess their performance by reflecting on their experience—an ability known as metacognition. For instance, people may assess how confident they are in a response based on the experience of responding[29]. As a secondary question, we asked whether frequent rewards for correct answers would also shape later confidence in the absence of feedback. To do so, we included a separate phase of the game that measured trial-by-trial confidence in responses in the absence of external feedback. We refer to this measure as *local confidence*.

Across both studies, we hypothesized that participants who received more rewards during the task would estimate higher previous accuracy, predict higher future accuracy, and rate themselves higher in competence in the game. We also predicted that receiving more rewards would lead to having higher local confidence when completing the task again without feedback. On a trial-by-trial level (Study 2), we hypothesized that rewards would increase the influence of accuracy feedback on self-evaluations.

Methods

Study 1

All manipulations and measures, sample size and exclusion criteria used in this study were preregistered on June 2, 2023. A preregistration document can be found at https://aspredicted.org/S2J_XSG. All data in Study 1 were collected in June 2023.

Participants. We employed a convenience sampling strategy and recruited 400 participants from CloudResearch with an approval rate of 95% or above. All participants were compensated $2.40 for their participation in the study, with bonuses beyond this base payment dependent on their performance in the game (these ranged from an additional 20 to 80 cents). 364 participants self-reported their age and gender. They ranged from 18 to 75 years ($M = 39.86$, $SD = 11.46$). 206 identified as men, 157 identified as women, and 1 as other. 360 reported their racial/ethnic affiliation (they could select multiple categories). The majority identified as White or Caucasian (68.6%), followed by Black or African American (10.8%), East Asian (8.2%), Hispanic or Latinx (6.4%), South Asian (2.3%), American Indian or Alaska Native (1.8%), Middle Eastern (0.8%), Other (0.8%), and Pacific Islander or Native Hawaiian (0.3%). The sample size was determined through a power analysis based on the exclusion rate and effect sizes observed in a pilot study (described in Supplementary Note 2). Participants were eliminated from analyses based on our preregistered rule: if they failed an attention check phase after the main task that included six rounds of the game at the easiest level (i.e., no noise in dot motion), defined as answering three or more rounds incorrectly. This left a total of 340 participants (85.0%) whose data were used in all Study 1 analyses (unless otherwise specified), with 176 participants in the high-reward condition and 164 in the low-reward condition. All participants completed informed consent in accordance with the University of Southern California Office of the Protection of Research Subjects, which approved the study procedures. We only included in our sample the participants who recorded task data; there were an additional 84 participants who started the Qualtrics survey (containing the study link) without participating in the experiment. In cases where multiple entries were recorded for the same participant, we only used their data from the first entry for analysis.

A post-hoc sensitivity analysis in G*Power 3.1[30] showed that the minimum detectable effect size at 80% power was $f^2 = 0.018$ for a single coefficient in a multiple linear regression model.

Procedure. We programmed the study using PsychoPy (version 2021.2.2)[31] and set up a Qualtrics survey, which directed participants to the study link on the online platform Pavlovia. Thus, participants completed the experiment online without a researcher present. All participants played 80 rounds of the dot motion game, in which they saw dots moving on the screen and indicated the direction in which the majority of the dots were moving. On each round, the dots appeared for 0.75 seconds, after which subjects had unlimited time to indicate the direction of motion by pressing the left or right arrow keys. To ensure similar accuracy levels across all participants, we used a "staircase" procedure to adapt task difficulty and keep accuracy at roughly 70%[32]. This algorithm makes the task harder when a participant answers several rounds correctly and easier after incorrect rounds. Specifically, this meant adjusting the amount of random noise in the motion of dots, making it harder or easier to tell which direction the majority of dots were moving. We implemented a QUEST staircase using JsQuest[33], which calculates the most probable Bayesian estimate of the difficulty level on each round that would yield the target accuracy rate of 70%[34]. Before starting the game, participants had 3 practice rounds, for which the task was at the easiest level (i.e., all dots moving in the same direction with no noise).

For every correct answer in the game, participants had a turn with a virtual slot machine; for incorrect answers, they got no turn with the slot machine. The virtual slot machine determined whether the participant received a token, and tokens translated to a monetary bonus on top of their base payment at a rate of 1 token = 1 cent. After each round in the game, participants received feedback both about their accuracy on that round as well as whether they earned a token in the case of a correct answer. This feedback appeared sequentially—first, accuracy feedback (i.e., "correct" or "incorrect" text) would display. In the event that participants had answered the round correctly, an image of a slot machine appeared underneath the accuracy feedback after 0.75 seconds. Then, reward feedback (i.e., an image of a token or no token) appeared 1.5 seconds after accuracy feedback, underneath the slot machine. All three items (accuracy feedback, slot machine, and reward feedback) would remain on screen for another second before the next trial, for a total of 2.5 seconds on this feedback page. If participants had answered the round incorrectly, only the accuracy feedback would display for the full 2.5 seconds.

Although answering correctly guaranteed participants a turn with the slot machine, it did not guarantee a reward. Participants were randomly assigned (via assignment logic within the experimental platform) to one of two between-subjects conditions. In the high-reward condition, the probability of the slot machine giving a token was 85%, while in the low-reward condition, the probability was 25%. We used these probabilities to generate strong and weak experiences of reward; specifically, a 70% accuracy rate from the staircase procedure would yield 56 trials on average with correct answers, and an 85% reward rate would yield approximately 48 rewarded trials, while a 25% rate would yield approximately 14 rewarded trials. In this way, reward delivery was probabilistic and unrelated to accuracy. After completing all rounds of the game, participants answered several questions about their previous performance. We used 3 measures to probe task-specific self-beliefs: estimated previous accuracy, predicted future accuracy, and self-evaluations of competence in the game. Participants rated their estimated previous accuracy and predicted future accuracy on continuous scales ranging from 0 to 100, corresponding to the percentage of rounds they believed they had answered correctly or would answer correctly in the future, respectively. For self-evaluations of competence, participants rated how good they were at the game on a continuous scale ranging from 0 to 100, with 0 as *Not good at all* and 100 as *Very good*. Additionally, participants also rated their enjoyment (i.e., "how much did you enjoy completing the dot motion game?"), perceived difficulty (i.e., "how difficult did you find the dot motion game?"), and motivation (i.e., "how motivated were you to do well at the dot motion game?"), all on continuous scales ranging from 0 to 100.

Local Confidence. We use the term "local confidence" in this study to refer to metacognitive judgments of confidence about one's immediate performance at the trial level. In order to measure local confidence in immediate responses, participants played 20 additional rounds of the game. Participants received neither accuracy nor reward feedback during these trials. Trials were presented at the difficulty level at which the staircase had ended for each participant and were not subject to further staircasing, as prior work has suggested that using a single difficulty level provides the most accurate measure of metacognitive ability[35]. After each response, participants rated their confidence level on a scale ranging from 1 = *Guessing* to 4 = *Very confident*. Thus, these rounds provided a trial-by-trial measure of participants' internal evaluations of confidence in a particular response while playing the game—as opposed to global evaluations of their performance—without influence from additional rewards or objective performance feedback. Past work has used this procedure for studying metacognitive judgments of confidence[36]. For this portion of the study, participants also learned that each correct answer earned one token; this instruction served both to incentivize performance as well as to establish a new reward structure in which previous reward contingencies became irrelevant. (See Supplementary Note 8 and Supplementary Fig. 2 for analyses of these data.)

Finally, participants played 6 additional rounds of the game at the easiest level. Like the previous trials, participants learned that each correct answer earned one token. We used performance on these trials to exclude inattentive participants. Since the easy trials occurred at the end of the experiment, we reasoned that performing at a chance level or below on these trials likely indicated a lack of learning or attention. We also had used a similar rule previously in a pilot study (reported in Supplementary Note 2). Further, to probe for knowledge about reward structure in the main phase of the game, participants estimated the probability with which they had previously received a token from the slot machine for a correct answer, forming a measure of perceived reward (see Supplementary Note 7 and Supplementary Table 11 for exploratory analyses associated with this measure). Since the 20 rounds probing local confidence and 6 additional rounds of the game did not have a slot machine that could give tokens for correct answers, we included these rounds before the questions probing perceived reward. However, if these additional rounds of the game had some biasing effect on estimates, this bias would have affected all participants regardless of the reward condition to which they were assigned. Finally, participants provided their demographic information and read a debriefing form regarding the study. See Fig. 1 for an overview of the experiment design.

Analytic Strategy. We analyzed all data using R version 4.3.3[37]. Graphs were generated using the ggplot2[38], ggdist[39], and cowplot[40] packages. We used multiple linear regression models to test whether reward condition (coded as −1 for low-reward and 1 for high-reward conditions) predicted estimations of previous accuracy, predictions of future accuracy, and competence ratings, controlling for objective accuracy. Even though the staircase procedure held accuracy at roughly 70%, we included objective accuracy as a covariate to account for any remaining variability in actual performance between participants and to confirm that participants had some veridical sense of how they had performed. For all multiple linear regression models reported, unless otherwise noted, residual plots showed no signs of major deviations from normality or evidence of heteroskedasticity, and variance inflation factors indicated no concerning multicollinearity, suggesting that model assumptions were met. In cases of deviation from normality or evidence of heteroskedasticity, we used the sandwich[41] and lmtest[42] packages to conduct regression with robust standard errors. For all *t*-tests reported, the data were visually inspected to be approximately normal. All statistical tests were two-tailed and used an alpha level of 0.05 to determine significance.

To test the effect of rewards on later local confidence, we fit a generalized estimating equation (GEE) model assuming a Poisson distribution (according to our preregistered plan) using the geepack package[43]. This approach assumes that the outcome is count-like, with equal spacing. Model diagnostics confirmed that the asymptotic and dispersion assumptions of the Poisson GEE were adequately met (dispersion ratio = 0.12). To ensure that the observed effect was robust to modeling assumptions, we also fit an exploratory cumulative link mixed model (including random intercepts for participants) using the ordinal package[44], which is better statistically suited for modeling ordered categorical outcomes (results of this model are reported in Supplementary Note 8).

Study 2

Experimental procedures, analyses, exclusion criteria for Study 2 were preregistered on September 8, 2023 on AsPredicted. A preregistration document and can be found at https://aspredicted.org/F92_NGB. All data in Study 2 were collected between September and October 2023.

Participants. We aimed to recruit 400 participants, but due to a platform error, data were collected from 399 unique participants. We employed a convenience sampling strategy. Of the recruited participants, 208 came from the University of Southern California subject pool and received course credit in exchange for participation in the study. 187 self-reported their age and gender. These respondents ranged from 18 to 44 years ($M = 20.31$, $SD = 2.19$), and 73 self-identified as men, 113 as women, and 1 as other. 184 participants also indicated their racial/ethnic affiliation, and they were allowed to select multiple categories. Of these participants, the most frequently selected categories were White or Caucasian (32.5%) and East Asian (31.1%), followed by Hispanic or Latinx (12.4%), South Asian (10.1%), Black or African American (4.8%), Middle Eastern (3.8%), American Indian or Alaska Native (1.4%), Pacific Islander or Native Hawaiian (1.0%), and Other (2.9%).

The remaining 191 participants came from CloudResearch and had an approval rate of 95% or above. These participants were compensated $2.40 for their participation in the study. 182 self-reported age, gender, and racial/ethnic affiliation; these respondents ranged from 19 to 72 years ($M = 36.81$, $SD = 11.44$), and 112 self-identified as men and 70 as women. The majority identified as White or Caucasian (60.1%), followed by Black or African American (16.6%), Hispanic or Latinx (8.8%), East Asian (8.3%), South Asian (3.6%), American Indian or Alaska Native (2.1%), and Middle Eastern (0.5%). No participants identified as Pacific Islander or Native Hawaiian or selected "Other".

We used the same preregistered exclusion rule as Study 1 (excluding participants who answered incorrectly on three or more of the easy dot

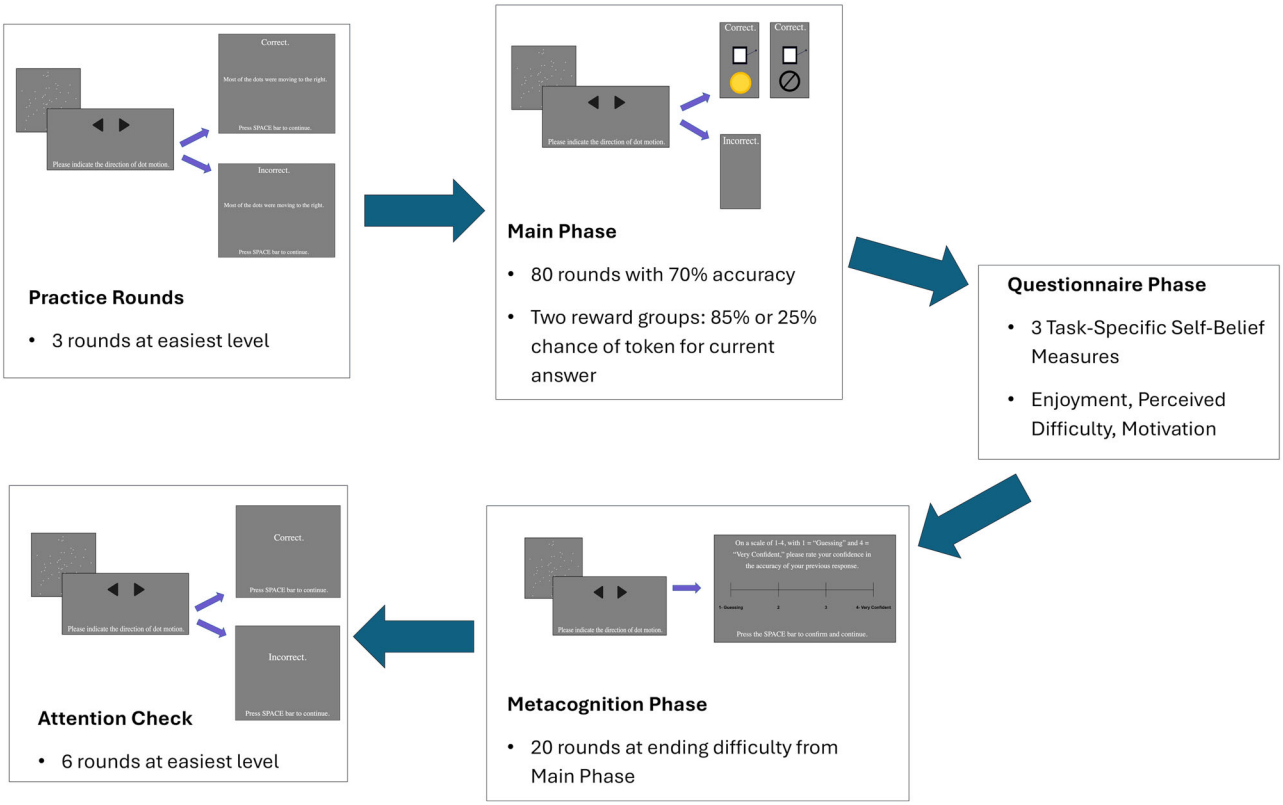

Fig. 1 | Full task structure. Participants first practiced the game at the easiest level for 3 rounds before proceeding to 80 rounds of the main phase. Then, they filled out questionnaires probing task-specific self-beliefs. They then completed 20 additional rounds of the game without accuracy or reward feedback at the ending difficulty level from the main phase, rating their confidence after each answer. Finally, participants completed 6 rounds at the easiest level, which served as an attention check.

motion trials at the end of the experiment). After exclusions, a total of 342 participants (85.5%) remained for analysis, with 174 participants in the high-reward condition and 168 in the low-reward condition. All Study 2 analyses used this sample, unless otherwise specified. All participants completed informed consent in accordance with the University of Southern California Office of the Protection of Research Subjects, which approved the study procedures. We only included in our sample the participants who recorded task data; there were an additional 135 participants who started the Qualtrics survey without participating in the experiment. In cases of duplicate data entries for the same participant, we only used data from the first entry.

A sensitivity analysis indicated that the minimum detectable effect size at 80% power for a single coefficient in a multiple linear regression was $f^2 = 0.018$. The smallest detectable effect size for a two-tailed paired samples t-test was $d = 0.16$.

Procedure. As in Study 1, we programmed the study on Psychopy (version 2021.2.2)[31] and created a Qualtrics survey that directed participants to the online study on Pavlovia. Participants completed the experiment online with no researcher present. The procedure in Study 2 was similar to Study 1, with the following changes. Before every round of the game, participants used the up or down arrow keys to indicate whether they expected that they would answer correctly or incorrectly on the next round, forming a measure of performance expectation. Additionally, because part of our sample in Study 2 came from a university subject pool, rewards did not translate directly to a bonus; instead, tokens for all participants served as raffle tickets for a $10 gift card draw at the end of the study. Moreover, in Study 2, each trial was associated with a unique image cue representing either a lottery ticket after a correct answer or an "incorrect receipt" after an incorrect answer. We used images of objects and scenes, taken from pre-existing image databases[45,46] and randomized per participant whether objects represented lottery tickets and scenes represented incorrect receipts, or vice versa. We also randomized the order in which the images appeared. Following the game, we probed memory for several images by asking participants to indicate "Yes" or "No" in response to the question, "Did you see this image during the game?" 30 were new images that had not been displayed during the game, while we aimed to test up to 10 images previously associated with incorrect trials, up to 10 associated with correct-unrewarded trials, and up to 10 associated with correct-rewarded trials. We included these memory probes to test a second possible mechanism, whereby rewarded trials may stand out in memory, which could in turn serve to bias self-evaluations at the conclusion of the game. However, due to a critical data recording error, memory probe data could not be accurately associated with specific trials, preventing us from testing this alternative hypothesis. Before finishing the study, participants also filled out a general self-esteem questionnaire[47] and the PHQ-9 to assess depressive symptoms[48]. (Analyses associated with the depression measure are reported in Supplementary Note 9). We added these measures to explore how individual differences may relate to task-specific self-belief updating.

Analytic Strategy. All analyses were implemented in R version 4.3.3[37], and graphs were generated using the ggplot2[38], ggdist[39], and cowplot[40] packages. We used a mixed-effects logistic regression using the lme4 package[49] to test how trial-by-trial performance expectations (measured as a binary variable, where 0 = incorrect and 1 = correct; total of 80 trials per participant) depend on accuracy (i.e., whether participants answered correctly or incorrectly) and reward (i.e., whether reward was presented) feedback on the preceding trial. This model, therefore, estimated whether participants increased their expectations of success in response to correct (versus incorrect) answers and whether they did so to a larger extent if a correct answer was rewarded. The model included random intercepts for participant and by-participant random slopes for both reward and accuracy feedback. The correlations between random effects were not modeled. Model diagnostics indicated no violations of key assumptions, including

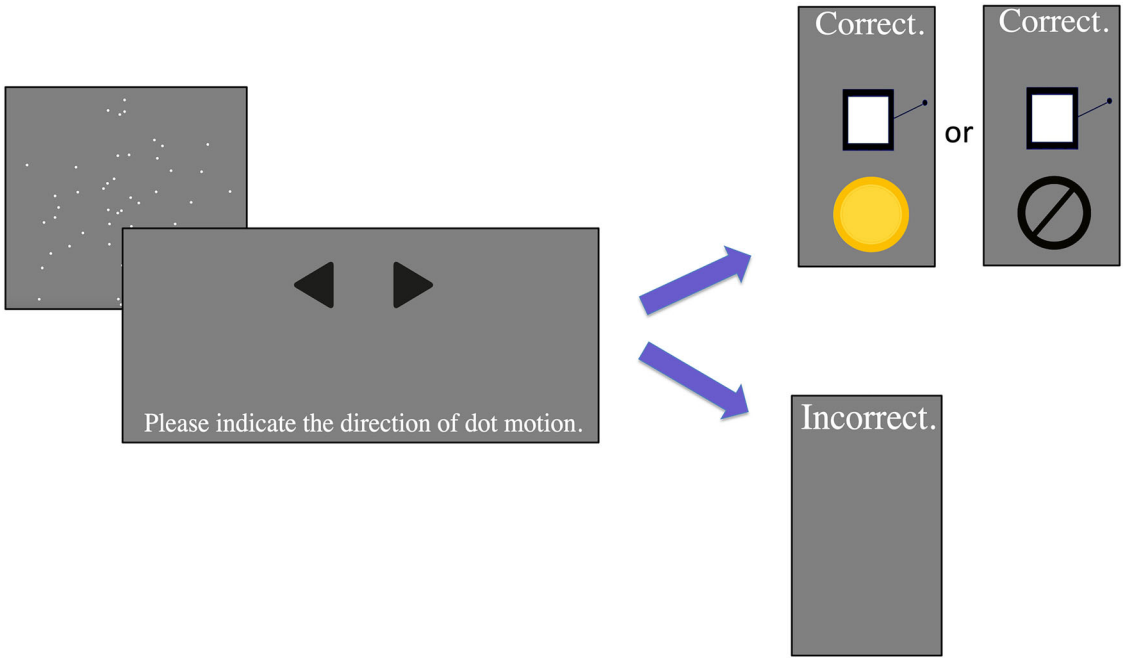

Fig. 2 | Schematic of Study 1 task flow. After making a response for a round of the dot motion game, accuracy feedback was displayed first, reading either "Correct" or "Incorrect." If the response was incorrect, nothing else would display before the next round began. If the response was correct, an image of a slot machine appeared underneath the accuracy feedback, followed by an image of a token or a circle with a slash, indicating whether a token was won or not.

linearity of the logit, multicollinearity, heteroscedasticity, or model convergence. The assumption of normality for *t*-tests was adequately met based on visual inspection of difference scores.

Identical procedures as Study 1 were used for modeling task-specific self-beliefs and trial-by-trial confidence. We confirmed that the assumptions of normality and absence of heteroskedasticity and multicollearity were met for multiple linear regression models. Model diagnostics for the GEE model predicting local confidence indicated that asymptotic and dispersion assumptions were met (dispersion ratio = 0.10).

Results

Study 1

We first investigated how the probability of receiving rewards during a task influences subsequent self-evaluations of ability. In a dot motion task described as a perceptual game, participants repeatedly judged whether dots were moving to the left or right. To hold performance approximately constant across participants (~70% accuracy), we implemented a staircase algorithm that increased or decreased the difficulty of the task, which was accomplished by increasing or decreasing random noise that made the direction of overall motion more or less difficult to detect[34]. After each round, participants received accuracy feedback, allowing them to track their performance objectively. In addition, participants were informed that each time they answered correctly, they would have the opportunity to play with a slot machine that could give them a token worth a monetary reward. On these rounds, they saw an image of a slot machine appear on screen, followed by a token or no token (Fig. 2). Thus, reward was probabilistic, and participants were randomly assigned to a "high-reward" condition, in which they received a token on 85% of correct trials, or a "low-reward" condition, in which they received a token on 25% of correct trials. Participants were instructed that, while correct responses always generated an *opportunity* to play with the slot machine, the outcome was unrelated to their response. After the game, they then made three self-evaluations about their own ability in the game: an estimate of the proportion of trials they had answered correctly, a prediction about the proportion of trials they would answer correctly if they played 100 additional rounds of the game, and a rating of how good (i.e., competent) they were at the game overall. We hypothesized that participants who received more rewards when completing a task would estimate higher previous accuracy, predict higher future accuracy, and rate themselves higher in competence in the game.

Although not specified in our preregistered plan, we examined whether the staircase procedure successfully held accuracy constant between reward conditions. The average accuracy was 69.32% (high reward: $M = 0.69$, SD = 0.09; low reward: $M = 0.70$, SD = 0.09), and a two-tailed independent samples *t*-test revealed no significant differences between reward conditions ($t(337.8) = -1.86$, $p = 0.06$, Cohen's $d = 0.20$, 95% CI [−0.036, 0.001]).

Task-specific self-beliefs. We first assessed whether self-beliefs about the task differed across reward conditions, adjusting for each participant's actual accuracy. The multiple linear regression models explained 11.1% of the variance in estimations of previous accuracy (*Adjusted* $R^2 = 0.111$, $F(2, 337) = 22.24$, $f^2 = 0.13$, $p < 0.001$) and 13.9% of the variance for predicted future accuracy (*Adjusted* $R^2 = 0.139$, $F(2, 337) = 28.37$, $f^2 = 0.17$, $p < 0.001$). Actual accuracy was significantly associated with both estimated previous accuracy ($b = 59.70$, SE = 9.81, $t(337) = 6.09$, $p < 0.001$, 95% CI [40.40, 78.98]) and predicted future accuracy ($b = 74.66$, SE = 10.38, $t(337) = 7.19$, $p < 0.001$, 95% CI [54.23, 95.08]), indicating that participants' estimates of their own accuracy generally reflected their objective accuracy.

Yet, in the same model, reward was also significantly associated with estimated previous accuracy ($b = 2.87$, SE = 0.87, $t(337) = 3.32$, $p < 0.001$, 95% CI [1.17, 4.57]; Fig. 3A) and predicted future accuracy ($b = 2.70$, SE = 0.91, $t(337) = 2.95$, $p = 0.003$, 95% CI [0.90, 4.50]; Fig. 3B), indicating that reward exerted an influence on self-evaluations even after taking into account the effect of objective accuracy. That is, participants who received more frequent rewards for correct answers thought they had performed more accurately and would perform more accurately in the future, even when accounting for their objective performance. Specifically, participants in the high-reward condition estimated higher accuracy ($M = 61.91$, SE = 0.89) than in the low-reward condition ($M = 57.22$, SE = 0.92). Although not preregistered, a two-tailed independent samples *t*-test assessing the absolute difference between conditions supported our main findings ($t(334.28) = 2.59$, $d = 0.28$, $p = 0.01$, 95% CI [1.12, 8.25]). Participants in the high-reward condition also predicted greater accuracy ($M = 65.19$, SE = 0.97) than those in the low-reward condition ($M = 61.11$, SE = 0.98), and a

(non-preregistered) two-tailed independent samples *t*-test also provided converging evidence for our primary regression results ($t(335.32) = 2.09$, $d = 0.23$, $p = 0.04$, 95% CI [0.24, 7.92]). Given that participants underestimated their own accuracy overall, those in the high-reward condition came closer to their true accuracy than those in the low-reward condition (refer to Supplementary Note 1). In the corresponding regression predicting participants' competence ratings, the two predictors explained 6.8% of the variance (*Adjusted* $R^2 = 0.068$, $F(2, 337) = 13.4$, $f^2 = 0.08$, $p < 0.001$). In this model, accuracy significantly predicted competence ratings ($b = 56.79$, SE = 11.40, $t(337) = 4.98$, $p < 0.001$, 95% CI [34.37, 79.21]), but the effect of reward was non-significant ($b = 1.90$, SE = 1.00, $t(337) = 1.89$, $p = 0.059$, 95% CI [−0.08, 3.87]; Fig. 3C). Although participants in the high-reward condition rated higher game competence ($M = 62.57$, SE = 1.00) than in the low-reward condition ($M = 59.78$, SE = 1.07), a (non-preregistered) two-tailed independent samples *t*-test also did not provide evidence of a significant difference between the two conditions ($t(331.44) = 1.35$, $d = 0.15$, $p = 0.20$, 95% CI [−1.28, 6.87]).

These findings replicated a pilot study that found similar effects of reward on evaluations of estimated previous accuracy and predicted future accuracy (see Supplementary Note 2). Moreover, in an exploratory (non-preregistered) analysis, we examined the relationship of self-beliefs with experienced reward rate, which may track more closely with beliefs. Although an exploratory manipulation check confirmed that participants in the high-reward condition received significantly higher average bonuses ($M = 66.89$ cents, SD = 8.51) than those in the low-reward condition ($M = 34.50$ cents, SD = 5.23, $t(294.0) = 42.59$, $p < 0.001$, $d = 4.55$, 95% CI [30.89, 33.88]), actual reward rates differed across individuals within the same condition due to the probabilistic nature of feedback. That is, different individuals in the same reward condition could have had relatively higher or lower reward rates compared to one another, and those different rates of reward might produce different self-beliefs. We therefore refit the linear regression models predicting task-specific self-beliefs using each participant's actual reward rate rather than reward condition. This approach provided a more precise metric of the actual rate of reward each participant experienced. In these models, reward rate was significantly associated with estimated accuracy ($b = 10.67$, SE = 2.85, $t(337) = 3.75$, $p < 0.001$, 95% CI [5.07, 16.28], predicted accuracy ($b = 9.45$, SE = 3.02, $t(337) = 3.12$, $p = 0.002$, 95% CI [3.50, 15.40]), and competence ratings ($b = 7.55$, $SE = 3.32$, $t(337) = 2.28$, $p = 0.024$, [1.02, 14.08]; see Supplementary Note 6 and Supplementary Table 9).

Although the staircase procedure held *accuracy* constant, different participants might have reached 70% accuracy at different difficulty levels of the task. If difficulty levels differed across conditions, then this difference might also explain self-beliefs across conditions. We therefore tested whether the *difficulty level* of the staircase impacted task-specific self-belief measures. This analysis was also not originally in our preregistered plan, but we aimed to ensure results remained the same when adjusting for the objective difficulty of the game, given that participants might have been aware of task difficulty and used difficulty as a cue to their performance. To do so, we refit the regression models predicting the three task-specific self-belief measures with an additional predictor that reflects the ending difficulty level of the task for each participant. The reward effect remained significant for estimated accuracy ($b = 2.89$, $SE = 0.87$, $t(336) = 3.33$, $p < 0.001$, 95% CI [1.18, 4.59]) and predicted accuracy ($b = 2.62$, SE = 0.91, $t(336) = 2.87$, $p = 0.004$, 95% CI [0.82, 4.41]; see Supplementary Table 2). The average ending difficulty level of the task also did not differ significantly by condition (see Supplementary Note 4 for details). Thus, we observe no credible evidence that an awareness of the difficulty of the game could account for the observed effect of reward on self-beliefs.

While we had preregistered an exclusion rule for our analyses, we also tested whether the main findings were robust to alternative exclusion criteria. For example, our preregistered rule used trials after the main phase, but it is possible that an exclusion rule relating to performance during the main 80 rounds of the games could capture inattention better. Specifically, in exploratory analyses, we excluded participants who spent most of the game at the easiest level (i.e., no noise in dot motion). For these analyses, the inferences remained the same (see Supplementary Note 4).

Game perceptions. Did the effects of reward extend to perceptions of the game itself, beyond self-evaluations of ability? In exploratory analyses, we examined this question by testing the effects of reward on self-reported enjoyment, difficulty, and motivation in the task using linear multiple regression. Reward did not significantly predict enjoyment ($b = 2.08$, robust SE = 1.48, $t(337) = 1.40$, $p = 0.16$, 95% CI [−0.83, 4.99]) or motivation ($b = 0.70$, SE = 1.27, $t(337) = 0.55$, $p = 0.59$, 95% CI [−1.80, 3.19]), but it did predict perceived difficulty ($b = -2.50$, SE = 1.23, $t(337) = 0$-2.03, $p = 0.043$, 95% CI [−4.92, −0.08]; see Supplementary Note 9 for model details). Specifically, receiving more rewards was associated with perceiving the game as less difficult. On the other hand, we observe little credible evidence that the influence of rewards on self-evaluations of ability was related to increased motivation or overall enjoyment of the game itself.

Local confidence. Finally, beyond overall performance judgments, we tested whether rewards during the main phase would impact later confidence when performing the task in the absence of accuracy or reward feedback. In a subsequent phase of the dot motion game, participants indicated their confidence (on a scale from 1 = *Guessing* to 4 = *Very confident*) after each trial of the game, forming a measure of local confidence. In this phase, the average accuracy was 73.60% (high reward: $M = 0.74$, SD = 0.14; low reward: $M = 0.73$, SD = 0.14), and a (non-preregistered) two-tailed independent samples *t*-test revealed no significant differences between reward conditions ($t(333.9) = 0.72$, $d = 0.08$, $p = 0.47$, 95% CI [−0.02, 0.04]). The high-reward condition rated an average local confidence across trials of 2.74 (SD = 0.71), while the low-reward condition rated an average local confidence across trials of 2.71 (SD = 0.62).

We predicted that receiving more rewards when playing the game in the first part of the experiment would lead participants to feel more confident that they would later answer correctly in the absence of any feedback. In a generalized estimating equation (GEE) model, we used accuracy on each trial and reward condition to predict trial-by-trial confidence ratings, allowing for an interaction between the two. Trial-by-trial accuracy (i.e., correct versus incorrect response) significantly predicted local confidence ratings ($b = 0.06$, SE = 0.005, $z = 161.13$, $p < 0.001$, 95% CI [0.05, 0.07]), indicating that participants were attuned to whether they had answered correctly. However, reward did not significantly predict local confidence overall ($b = 0.0001$, SE = 0.01, $z = 0.00$, $p = 0.99$, 95% CI [−0.03, 0.03]). Accuracy and reward did show a significant interaction ($b = 0.01$, SE = 0.005, $z = 4.72$, $p = 0.03$, 95% CI [0.001, 0.02]), suggesting that participants in the high reward (versus low reward) condition showed more differentiated confidence between correct and incorrect answers. Altogether, however, we did not observe credible evidence that rewards boosted overall later local confidence. In addition, the interaction effect suggesting sharpened metacognitive accuracy was not replicated in Study 2 and therefore should be interpreted with caution.

Study 2

Study 1 showed that rewards can bias self-evaluations of ability, particularly estimations of previous accuracy and predictions of future accuracy. Although participants observed exactly how they performed on each trial, those who received more frequent rewards from the slot machine following correct answers thought they had performed better. This finding suggests that reward had an immediate effect on information processing, though we did not observe credible evidence that this effect boosted later confidence in new responses made in the absence of reward. What mechanism might support the bias in task-specific self-beliefs? Study 2 investigated the trial-by-trial influence of rewards on expectations of accuracy as people develop a sense of their ability in the game. In particular, we asked whether rewards amplify the extent to which people learn about their own ability from good performance, leading to larger updates

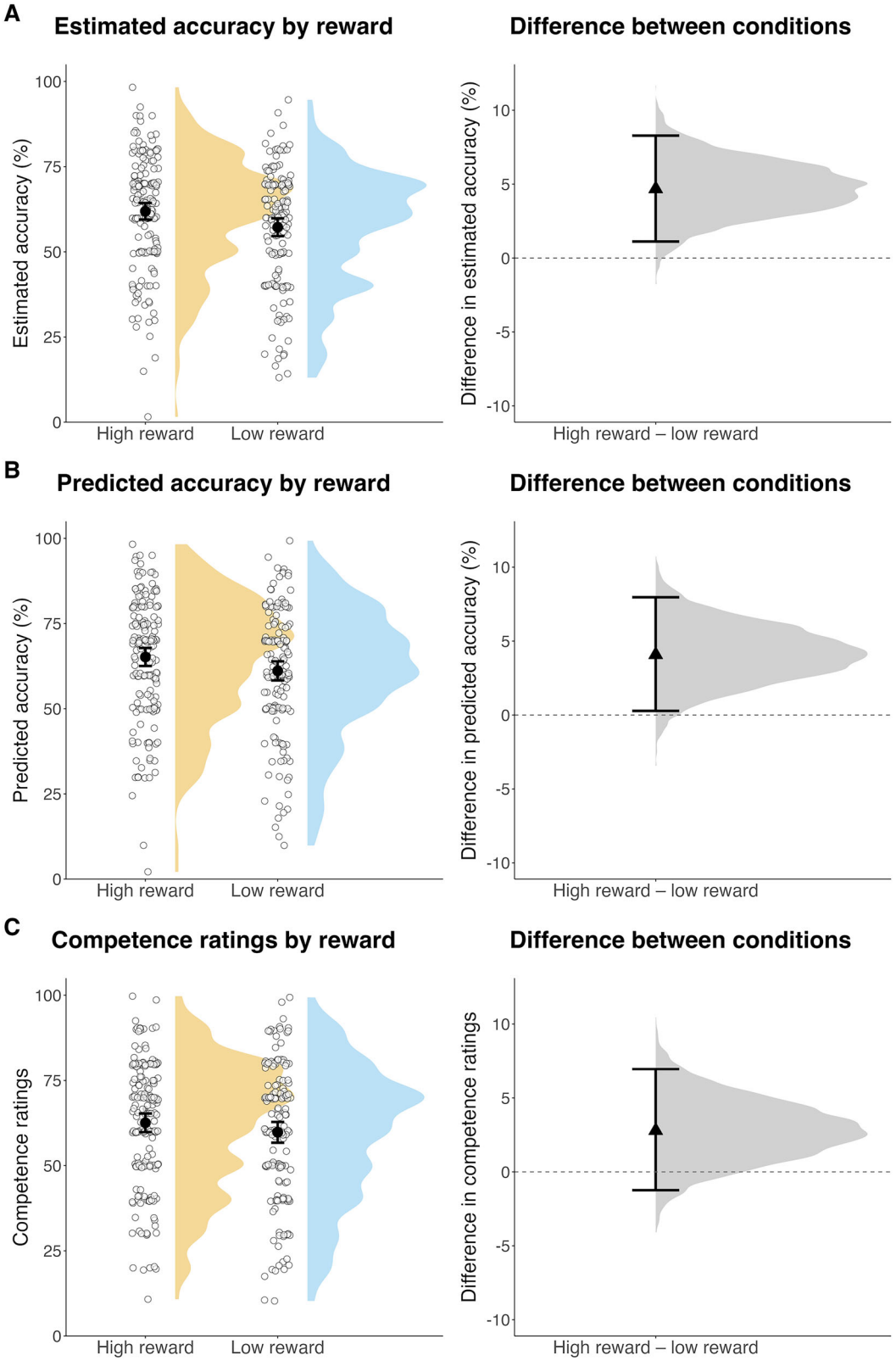

Fig. 3 | Comparison of task-specific self-belief measures between reward conditions. In each panel, individual data points are shown for each reward condition (on the left). $N = 340$ participants. Difference density plots (on the right) were generated by bootstrapped resampling (5000 times) to create an empirical distribution of the difference in means between high and low-reward conditions. Circles (on left) and triangles (on right) indicate means. Error bars represent ± 1 SE from the mean. **A** Distribution of estimated accuracy by reward condition. **B** Distribution of predicted accuracy by reward condition. **C** Distribution of competence ratings by reward condition.

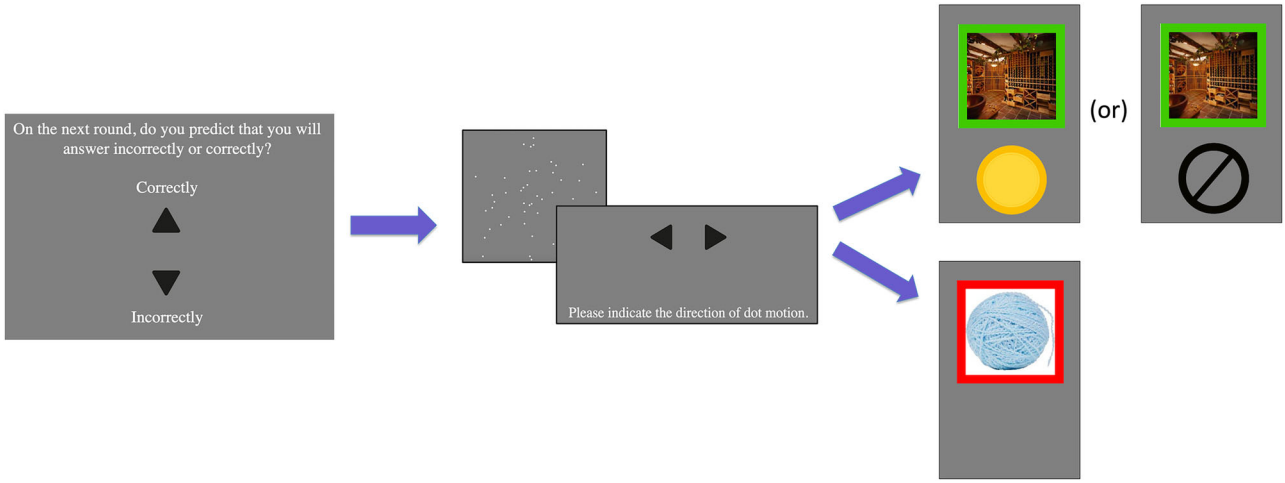

Fig. 4 | Schematic of Study 2 task flow. After each round, participants saw a trial-unique image. For correct responses, this image had a green frame, and for incorrect responses, the image had a red frame. If the response was incorrect, nothing else was displayed before the next round began. If the response was correct, either a token or a circle with a slash appeared underneath the first image. In this example, object images (e.g., a ball of yarn) served as "incorrect receipts" while scene images (e.g., a wine cellar) represented lottery tickets that could give a token (similar to the slot machine in Study 1). Object and scene images shown are reprinted with permission from the authors of the original publications[45,46].

in self-beliefs. That is, after a correct response, people can increase their estimate of their performance level, but this update may be greater when feedback is accompanied by a reward.

In Study 2, we used the same task as in Study 1 with some key additions. First, participants were now asked to indicate whether they expected to answer correctly or incorrectly before each trial of the game, allowing us to ask how rewards change perceptions of ability on a trial-by-trial basis. This procedure allowed us to test dynamic belief updating, asking how participants learned from each instance of feedback and updated their expectations accordingly. In addition, instead of seeing a slot machine, each trial was now associated with a trial-unique image representing a lottery ticket that could give a token for a correct answer or an incorrect receipt for an incorrect answer (Fig. 4).

Trial-by-trial performance expectation. We adapted reinforcement learning models to the task in order to model participants' beliefs about their likelihood of answering correctly on each trial. We compared models with and without a reward bias on updating. Although the reward model specified in our preregistration fit best (protected exceedance probability = 1), model recovery suggested that the models were not sufficiently identifiable (that is, the competing models could not be consistently distinguished from one another when fit to simulated data). Thus, any interpretation of model selection or parameter estimates should be made with caution. We report detailed results from this computational modeling in Supplementary Note 10.

Although not included in our preregistration, we therefore conducted a mixed-effects logistic regression to predict trial-by-trial participant expectations using accuracy and reward feedback on the immediately preceding trial. This regression approach offers a more robust and interpretable framework and approximates the preregistered computational model with a learning rate of 1 (i.e., modeling how participants update their ratings based on feedback on the most recent trial). We predicted that reward feedback would increase expectations of accuracy above and beyond the influence of accuracy feedback, such that participants' correct responses would have greater uptake into self-beliefs when accompanied by reward than not. Both accuracy ($b = 0.66$, SE = 0.05, $z = 14.59$, $p < 0.001$, 95% CI [0.57, 0.75]) and reward ($b = 0.10$, SE = 0.04, $z = 2.65$, $p = 0.01$, 95% CI [0.03, 0.18]) on the previous trial emerged as significant predictors, demonstrating that reward boosted expectations about performance beyond the effect of accuracy. Specifically, participants updated their expectations more in the positive direction when they had been rewarded for correct answers ($M = 0.03$, SE = 0.004) versus unrewarded ($M = 0.02$, SE = 0.0009; Fig. 5). As an exploratory analysis, two-tailed paired samples t-tests provided converging evidence that participants increased their trial-by-trial expectations significantly more not only after correct than incorrect responses ($t(341) = 9.73$, $d = 0.53$, $p < 0.001$, 95% CI [0.07, 0.10]), but also after correct-rewarded trials than correct-unrewarded trials ($t(341) = 2.26$, $d = 0.12$, $p = 0.02$, 95% CI [0.001, 0.03]). These results suggest that rewards shaped the extent to which people updated their expectations of future success, leading them to integrate recent accuracy feedback more into self-beliefs.

Task-specific self-beliefs. We next examined the same task-specific self-belief measures as in Study 1. Validating the staircase procedure, objective accuracy was 69.0% (high reward: $M = 0.70$, SD = 0.08; low reward: $M = 0.71$, SD = 0.07). An exploratory two-tailed independent samples t-test did not reveal significant differences between reward conditions ($t(332.21) = -1.92$, $d = 0.21$, $p = 0.06$, 95% CI [−0.03, 0.0004]).

For estimations of previous accuracy, we again found that both accuracy ($b = 84.64$, SE = 10.03, $t(339) = 8.44$, $p < 0.001$, 95% CI [64.91, 104.37]) and reward ($b = 1.71$, SE = 0.77, $t(337) = 2.23$, $p = 0.03$, 95% CI [0.20, 3.22]) predicted estimates of accuracy. Thus, participants in the high-reward condition again believed they had performed more accurately compared to those in the low-reward condition. This model explained 17.3% of the variance in estimations of accuracy ($Adjusted\ R^2 = 0.173$, $F(2, 346) = 36.54$, $f^2 = 0.22$, $p < 0.001$). However, reward was not significantly associated with either predictions of future accuracy ($b = 1.41$, SE = 0.85, $t(337) = 1.65$, $p = 0.10$, 95% CI [−0.27, 3.09]) or self-evaluations of competence ($b = 1.12$, SE = 0.91, $t(337) = 1.23$, $p = 0.22$, 95% CI [−0.67, 2.91]; see Supplementary Table 1). These results were consistent with exploratory analyses using actual reward rate instead of reward condition as a predictor (Supplementary Note 6 and Supplementary Table 10). These findings thus partially replicate the reward effects on task-specific self-beliefs observed in Study 1. Smaller effects on these measures in Study 2 may be due to trial-by-trial expectation questions drawing greater attention to accuracy, as well as more abstract and distal rewards than Study 1 (i.e., tokens that translated to raffle tickets for a lottery as opposed to a direct monetary bonus), which we consider further in the Discussion.

As in Study 1, we tested whether the results were robust to exclusion criteria. We fit exploratory (non-preregistered) models predicting each measure of task-level self-belief after excluding participants who spent most of the 80 rounds of the game at the easiest level. In these analyses, we

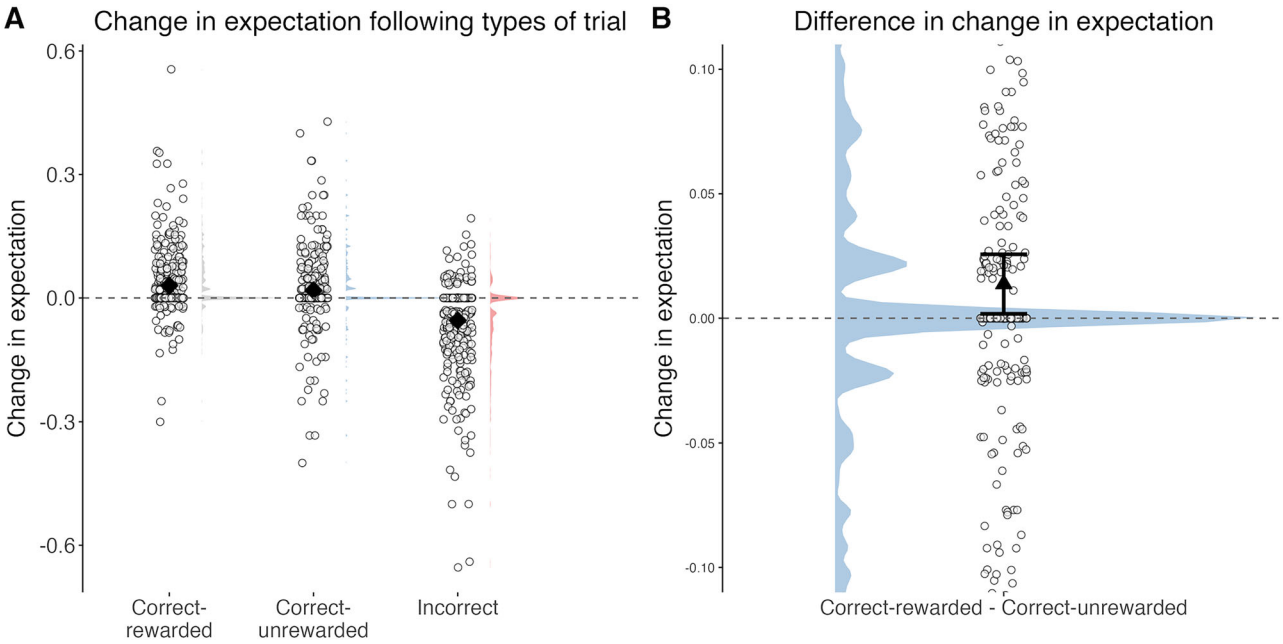

Fig. 5 | Comparison of changes in performance expectations following different feedback. A Each dot represents one participant's mean change in performance expectation following each type of trial. $N = 342$ participants. Diamonds indicate means. **B** Each dot represents the difference between one participant's mean change in performance expectation following correct-rewarded trials and their mean change in performance expectation following correct-unrewarded trials. $N = 342$ participants. The triangle indicates the mean, and the error bars represent the 95% confidence interval around the mean.

observed stronger effects of reward on task-specific self-beliefs, across estimated accuracy ($b = 2.53$, SE = 0.73, $t(383) = 3.47$, $p < 0.001$, 95% CI [1.10, 3.97]), predicted accuracy ($b = 2.23$, SE = 0.79, $t(383) = 2.84$, $p = 0.005$, 95% CI [0.69, 3.78]), and competence ratings ($b = 1.95$, SE = 0.85, $t(383) = 2.29$, $p = 0.02$, 95% CI [0.28, 3.63]; see Supplementary Note 5 and Supplementary Table 7). We also refit multiple regression models with an additional predictor that captures the ending difficulty level of the task for each participant, and the reward effect on estimated accuracy remained significant ($b = 1.68$, SE = 0.77, $t(336) = 2.19$, $p = 0.03$, 95% CI [0.17, 3.20]; see Supplementary Note 4 and Supplementary Table 3).

Local confidence. As in Study 1, we also tested whether rewards during the main phase of the game would affect later confidence in a separate phase without accuracy or reward feedback. The average accuracy in these trials was 73.40% (high reward: $M = 0.72$, SD = 0.13; low reward: $M = 0.74$, SD = 0.14), and a (non-preregistered) two-tailed independent samples t-test revealed no significant differences between reward conditions ($t(339.34) = -1.34$, $d = 0.14$, $p = 0.18$, 95% CI [−0.05, 0.01]). The high-reward condition rated an average local confidence across trials of 2.86 (SD = 0.65), while the low-reward condition rated an average local confidence across trials of 2.91 (SD = 0.57).

Although trial-by-trial accuracy significantly predicted local confidence ($b = 0.05$, SE = 0.004, $z = 175.20$, $p < 0.001$, 95% CI [0.05, 0.06]), reward did not ($b = -0.005$, SE = 0.01, $z = 0.18$, $p = 0.67$, 95% CI [−0.03, 0.02]). The interaction between the two was also non-significant ($b = -0.004$, SE = 0.004, $z = 1.07$, $p = 0.30$, 95% CI [−0.01, 0.004]). The interaction was somewhat sensitive to model assumptions, such that an alternative model specification suggested a *negative* interaction between reward and accuracy ($b = -0.02$, SE = 0.002, $z = -11.80$, $p < 0.001$, 95% CI [−0.03, −0.02]), unlike Study 1 (see Supplementary Note 8 for details). Overall, we observe no credible evidence that rewards in the main phase of the game boosted local confidence in the absence of feedback, and we did not observe consistent interaction effects across studies.

Game perceptions. We again examined whether rewards influenced perceptions of the game beyond self-evaluations of ability. In multiple regression models, reward was not significantly associated with enjoyment ($b = -0.21$, SE = 1.36, $t(339) = -0.15$, $p = 0.88$, 95% CI [−2.90, 2.47]), perceived difficulty ($b = -1.60$, SE = 1.15, $t(339) = -1.39$, $p = 0.17$, 95% CI [−3.86, 0.66]), or self-reported motivation ($b = -0.64$, SE = 1.32, $t(339) = -0.49$, $p = 0.63$, 95% CI [−3.22, 1.93]; see Supplementary Note 9 and Supplementary Table 14).

Link to self-esteem. Finally, in order to assess links to broader self-views, we added a measure of general self-esteem in Study 2 and examined the link between this measure and task-specific self-beliefs. In a multiple linear regression model with reward and self-esteem as predictors, self-esteem was significantly associated with competence ratings ($b = 3.37$, SE = 1.43, $t(314) = 2.36$, $p = 0.02$, 95% CI [0.55, 6.19]), which reflect generalized judgments about ability in the task, though not with estimations of prior accuracy ($b = 1.85$, SE = 1.21, $t(314) = 1.52$, $p = 0.13$, 95% CI [−0.54, 4.24]) or predictions of future accuracy ($b = 1.66$, SE = 1.36, $t(314) = 1.22$, $p = 0.22$, 95% CI [−1.02, 4.33]). These results suggest a link between task-specific competence judgments and general views about the self (see Supplementary Note 9 and Supplementary Tables 15 and 16 for more details).

Discussion

Summary

Across two behavioral experiments, we examined the impact of rewards on self-evaluations of ability in a task by decoupling rewards from objective performance feedback. Results showed that rewards biased people's understanding of their prior performance, even when they had access to objective performance feedback. We then probed the learning dynamics behind this bias, finding that rewards increase the extent to which people update their beliefs following good performance in a task.

These findings expand our understanding of the links between reward learning and self-beliefs. Prior research has already established several links between rewards and self-beliefs. First, rewards influence motivation, which can shape future performance, and, in turn, impact self-beliefs[8,50–52]. Second, in many cases, rewards are reflective of ability to some extent, and in these situations, people may use rewards as diagnostic indicators of performance

to inform self-evaluations. The current findings extend prior work by demonstrating that rewards may still exert a direct influence on self-evaluations beyond motivation or any informative value, potentially by virtue of their affective qualities. This account would be in line with prior work showing affect-related biases in forming and updating self-beliefs[14,53] and reward-driven biases in belief updating more generally[54,55]. Our findings also complement recent work showing that rewards bias people's impressions of *others*' competence even when objective performance is known[20], and work demonstrating that monetary incentives for accurate confidence judgments biased these very judgments[18]. The present work demonstrates that reward biases people's overall impressions of their own performance and abilities in a task domain, even when accuracy feedback is provided. In contrast, we did not detect consistent evidence that rewards for correct responses influenced metacognitive confidence in later responses in the absence of reward.

The present studies also add to a body of work on inaccuracies in people's self-evaluations of ability. Perceptions of skill often show only small correlations with objective performance[56]. The underestimation of performance in the current work aligns with previous work showing that people update their estimates of their performance more after negative feedback when they perceive an opportunity to improve, as when learning a new skill[28]. Given that self-evaluations then shape goal persistence and future outcomes[1], understanding the nature of these flawed self-assessments can help inform interventions to mitigate their detrimental effects. Here, we suggest an additional way in which self-beliefs can fail to reflect reality: They can be constructed through the experience of reward, which may reflect influences beyond objective performance, such as the availability of resources within an institution.

Limitations

Although our current studies afforded several benefits pertaining to experimental control, they are also limited in generalizability to more complex, real-world learning scenarios. By using perceptual games, we presented participants with a task in which they were unlikely to have an existing self-concept. This approach also allowed us to use a staircase procedure to control average performance, ensuring that differences in reward did not promote differences in rates of correct responses between conditions on average. On the other hand, learning a new skill in the real world is likely to last much longer than the length of our current experiments and typically occurs over many repeated experiences. Moreover, the rewards used in the present studies were small; in Study 1, tokens translated to monetary bonuses of under a dollar, and Study 2 used an indirect reward structure, in which tokens translated to raffle tickets in a lottery for a 10-dollar gift card. These reward manipulations may blunt the positive feelings that accompany rewards in real life. The change to a less direct reward structure in Study 2 may have contributed to the smaller effect sizes observed in Study 2 compared to Study 1 on the task-specific self-belief measures. In addition, drawing more attention to trial-by-trial accuracy in Study 2 may have dampened the influence of reward on overall beliefs as compared to Study 1 and deviated more from the way in which people spontaneously track their own ability. Nonetheless, this approach allowed us to assess moment-to-moment changes in self-beliefs during learning as opposed to retroactive judgment, informing how rewards change information uptake.

In the present study, rewards were extrinsic. This approach was used to dissociate performance and reward, such that correct performance could be rewarded or unrewarded. However, similar psychological processes may apply to intrinsic reward. Good performance can be intrinsically rewarding and elicit feelings of pride[11], and there is overlap in neural representation of reward value, accuracy, and confidence[22,25]. Even in these cases, reward could have an affective impact that boosts self-evaluations of ability above and beyond its informative properties. Indeed, beliefs about the self tend to hold intrinsic value and induce emotions that can bias learning[10,14]. Accordingly, the current results do not necessarily suggest that extrinsic reward is needed to generate self-confidence; in fact, extrinsic reward can also have negative effects on long-term motivation[57]. Rather, they highlight a direct role of reward in shaping self-beliefs. Future work can test how intrinsic affective responses relate to these shifts in self-beliefs.

In the current study, rewards were delivered probabilistically (via a slot machine or lottery ticket), but only after correct answers. It therefore remains an open question how the observed reward bias may change in situations when reward feedback is either entirely unrelated to accuracy or even more tightly linked to performance. When it is clear exactly why a reward was or was not garnered for a particular response, it is possible that people can more easily determine whether they earned it, which might shape the extent to which rewards influence self-beliefs. Future work could examine how uncertainty around rewards affects the impact of feedback on self-perceptions. In addition, the current study used extrinsic *monetary* rewards, but rewards also often come in the form of praise or verbal reinforcement. Future work can test the role of social rewards, which rely on similar brain processes to monetary rewards[58] but may have different effects on intrinsic motivation[59].

Nonetheless, given that people often receive extrinsic rewards for performance in the real world, future work can test how the current findings may apply to scholastic achievement. Students' self-beliefs directly influence performance and future achievement, informing their decisions about which challenges to take on and what paths to pursue to maximize chances of success[60].

Conclusion

In sum, we identify an influence of rewards on self-belief formation, above and beyond objective feedback about performance. This work illuminates the computations through which people learn about their abilities over time and highlights influences from reward learning that can complement more conceptual forms of learning[61]. By dissociating affective influences from objective information, these findings can inform how individuals form expectations of success that guide significant decisions about when to persevere and when to give up.

Data availability

The raw, trial-by-trial subject-level data that support the current findings are available in the study's online repository in CSV format: https://doi.org/10.17605/OSF.IO/Z4J98[62].

Code availability

The code for running the experiment (Psychopy file and dependencies) and for all analyses of the data (R files) that appear in both the main text and supplemental information can be found in the study's online repository: https://doi.org/10.17605/OSF.IO/Z4J98[62].

Received: 17 January 2025; Accepted: 24 June 2025;
Published online: 01 October 2025

References

1. Fouad, N. A., Smith, P. L. & Zao, K. E. Across academic domains: extensions of the social-cognitive career model. *J. Counsel. Psychol.* **49**, 164–171 (2002).
2. Oyserman, D. Identity-based motivation. *Emerg. Trends Soc. Behav. Sci.* **38**, 1–11 (2015).
3. Markus, H. & Wurf, E. The dynamic self-concept: a social psychological perspective. *Annu. Rev. Psychol.* **38**, 299–337 (1987).
4. Markus, H. & Kunda, Z. Stability and malleability of the self-concept. *J. Personal. Soc. Psychol.* **51**, 858–866 (1986).
5. Gore, J. S. & Cross, S. E. Who am i becoming? A theoretical framework for understanding self-concept change. *Self Identity* **13**, 740–764 (2014).

6. Condly, S. J., Clark, R. E. & Stolovitch, H. D. The effects of incentives on workplace performance: a meta-analytic review of research studies. *Perform. Improv. Q.* **16**, 46–63 (2003).
7. Jenkins Jr, G. D., Mitra, A., Gupta, N. & Shaw, J. D. Are financial incentives related to performance? A meta-analytic review of empirical research. *J. Appl. Psychol.* **83**, 777 (1998).
8. Bem, D. J. *Advances in Experimental Social Psychology*, Vol. 6. (ed. Berkowitz, L.) 1–62 (Academic Press, 1972).
9. Melnikoff, D. E. & Strohminger, N. Bayesianism and wishful thinking are compatible. *Nat. Hum. Behav.* **8**, 692–701 (2024).
10. Bromberg-Martin, E. S. & Sharot, T. The value of beliefs. *Neuron* **106**, 561–565 (2020).
11. Müller-Pinzler, L. et al. Neurocomputational mechanisms of affected beliefs. *Commun. Biol.* **5**, 1241 (2022).
12. Storbeck, J. & Clore, G. L. Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Soc. Personal. Psychol. Compass* **2**, 1824–1843 (2008).
13. Koban, L. et al. Social anxiety is characterized by biased learning about performance and the self. *Emotion* **17**, 1144–1155 (2017).
14. Sharot, T. & Garrett, N. Forming beliefs: why valence matters. *Trends Cogn. Sci.* **20**, 25–33 (2016).
15. Schwarz, N. *Handbook of Theories of Social Psychology* Volume 1 (eds Van Lange, P. A. M., Kruglanski, A. & Higgins, E. T.) 289–308 (SAGE Publications Ltd, 2012).
16. Schwarz, N. & Clore, G. L. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Personal. Soc. Psychol.* **45**, 513–523 (1983).
17. Song, H. & Schwarz, N. If it's hard to read, it's hard to do: processing fluency affects effort prediction and motivation. *Psychol. Sci.* **19**, 986–988 (2008).
18. Lebreton, M. et al. Two sides of the same coin: monetary incentives concurrently improve and bias confidence judgments. *Sci. Adv.* **4**, eaaq0668 (2018).
19. Rouault, M., Dayan, P. & Fleming, S. M. Forming global estimates of self-performance from local confidence. *Nat. Commun.* **10**, 1141 (2019).
20. Hackel, L. M., Kogon, D., Amodio, D. M. & Wood, W. Group value learned through interactions with members: a reinforcement learning account. *J. Exp. Soc. Psychol.* **99**, 104267 (2022).
21. Daniel, R. & Pollmann, S. Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* **59**, 3457–3467 (2012).
22. Foerde, K. & Shohamy, D. Feedback timing modulates brain systems for learning in humans. *J. Neurosci.* **31**, 13157–13167 (2011).
23. Rouault, M. & Fleming, S. M. Formation of global self-beliefs in the human brain. *Proc. Natl Acad. Sci. USA* **117**, 27268–27276 (2020).
24. Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32**, 6117–6125 (2012).
25. Hoven, M. et al. Motivational signals disrupt metacognitive signals in the human ventromedial prefrontal cortex. *Commun. Biol.* **5**, 244 (2022).
26. Wittmann, M. K. et al. Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91**, 482–493 (2016).
27. Wittmann, M. K. et al. Causal manipulation of self-other mergence in the dorsomedial prefrontal cortex. *Neuron* **109**, 2353–2361.e11 (2021).
28. Müller-Pinzler, L. et al. Negativity-bias in forming beliefs about own abilities. *Sci. Rep.* **9**, 14416 (2019).
29. Fleming, S. M. Metacognition and confidence: a review and synthesis. *Annu. Rev. Psychol.* **75**, 241–268 (2024).
30. Faul, F., Erdfelder, E., Buchner, A. & Lang, A. Statistical power analyses using g* power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
31. Peirce, J. W. Psychopy-psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).
32. Leek, M. R. Adaptive procedures in psychophysical research. *Percept. Psychophys.* **63**, 1279–1292 (2001).
33. Kuroki, D. & Pronk, T. jsquestplus: a Javascript implementation of the quest+ method for estimating psychometric function parameters in online experiments. *Behav. Res. Methods* **55**, 3179–3186 (2023).
34. Watson, A. B. & Pelli, D. G. Quest: a Bayesian adaptive psychometric method. *Percept. Psychophys.* **33**, 113–120 (1983).
35. Rahnev, D. & Fleming, S. M. How experimental procedures influence estimates of metacognitive ability. *Neurosci. Conscious.* **2019**, niz009 (2019).
36. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (R Core Team, 2024).
38. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org (Springer-Verlag New York, 2016).
39. Kay, M. *ggdist: Visualizations of Distributions and Uncertainty*). https://mjskay.github.io/ggdist/. R package version 3.3.2 (2024).
40. Wilke, C. O. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot. R package version 1.1.3 (2024).
41. Zeileis, A. Object-oriented computation of sandwich estimators. *J. Stat. Softw.* **16**, 1–16 (2006).
42. Zeileis, A. & Hothorn, T. Diagnostic checking in regression relationships. *R. N.* **2**, 7–10 (2002).
43. Yan, J. geepack: Yet another package for generalized estimating equations. *R.-N.* **2/3**, 12–14 (2002).
44. Christensen, R. H. B. *Ordinal—Regression Models for Ordinal Data*. https://CRAN.R-project.org/package=ordinal. R package version 2023.12-4.1 (2023).
45. Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol.: Gen.* **139**, 558 (2010).
46. Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol. Sci.* **21**, 1551–1556 (2010).
47. Rosenberg, M. Rosenberg self-esteem scale (rse). *Accept. Commit. Ther. Meas. Package* **61**, 18 (1965).
48. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
49. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
50. Skinner, B. F. The generic nature of the concepts of stimulus and response. *J. Gen. Psychol.* **12**, 40–65 (1935).
51. Hendijani, R., Bischak, D. P., Arvai, J. & Dugar, S. Intrinsic motivation, external reward, and their effect on overall motivation and performance. *Hum. Perform.* **29**, 251–274 (2016).
52. Eisenberger, R., Pierce, W. D. & Cameron, J. Effects of reward on intrinsic motivation-negative, neutral, and positive: comment on Deci, Koestner, and Ryan (1999). *Psychol. Bull.* **125**, 677–691 (1999).
53. Koellinger, P. & Treffers, T. Joy leads to overconfidence, and a simple countermeasure. *PLoS ONE* **10**, e0143263 (2015).
54. Babür, B. G., Leong, Y. C., Pan, C. X. & Hackel, L. M. Neural responses to social rejection reflect dissociable learning about relational value and reward. *Proc. Natl Acad. Sci. USA* **121**, e2400022121 (2024).
55. Fischer, A. G., Bourgeois-Gironde, S. & Ullsperger, M. Short-term reward experience biases inference despite dissociable neural correlates. *Nat. Commun.* **8**, 1690 (2017).
56. Dunning, D., Heath, C. & Suls, J. M. Flawed self-assessment: implications for health, education, and the workplace. *Psychol. Sci. Public Interest* **5**, 69–106 (2004).

57. Cameron, J. Negative effects of reward on intrinsic motivation-a limited phenomenon: comment on Deci, Koestner, and Ryan (2001). *Rev. Educ. Res.* **71**, 29–42 (2001).
58. Bhanji, J. P. & Delgado, M. R. The social brain and reward: social information processing in the human striatum. *WIREs Cogn. Sci.* **5**, 61–73 (2014).
59. Cameron, J. & Pierce, W. D. Reinforcement, reward, and intrinsic motivation: a meta-analysis. *Rev. Educ. Res.* **64**, 363–423 (1994).
60. Felson, R. B. The effect of self-appraisals of ability on academic performance. *J. Personal. Soc. Psychol.* **47**, 944–952 (1984).
61. Elder, J., Davis, T. & Hughes, B. L. Learning about the self: motives for coherence and positivity constrain learning from self-relevant social feedback. *Psychol. Sci.* **33**, 629–647 (2022).
62. Luo, J., Mende-Siedlecki, P. & Hackel, L. M. Rewards Shape Self-evaluations of Ability. *Open Sci. Framework.* osf.io/z4j98 (2025).

Acknowledgements

We would like to thank members of the Social Learning and Choice Lab for helpful comments and feedback.

Author contributions

J.L., L.H., and P.M.S. conceptualized the study. J.L. programmed the task, collected the data, conducted statistical analyses, and drafted the manuscript. L.H. and P.M.S. supported analyses, provided supervision, offered critical feedback on drafts, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-025-00286-7.

Correspondence and requests for materials should be addressed to Jean Luo.

Peer review information *Communications Psychology* thanks Niv Reggev and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editors: Troby Ka-Yan Lui. [A peer review file is available].