nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02253-y

Systematic review and meta-analysis of educational approaches to reduce cognitive biases among students

Received: 19 February 2024

Accepted: 23 May 2025

Published online: 26 August 2025

Check for updates

Ghassani Swaryandini[1], Jessica Graham[1], Shantell Griffith[1], Vasco Grilo[2], Federica Ruzzante[3], Xingruo Zhang[4], Siu Kit Yeung[5], Marta Mangiarulo[6], Geetanjali Basarkod[7], Clarence Ng[8], Philip Parker[7], Jason Tangen[1], Alexander Saeri[1], Emily Grundy[9], Peter Slattery[9] & Michael Noetel[1]

Resistance to cognitive biases is a crucial element of rationality that influences judgement and decision-making. Here we synthesized the effects of debiasing training in educational settings. Our systematic review found 54 randomized controlled trials consisting of 383 effect sizes and 10,941 participants. Our meta-analysis of educational interventions showed a small, yet significant, improvement in reducing the likelihood of committing biases compared with control conditions ($g = 0.26$, 95% confidence interval 0.14 to 0.39), 160 effects from 41 studies, $P < 0.001$). Most studies focused on reducing the likelihood of committing biases (for example, confirmation bias) using cognitive strategies. Some biases seemed difficult to overcome (for example, representativeness heuristic), and questions remain about the depth and transferability of learning beyond classroom settings. All studies had unclear or high risk of bias and there was some risk of publication bias. While evidence suggests that educational interventions can reduce bias on targeted tasks, more research is needed to determine whether these improvements translate to meaningful changes in real-world decision-making and to identify which paedagogical approaches are most effective for reducing the influence of cognitive biases.

To navigate a complex world, young people need to learn how to make good decisions[1]. While many assume good decision-making stems from intelligence, research shows intelligence and rationality are only modestly correlated[2]. Intelligence refers to cognitive abilities to process information and acquire knowledge, whereas rationality describes our ability to make good decisions with available information[3,4]. One key to rational decision-making—and a major component of rational thinking tests—is our ability to overcome cognitive biases[4]. Given how influential cognitive biases are in judgement and decision-making, it is important to understand whether young people can learn to overcome those biases[5].

It is not obvious that we can overcome bias. Daniel Kahneman once thought "training was hopeless for all kinds of judgements"[6], yet systematic reviews have shown debiasing interventions can reduce professional biases in social work[7], medical diagnostics[8] and health-related judgements[9]. These reviews focused on specific contexts—where

[1]School of Psychology, The University of Queensland, Brisbane, Queensland, Australia. [2]Frente Animal, Porto, Portugal. [3]MoMiLab Research Unit, IMT School for Advanced Studies, Lucca, Italy. [4]ILR School, Cornell University, Ithaca, NY, USA. [5]Department of Psychology, Chinese University of Hong Kong, Hong Kong, China. [6]School of Psychology and Vision Sciences, University of Leicester, Leicester, UK. [7]Institute for Positive Psychology and Education, Australian Catholic University, Sydney, New South Wales, Australia. [8]Institute for Learning Sciences and Teacher Education, Australian Catholic University, Brisbane, Queensland, Australia. [9]Massachusetts Institute of Technology, Cambridge, MA, USA. e-mail: m.noetel@uq.edu.au

https://doi.org/10.1038/s41562-025-02253-y

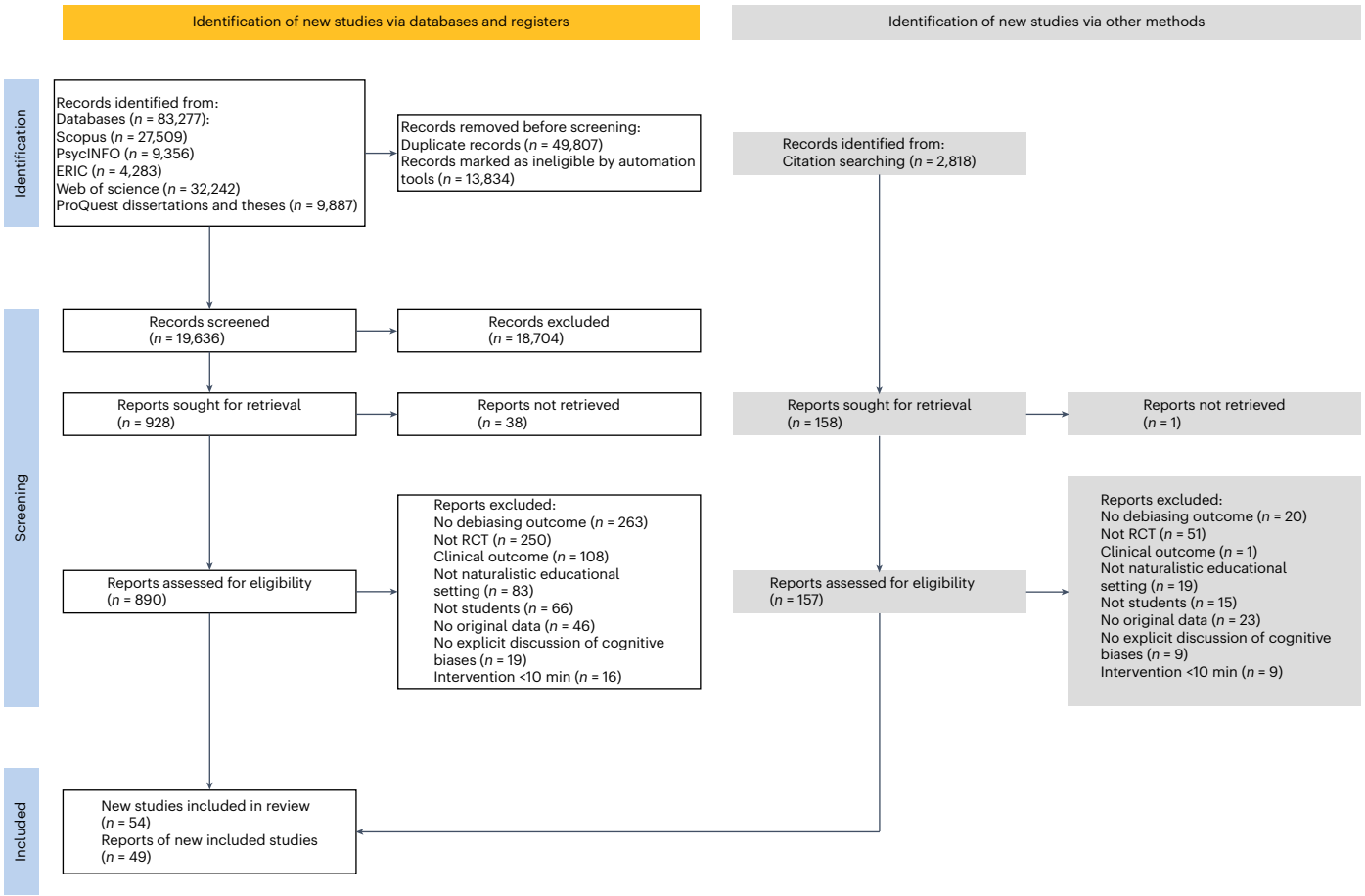

Fig. 1 | PRISMA flow diagram for study selection. The PRISMA diagram shows systematic review selection process: 19,636 records were screened, 1,086 full-text articles were assessed and 54 studies were included in the qualitative synthesis (41 in the meta-analysis) with a total sample of 10,941 participants.

participants may be particularly motivated for success. It may be harder to teach students judgement and decision-making due to the wide range of decisions they face. So, it is unclear whether educational interventions can decrease bias when taught at schools and universities. These institutions are important vehicles for teaching rational thinking, so this review aims to see whether educators can help students overcome bias.

The influential heuristics and biases framework[10] posits that people often rely on mental shortcuts (heuristics), which can lead to systematic errors (cognitive biases) in judgement and decision-making[10]. While some researchers argue that heuristics can lead to efficient and accurate intuitive decision-making[11], especially in practised, predictable situations[12] or social judgements[13], heuristics can also lead to decisions that hinder our goals, potentially causing various societal issues[14–17]. Our study acknowledges this debate while focusing on the potential benefits of mitigating cognitive biases that hinder optimal judgement and decision-making.

Cognitive biases occur when one's construction of reality does not match the truth[18]. For example, availability heuristic helps people make numerical judgements (for example, risks) based on how easily something comes to mind[10], which leads more people to fear flying over driving despite higher mortality rates from car crashes[19]. Failure to override cognitive biases can lead to poor judgement when weighing uncertain outcomes, leading to poor decision-making[4,20]. The ability to overcome cognitive biases is a part of broader assessments of rationality because rationality predicts good decision-making more than intelligence[4,21–23].

Traditional intelligence tests (for example, the Weschler Adult Intelligence Scale)[24] do not capture most rationality components, prompting Stanovich et al.[4] to develop the Comprehensive Assessment of Rational Thinking (CART) as a gold-standard operationalization of rationality. Almost half of the CART subtests measure one's ability to override cognitive biases—a core component not typically included in educational curricula; other rationality components (that is, probabilistic and scientific reasoning[4,25]) are more commonly taught[25,26]. With tools such as CART, researchers can now explore whether and how we can train people to mitigate cognitive biases.

Mitigating cognitive biases exists on a continuum—one can always strive to be less influenced by biases but never be fully free from them. As cognitive biases are complex to measure, researchers use various tasks as proxies. For example, the belief bias task requires participants to evaluate arguments based on logic rather than the plausibility of its conclusion[4,27–29]. Critics note that prior exposure to heuristics and biases tasks could have improved their performance simply because they have completed the task before rather than showing genuine improvement[30,31]. However, measures such as the decision-making Competence[32] and CART[4] have been linked to real-world outcomes, such as lower risk-taking behaviours in financial decisions or substance use[29,33]. While most cognitive biases cannot be fully overcome, heuristics and biases tasks appear to be relatively valid proxies for cognitive bias mitigation.

'Debiasing' education aims at reducing the impact of cognitive biases[5]. Global companies such as Google and Starbucks have spent billions of dollars on bias reduction training, such as racial and gender biases[34–36]. However, attempts to train people to mitigate cognitive biases have shown mixed results. Some interventions were successful in reducing framing and bandwagoning biases, but not for anchoring and halo bias[37]. Another study managed to mitigate confirmation bias (CB),

fundamental attribution error (FAE), bias blind spot (BBS) and social projection bias, but not anchoring bias[38]. This suggests different biases may have different underlying mechanisms, so we need to explore which biases can be most effectively addressed through training.

Critical thinking and rational thinking are closely connected, with critical thinking considered part of rationality[39]. Ennis' critical thinking dispositions (for example, seek and offer clear reasons and consider other points of view) and abilities (for example, analyse arguments and judge source credibility) parallel skills that make us resistant to cognitive biases[40,41]. While critical thinking education is embedded in educational standards such as the Common Core State Standards[42] and Next Generation Science Standards[43] in the USA, these standards rarely address cognitive biases, plausibly due to doubts about whether we can learn to overcome them[6]. Hence, research in critical thinking education can inform debiasing interventions.

A meta-analysis showed that successful critical thinking interventions typically provide opportunities for dialogue and discussion, expose students to real-world problems and incorporate mentoring[44]. Combining these components produces larger effects ($g = 0.57$, $P < 0.001$) compared with just using authentic instruction ($g = 0.25$), dialogue ($g = 0.23$) or combining authentic instruction and dialogue without mentoring ($g = 0.32$)[44]. Similarly, game-based learning and collaborative problem-solving methods produced larger effect sizes (that is, $g_{\text{game based}} = 0.86$; standardized mean difference(SMD)$_{\text{collaborative problem solving}} = 0.82$)[45,46] than class-based instruction ($g = 0.39$)[44]. Teaching clear critical thinking principles while having students apply them to specific contexts (that is, the 'mixed approach')[47] yielded stronger outcomes ($g = 0.38$) than teaching abstract principles ($g = 0.29$) or subject-specific applications ($g = 0.23$) in isolation[44]. As critical thinking and rationality share characteristics and underlying philosophies[48], teaching methods probably influence debiasing training effectiveness as well.

This review aims to synthesize the literature on debiasing training in educational settings. First, we examined whether educational interventions reduce biases overall. Second, we hypothesized that some biases would be more amenable to change than others, as found in other domains[37,38]. We tested whether effects of educational interventions differed on the bias being targeted (for example, availability heuristic versus CB) and how it was measured (for example, knowledge of bias versus likelihood of committing it). We deemed the likelihood of committing a bias more important than mere knowledge of various biases, because knowledge does not automatically decrease the likelihood of committing the bias. Third, as is true for critical thinking[44] and other learning areas[49], some methods for teaching students about biases would probably be more effective than others. We examined whether educational design moderated the effects (for example, online versus face to face). We explored whether demographics (for example, age and gender) and contexts (for example, school versus university) influenced outcomes. We assessed whether the risk of bias in the studies affected the results. Ultimately, we sought to determine whether students can be taught to mitigate biases and if so, when and how they learn best.

Results

Study characteristics

The PRISMA (preferred reporting items for systematic reviews and meta-analyses) flow diagram (Fig. 1) shows our study selection process[50,51]. Following a database search, duplicate removal and filtering studies through RobotSearch, we screened 19,636 titles and abstracts, plus 2,818 studies from the reference lists of included studies. After screening, we assessed 1,086 full-text articles against eligibility criteria. Of those, we excluded 1,037 studies, usually because they were not randomized or because there was no measure of cognitive bias (full reasons in Supplementary Table 1). This further selection yielded 54 studies across 49 papers and a pooled sample size of 10,941 participants. The characteristics of each included study are presented in Table 1.

Effects of teaching debiasing in education

Overall, teaching debiasing had a significant positive effect on reducing the likelihood of committing biases when compared with control conditions (no intervention or active control conditions; $g = 0.26$, 95% confidence interval (CI) 0.14 to 0.39, $n_{\text{studies}} = 41$, $k_{\text{effect sizes}} = 160$, $P < 0.001$; Fig. 2). There was a high level of heterogeneity not explained by sampling error alone ($I^2 = 74.94$). We explored the heterogeneity effects by assessing the likely distribution of true effect sizes that would be meaningfully helpful (that is, $g > 0.2$) or harmful (that is, $g < -0.2$). Teaching debiasing to students may be helpful in over half of the interventions (59% of true effects, 95% CI 46% to 72%), whereas only a small proportion of interventions may be harmful (4% of true effects, 95% CI 1% to 10%). To better explain the heterogeneity, we conducted several moderator analyses, which are outlined below. In response to a reviewer's comment, we made some changes to the pre-registered moderation analyses. The results of the moderation analyses for age, gender and educational setting, as well as the associations between moderators, are available in Supplementary Note 1.

Moderation analyses by cognitive biases

Different cognitive biases were a significant moderator when comparing the debiasing intervention with the control condition ($F(7,126) = 2.25$, $n_{\text{studies}} = 37$, $k_{\text{effect sizes}} = 134$, $P = 0.03$). As shown in Fig. 3, debiasing training successfully attenuated mixed measures of bias (that is, where the tasks measured multiple cognitive biases; $g = 0.59$, 95% CI 0.35 to 0.84, $n_{\text{studies}} = 7$, $k_{\text{effect sizes}} = 43$, $P < 0.001$) and overconfidence ($g = 0.28$, 95% CI 0.05 to 0.50, $n_{\text{studies}} = 9$, $k_{\text{effect sizes}} = 18$, $P = 0.02$). Both mixed measures of bias ($P_{\text{adj}} < 0.001$) and overconfidence ($P_{\text{adj}} = 0.03$) remained significant after correcting for the false discovery rate (FDR). As shown in Fig. 3, effects were not significant for the FAE, causal illusions, judgement accuracy, anchoring, representativeness heuristic and CB.

These biases could be naturally grouped into those associated with social group membership (that is, FAE, stereotyping, halo bias and weight bias) and errors due to cognitive heuristics (for example, anchoring and causal illusions). Pooled effects for these two groups of biases were not significantly different from each other ($F(1,158) = 2.55$, $P = 0.11$; social group membership biases: $g = 0.41$, 95% CI 0.12 to 0.71, $n_{\text{studies}} = 7$, $k_{\text{effect sizes}} = 21$, $P = 0.008$; cognitive heuristics: $g = 0.25$, 95% CI 0.12 to 0.39, $n_{\text{studies}} = 36$, $k_{\text{effect sizes}} = 139$, $P < 0.001$).

Moderation analyses by intervention format and duration

The delivery format of the intervention (for example, text/reading or video based) was a significant moderator ($F(4,155) = 2.96$, $n_{\text{studies}} = 41$, $k_{\text{effect sizes}} = 160$, $P = 0.02$). Educational games had the largest effect size ($g = 0.60$, 95% CI 0.33 to 0.87, $n_{\text{studies}} = 4$, $k_{\text{effect sizes}} = 11$, $P < 0.001$, $P_{\text{adj}} < 0.001$), followed by traditional training ($g = 0.21$, 95% CI 0.03 to 0.38, $n_{\text{studies}} = 18$, $k_{\text{effect sizes}} = 76$, $P = 0.02$, $P_{\text{adj}} = 0.04$). Text/reading ($g = 0.26$, 95% CI 0.02 to 0.50, $n_{\text{studies}} = 8$, $k_{\text{effect sizes}} = 17$, $P = 0.03$, $P_{\text{adj}} = 0.056$) and video-based training ($g = 0.44$, 95% CI 0.03 to 0.86, $n_{\text{studies}} = 3$, $k_{\text{effect sizes}} = 26$, $P = 0.04$, $P_{\text{adj}} = 0.06$) were not statistically significant following FDR correction. Mixed delivery showed a non-significant effect (a combination of different formats; $g = 0.18$, 95% CI −0.03 to 0.38, $n_{\text{studies}} = 10$, $k_{\text{effect sizes}} = 30$, $P = 0.09$, $P_{\text{adj}} = 0.13$).

The length of the intervention ($F(1,143) = 0.86$, $n_{\text{studies}} = 35$, $k_{\text{effect sizes}} = 145$, $P = 0.35$), timing of measurement (that is, with or without follow-up period; $F(1,158) = 1.04$, $n_{\text{studies}} = 41$, $k_{\text{effect sizes}} = 160$, $P = 0.31$) and the number of sessions (that is, one-off or multiple sessions; ($F(1,158) = 1.31$, $n_{\text{studies}} = 41$, $k_{\text{effect sizes}} = 160$, $P = 0.25$) were not significant moderators.

Moderation analyses by conditions

When comparing debiasing interventions to control conditions, there were no significant moderation effects of the different types of intervention focus (that is, specific bias mitigation, reasoning improvement or calibration training; $F(1,145) = 0.00$, $n_{\text{studies}} = 39$, $k_{\text{effect sizes}} = 147$, $P = 0.96$), different delivery modes (for example,

https://doi.org/10.1038/s41562-025-02253-y

Table 1 | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Abendroth & Richter[92], study 1	Germany	39 psychology students	23.59 (6.62)	CB	Verification task: participants decided whether or not three different types of test item matched the situation described by the texts[140]	Video-based	Ten video scenes were viewed in the metacognitive strategy training condition. Then, participants received an outline that summarized the strategies or steps and were instructed to work on two texts with opposing stances on the issue of man-made causes of global warming	Active: twelve video scenes for PQ4R training; strategies to increase receptive processing of information; elaboration and integration of information
Abendroth & Richter[92], study 2	Germany	46 psychology students	21.2 (4.34)	CB	Verification task: participants decided whether or not three different types of test items matched the situation described by the texts[141]	Mixed	Computer-based metacognitive strategy training: awareness of prior beliefs, monitor intertextual inconsistencies, using prior knowledge to evaluate arguments, and regulation of cognition	Active: computer-based PQ4R training; strategies to increase receptive processing of information; elaboration and integration of information
Aczel et al.[5]	Hungary	154 university students	21.73 (3.61)	Anchoring; base rate neglect; hindsight bias; overconfidence; sunk-cost fallacy; mixed niases	Custom measure developed as part of a previous unpublished study: the participants' susceptibility to nine biases based on tasks adapted from the heuristics and biases literature. One task measured each bias with four possible answer options per item. Only one option was the normatively correct one	Traditional training	Awareness training: introduction to heuristics and biases, then a presentation of each bias (real-life examples, an explanation, and techniques to avoid the bias). Analogical training: participants completed different tasks to detect structural similarities between situations containing the same bias. Participants had a discussion and received a presentation on the biases. Participants were asked to think of situations when they might commit the bias, suggest strategies to mitigate the biases, and a presentation of research-based coping techniques	No intervention
Adame[140]	USA	356 university students	20.13 (3.6)	Anchoring	Custom measure: the experimenter created the stimuli (pictures of a clear jar filled with a set of small objects)	Text/reading	Three training modules on cognitive biases and heuristics in decision-making processes Self-generated condition: consider reasons why the anchor is not associated with the estimation target, and then write at least three reasons Pre-generated condition: read from a provided list of reasons why the anchor is not associated with the estimation target	Active: a reading task on the history of the US state where the research took place
Almashat et al.[142]	USA	102 psychology students	19.77 (1.46)	Anchoring	Decision-making questions to choose a treatment for the patients in the vignettes	Text/reading	Participants read the background information, then the vignettes with outcome information (in terms of cumulative probability, interval probability, or life expectancy). They then answered some questions and chose a treatment	Active: three vignettes with written instructions (cumulative probability/ interval probability/life expectancy) and control questionnaire (related to stress, dental hygiene, and physical fitness)
Barberia et al.[143]	Spain	60 secondary school students	14.55 (2.32)	Causal illusions	Contingency learning task: assessing a causal relation between taking a drug and healing from a disease[144]	Traditional training	Conventional contingency learning paradigm: staging (to induce the causal illusion) and explanation of how to approach everyday causal inference. The participants were introduced to the concept of contigency as the correct way to infer causality from empirical information	No intervention

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Barberia et al.[145]	Barcelona	106 university students	21.2 (3.07)	Causal illusions	Contingency learning task: assessing a causal relation between taking a drug and healing from a disease[144]. Revised paranormal beliefs scale (Spanish adaptation): a global score of paranormal beliefs; for example, witchcraft, extraterrestrial life, superstition[146]	Traditional training	A staging phase (to induce the causal illusion) followed by a debriefing phase (theoretical explanations of the Barnum effect and performance in the 2-4-6 task). Then, they moved to the Wason study, and showed the typical confirmatory strategy and the 'consider-the-opposite' strategy. This was completed with a description of the CB	No intervention
Bessarabova et al.[147]	USA	703 college students	22.03 (5.34)	BBS	Three multiple-choice questions with a scenario describing BBS with four response options; they had to determine what bias each scenario represented. Nine of the nineteen scenarios from Pronin et al.[148]. BBS was measured by subtracting the peers' perceived susceptibility score	Educational games	A serious game called MACBETH, in which players take on an intelligence analyst role and are given a series of impending terrorist attack scenarios. The game addressed BBS, CB, and FAE. In the hybrid condition, players were explicitly instructed and tested on biases as part of their gameplay	Alternative methods: in the implicit training condition, players pursued scenarios in which making biased decisions cost them points and sometimes loss of a mission, but they were not explicitly trained on biases
Botta[149]	USA	93 school students	NR	Mixed biases	Heuristic measure created for the study; modified versions of tasks to measure representativeness heuristics, availability bias, adjustment and anchoring, the conjunction fallacy, effect of sample size, and the outcome approach[149]	Traditional training	Classroom instruction on probability and statistics using activity-based methodology or traditional lecturing methodology. Heuristical content was included	Active: class lecture, discussion, and completion of assigned homework for practice on statistics
Bou Khalil et al.[150]	Lebanon	86 medical residents	27 (2)	Anchoring	Framing score questionnaire. If the participant responded differently, the answer was considered discrepant (framing score = 1). If not, the answer was non-discrepant (framing score = 0). There were three vignettes, so the score ranges from 0 (no framing effect) to 3 (positive framing effect)	Text/reading	Access to three tests independent from the medical context to demonstrate the presence of cognitive illusions. The tests consisted of solving logical reasoning problems and showing that the answer could change by expanding the frame used in the data formulation	Active: access to six general knowledge questions that were unrelated to the framing effect or decontextualization

https://doi.org/10.1038/s41562-025-02253-y

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Calvillo et al. [52], study 1	USA	246 college students	NR	Cognitive reflection; mixed biases	CRT[58], the heuristics-and-biases tasks: "causal base rate, sample size, regression to the mean, gambler's fallacy, the conjunction problem, covariation detection, methodological reasoning, Bayesian reasoning, denominator neglect and probability matching"[151]	Text/reading	Consider-the-opposite group: read instructions that said "the first answer most participants think of is incorrect" and that they need to realise this to solve these problems correctly Feedback group: shown their answers to the first 10 items and the correct answers with an explanation of why the answer was correct or incorrect	No intervention
Calvillo et al. [52], study 2	USA	271 college students	NR	Cognitive reflection;CB	CRT[58], belief bias items had either believable and invalid conclusions or unbelievable and valid conclusions[152]	Text/reading	Feedback group: shown their answers to the first 10 items and the correct answers with an explanation of why the answer was correct or incorrect Consider-the-opposite group: read instructions that said "the first answer most participants think of is incorrect" and that they need to realise this to solve these problems correctly	No intervention
Clegg et al.[153]	USA	398 college students	NR (NR)	Anchoring; BBS; judgement accuracy; representativeness heuristic	Questionnaires developed to test declarative knowledge of biases and items to elicit the biases. Scores ranged from 0 (completely biased) to 1 (completely unbiased)	Educational games	Game-based intervention to train the recognition and mitigation of anchoring bias, projection bias, and representativeness bias	Alternative methods: watched a 30-min video with people exhibiting various cognitive biases. A professor then explained the interaction, the causes of bias, and ways to mitigate the bias
Dunbar et al.[62]	USA	703 college students	22.03 (5.34)	CB; FAE; mixed biases	A 'drag and drop' exercise to match the definitions of cognitive biases. New cognitive bias scale: modeled after Rassin's[154] test strategy scale with modified wording "so all items had more certainty in why a participant was choosing something as a confirmation"[62]. The items had four possible answers. A second measure based on Wason's[155] CBAM with nine items was used. Each CBAM offers a brief scenario followed by four response options	Educational games	A serious game called MACBETH in which players try to detect and prevent terrorist threats by gathering and assessing intelligence data (implicit and explicit conditions)	Alternative methods: an instructional video designed to inform viewers about the nature of cognitive biases by using entertaining vignettes
Dunbar et al. [63], study 1	USA	508 college students	21.3 (4.94)	CB	Cognitive bias scale: modeled after Rassin's[154] test strategy scale	Educational games	Repeat-play of a serious game called MACBETH, in which players try to detect and prevent terrorist threats by gathering and assessing intelligence data.	Alternative methods: watched a traditional instructional video explaining FAE and CB and non-repeat game condition (combined)

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Dunbar et al.[63], study 2	USA	558 college students	21.61 (4.89)	CB; FAE	Cognitive bias scale: modeled after Rassin's[154] test strategy scale, Ron's bad day scenario, to test reliance on situational cues[156]. Participants saw two scenarios (one positive and one negative) and evaluate what factors could contribute to the events	Educational games	A serious game called MACBETH, in which players try to detect and prevent terrorist threats by gathering and assessing intelligence data (multi-player vs. single-player versions)	Alternative methods: watched a traditional instructional video explaining FAE and CB and non-repeat game condition (combined)
Dwyer et al.[56]	Ireland	81 college students	NR	Cognitive reflection	LJRA[59]: "seven open-ended questions about real-world problems, which measures reflective judgement by: (1) identification of scorable units of text and (2) the examination of the structure of both the unit and conceptual elements within by (3) mapping the responses for purposes of assessing the hierarchical complexity of reasoning performance"[56]	Mixed	Argument-mapping critical thinking training focused on analysis, evaluation, and inference Hierarchical outlines-infused critical thinking training focused on analysis, evaluation, and inference	No intervention
Emory & Luo[157]	USA	22 college students	23.64 (NR)	Overconfidence	JOL scale[158]. "Calibration accuracy was calculated by comparing JOL estimates to assessment scores. Both absolute and relative values of calibration were evaluated in addition to any change based upon the intervention"[157]	Mixed	Three hours of module with additional direct instruction training to self-monitor learning tasks and choosing appropriate strategy selection. Then, an additional e-training intervention (written study strategies and a video instruction of the metacognitive monitoring process). In the video, an instructor explained the concepts of calibration (over/under confidence), metacognition, and how these processes can guide students to choose study strategies	Active: 'read an excerpt and watched a companion lecture from a popular and well-adopted psychology textbook'[157]
Fitterman-Harris & Vander Wal[159]	USA	101 first-year medical students	23.55 (1.65)	Weight bias	AFAT[160]: a 47-item instrument that measures explicit bias against fat people. Universal measure of bias-fat: negative judgement subscale[161]: a 20-item instrument to measure explicit weight bias. Weight IAT[162] to assess implicit weight bias	Mixed	Participants were presented with case examples (including patients' presenting problems and demographic information such as age, biological sex, and BMI) and watched a 17-minute video ('*Stigma – The Human Cost of Obesity*'). The video includes educational content and weight bias anecdotes from individuals with obesity. Participants then did a role-play activity designed to cause cognitive dissonance. Then, the group had a discussion of how mitigating weight bias can improve their medical practice	Alternative methods: participants in the control group had an educational PowerPoint presentation to provide an overview of obesity prevalence using data from the National Health & Nutrition Examination survey. At the end of the presentation, the group leaders initiated a brief question and answer period

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Gagne[163]	USA	63 university students	19.18 (1.39)	Weight bias	AFAT[160]: a 47-item instrument that measures explicit bias against fat people. Weight IAT[162] to assess implicit weight bias	Traditional training	Education and training session designed to reduce explicit and implicit weight bias. The education component of this session focused on the nature and development of implicit bias and the idea of prejudice as a habit. "The training section informed participants of five strategies to reduce implicit bias: stereotype replacement, counter-stereotypic imaging, individuation, perspective taking, and increasing contact with out-group members"[163]	Active: an education and training session on stress and stress-reduction practices.
Gutierrez[164]	USA	160 university students	22.76 (7.15)	Judgement accuracy	"Accuracy was evaluated by calculating the continuous difference score between the confidence judgment and actual performance (that is, squared difference)"[164]	Traditional training	Strategy training: "instruction regarding more sophisticated and adaptive strategies that are more conducive to enhancing calibration accuracy with respect to performance. Training included direct instruction and individual practice in using strategies with scaffolded feedback in a face-to-face lecture format"[164]; participants either receive incentive if >80% items answered correctly or receive no incentives	Active: incentive and watch a psychology-related film
Gutierrez[66]	USA	125 graduate students	NR	Stereotyping	IAT[162] to measure an individual's implicit bias (the difference in response times when matching stereotype-congruent presentations and the stereotype-incongruent presentations)	Educational games	'Fair Play': a point-and-click adventure game with various challenging scenarios in academic settings. "The main character, Jamal Davis, experiences subtle unintentional racial discrimination that results from implicit bias. When Jamal encounters an example of implicit bias, such as being mistaken for a caterer (status leveling) or being asked to speak on behalf of all Black graduate students (tokenism), the player is alerted to the bias by the appearance of an Almanac for 'just-in-time learning'. The Almanac includes definitions of the specific bias encountered, in-game examples, and citations to supporting research"[66]	Alternative methods: participants went to a website with an image of Jamal and a narrative of the experiences he has in 'Fair Play'. Participants were asked to "imagine yourself as Jamal"
Heijltjes et al.[165], study a	Netherlands	183 part-time Economics students	29.3 (6.5)	Mixed biases	Critical thinking skills tests: two causal base-rate tasks[166]; two non-causal base-rate tasks[167]; conjunction tasks[88]; two framing tasks[168], two Wason selection tasks[169]; two syllogistic reasoning tasks[170]	Video	"The critical thinking instructions in the experimental conditions consisted of computer-based presentation which the features of critical thinking, its importance, the required reasoning skills, the dispositions, and the risk of biased thinking and fallacies in thinking were explained. Examples and demonstrations of all task categories were provided, referring back to the tasks seen in the pre-test, which could have allowed participants to mentally correct initially erroneous responses. The critical thinking instructions were followed by self-explanation prompts during subsequent practice"[165]	Active: the control group received a 15-min video (that is, what happens in your brain when you are in love).

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Heijltjes et al.[165], study b	Netherlands	141 university student	20.81 (1.57)	Mixed biases	Critical thinking skills tests (sixteen tasks: two causal base-rate tasks, two non-causal base-rate tasks, two conjunction tasks, two framing tasks, two covariation tasks, two Wason selection tasks, syllogistic reasoning tasks)	Mixed	Explicit instruction and practice; activation prompts with hints for the tasks (20-min), critical thinking instruction, 15-min computer-based video presentation with general information about critical thinking; self-explanation prompts, critical thinking instruction, 15-min computer-based video presentation with general information about critical thinking	Active: task practice + implicit instruction during a 7-week regular economics course about argumentation and negotiation skills (no explicit instructions); only implicit instruction during a 7-week regular economics course about argumentation and negotiation skills
Huff & Nietfeld[171]	USA	118 fifth grade students	NR	Overconfidence	The calibration bias score, which consisted of the signed difference between the average confidence and average performance scores on each test; positive scores indicate overconfidence while negative scores indicate underconfidence	Traditional training	Comprehension monitoring and monitoring accuracy training. Students read the 12 practice passages and provided confidence judgments. They received instructions on comprehension monitoring strategies during the daily lessons. They also sit in the think-aloud and discussion component, in which they discuss why is it important to think about their confidence and how can a confidence judgment help them monitor their comprehension. Students were also encouraged to compare their confidence judgments with their actual performance	No intervention
Jacobson et al.[117]	USA	278 high-school students	NR	Decision-making competence	Decision-making competence test [32], which includes "resistance to framing, recognising social norms, under/overconfidence, decision rules, risk perception, and resistance to sunk costs"[117].	Traditional training	"In addition to the district-approved US History curriculum, Experimental-group participants received an integrated curriculum that included instruction in the Decision Quality (DQ) model. Exposure to the DQ material occurred in an introductory unit and was reinforced through classroom simulations, lectures, and writing assignments"[117]	Standard US history curriculum
Joy-Gaba[97], study 3	USA	212 university students	19.02 (NR)	BBS	BBA questionnaire[172] where participants had to rate their own tendency or the average person's tendency to commit cognitive biases	Mixed	Completing IAT and automatic bias education: participants viewed a 10-min clip of the bias education lecture for both experiential and education	Active: BBS manipulation - participants only watched the 10-min video or complete the IAT; no intervention
Kolić-Vehovec[173], study 2	Croatia	223 graduate students	22.52 (1.59)	Judgement accuracy; overconfidence	The discrimination index measured the students' ability to distinguish between their confidence for correct and incorrect performance; the bias index was used to measure whether students were accurate, overconfident, or underconfident[174]	Traditional training	The lecturer explained operant conditioning and provided real-life examples, then the participants had four small-group exercises. Students participated in a text reading task: 360-word passage about the detrimental effects of overconfident JOL and performance. The collaborative informed condition participants worked on 11 examples in small groups of 4 or 5 students, whereas individual condition participants worked on 11 examples individually	Active: attended the same lecture, then they read a 318-word passage about classical conditioning and worked on 11 examples in small groups or individually

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Lee et al.[67]	USA	335 university students	NR	Anchoring; base rate neglect; representativeness heuristic; stereotyping	Custom measure: anchoring bias (estimated quantities after exposure to irrelevant numerical anchors) and three forms of representativeness bias adapted from Cox & Mouw[175]: gambler's fallacy (misunderstanding probability in random events), base-rate fallacy (ignoring statistical base rates in favor of stereotypical information), and insufficient data bias (making judgments with limited information). The study also introduced a new scale (NewRep) specifically designed to assess stereotype-based representativeness bias through diagnostic versus stereotypical response options	Mixed	Combined condition: participants watched the 14-min slideshow lecture before participating in the game	Alternative methods: game-only condition: an educational game with four scenarios about the biases and bias mitigation. The game provides a short description of the biases at the beginning. Slideshow condition: watched a 14-min presentation explaining the biases and mitigation strategies
Legaki & Assimakopoulos[176]	Greece	261 college students	NR	Judgement accuracy	The evaluation experiment was composed of 30 questions of the same type. Students' performance was calculated as the sum of right answers (100 as maximum value)	Educational games	F-LauReL is a web-based platform with information about forecasting, recent research findings, and the gamified applications with instructions. It is composed of three web-gamified applications (Horses for Courses, JudgeIt, and Metrics to Escape)	No intervention
Legaki et al.[65]	Greece	285 university students	NR	Judgement accuracy	Evaluation composed of 30 multiple-choice or true/false questions to test the students' knowledge of biases. Students' performance was calculated as the sum of right answers (normalized to have 100 as maximum value)	Mixed	Read and play: reading research by Tversky & Kahneman[10] about heuristics and biases and playing JudgeIt game that takes them through a narrative of an explorer who is challenged to discover and collect elements in a series of imaginary destinations (Dreamland: representativeness bias; Neverland: availability heuristic; Amnesialand: adjustment and anchoring); read or play only conditions	No intervention

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Martínez et al.[70]	Spain	234 undergraduate students	20.23 (1.78)	Causal illusions	Standard contingency judgment task[114]. Participants were asked to imagine being a medical doctor and they had to determine whether a fictitious drug was effective in providing relief to a series of fictitious patients suffering from a fictitious disease	Traditional training	Induction and training: induction phase includes a 'staging' about a fake psychological theory (modes of thought). Participants then completed a computerised version of Wason's 2-4-6 task[155] (finding a rule that determines the relationship between three numbers). Training phase consisted of explanation of the cognitive biases involved in the induction phase, followed by mitigation strategies (considering-the-opposite strategy and a training-in-rules methodology) Training: participants completed the training phase, consisted on explaining the two cognitive biases that were induced in the induction phase. They were given the considering-the-opposite strategy and a training-in-rules methodology	No intervention
Morsanyi et al.[177]	UK	108 university students	20.5 (NR)	Representativeness heuristic	Probabilistic reasoning questionnaire consisting of six tasks measuring the equiprobability bias[10,178,179]. "For each problem, there were three response options: a heuristic one, a normatively correct one, and a third response that was neither normatively correct nor heuristic"[177]	Traditional training	Four exercises to demonstrate the concept of randomness through generating sequences of probabilistic outcomes. "Combined examples training (participants had to generate series of probabilistic outcomes by tossing coins, throwing dice, etc) with training in rules"[177]	No intervention
Onal & Kumkale[180]	Turkey	126 undergraduate students	NR	Halo bias; judgement accuracy; overconfidence	Strong correlations between ratings for the irrelevant dimension and ratings for the other dimensions imply halo error. Rating accuracy: difference between participants' ratings and the subject-matter experts' true score estimates. Lower ratings compared to the true scores suggested negativity bias[181]	Traditional training	Participants watched a video of a fictitious instructor. After the video, participants received instructions on how to distinguish remember judgments from know judgments	Active: completed various measures of individual differences
Ramdass[182]	USA	88 elementary school students	NR	Judgement accuracy; overconfidence	This measure was calculated based on self-efficacy ratings and the math scores. The measure was calculated by subtracting the absolute value of each bias score from 100. Bias is in the direction of the errors in judgment	Traditional training	Students in the strategy training groups were taught a step-by-step strategy to solve various math problems (calibracy training) Self-reflection group only: an instruction of "It is important to read each problem very carefully before trying to solve it"	No intervention

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Rhodes et al.[55]	USA	732 university students	NR	BBS; mixed biases	The recognition and discrimination scale of the ABC consisted of nine items to test the participants' factual knowledge of biases. The behavioural elicitation scale of the ABC tested the participants' ability to avoid cognitive biases in a set of judgment tasks. There are six subscales: CB; FAE; BBS; anchoring; representativeness; and projection[183]	Educational games	Heuristica: 3D first person science fiction game that featured various learning opportunities to learn about cognitive biases. Cycles: 2D game consisting of a collection of different puzzles with an overarching science fiction theme[55]. Missing: educational game with mystery narratives and mini-challenges that taught various cognitive biases. Participants had to find and use information related to a crime, and after-action reviews that elaborated upon the content taught during play episodes[55]	Active: unrelated video: a clip from '*This Emotional Life*' that taught content unrelated to cognitive biases. Related video: a video with a 'teacher' explaining the same cognitive biases as the games
Rodríguez-Ferreiro et al.[184]	Spain	72 undergraduate students	23.26 (3.26)	Causal illusions	Standard contingency judgment task[185]. Participants imagined themselves as a doctor and determine whether a fictitious drug was effective in relieving a fictitious disease. Paranormal beliefs scale Spanish adaptation[146] to assess participants' beliefs in withcraft, superstitions, and other paranormal beliefs	Mixed	The instructor provided information regarding the two biases and how to overcome them as part of a training-in-rules approach	No intervention
Roelle et al.[68], study 1	Germany	42 university students	25.48 (NR)	Overconfidence	JOL: the participants judged their comprehension level of the content on a six-point rating scale	Mixed	A multimedia presentation about the pitfalls of making overconfident JOLs: how it is common to overconfidently judge one's level of knowledge and the detrimental consequences of making overconfident JOLs	Active: a task to read and understand an expository text on the diathesis-stress model
Roelle et al.[68], study 3	Germany	97 high-school students, 8th grade	13.53 (NR)	Overconfidence	JOL: the participants judged their comprehension level of the content on a six-point rating scale	Mixed	A multimedia presentation about the pitfalls of making overconfident JOLs: how it is common to overconfidently judge one's level of knowledge and the detrimental consequences of making overconfident JOLs. Then, participants were given effective strategies to use if they encounter comprehension difficulties: explaining content using their own words and thinking about the content using their own examples	Active: watched a video about optical illusions and received a multimedia presentation about e-learning

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Rowland[186]	USA	21 master of counselling students	NR	FAE	The dispositional-situational index from the attributional questionnaire [187] to assess the participants' attributions. The ratings were on a nine-point scale	Traditional training	A 1.5-h lesson examining the different perceptions which various persons may have of a single event. Participants viewed the film '*Eye of the Beholder*'. A discussion of the contextual cues and intrapersonal dynamics of divergent perspectives of common events followed the film. Finally, participants were required through a homework assignment to focus on situational cues during their counseling sessions throughout the following week and to report written observations to the instructor the following week	No intervention
Salvatore & Morton[53], study 2a	UK	258 psychology students	NR	Representativeness heuristic	Evaluations of science (indicative of identity-based responding): 7-point response (1 = strongly disagree, 7 = strongly agree) to assess the participants' reactions to scientific research[188]	Text/reading	A full-page tutorial on ecological fallacy, including a paragraph explaining the fallacy in simple terms, using an example (the height of men versus women, on average), what could be a possible conclusion from their observation, and followed by a second example and a short quiz to gauge the participants' understanding of the fallacy. Both studies explicitly reminded participants that "what might be true in general is not true of every individual person"	Active: read a summary of an ostensible scientific article about a topic they were unfamiliar with
Salvatore & Morton[53], study 2b	UK	234 psychology students	NR	Representativeness heuristic	Evaluations of science (indicative of identity-based responding): 7-point response (1 = strongly disagree, 7 = strongly agree) to assess the participants' reactions to scientific research[188]	Text/reading	A full-page tutorial on ecological fallacy, including a paragraph explaining the fallacy in simple terms, using an example (the height of men versus women, on average), what could be a possible conclusion from their observation, and followed by a second example and a short quiz to gauge the participants' understanding of the fallacy. Both studies explicitly reminded participants that "what might be true in general is not true of every individual person"	Active: read a summary of an ostensible scientific article about a topic they were unfamiliar with
Sellier et al.[57]	France	316 graduate business students	28.24 (3.69)	BBS; cognitive reflection; FAE	BBS scale[98]: each item described a bias and asked respondents to indicate the incident of the bias for themselves and the average American. Three-item CRT[58]. The neglect of external demands scale[189] to measure the propensity to make correspondent inferences	Educational games	The one-shot debiasing intervention was conducted through a serious video game, '*Missing: The Pursuit of Terry Hughes*' "Game players act as amateur detectives and search for a missing neighbor, who is embroiled in fraud committed by her employer, a pharmaceutical company. Players make bias-eliciting judgments and decisions during game play. At the end of each episode, participants receive training in the 'teach' portion of the game via an after-action review. In the review, experts define the three biases targeted by the game and provide strategies to mitigate each bias"[57]	No intervention

https://doi.org/10.1038/s41562-025-02253-y

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Shaw et al.[64]	USA	234 university students	22 (NR)	BBS; CB; FAE	The likelihood of their own responses and their estimates of "an average student at their university", which reflected the bias illusion of superiority; the BBS scale had seven questions; the CB scale presented scenarios and asked participants to choose a hypothesis and rate the evidence on a seven-point Likert scale (unimportant to extremely important) to evaluate their hypothesis[190]; rate the protagonist in scenarios[148]	Educational games	30-min digital game designed to teach people how to recognize and mitigate three cognitive biases: FAE, CB and BBS. "In the assigned character condition, players were given a predesigned androgynous character. This avatar wore a space-suit that obscured all but the character's body type, to ensure that we minimized the extent to which disidentification on the basis of age, race, and other identifiers might occur"[64]	Alternative methods: video condition: professionally produced video that covered cognitive biases by watching a series of scenarios involving a college student during his interactions with friends who engaged in cognitive biases
Sibbald et al.[93]	Canada, USA, Netherlands	40 medicine residents	NR	CB	Errors in ECG (10 cases designed to suggest search satisficing and CB)	Traditional training	"A 'debiasing' condition involving instruction on identifying common biases coupled with a checklist that prompts learners to identify any cognitive biases that might lead to errors in interpretation"[93]	Active: control: participants were given general instructions to review their diagnoses carefully. Content checklist condition: "participants received instruction on use of the content checklist which drew attention to various aspects of the ECG, including rate, rhythm, axis, hypertrophy, ischemia, and intervals"[93]
Silver[191]	USA	89 university students	NR	Halo bias	IPT differentiation: how correlated are ratings of "implicit personal theory", where lower correlations show lower halo effect, which indicates that the participants were more biased[192].	Traditional training	Each group was asked to present their ratings and to explain the rationale they used. A discussion of halo error and its effect on rating accuracy then followed. The training program combined accuracy and error training. The experimenter stressed that accuracy is the ultimate goal, not simply increasing the variance in one's ratings. Participants were next asked to think of examples in their lives when they had experienced halo bias from a supervisor or had exhibited halo bias themselves	Active: control: asked to rate one of the videotaped sketches of a psychology instructor used as stimulus material in this study. Participants were only retested for cognitive complexity and IPT differentiation and given follow-up questions during the second sessions

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Swift et al.[193]	UK	43 university students	22.9 (4)	Weight bias	Fat phobia scale: a 14-item questionnaire about beliefs and feelings towards people who are obese[194]; BOAP scale: eight items that measure beliefs about the degree of controllability of obesity on a six-point scale[195]; Willpower subscale of the AFAT[160] that assesses the belief that being overweight is a matter of personal control or lack thereof; IAT[162] to measure the reaction time of automatic memory-based associations.	Video	Two 17-min films were shown ('*Weight Prejudice: Myths and Facts*' and '*Weight Bias in Healthcare*'). In the films, there were a range of practical strategies to improve the healthcare experience for obese patients, including attributions of weight controllability, empathy induction, and debunking weight-based stereotypes	Active: watch an extract from an episode of a popular historical documentary series (unrelated to body weight or food)
Testa et al.[196]	Italy	408 high-school students	NR	Overconfidence	Confidence bias calculated using a 1D Rasch model: intervals of differential persons functioning measures of calibration[197]	Traditional training	Teaching-learning sequence followed a conceptual sequence so students build increasingly sophisticated models and also pedagogical countermeasures for overconfidence, such as critical and regulatory feedback strategy and the 'consider-the-opposite' strategy	Active: a 4-week teaching sequence that followed the Italian national guidelines about introductory quantum mechanics in both chemistry and physics
Van Bockstaele et al.[198]	Netherlands	39 high-school students	14.03 (1.22)	Hostile attribution bias	A variant of the interpretation recognition task[199]. There were ten ambiguous scenarios presented, in which the motives of others could be interpreted both positively and negatively. The participants rated the likelihood of them interpreting the motives on two separate four-point Likert scales	Traditional training	We presented ambiguous scenarios in which the motives or behaviours of others could be interpreted both positively and negatively. After completing the word fragment, participants responded to a yes/no question to show their level of comprehension of the scenario content. They were then given feedback on the comprehension questions	No intervention
van Brussel et al.[54]	Netherlands	141 student teachers from a Dutch primary teacher education institution	19.9 (4.59)	CB	CB tasks: hypothesis testing tasks and Wason's four-card selection tasks. CB transfer test[200]: participants had to determine whether an individual is an extravert or introverted person and ask questions that tie in with the opposite of the activated stereotype	Mixed	A text instruction on CB and how to mitigate this bias using 'consider-the-opposite' strategy, including examples in the educational context. Participants in the teaching video condition prepared a lesson using generated self-explanations of the instruction and practice materials. They then recorded a video to teach the prepared lesson to a fictitious audience	Active: control (re-study), preparing to teach

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
van Peppen et al.[69]	Netherlands	157 university students	NR	Mixed biases	Nine heuristics-and-biases items ((1) base-rate items; (2) conjunction items; (3) syllogistic reasoning items)[201] Custom measure: four items of two task-categories with similar features to the learning items, requiring knowledge and rules of logic or statistics	Text/reading	Worked examples: participants received a worked-out solution to each problem. Participants first viewed the 10-min video instructions (a general instruction on critical thinking and explicit instructions on three heuristics-and-biases tasks). Participants then received explicit instructions on how to avoid base-rate fallacies, conjunction fallacies, and biases in syllogistic reasoning. Participants had to practice either in an interleaved or blocked schedule, and with worked examples	Alternative methods: practice problems: similar video-based instruction on how to avoid base-rate fallacies, conjunction fallacies, and biases in syllogistic reasoning. They received a prompt after each of the tasks in which they were asked to explain how the answer was obtained. After that, participants were given feedback indicating whether their answer was correct or not
van Peppen et al.[85]	Netherlands	182 university students	19.53 (1.91)	Mixed biases	Three Wason selection items measured students' tendency to disprove a hypothesis by verifying rules rather than falsifying them[202]. Three base-rate items examined students' tendency to incorrectly judge the likelihood of individual-case evidence by not considering all relevant statistical information[4,10,166,203]	Text/reading	(1) All participants received a 12-min video instruction emphasising the importance of critical thinking in general and explaining which skills and attitudes are needed to think critically. They were given explicit instructions on how to avoid biases in syllogistic reasoning and conjunction fallacies followed by two worked examples with correct reasoning Contrasting examples: comparing a fictitious students' correct and incorrect solutions to the problem, and to indicate the wrong solution and the steps (2) Correct examples: provided with a fictitious student's correct solution and explanation to the problem, and were prompted to explain why these steps were correct. (3) Erroneous examples: provided with a fictitious student's wrong solutions. They were prompted to indicate the wrong step and provide the correct solution themselves	Alternative methods: similar intervention, but with practice problems condition: participants solved the problems themselves, instructed to choose the best answer and explain how to get to the answer
Veinott et al.[204]	USA	169 university students	21 (NR)	Mixed biases	Custom measure: each form of the cognitive bias test. Fifteen items to assess the participant's knowledge and recognition of the three biases, and 26 items assessed their ability to mitigate the biases in different contexts	Educational games	Game condition: first person point of view of the heuristica game to train participants to learn to mitigate biases but only tutorial and training phase of the game	Alternative methods: decision video condition: an entertaining host and vignettes of social situations where these cognitive biases might occur

https://doi.org/10.1038/s41562-025-02253-y

https://doi.org/10.1038/s41562-025-02253-y

Table 1 (continued) | Characteristics of studies examining debiasing interventions in educational settings

Citation	Country	*N*	Age mean (s.d.)	Bias	Bias/reasoning measure	Delivery format	Intervention	Comparison
Whitaker et al.[205]	USA	94 university students	21 (NR)	Mixed biases	Custom measure: each form of the cognitive bias test. Fifteen items to assess the participant's knowledge and recognition of the three biases, and 26 items assessed participant ability to mitigate bias across different situational contexts	Educational games	Heuristica. Student model plus worked-out examples: tailored gameplay and extra worked-out examples when the student shows misconceptions in their reasoning	Alternative methods: fixed-order: all participants saw the same gameplay order (no tailored learning opportunities). Student model: participants' gameplay is guided by the student modeler, which selects and orders learning opportunities customized to the students' mastery levels (without worked-out examples)

This table summarizes the key characteristics of studies included in our systematic review (53 studies across 49 papers, n = 10,941). ABC, assessment of biases in cognition; AFAT, anti-fat attitudes test; BOAP, beliefs about obese persons; CBAM, confirmation bias application measure; ECG, electrocardiogram; IAT, implicit association test; JOL, judgment of learning; MACBETH, mitigating analyst cognitive bias by eliminating task heuristics; NR, not reported; PQ4R, preview, question, read, reflect, recite, and review.

in-person with a teacher or online with computer-generated programmes; $F(1,150) = 0.53$, $n_{studies} = 40$, $k_{effect\ sizes} = 152$, $P = 0.47$), the type of biases (for example, attentional versus encapsulated; $F(2,157) = 0.86$, $n_{studies} = 41$, $k_{effect\ sizes} = 160$, $P = 0.42$) and the strategy used to teach about biases (for example, cognitive versus cognitive + motivational; $F(1,145) = 1.76$, $n_{studies} = 37$, $k_{effect\ sizes} = 134$, $P = 0.19$).

Moderation analyses by comparison conditions

For the likelihood of committing specific biases, the comparison condition researchers used influenced the size of the effects ($F(2,307) = 13.38$, $n_{studies} = 52$, $k_{effect\ sizes} = 310$, $P < 0.001$). Interventions were significantly better than no-intervention control groups ($g = 0.30$, 95% CI 0.18 to 0.42, $P < 0.001$) and active control groups ($g = 0.31$, 95% CI 0.20 to 0.41, $P < 0.001$). There were not enough effect sizes for other outcomes (that is, specific knowledge of biases or improved general reasoning) to moderate those by comparison group, but interventions were no better when compared with alternative methods of debiasing (for example, educational games versus video interventions; $g = 0.07$, 95% CI −0.05 to 0.18, $P = 0.25$). These alternate methods of debiasing are discussed in a qualitative synthesis below.

Sensitivity analyses

Risk of bias within studies. The consensus ratings of risk of bias for all included studies are presented in Supplementary Table 2. As shown in the summary plot (Fig. 4), no studies had a low overall risk of bias. This is mainly because only five studies had pre-registered their methods (studies 1 and 2 in ref. 52; studies 2a and 2b in ref. 53, and ref. 54) or sufficiently reported their randomization process.

As a sensitivity analysis (Fig. 5), we analysed the studies assessed as 'low risk of bias' on each criterion (with at least three studies) compared with the overall analysis. These effects were not significantly different from the overall estimate including all studies (P values between 0.11 and 0.65). For most criteria, effect sizes were similar to the main analysis. The largest absolute difference was for studies at low risk of selective reporting (usually because they were pre-registered; $g = 0.02$, 95% CI −0.12 to 0.16, $P = 0.71$), but the difference in pooled estimates was not significant ($\beta_{low\ risk\ versus\ all\ studies} = -0.25$, 95% CI −0.57 to 0.08, $P = 0.13$).

Outliers. Only one effect size was greater than $g = 2.5$. Removing the row from the analysis ($g = 0.26$, 95% CI 0.14 to 0.39, $n_{studies} = 41$, $k_{effect\ sizes} = 159$, $P < 0.001$) did not substantially change the model results (with outliers included: $g = 0.26$, 95% CI 0.14 to 0.39, $n_{studies} = 41$, $k_{effect\ sizes} = 160$, $P < 0.001$). Thus, we decided to keep all effect sizes in the analysis.

Risk of bias across studies. We plotted effect sizes against standard errors using the funnel plot shown in Fig. 6. The multilevel Egger's test was significant ($F(1,158) = 13.71$, $n_{studies} = 41$, $k_{effect\ sizes} = 160$, $P < 0.001$), indicating a small study effect. By contrast, the three-parameter selection model (3PSM) likelihood ratio test indicated no pattern of effect sizes consistent with publication bias ($\chi^2(1) = 0.63$, $P = 0.42$): both affirmative and non-affirmative studies were equally likely to be published. The s value showed that studies need to be 3.46 times more likely to be published when significant (compared with when the results are nonsignificant) to reduce the effects of teaching debiasing to zero. Overall, there was modest evidence consistent with publication bias.

Other analyses and qualitative synthesis

Effects of teaching debiasing on other outcomes. Only one study measured participants' knowledge of specific cognitive biases[55], finding significant improvements following educational game interventions compared with non-debiasing video control ($P < 0.001$, Cohen's d of 0.79–1.56). However, one game ('Missing') performed

https://doi.org/10.1038/s41562-025-02253-y

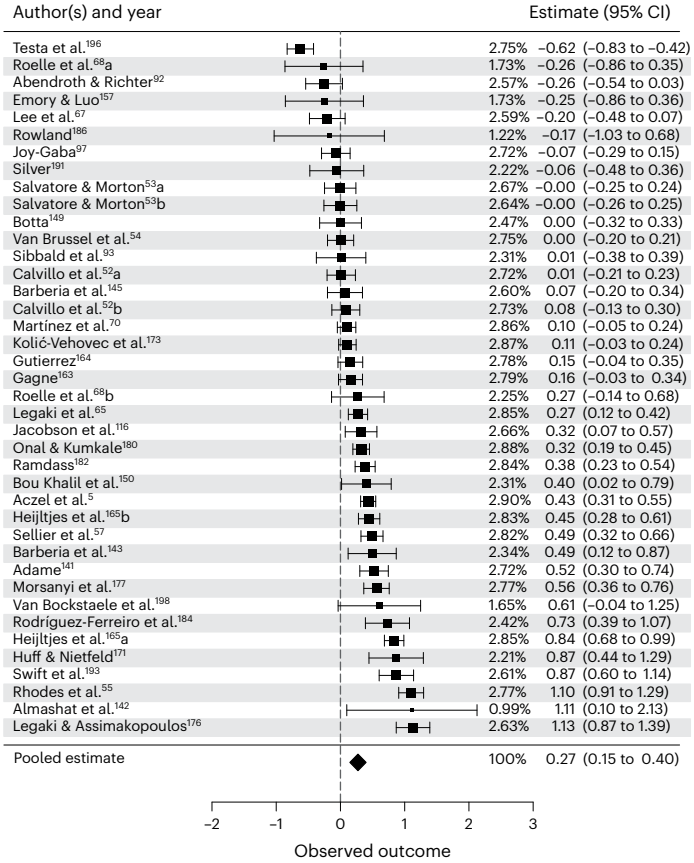

Fig. 2 | Forest plot of pooled results for educational intervention effects on cognitive bias. Hedges' *g* effect sizes (centre points) with 95% CIs (error bars). This analysis includes 41 studies with 160 effect sizes and a total sample of 10,941 participants. Studies are ordered by effect size magnitude. The diamond at the bottom represents overall pooled effect size ($g = 0.26$, 95% CI 0.14 to 0.39, $P < 0.001$, $I^2 = 74.94\%$). Studies with a and b refer to separate studies in one publication. The full characteristics of individual studies are presented in Supplementary Table 2.

significantly worse than the video of cognitive biases ($t(243) = -2.67$, $P < 0.01$). The other educational games ('Cycles' and 'Heuristica') did not perform better than the cognitive bias video condition. The results highlighted that video-based intervention could be more effective than some educational games when the goal is to improve participants' declarative knowledge about cognitive biases.

Four studies measured improvements in reflective cognition or reasoning (studies 1 and 2 in ref. 52 and refs. 56,57), using instruments such as the cognitive reflection task (CRT)[58] or lectical reflective judgement assessment (LRJA)[59]. The CRT measures one's ability to inhibit automatic—possibly biased—responses and engage in reflective thinking before responding[40,60], whereas the LRJA measures the skills to recognize biases, work with different perspectives, evaluate evidence and frame decisions[61]. While these measures do not directly assess the likelihood of committing specific biases, they are designed to evaluate a person's general ability to overcome cognitive biases. Debiasing interventions improved CRT performance, although this did not transfer to heuristics-and-biases tasks or belief bias items[52,57]. Conversely, a 6 week classroom-based intervention improved reflective judgement performance in specific conditions[56]. Participants with higher dispositions to critical thinking (that is, open mindedness, analytic and systematic way of thinking) benefitted from a text-based learning strategy, whereas those with lower dispositions benefited from a visual learning strategy (for example, visual diagrams).

Comparison between alternate methods of debiasing. In head-to-head comparisons, some types of intervention worked better than others. Most studies compared educational games with video-based interventions[55,62,63,64]. Educational games reduced cognitive biases more effectively than video-based interventions ($g = 0.37$, 95% CI 0.07 to 0.67, $n_{\text{studies}} = 6$, $k_{\text{effect sizes}} = 43$, $P = 0.02$). Two studies compared educational games with text-based intervention[65,66] and found that participants in the game-based group performed significantly better[66]. However, educational games could reduce racial bias only when it evoked higher empathy; more participants abandoned the game than the text, suggesting the text was more engaging[66].

Combining either traditional lectures with educational games[67] or with a multimedia presentation[68] showed promise. One study found that the combination of educational games and traditional lectures led to better learning and retention than either method alone[67]. This combination allowed participants to learn concepts through slides and practice mitigation techniques through educational games.

Some studies compared different approaches, such as written feedback versus written instructions[52], or practice examples versus worked examples with explicit instructions[69]. Calvillo et al.[52] found both written instructions and elaborative feedback increased CRT performance[58], but elaborative feedback did so without increasing response time. This suggests that written instructions increased deliberation but feedback improved intuitive decision-making. van Peppen and colleagues[69] found that practice improved the effects of explicit instructions: practising unbiased reasoning helped, but worked examples helped people learn more effectively. Both studies found learning was limited to explicitly taught tasks and it did not transfer well to other heuristics-and-biases tasks.

One study compared traditional debiasing training (that is, explanation of biases, discussion and the 'consider-the-opposite' strategy) with a similar training where they induced a bias among participants[70]. The researchers taught participants a fake psychological theory to induce cognitive biases among participants before the debiasing training. They found that the induction generated a healthy scepticism among participants, which reduced causal illusions.

Discussion

This review synthesized the literature on debiasing in formal educational settings (for example, schools, universities). We found 54 studies with 383 effect sizes. Most studies included a measure of whether students were less likely to commit cognitive biases (41 studies with 160 effect sizes). Our meta-analyses showed that students could learn to reduce their cognitive biases when interventions taught a range of cognitive biases compared with control conditions. Teaching debiasing showed meaningful benefits ($g > 0.2$) in around half of the interventions.

Evidence suggests that people can learn to override their intuitions to engage in more rational thinking[71,72]. We saw the strongest effects on mixed biases, where outcome measures comprised several tasks representing multiple cognitive biases, followed by overconfidence. Our findings on mixed biases suggest it might be easier to modestly decrease the influence of multiple biases than to meaningfully reduce the effects of one (for example, CB). This suggests there may be 'low hanging fruit' for educators who want to increase awareness across a range of biases, but different strategies are required to make a meaningful difference on any individual bias.

One such strategy might be giving students rapid feedback about their biased judgements[73], especially for quantifiable biases such as overconfidence[74]. It is harder to show errors in theFAE: educators can point to how we prefer dispositional causes of others' mistakes, but other's mistakes are probably caused by both situational and dispositional characteristics. This may mean educator feedback is less compelling, making it harder to shift biases such as representativeness heuristic and CB. When we can provide clear and unambiguous feedback, it

https://doi.org/10.1038/s41562-025-02253-y

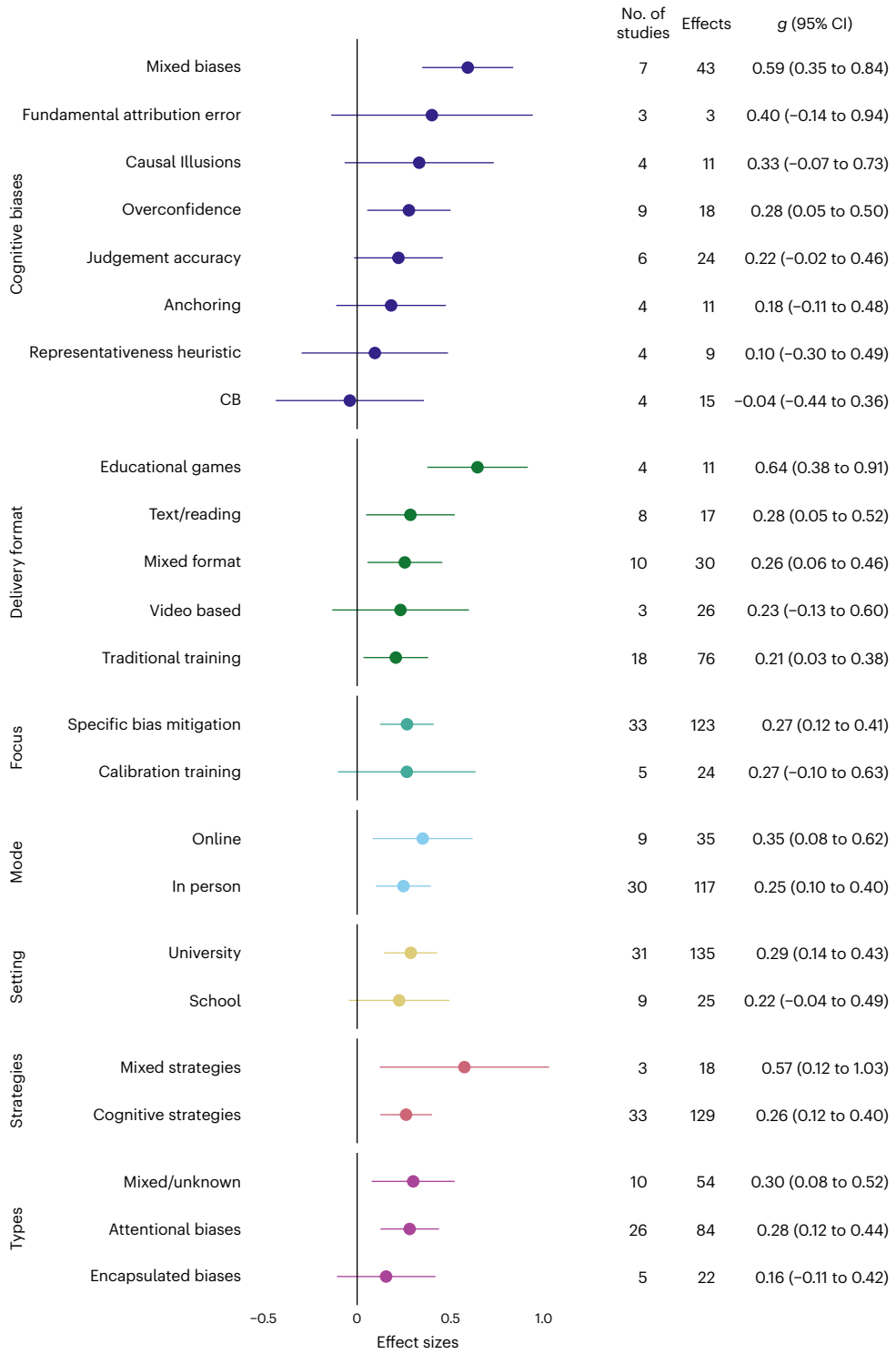

Fig. 3 | Plot of overall moderation analyses. Hedges' *g* effect sizes (centre points) with 95% CIs (error bars) for each moderator level. Different colours represent different moderator variables. Cognitive biases and delivery format were the only significant moderators of effects. The full statistical results for all moderator analyses are in 'Results' section.

might be easier to reduce biases. Our qualitative synthesis supports this, showing that conditions with games or elaborative feedback typically outperformed those without (for example, lectures). This important role of feedback is consistent with meta-analyses on feedback for learning[73,75], critical thinking[44] and arguments for developing intuitive expertise[12].

Across all biases, interventions performed well against 'no intervention' and 'active control' groups. This finding is encouraging, suggesting that debiasing effects are not merely due to demand characteristics. Effects were consistent across education settings and delivery modes (online or in person). We expect this was due to the heterogeneity within levels, where the variation in the content made the process of learning less consequential. Educational games outperformed other formats, suggesting evidence-based teaching strategies—such as interactivity[76,77], management of cognitive load[78,79], authentic assessments[80] and need-supportive learning environments[81,82]—remain

https://doi.org/10.1038/s41562-025-02253-y

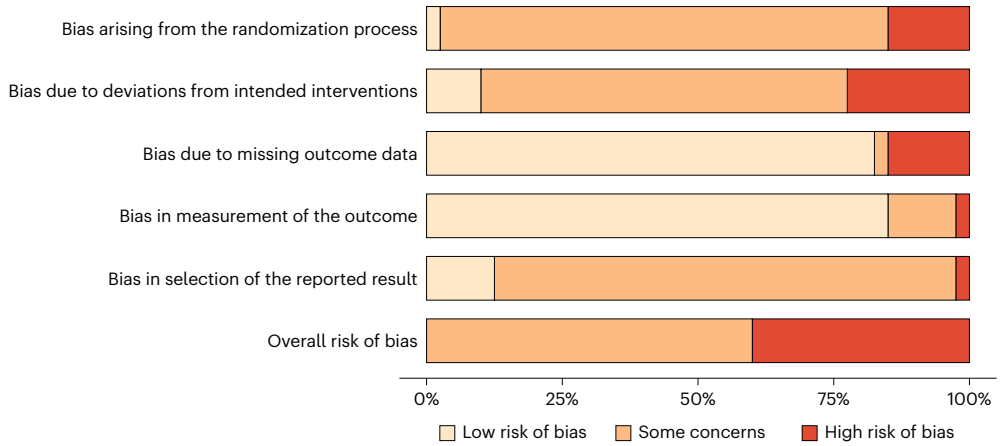

Fig. 4 | Risk of bias summary for included studies. The proportion of studies assessed as having low risk (cream), some concerns (orange) or high risk (red) for each domain of the Cochrane Risk of Bias tool 2.0. Full risk of bias assessments for all studies are available in Supplementary Table 2.

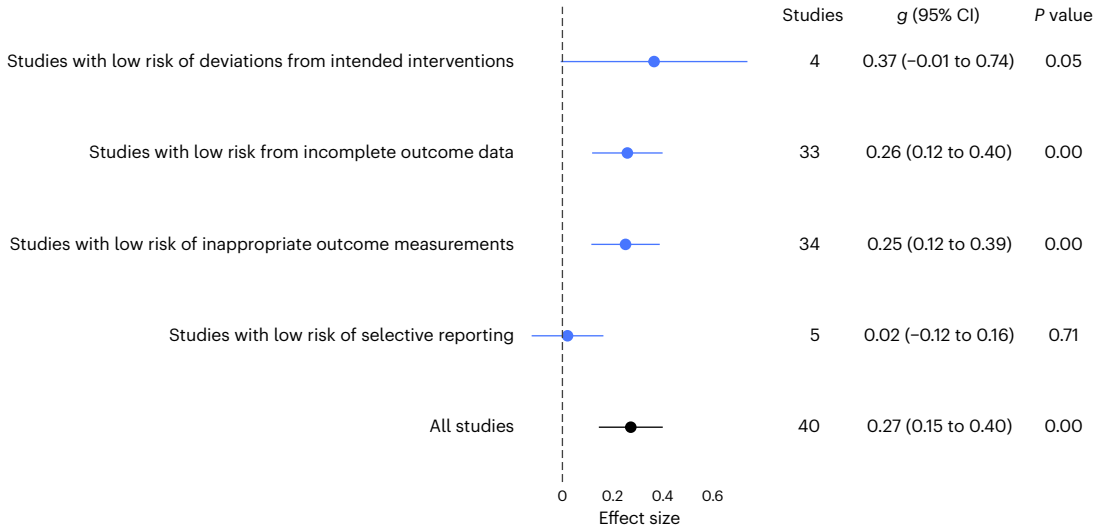

Fig. 5 | Sensitivity analysis by risk of bias criteria. Hedges' *g* effect sizes (centre points) with 95% CIs (error bars) for different risk of bias criteria. Main analysis effect size ($g = 0.26$, 95% CI 0.14 to 0.39, $P < 0.001$) compared with effect sizes from studies meeting specific risk of bias criteria. There were no statistically significant differences between the pooled estimate for all studies (in black) and estimates including only low-risk studies for any criteria (in blue): deviations from intervention ($\beta_{\text{low risk versus all studies}} = 0.09$, 95% CI −0.29 to 0.48, $P = 0.62$), incomplete outcome data ($\beta_{\text{low risk versus all studies}} = -0.015$, 95% CI −0.07 to 0.04, $P = 0.61$), inappropriate outcome measurement ($\beta_{\text{low risk versus all studies}} = -0.025$, 95% CI −0.08 to 0.03, $P = 0.37$) or selective reporting ($\beta_{\text{low risk versus all studies}} = -0.25$, 95% CI −0.57 to 0.08, $P = 0.13$).

relevant for teaching debiasing. Most studies focused on the effects of training versus no training rather than on learning design manipulations. Therefore, despite nonsignificant moderation effects, we maintain that learning design probably matters.

Bias mitigation interventions are only valuable if they are retained over time and can improve practical decision-making[83]. The longer the gap between the first and second exposure to heuristics and biases task, the more likely participants would show their 'true' responses once the 'experience effect' wore off[30]. While we found no significant differences between immediate and delayed post-test outcomes, only eight studies reported follow-up measurements, so a nonsignificant difference may reflect a low power for the analysis. Future research should prioritize assessing whether these effects are retained for a meaningful period.

While our meta-analysis shows positive effects of debiasing interventions, questions remain about the depth of learning achieved. Many interventions incorporated explanatory components and strategic training (for example, consider-the-opposite strategies and feedback)[52] rather than just providing correct answers. However, most outcome measures focused on performance on similar bias tasks rather than assessing deeper understanding or changes in thinking dispositions. This measurement limitation makes it difficult to determine whether participants developed genuine insight into why certain responses reflect biased thinking versus simply learning to recognize and avoid specific response patterns.

In educational interventions that teach students how to think, transfer of gains to other subject areas is crucial[84]. Some included studies found that although intervention groups outperformed control groups on near-transfer items, there were no significant differences in far-transfer items[52,85], suggesting superficial rather than deep learning. This pattern appears in systematic reviews of games to reduce bias[83] and critical thinking interventions[86], where benefits remained limited to domain-specific measures rather than extending to generic critical thinking measures. This transfer challenge is pervasive throughout education[87], and we lack confidence that effective transfer occurs in debiasing education. Future research should incorporate measures that better distinguish between surface-level pattern recognition and deeper

https://doi.org/10.1038/s41562-025-02253-y

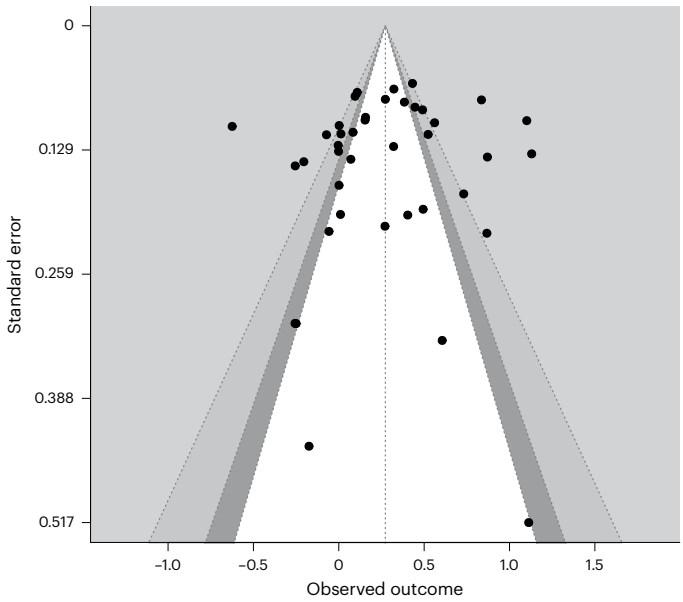

Fig. 6 | Funnel plot for publication bias assessment. The relationship between effect sizes (*x* axis) and standard errors (*y* axis) for all included studies. Plot asymmetry suggests potential publication bias. Full publication bias assessment results are in 'Results' section.

conceptual understanding that enables application across contexts. Until we identify methods of promoting transfer, educators should focus on biases in decisions relevant to students (for example, course selection or university applications) rather than on less-relatable examples (for example, 'is Linda a bank teller?'[88]). These authentic curricula may enhance motivational engagement with cognitive strategies[80], a recommendation also supported in critical thinking education literature[39,41,44].

These considerations suggest several priorities for improving debiasing interventions. First, incorporating more varied practice contexts and explicit discussion of underlying principles may help promote deeper understanding versus mere recognition of correct responses. Second, assessment methods should go beyond performance on standard bias tasks to evaluate whether students can articulate why certain thinking patterns reflect bias and demonstrate flexible application of debiasing strategies. Finally, connecting abstract principles to authentic decisions relevant to students' lives may promote engaged learning needed for lasting changes in thinking dispositions rather than just task-specific improvements.

We should test whether debiasing interventions interfere with functional heuristics. Our study focused on reducing cognitive biases, aligned with the heuristics and biases framework, which has faced criticism[13,89,90]. Some researchers argue that 'errors' in laboratory tasks may reflect strategies well adapted for real-world decision-making[11]. For instance, CB might be a side-effect of efficient Bayesian reasoning[91]. Four studies targeting CB (study 2 in ref. 52 and refs. 54,92,93) showed no significant improvement over control groups, underscoring how deeply ingrained this bias is. As CB is often intertwined with personal beliefs and motivations, classroom training may not automatically translate to other domains without additional supports—such as repeated practice in varied contexts, prompts to 'check for contrary evidence' or real-life incentives. Future research should explore whether reducing biases in educational settings improves real-world decision-making or potentially hinders the development of adaptive heuristics or causes other adverse effects. This highlights the need for more ecologically valid measures of decision quality.

Limitations of included studies

No studies in this review were judged to have an overall 'low' risk of bias, mainly due to lack of pre-registration or explicit blinding procedures. Education studies seldom meet the Cochrane 'low risk of bias' criteria as blinding participants or teachers can be difficult[78,94]. Our ratings may be harsh because many studies pre-dated reporting standards such as CONSORT[95] and JARS[96]; the studies may have used sound methods but just failed to report details. These ratings did not significantly influence the pooled effect sizes. Future researchers can mitigate experimental biases by pre-registering studies (for example, Open Science Framework), allowing transparent research practices. More low-risk studies are needed to verify whether findings are robust against common experimental biases.

The funnel plot and multilevel Egger's test indicated a small study effect in the included studies, indicating tendencies for small studies only to be published when they are significant. Prospective registration could help with controlling publication bias. Nevertheless, the *s* value indicated that publication bias would need to be relatively strong to nullify findings. Similarly, the 3PSM likelihood ratio test suggested no pattern of effect sizes consistent with publication bias. The findings presented here may be somewhat inflated by publication bias, but the key findings appear relatively robust to this bias.

There are many cognitive biases and the field is yet to deeply explore them. For example, only two studies explored effects to reduce the BBS[57,97], which is the belief that we are less biased compared with others. This meta-bias can cause people to be less receptive towards debiasing advice or engage in bias mitigation strategies because they are less likely to accept that their judgement could be wrong[98]. Hence, BBS could be an important factor that influences the success rate of debiasing training[98]. There is also a range of other influential biases (for example, base-rate neglect and sunk-cost fallacy)[99,100] that need further exploration.

Limitations of this review

We selected only randomized controlled trials (RCTs) for this review so we could draw stronger causal estimates of treatment effects. While RCTs have high internal validity, they often have low external validity. Many educators work with students for up to 40 weeks of the year and most of the included studies lasted less than 2 h. By only including RCTs, we may be missing more ecologically valid studies. Future researchers could review debiasing studies using observational studies, quasi-experimental and qualitative designs.

Another limitation of our methods is a confound between the ability for a bias to be taught and ease of measurement. Some biases (for

example, anchoring) are relatively easy to measure[101]. Others may be more laborious (for example, CB)[102]. Higher effects may reflect measurement accuracy rather than greater teachability. Our review was not designed to assess the reliability and validity of the measures used in this domain, but future reviews could compare different measures. The variety of tasks used to measure cognitive biases can be problematic, as wording and response format can influence participants' responses (for example, open-ended questions versus multiple choice)[103]. Conversely, overuse of similar measures risks invalidity if participants know the answers (for example, 'A bat and ball cost US$1.10…')[58]. Therefore, in addition to standardized measures of bias (for example, CART)[25,104], future research should develop consistent and accurate methods for measuring cognitive biases.

As our review only assessed outcomes that can be easily measured, we could not make causal inferences about how these interventions will improve practical decision-making in day-to-day lives. Simply knowing that you are being tested gives enough cue for anyone to immediately override their system I thinking[105]. However, we often miss these cues in real-life settings. As such, it is difficult to say how well benefits transfer to real-life situations, important decision-making and meaningful outcomes such as decision quality[106].

Our findings show that education reduces bias on average, though substantial unexplained heterogeneity remains. There are probably other influential moderators that we did not code for or could not discern from the methods alone. Student-level moderators (for example, gender and age) were not adequately addressed, as they relied on study averages rather than individual effects. This approach risks ecological fallacy, where aggregate-level treatment–covariate interactions may not reflect individual-level interactions[107]. Although we pre-registered age and gender as moderators, these moderation analyses were underpowered and potentially confounded by study-level characteristics. Future research could consider individual participant data meta-analyses to investigate which individual factors influence the effects of debiasing interventions and identify which individuals benefit the most from these interventions.

The large number of moderators we registered may have increased the chance of type I errors. While many significant moderators had P values well below critical values, some that were close to significance ($P = 0.04$) warrant caution. Additionally, some moderators were confounded, complicating interpretation. For example, interventions focusing on reasoning improvement exclusively addressed mixed types of biases, whereas encapsulated biases were only addressed in interventions focusing on specific bias mitigations. When both confounded moderators were entered into the multivariate meta-analysis model, there were no significant moderating effects, making it unclear which factor drove larger effect sizes. Readers should note that moderation analyses are observational and susceptible to unmeasured confounding.

Conclusions

Debiasing interventions can be effective in mitigating a range of cognitive biases, improving important components of rationality. Some biases (for example, anchoring and FAE) can be more challenging to mitigate. The differences between cognitive biases may stem from how easily students can get rapid, accurate feedback.

The findings of this review have important implications for both researchers and educators. Researchers should continue to examine and validate measures for cognitive biases (for example, CART)[25] to ensure consistency in cognitive bias measurement. Further exploration is needed for overcoming hard-to-shift biases (for example, CB) and on methods for deep learning and transfers to real-world decision-making. While our review highlights the effectiveness of debiasing training in formal educational settings, more research is needed before strong policy recommendations can be made. Educators might consider these findings when designing critical thinking programmes, but should be aware of the current limitations in the evidence base.

Methods

We prospectively registered this systematic review via the Open Science Framework (OSF)[108] on 26 July 2022. We registered the protocol before conducting the search, but after validating the search terms on the 'target set' of 13 papers. Since we registered the protocol, we have made a small number of changes to our methods following reviewer recommendations. We conducted additional moderation analyses in response to reviewer suggestions (that is, biases focused on social group membership and outcomes measured at immediate post-test versus follow-up). Also, rather than moderating for outcome, we conducted our analyses separately for each outcome to ensure other moderation analyses reflected homogeneous outcomes. We also added sensitivity analyses that were not in the pre-registration, specifically testing whether effect sizes differed when using only studies judged to be low risk of bias and when removing outlier effect sizes. Additionally, we implemented the Benjamini–Hochberg[109] procedure to control the FDR across all moderator analyses, which was not specified in our original protocol but added to strengthen our statistical approach and address concerns about multiple comparisons. We adhered to the PRISMA statement[110] and the Reporting Standards for Research in Psychology[96] to present our method and results.

Inclusion and ethics statement

As this study is a systematic review and meta-analysis rather than primary research, some aspects of the standard inclusion and ethics framework are not directly applicable. We made efforts to include studies from diverse geographical regions in our search strategy, including papers that were published in languages other than English, and applied consistent inclusion criteria regardless of study origin. Our author team includes researchers from Australia, Indonesia, Italy, Portugal and Hong Kong. We have made deliberate efforts to cite relevant research from all regions represented in our analysis, avoiding citation bias towards high-income country publications. No new participant data were collected, and no biological materials or cultural artifacts were transferred as part of this work. Our review was exempt from ethics committee approval as it involved analysis of published data only.

Eligibility criteria

We selected studies based on these pre-specified inclusion criteria:

(1) Design: we only included studies with randomized assignment into groups (either at an individual or cluster level). We only included RCTs because they are among the most robust methods of assessing educational interventions[111]. We excluded nonrandomized trials or cohort studies. While nonrandomized designs are pragmatic in many educational settings, they carry higher risks of confounding unless they use sophisticated statistical methods to control for all systematic differences between the intervention and control groups[112]. Using only RCTs enabled us to draw stronger causal claims on treatment effects[112].

(2) Participants: we included students in any kind of formal educational setting (for example, school and university). We excluded on-the-job learning, or where the purpose of the training is to alleviate mental illness (for example, interpretation bias modification for participants with health anxiety)[113] rather than to reduce bias/improve rationality.

(3) Interventions: we included any intervention with an explicit focus on explaining or providing strategies to overcome cognitive biases. Merely presenting content that would probably reduce bias (for example, courses on statistics and probability and critical thinking interventions) were not eligible unless the discussion of cognitive biases was explicit. This included:

- Cognitive bias modification training
- Cognitive bias mitigation (game based)
- Education interventions on heuristics and biases

- Motivational debiasing strategies
- Calibration training (for over-/underconfidence)
- Affective strategies to overcome biases

Interventions had to be 'educational' and not merely 'nudges' (that is, subtly altering the environment to influence decision-making). We operationalized educational interventions as interventions with curricular aims and objectives, usually guided by a teacher or instructor. We included studies where the educational intervention was ≥10 min. Brief intervention designs or short interventions (that is, 9 min or less) were excluded. We determined that interventions shorter than 10 min were closer to nudging than educational, where creating a minor change was not 'teaching' students the skills to mitigate cognitive biases, but rather served as a guide to shape people's behaviours as intended.

(1) Comparisons: we included comparisons against any other types of control, educational intervention or debiasing training. We grouped the following comparisons for moderation analyses:
   - No intervention or placebo interventions
   - Other active educational interventions (for example, mathematics training)
   - Alternative methods of debiasing (for example, game-based versus direct instruction)

(2) Outcome: we included learning outcomes related to knowledge or skills involved in overcoming biases. These outcomes were be analysed separately as follows:
   - The likelihood of committing specific biases (for example, standard contingency judgement task measuring causal illusions[114] and heuristics and biases scale[115])
   - Specific knowledge of biases (for example, determining what bias was represented in some scenarios)
   - Improved reasoning or cognition generally (for example, improved rationality assessed through CART[4] or the alternative form of cognitive reflection test[116])
   - Improved decision quality (for example, the youth decision-making competence scale[33])

(3) Finally, we included studies in any language, using either Google Translate or help from reviewers who spoke multiple languages. We included both published and unpublished studies from any time period. The studies needed to present original data (that is, we excluded review articles and research protocols where data were not yet available).

Information sources and search strategy

We generated a search strategy using the titles and abstracts of an initial sample of papers. We derived the strategy from existing reviews[9,83] and primary studies[5,117]. These papers helped us generate a list of terms that identified the target papers (that is, 'objective approach')[118]. This strategy ensured that we optimised sensitivity while maintaining specificity. The list of search terms was as follows:

- Participants: (educati* or student* or adolescen* or undergraduate* or universit* or college or participant* or teen* or child* or school or youth) and
- Intervention: (judgment* or judgement* or 'decision competence' or bias* or debias* or fallac* or rational* or intuition or heuristic* or 'decision making' or 'decision-making') and
- (train* or strateg* or intervention or teach* or technique or program* or *game* or curriculum) and
- Comparison: (randomised or randomized or experiment* or control or condition*) and
- Outcome: (skill* or learn* or reduc* or improve* or awareness or mitigat*)

We entered these search terms onto Scopus, PsycINFO, ERIC, Web of Science and Proquest Dissertations and Theses. We conducted searches on 28 July 2022 until 4 August 2022, with additional searches using the references from included studies conducted on 6 April 2023.

Study selection

We used EndNote X9 (ref. 119) to find and remove duplicated studies from the search. To accelerate our title/abstract searching, we first filtered the deduplicated studies through RobotSearch to remove the non-RCT studies[120]. RobotSearch is a pre-trained algorithm for identifying randomized experiments from titles and abstracts. It is highly accurate, with 99.1% sensitivity and 77.2% specificity[120]. We then used Covidence[121] to screen the title and abstract of the remaining studies, which was done independently and in duplicate across two reviewers. We sought full texts for all articles that passed title and abstract screening. The full-text screening was also conducted independently and in duplicate. In both screening stages, any conflicts were resolved through discussion or a consultation with a third reviewer. Finally, one reviewer searched the reference lists of included studies for papers that may have been missed in the first search[122]. All studies were included in our systematic review. We included studies in the meta-analysis when we were able to extract or impute sufficient data for effect size calculations using guidelines from the Cochrane *Handbook for Systematic Reviews of Interventions*[123].

Data items and collection process

One reviewer developed a data extraction form, which then was piloted and revised alongside another reviewer. The form extracted details regarding participants (for example, age and education level), outcome information (description of measures and construct type), intervention (description, duration, delivery format, intervention focus and strategies taught), comparison (type of comparison, description and duration) and any metric that could be used to calculate an effect size (for example, means, standard deviations, CIs and *P* values). When the included study did not present exact values but presented figures, we extracted data from figures using WebPlotDigitizer[124]. Reviewers independently extracted all items in duplicate and resolved disagreements via discussion or consultation with a third author when necessary.

Risk of bias in individual studies

We assessed the risk of bias in individual studies using the Cochrane revised tool to assess risk of bias in randomized trials (RoB 2)[125]. This tool assesses whether studies have sufficiently mitigated the risk in five domains of biases: bias arising from the randomization process, bias due to deviations from intended interventions, bias due to missing outcome data, bias in measurement of the outcome and bias in selection of the reported result. We chose this standard because RoB 2 has shown better sensitivity, specificity and validity compared with other quality assessment tools and its predecessor (RoB 1). Compared with RoB 1, RoB 2 has more guidance from signalling questions embedded in the tool[126]. Following the guidance in the Cochrane *Handbook*[123], the overall risk of bias for a study was considered low risk if the study was judged as low risk in all domains. We analysed risk as 'intention to treat', assessing the effect of 'assignment' (that is, including all randomized participants) rather than 'adherence' to the intervention. As per AMSTAR 2 (ref. 127), we conducted quality assessment independently and in duplicate until reviewers reached sufficient agreement. Specifically, we conducted quality assessments of the first 17 included studies (~35%) independently and in duplicate using the RoB 2 Excel tool. We resolved disagreements through discussion and consultation with a third author where necessary. Once we reached over 80% agreement on all domains, one reviewer conducted the risk of bias individually for the remaining studies.

Summary measures and synthesis of results

We extracted data for each eligible effect size reported within each study. We used the post-test, between-groups standardized mean difference as the principal summary measure. We used Hedges's *g* to correct for biases in small sample sizes[128]. We can interpret Hedges's *g* in a similar convention as Cohen's *d* (that is, 0.2 = small, 0.5 = moderate and 0.8 = large). This measure was calculated using the metafor[129] package in R[130]. When means and standard deviations were not reported in the original paper, we used other statistics to calculate effect sizes (for example, *P* values and CIs) or imputed estimates using recommendations from the Cochrane *Handbook*.

We conducted multilevel meta-analyses, nesting effect sizes within studies using the rma.mv function in the metafor package. All statistical tests conducted in our meta-analysis, including moderator analyses using *F* tests, were interpreted using the conventional approach for these analyses, which examines the significance of variance explained by moderators without directional hypotheses. For correlation analyses examining relationships between moderators, two-tailed tests were used as no directional hypotheses were pre-registered. For each analysis, we assessed heterogeneity using I^2 at level 2 and level 3 (within and between studies, respectively). This shows how much variance is not explained by sampling error[131]. The results of these analyses were visualised using the forest function in the meta package[132], producing a forest plot of aggregated results (effects nested within studies). We also assessed heterogeneity using processes outlined by Mathur & VanderWeele[133]. These processes assessed the proportion of true effects (that is, those not due to chance, artefacts or sampling error) that are likely to be helpful (we defined as a small, positive effect; Hedges' $g > 0.2$) or harmful (Hedges' $g < -0.2$).

Additional analyses

We ran a series of moderation analyses to explore different possible sources of heterogeneity. These analyses included (1) age (continuously moderated; using mean/median); (2) educational setting (for example, primary/elementary school versus middle/high school versus university versus holiday camps); (3) gender (continuously moderated; using percentage female, where reported); (4) specific bias addressed by the intervention (that is, specific bias mitigation, reasoning improvement or calibration training); (5) types of biases (specific cognitive biases, encapsulated versus attentional biases); (6) debiasing strategies taught (that is, cognitive-oriented strategies versus affective-oriented strategies versus motivational-oriented strategies); (7) length of intervention (for example, one session versus multiple sessions, reported length of intervention); (8) delivery format (that is, reading versus game-based versus traditional training versus video-based) and (9) delivery mode (that is, person versus computer-aided). These analyses were visually plotted using the ggplot2 (ref. 134) package in R.

We also conducted a range of sensitivity analyses that were not in the pre-registration following recommendations from a reviewer. We tested whether effect sizes were smaller than in our main analysis when only using studies judged to be low risk of bias. We also tested whether effect sizes were different when removing outlier effect sizes. We checked the correlations between these moderators to assess whether any of them were confounded. We ran a pairwise chi-square test for our categorical moderators, correlational analysis for our continuous moderators and a one-way analysis of variance for both categorical and continuous moderators. To account for multiple comparisons across our moderator analyses, we applied the Benjamini–Hochberg procedure[109] to control the FDR. This adjustment was implemented across all moderator analyses simultaneously to maintain a more conservative approach to significance testing[135–137]. The FDR control procedure helps maintain the expected proportion of false positives among all rejected null hypotheses, providing a more balanced approach between type I error control and statistical power compared to traditional family-wise error rate corrections (for example, the Bonferroni correction)[138].

Risk of bias across studies

To assess publication bias, we conducted a multilevel meta-analytic Egger's test. This test assesses whether effect sizes and standard errors are associated, which indicates publication bias. This test also controls for clustering. As a second test of publication bias, we also conducted a selection model[139] using the weightr function in R, after aggregating effect sizes within studies (using the aggregate function in metafor).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the findings of this systematic review and meta-analysis are openly available via the OSF repository at https://osf.io/7x5z6. This repository contains the complete dataset extracted from the included studies, including effect sizes, moderator variables and risk of bias assessments. The minimum dataset necessary to interpret, verify and extend the research includes the coded effect sizes from each study, sample sizes and moderator variables used in our analyses. All data are provided in accessible formats.

Code availability

The R code used to conduct all analyses and generate the figures in this systematic review and meta-analysis is available via the same OSF repository at https://osf.io/7x5z6. The code includes all scripts used for data preparation, meta-analytic models, moderation analyses, sensitivity analyses and visualization. All analyses were conducted using R (version 4.4.1) with the metafor, meta and ggplot2 packages. The repository includes documented code with comments and a README file to facilitate reproducibility of all results presented in this Article.

References

1. Beyth-Marom, R., Fischhoff, B., Quadrel, M. J. & Furby, L. in *Teaching Decision Making to Adolescents* (eds Baron, J. & Brown, R. V.) (Routledge, 1991).
2. Stanovich, K. E. & West, R. F. What intelligence tests miss. *Psychologist* **27**, 80–83 (2014).
3. Toplak, M. E., Sorge, G. B., Benoit, A., West, R. F. & Stanovich, K. E. Decision-making and cognitive abilities: a review of associations between Iowa Gambling Task performance, executive functions, and intelligence. *Clin. Psychol. Rev.* **30**, 562–581 (2010).
4. Stanovich, K. E., West, R. F. & Toplak, M. E. *The Rationality Quotient: Toward a Test of Rational Thinking* (MIT Press, 2016).
5. Aczel, B., Bago, B., Szollosi, A., Foldes, A. & Lukacs, B. Is it time for studying real-life debiasing? Evaluation of the effectiveness of an analogical intervention technique. *Front. Psychol.* **6**, 1120 (2015).
6. Yagoda, B. The cognitive biases tricking your brain. *The Atlantic* (September 2018).
7. Featherston, R. et al. Interventions to mitigate bias in social work decision-making: a systematic review. *Res. Soc. Work Pract.* **29**, 741–752 (2019).
8. Prakash, S., Sladek, R. M. & Schuwirth, L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. *Med. Teach.* **41**, 517–524 (2019).
9. Ludolph, R. & Schulz, P. J. Debiasing health-related judgments and decision making: a systematic review. *Med. Decis. Making* **38**, 3–13 (2018).
10. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).

11. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
12. Kahneman, D. & Klein, G. Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* **64**, 515–526 (2009).
13. Funder, D. C. Errors and mistakes: evaluating the accuracy of social judgment. *Psychol. Bull.* **101**, 75–90 (1987).
14. Pronin, E. Perception and misperception of bias in human judgment. *Trends Cogn. Sci.* **11**, 37–43 (2007).
15. Saposnik, G., Redelmeier, D., Ruff, C. C. & Tobler, P. N. Cognitive biases associated with medical decisions: a systematic review. *BMC Med. Inform. Decis. Mak.* **16**, 138 (2016).
16. Featherston, R., Downie, L. E., Vogel, A. P. & Galvin, K. L. Decision making biases in the allied health professions: a systematic scoping review. *PLoS ONE* **15**, e0240716 (2020).
17. Edmond, G. & Martire, K. A. Just cognition: scientific research on bias and some implications for legal procedure and decision-making. *Modern L. Rev.* **82**, 633–664 (2019).
18. Blanco, F. in *Encyclopedia of Animal Cognition and Behavior* (eds Vonk, J. & Shackelford, T. K.) (Springer, 2022).
19. Odds of dying. Injury facts. *National Safety Council* https://injuryfacts.nsc.org/all-injuries/preventable-death-overview/odds-of-dying/ (2017).
20. Stanovich, K. E. & West, R. F. The assessment of rational thinking: IQ ≠ RQ. *Teach. Psychol.* **41**, 265–271 (2014).
21. Bruine de Bruin, W., Parker, A. M. & Fischhoff, B. Decision-making competence: more than intelligence? *Curr. Dir. Psychol. Sci.* **29**, 186–192 (2020).
22. Ghazal, S., Cokely, E. T., Garcia-Retamero, R. & Feltz, A. *Cambridge Handbook of Expertise and Expert Performance* (Cambridge Univ. Press, 2018).
23. Primi, C., Donati, M. A., Chiesi, F. & Panno, A. in *Individual Differences in Judgement and Decision-Making* (eds Toplak, M. E. & Weller, J.) 58–76 (Psychology Press, 2016).
24. Wechsler, D. *WAIS-IV: Wechsler Adult Intelligence Scale—Fourth Edition* (Pearson, 2008).
25. Stanovich, K. E. The comprehensive assessment of rational thinking. *Educ. Psychol.* **51**, 23–34 (2016).
26. Todd, B. Notes on good judgement and how to develop it. *80,000 Hours* https://80000hours.org/2020/09/good-judgement/ (2020).
27. Kahneman, D. *Thinking, Fast and Slow* (Penguin, 2011).
28. Stanovich, K. E. Miserliness in human cognition: the interaction of detection, override and mindware. *Think. Reason.* **24**, 423–444 (2018).
29. Toplak, M. E., West, R. F. & Stanovich, K. E. Real-world correlates of performance on heuristics and biases tasks in a community sample. *J. Behav. Decis. Mak.* **30**, 541–554 (2017).
30. Chandler, J., Paolacci, G., Peer, E., Mueller, P. & Ratliff, K. A. Using nonnaive participants can reduce effect sizes. *Psychol. Sci.* **26**, 1131–1139 (2015).
31. Haigh, M. Has the standard cognitive reflection test become a victim of its own success? *Adv. Cogn. Psychol.* **12**, 145–149 (2016).
32. Bruine de Bruin, W., Parker, A. M. & Fischhoff, B. Individual differences in adult decision-making competence. *J. Pers. Soc. Psychol.* **92**, 938–956 (2007).
33. Parker, A. M. & Fischhoff, B. Decision-making competence: external validation through an individual-differences approach. *J. Behav. Decis. Mak.* **18**, 1–27 (2005).
34. Ro, C. The complicated battle over unconscious-bias training. *BBC* (29 March 2021).
35. Sukhera, J. Starbucks and the impact of implicit bias training. *The Conversation* (27 May 2018).
36. Walker, T. B. & Feloni, R. Here's the presentation Google gives employees on how to spot unconscious bias at work. *Business Insider* (2020).
37. Cantarelli, P., Belle, N. & Belardinelli, P. Behavioral public HR: experimental evidence on cognitive biases and debiasing interventions. *Rev. Public Pers. Adm.* **40**, 56–81 (2020).
38. Morewedge, C. K. et al. Debiasing decisions: improved decision making with a single training intervention. *Policy Insights Behav. Brain Sci.* **2**, 129–140 (2015).
39. Davies, M. in *Higher Education: Handbook of Theory and Research* (ed. Perna, L. W.) 41–92 (Springer, 2015).
40. Stanovich, K. E. *The Oxford Handbook Of Thinking And Reasoning* (Oxford Univ. Press, 2012).
41. Ennis, R. H. *The Palgrave Handbook of Critical Thinking in Higher Education* (Palgrave Macmillan, 2015).
42. Common core state standards. *National Governors Association* https://preview.fadss.org/resources/webinars/webinar2/FSBAPresentationforCommunities_transcribed.pdf (2010).
43. *Next Generation Science Standards: for States, by States* (National Academies Press, 2013).
44. Abrami, P. C. et al. Strategies for teaching students to think critically: a meta-analysis. *Rev. Educ. Res.* **85**, 275–314 (2015).
45. Mao, W., Cui, Y., Chiu, M. M. & Lei, H. Effects of game-based learning on students' critical thinking: a meta-analysis. *J. Educ. Comput. Res.* **59**, 1682–1708 (2022).
46. Xu, E., Wang, W. & Wang, Q. The effectiveness of collaborative problem solving in promoting students' critical thinking: a meta-analysis based on empirical literature. *Humanit. Soc. Sci. Commun.* **10**, 16 (2023).
47. Ennis, R. H. Critical thinking and subject specificity: clarification and needed research. *Educ. Res.* **18**, 4 (1989).
48. Siegel, H. *Education's Epistemology: Rationality, Diversity, and Critical Thinking* (Oxford Univ. Press, 2017).
49. Hattie, J. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement* (Routledge, 2008).
50. Page, M. J. et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *Br. Med. J.* **372**, n160 (2021).
51. Haddaway, N. R., Page, M. J., Pritchard, C. C. & McGuinness, L. A. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst. Rev.* **18**, e1230 (2022).
52. Calvillo, D. P., Bratton, J., Velazquez, V., Smelter, T. J. & Crum, D. Elaborative feedback and instruction improve cognitive reflection but do not transfer to related tasks. *Think. Reason.* **29**, 276–304 (2022).
53. Salvatore, J. & Morton, T. A. Evaluations of science are robustly biased by identity concerns. *Group Process. Intergr. Relat.* **24**, 568–582 (2021).
54. van Brussel, S., Timmermans, M., Verkoeijen, P. & Paas, F. Teaching on video as an instructional strategy to reduce confirmation bias—a pre-registered study. *Instr. Sci.* **49**, 475–496 (2021).
55. Rhodes, R. E. et al. Teaching decision making with serious games: an independent evaluation. *Games Cult.* **12**, 233–251 (2017).
56. Dwyer, C. P., Hogan, M. J. & Stewart, I. The effects of argument mapping-infused critical thinking instruction on reflective judgement performance. *Think. Skills Creat.* **16**, 11–26 (2015).
57. Sellier, A. L., Scopelliti, I. & Morewedge, C. K. Debiasing training improves decision making in the field. *Psychol. Sci.* **30**, 1371–1379 (2019).
58. Frederick, S. Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005).
59. Dawson, T. L. *Metacognition and Learning in Adulthood. Prepared in Response to Tasking from ODNI/CHCO/IC Leadership Development Office* (Developmental Testing Service LLC, 2008).

60. Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Hambrick, D. Z. & Engle, R. W. Understanding the relationship between rationality and intelligence: a latent-variable approach. *Think. Reason.* **29**, 1–42 (2023).
61. Lectical reflective judgment assessment. *LecticaLive* https://lecticalive.org/about/lrja (2024).
62. Dunbar, N. E. et al. Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game. *Comput. Hum. Behav.* **37**, 307–318 (2014).
63. Dunbar, N. E. et al. Mitigation of cognitive bias with a serious game: two experiments testing feedback timing and source. *Int. J. Game Based Learn.* **7**, 86–100 (2017).
64. Shaw, A. et al. Serious efforts at bias reduction: the effects of digital games and avatar customization on three cognitive biases. *J. Media Psychol.* **30**, 16–28 (2018).
65. Legaki, N.-Z., Karpouzis, K., Assimakopoulos, V. & Hamari, J. Gamification to avoid cognitive biases: an experiment of gamifying a forecasting course. *Technol. Forecast. Soc. Change* **167**, 120725 (2021).
66. Gutierrez, B. *Fair Play: A Video Game Designed to Reduce Implicit Racial Bias* (Univ. Wisconsin, 2013).
67. Lee, Y.-H. et al. Training anchoring and representativeness bias mitigation through a digital game. *Simul. Gaming* **47**, 751–779 (2016).
68. Roelle, J., Schmidt, E. M., Buchau, A. & Berthold, K. Effects of informing learners about the dangers of making overconfident judgments of learning. *J. Educ. Psychol.* **109**, 99–117 (2017).
69. van Peppen, L. M. et al. Learning to avoid biased reasoning: effects of interleaved practice and worked examples. *J. Cogn. Psychol.* **33**, 304–326 (2021).
70. Martínez, N., Rodríguez-Ferreiro, J., Barberia, I. & Matute, H. A debiasing intervention to reduce the causality bias in undergraduates: the role of a bias induction phase. *Curr. Psychol.* **42**, 32456–32468 (2023).
71. Evans, J. S. B. T. & Stanovich, K. E. Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* **8**, 223–241 (2013).
72. Alter, A. L., Oppenheimer, D. M., Epley, N. & Eyre, R. N. Overcoming intuition: metacognitive difficulty activates analytic reasoning. *J. Exp. Psychol. Gen.* **136**, 569–576 (2007).
73. Wisniewski, B., Zierer, K. & Hattie, J. The power of feedback revisited: a meta-analysis of educational feedback research. *Front. Psychol.* **10**, 3087 (2019).
74. Muehlhauser, L. New web app for calibration training. *Open Philanthropy* https://www.openphilanthropy.org/research/new-web-app-for-calibration-training/ (2018).
75. Swart, E. K., Nielen, T. M. J. & Sikkema-de Jong, M. T. Supporting learning from text: a meta-analysis on the timing and content of effective feedback. *Educ. Res. Rev.* **28**, 100296 (2019).
76. Chi, M. T. & Wylie, R. The ICAP Framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **49**, 219–243 (2014).
77. Tomcho, T. J. & Foels, R. Meta-analysis of group learning activities: empirically based teaching recommendations. *Teach. Psychol.* **39**, 159–169 (2012).
78. Noetel, M. et al. Video improves learning in higher education: a systematic review. *Rev. Educ. Res.* **91**, 204–236 (2021).
79. Noetel, M. et al. Multimedia design for learning: an overview of reviews with meta-meta-analysis. *Rev. Educ. Res.* **92**, 413–454 (2022).
80. Chernikova, O. et al. Simulation-based learning in higher education: a meta-analysis. *Rev. Educ. Res.* **90**, 499–541 (2020).
81. Ahmadi, A. et al. A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. *J. Educ. Psychol.* **115**, 1158–1176 (2023).
82. Bureau, J., Howard, J. L., Chong, J. X. Y. & Guay, F. Pathways to student motivation: a meta-analysis of antecedents of autonomous and controlled motivations. *Rev. Educ. Res.* **92**, 46–72 (2022).
83. Korteling, J. E. H., Gerritsma, J. Y. J. & Toet, A. Retention and transfer of cognitive bias mitigation interventions: a systematic literature study. *Front. Psychol.* **12**, 629354 (2021).
84. Halpern, D. F. Teaching critical thinking for transfer across domains: disposition, skills, structure training, and metacognitive monitoring. *Am. Psychol.* **53**, 449–455 (1998).
85. van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M. & van Gog, T. Enhancing students' critical thinking skills: is comparing correct and erroneous examples beneficial? *Instr. Sci.* **49**, 747–777 (2021).
86. Tiruneh, D. T., Verburgh, A. & Elen, J. Effectiveness of critical thinking instruction in higher education: a systematic review of intervention studies. *High. Educ. Stud.* **4**, 1–17 (2014).
87. Willingham, D. T. in *Critical Thinking: Why It Is So Hard to Teach?* 8–19 (American Federation of Teachers, 2007).
88. Tversky, A. & Kahneman, D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* **90**, 293–315 (1983).
89. McKenzie, C. R. M. Rational models as theories—not standards—of behavior. *Trends Cogn. Sci.* **7**, 403–406 (2003).
90. Szollosi, A. & Newell, B. R. People as intuitive scientists: reconsidering statistical explanations of decision making. *Trends Cogn. Sci.* **24**, 1008–1018 (2020).
91. Alexander, S. Confirmation bias as misfire of normal Bayesian reasoning. *Slate Star Codex* https://slatestarcodex.com/2020/02/12/confirmation-bias-as-misfire-of-normal-bayesian-reasoning/ (2020).
92. Abendroth, J. & Richter, T. How to understand what you don't believe: metacognitive training prevents belief-biases in multiple text comprehension. *Learn. Instr.* **71**, 101394 (2021).
93. Sibbald, M. et al. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. *Adv. Health Sci. Educ. Theory Pract.* **24**, 427–440 (2019).
94. Kyaw, B. M. et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. *J. Med. Internet Res.* **21**, e12959 (2019).
95. Schulz, K. F., Altman, D. G. & Moher, D. & CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.* **8**, 18 (2010).
96. Cooper, H. & Cooper, H. M. *Reporting Quantitative Research in Psychology: How to Meet APA Style Journal Article Reporting Standards* (American Psychological Association, 2020).
97. Joy-Gaba, J. A. *From Learning to Doing: The Effects of Educating Individuals on the Pervasiveness of Bias* (Univ. Virginia, 2011).
98. Scopelliti, I. et al. Bias blind spot: structure, measurement, and consequences. *Manage. Sci.* **61**, 2468–2486 (2015).
99. Oeberst, A. & Imhoff, R. Toward parsimony in bias research: a proposed common framework of belief-consistent information processing for a set of biases. *Perspect. Psychol. Sci.* **18**, 1464–1487 (2023).
100. Roth, S., Robbert, T. & Straus, L. On the sunk-cost effect in economic decision-making: a meta-analytic review. *Bus. Res.* **8**, 99–138 (2015).
101. Stanovich, K. E. & West, R. F. On the relative independence of thinking biases and cognitive ability. *J. Pers. Soc. Psychol.* **94**, 672–695 (2008).
102. Stanovich, K. E., West, R. F. & Toplak, M. E. Myside bias, rational thinking, and intelligence. *Curr. Dir. Psychol. Sci.* **22**, 259–264 (2013).

103. Aczel, B., Bago, B., Szollosi, A., Foldes, A. & Lukacs, B. Measuring individual differences in decision biases: methodological considerations. *Front. Psychol.* **6**, 1770 (2015).
104. Toplak, M. E. & Stanovich, K. E. Measuring rational thinking in adolescents: the assessment of rational thinking for youth (ART-Y). *J. Behav. Decis. Mak.* **37**, e2381 (2024).
105. Kahneman, D. *Thinking, Fast and Slow* (Macmillan, 2011).
106. Di Battista, A., Grayling, S. & Hasselaar, E. *Future of Jobs Report 2023* (World Economic Forum, 2023).
107. Fisher, D. J., Carpenter, J. R., Morris, T. P., Freeman, S. C. & Tierney, J. F. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *Br. Med. J.* **356**, j573 (2017).
108. Does debiasing training improve rationality? A systematic review and meta-analysis of randomised trials in educational settings. *OSF* https://osf.io/xrm4g (2022).
109. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
110. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int. J. Surg.* **8**, 336–341 (2010).
111. Torgerson, C. J. & Torgerson, D. J. *Randomised Trials in Education: An Introductory Handbook* (Education Endowment Foundation, 2013).
112. Kuss, O., Blettner, M. & Börgermann, J. Propensity score: an alternative method of analyzing treatment effects. *Dtsch. Arztebl. Int.* **113**, 597–603 (2016).
113. Antognelli, S. L., Sharrock, M. J. & Newby, J. M. A randomised controlled trial of computerised interpretation bias modification for health anxiety. *J. Behav. Ther. Exp. Psychiatry* **66**, 101518 (2020).
114. Matute, H. et al. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Front. Psychol.* **6**, 888 (2015).
115. Sklad, M. & Diekstra, R. The development of the heuristics and biases scale (HBS). *Procedia Soc. Behav. Sci.* **112**, 710–718 (2014).
116. Thomson, K. S. & Oppenheimer, D. M. Investigating an alternate form of the cognitive reflection test. *Judgm. Decis. Mak.* **11**, 99–113 (2016).
117. Jacobson, D. et al. Improved learning in US history and decision competence with decision-focused curriculum. *PLoS ONE* **7**, e45775 (2012).
118. Hausner, E., Guddat, C., Hermanns, T., Lampert, U. & Waffenschmidt, S. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *J. Clin. Epidemiol.* **77**, 118–124 (2016).
119. EndNote (The EndNote Team, 2013).
120. Marshall, I. J., Noel-Storr, A., Kuiper, J., Thomas, J. & Wallace, B. C. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Res. Synth. Methods* **9**, 602–614 (2018).
121. Covidence Systematic Review Software (Veritas Health Innovation, 2023).
122. Pigott, T. D. & Polanin, J. R. Methodological guidance paper: high-quality meta-analysis in a systematic review. *Rev. Educ. Res.* **90**, 24–46 (2020).
123. Higgins, J. P. T. et al. *Cochrane Handbook for Systematic Reviews of Interventions* (Wiley, 2019).
124. Rohatgi, A. WebPlotDigitizer. https://automeris.io/WebPlotDigitizer/ (2022).
125. Sterne, J. A. C. et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Br. Med. J.* **366**, l4898 (2019).
126. Flemyng, E. et al. Using Risk of Bias 2 to assess results from randomised controlled trials: guidance from Cochrane. *BMJ Evid. Based Med.* **28**, 260–266 (2023).
127. Shea, B. J. et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Br. Med. J.* **358**, j4008 (2017).
128. Hedges, L. V. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Behav. Stat.* **6**, 107–128 (1981).
129. Viechtbauer, W. The metafor package ver. 4.6-0. (2017).
130. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
131. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. *Introduction to Meta-Analysis* (Wiley, 2011).
132. Schwarzer, G. meta: an R package for meta-analysis (ver. 7.0-0). *R News* **7**, 40–45 (2007).
133. Mathur, M. B. & VanderWeele, T. J. New metrics for meta-analyses of heterogeneous effects. *Stat. Med.* **38**, 1336–1342 (2019).
134. Wickham, H. Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
135. Krzywinski, M. & Altman, N. Comparing samples—part II. *Nat. Methods* **11**, 355–356 (2014).
136. Glickman, M. E., Rao, S. R. & Schultz, M. R. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* **67**, 850–857 (2014).
137. Polanin, J. R. & Pigott, T. D. The use of meta-analytic statistical significance testing. *Res. Synth. Methods* **6**, 63–73 (2015).
138. Shaffer, J. Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584 (1995).
139. Hedges, L. V. & Vevea, J. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (eds Rothstein, H. R. et al.) (John Wiley & Sons, 2005).
140. Adame, B. J. Training in the mitigation of anchoring bias: a test of the consider-the-opposite strategy. *Learn. Motiv.* **53**, 36–48 (2016).
141. Schmalhofer, F. & Glavanov, D. Three components of understanding a programmer's manual: verbatim, propositional, and situational representations. *J. Mem. Lang.* **25**, 279–294 (1986).
142. Almashat, S., Ayotte, B., Edelstein, B. & Margrett, J. Framing effect debiasing in medical decision making. *Patient Educ. Couns.* **71**, 102–107 (2008).
143. Barberia, I., Blanco, F., Cubillas, C. P. & Matute, H. Implementation and assessment of an intervention to debias adolescents against causal illusions. *PLoS ONE* **8**, e71303 (2013).
144. Blanco, F., Matute, H. & Vadillo, A. M. Mediating role of activity level in the depressive realism effect. *PLoS ONE* **7**, e46203 (2012).
145. Barberia, I., Tubau, E., Matute, H. & Rodríguez-Ferreiro, J. A short educational intervention diminishes causal illusions and specific paranormal beliefs in undergraduates. *PLoS ONE* **13**, e0191907 (2018).
146. Díaz-Vilela, L. & Álvarez-González, C. J. Differences in paranormal beliefs across fields of study from a Spanish adaptation of Tobacyk's RPBS. *J. Parapsychol.* **68**, 405–421 (2004).
147. Bessarabova, E. et al. Mitigating bias blind spot via a serious video game. *Comput. Hum. Behav.* **62**, 452–466 (2016).
148. Pronin, E., Lin, D. Y. & Ross, L. The bias blind spot: perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).
149. Botta, V. A. *The Effect of Instructional Method on use of Heuristics and Statistics Comprehension* (Georgia State Univ., 1998).
150. Bou Khalil, R., Sleilaty, G., Kassab, A. & Nemr, E. Decontextualisation for framing effect reduction. *Clin. Teach.* **19**, 121–128 (2022).

151. Toplak, M. E., West, R. F. & Stanovich, K. E. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cogn.* **39**, 1275–1289 (2011).
152. Baron, J., Scott, S., Fincher, K. & Emlen Metz, S. Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *J. Appl. Res. Mem. Cogn.* **4**, 265–284 (2015).
153. Clegg, B. A. et al. Effective mitigation of anchoring bias, projection bias, and representativeness bias from serious game-based training. *Procedia Manuf.* **3**, 1558–1565 (2015).
154. Rassin, E. Blindness to alternative scenarios in evidence evaluation. *J. Investig. Psychol. Offender Profil.* **7**, 153–163 (2010).
155. Wason, P. C. Reasoning about a rule. *Q. J. Exp. Psychol.* **20**, 273–281 (1968).
156. Riggio, H. R. & Garcia, A. L. The power of situations: Jonestown and the fundamental attribution error. *Teach. Psychol.* **36**, 108–112 (2009).
157. Emory, B. & Luo, T. Metacognitive training and online community college students' learning calibration and performance. *Community Coll. J. Res. Pract.* **46**, 240–256 (2022).
158. Morrison, J. R., Bol, L., Ross, S. M. & Watson, G. S. Paraphrasing and prediction with self-explanation as generative strategies for learning science principles in a simulation. *Educ. Technol. Res. Dev.* **63**, 861–882 (2015).
159. Fitterman-Harris, H. F. & Vander Wal, J. S. Weight bias reduction among first-year medical students: a quasi-randomized, controlled trial. *Clin. Obes.* **11**, e12479 (2021).
160. Lewis, R. J., Cash, T. F., Jacobi, L. & Bubb-Lewis, C. Prejudice toward fat people: the development and validation of the antifat attitudes test. *Obes. Res.* **5**, 297–307 (1997).
161. Latner, J. D., O'Brien, K. S., Durso, L. E., Brinkman, L. A. & MacDonald, T. Weighing obesity stigma: the relative strength of different forms of bias. *Int. J. Obes.* **32**, 1145–1152 (2008).
162. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
163. Gagne, D. A. *Evaluation of an Obesity Stigma Intervention in Reducing Implicit and Explicit Weight Bias* (Saint Louis Univ., 2014).
164. Gutierrez, A. P. *Enhancing the Calibration Accuracy of Adult Learners: A Multifaceted Intervention* (Univ. Nevada, 2012).
165. Heijltjes, A., van Gog, T., Leppink, J. & Paas, F. Improving critical thinking: effects of dispositions and instructions oneconomics students' reasoning skills. *Learn. Instr.* **29**, 31–42 (2014).
166. Fong, G. T., Krantz, D. H. & Nisbett, R. E. The effects of statistical training on thinking about everyday problems. *Cogn. Psychol.* **18**, 253–292 (1986).
167. De Neys, W. & Glumicic, T. Conflict monitoring in dual process theories of thinking. *Cognition* **106**, 1248–1299 (2008).
168. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
169. Stanovich, K. E. in *In Two Minds: Dual Processes and Beyond* (ed. Evans, J.) Vol. 369, 55–88 (Oxford Univ. Press, 2009).
170. Evans, J. S. B. T. In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.* **7**, 454–459 (2003).
171. Huff, J. D. & Nietfeld, J. L. Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacogn. Learn.* **4**, 161–176 (2009).
172. Pronin, E. & Kugler, M. B. Valuing thoughts, ignoring behavior: the introspection illusion as a source of the bias blind spot. *J. Exp. Soc. Psychol.* **43**, 565–578 (2007).
173. Kolić-Vehovec, S., Pahljina-Reinić, R. & Rončević Zubković, B. Effects of collaboration and informing students about overconfidence on metacognitive judgment in conceptual learning. *Metacogn. Learn.* **17**, 87–116 (2022).
174. Schraw, G. A conceptual analysis of five measures of metacognitive monitoring. *Metacogn. Learn.* **4**, 33–45 (2009).
175. Cox, C. & Mouw, J. T. Disruption of the representativeness heuristic: can we be perturbed into using correct probabilistic reasoning? *Educ. Stud. Math.* **23**, 163–178 (1992).
176. Legaki, N. Z. & Assimakopoulos, V. F-LaurelXP: a gameful learning experience in forecasting. In *Proc. 2nd International GamiFIN Conference* Vol. 2186 (CEUR, 2018).
177. Morsanyi, K., Handley, S. J. & Serpell, S. Making heads or tails of probability: An experiment with random generators. *Br. J. Educ. Psychol.* **83**, 379–395 (2013).
178. Fox, C. R. & Levav, J. Partition-edit-count: naive extensional reasoning in judgment of conditional probability. *J. Exp. Psychol. Gen.* **133**, 626–642 (2004).
179. Green, D. R. *Probability Concepts in School Pupils Aged 11–16 Years* (Loughborough Univ, 1982).
180. Onal, I. G. & Kumkale, G. T. Effectiveness of source-monitoring training in reducing halo error and negativity bias in a performance appraisal setting. *Appl. Psychol.* **71**, 1635–1653 (2022).
181. Martell, R. F. & Evans, D. P. Source-monitoring training: toward reducing rater expectancy effects in behavioral measurement. *J. Appl. Psychol.* **90**, 956–963 (2005).
182. Ramdass, D. H. *Improving Fifth Grade Students' Mathematics Self-Efficacy Calibration and Performance through Self -Regulation Training* (City Univ. of New York, 2009).
183. Gertner, A., Zaromb, F., Schneider, R., Roberts, R. D. & Matthews, G. The assessment of biases in cognition: development and evaluation of an assessment instrument for the measurement of cognitive bias. *MITRE Technical Report MTR160163* https://www.mitre.org/news-insights/publication/assessment-biases-cognition (2016).
184. Rodríguez-Ferreiro, J., Vadillo, M. A. & Barberia, I. Debiasing causal inferences: over and beyond suboptimal sampling. *Teach. Psychol.* **50**, 230–236 (2023).
185. Matute, H., Yarritu, I. & Vadillo, M. A. Illusions of causality at the heart of pseudoscience. *Br. J. Psychol.* **102**, 392–405 (2011).
186. Rowland, K. *Counselor Attributional Bias* (Ball State Univ., 1981).
187. Storms, M. D. Videotape and the attribution process: reversing actors' and observers' points of view. *J. Pers. Soc. Psychol.* **27**, 165–175 (1973).
188. Morton, T. A., Haslam, S. A., Postmes, T. & Ryan, M. K. We value what values us: the appeal of identity-affirming science. *Polit. Psychol.* **27**, 823–838 (2006).
189. Scopelliti, I., Min, H. L., McCormick, E., Kassam, K. S. & Morewedge, C. K. Individual differences in correspondence bias: measurement, consequences, and correction of biased interpersonal attributions. *Manag. Sci.* **64**, 1879–1910 (2018).
190. Cook, M. B. & Smallman, H. S. Human factors of the confirmation bias in intelligence analysis: decision support from graphical evidence landscapes. *Hum. Factors* **50**, 745–754 (2008).
191. Silver, E. M. *Cognitive Style as a Moderator Variable in Rater Training to Reduce Illusory Halo* (Kansas State Univ., 1986).
192. Silver, E. M. *Halo Bias, Implicit Personality Theory, and Cognitive Complexity: Possible Relationships and Implications for Improving the Psychometric Quality of Ratings* (Kansas State Univ., 1982).
193. Swift, J. A. et al. Are anti-stigma films a useful strategy for reducing weight bias among trainee healthcare professionals? Results of a pilot randomized control trial. *Obes. Facts* **6**, 91–102 (2013).
194. Allison, D. B., Basile, V. C. & Yuker, H. E. The measurement of attitudes toward and beliefs about obese persons. *Int. J. Eat. Disord.* **10**, 599–607 (1991).
195. Crandall, C. S. Prejudice against fat people: ideology and self-interest. *J. Pers. Soc. Psychol.* **66**, 882–894 (1994).

196. Testa, I. et al. Effects of instruction on students' overconfidence in introductory quantum mechanics. *Phys. Rev. Phys. Educ. Res.* **16**, 010143 (2020).
197. Boone, W. J., Staver, J. R. & Yale, M. S. *Rasch Analysis in the Human Sciences* (Springer, 2016).
198. Van Bockstaele, B., van der Molen, M. J., van Nieuwenhuijzen, M. & Salemink, E. Modification of hostile attribution bias reduces self-reported reactive aggressive behavior in adolescents. *J. Exp. Child Psychol.* **194**, 104811 (2020).
199. Houtkamp, E. O., van der Molen, M. J., de Voogd, E. L., Salemink, E. & Klein, A. M. The relation between social anxiety and biased interpretations in adolescents with mild intellectual disabilities. *Res. Dev. Disabil.* **67**, 94–98 (2017).
200. Snyder, M. & Swann, W. B. Hypothesis-testing processes in social interaction. *J. Pers. Soc. Psychol.* **36**, 1202–1212 (1978).
201. West, R. F., Toplak, M. E. & Stanovich, K. E. Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* **100**, 930–941 (2008).
202. Stanovich, K. E. *Rationality and the Reflective Mind* (Oxford Univ. Press, 2011).
203. Stanovich, K. E. & West, R. F. Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* **23**, 645–665 (2000).
204. Veinott, E. S. et al. The effect of camera perspective and session duration on training decision making in a serious video game. In *International Games Innovation Conference* (IEEE, 2013).
205. Whitaker, E. et al. The effectiveness of intelligent tutoring on training in a video game: an experiment in student modeling with worked-out examples for serious games. In *International Games Innovation Conference* (IEEE, 2013).

Acknowledgements

The authors received no specific funding for this work.

Author contributions

Conceptualization by G.S., M.N., P.P., C.N., G.B., A.S., E.G. and P.S. Methodology by G.S., M.N., J.G., S.G., V.G., F.R., M.M., S.K.Y. and X.Z. Analysis by G.S. and M.N. Writing—original draft preparation by G.S. and M.N. Writing—review and editing by G.S., M.N., J.T., P.P., C.N., G.B., V.G., F.R., M.M., S.K.Y. and X.Z. Supervision by M.N., P.P., C.N., G.B. and J.T.

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/s41562-025-02253-y.

Correspondence and requests for materials should be addressed to Michael Noetel.

Peer review information *Nature Human Behaviour* thanks Ioana Cristea and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Michael Noetel

Last updated by author(s): 8 April 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We searched the following databases: Scopus, PsycINFO, ERIC, Web of Science, and Proquest Dissertations and Theses. We used EndNote X9 to remove duplicates and RobotSearch (v0.7) to filter abstracts for randomised trials. We used Covidence (www.covidence.org) for title and abstract screening and full-text screening. We used Google Sheets for data extraction. References for RobotSearch and Covidence below: Marshall IJ, Kuiper J, Banner E, Wallace BC. "Automating Biomedical Evidence Synthesis: RobotReviewer." Proceedings of the Conference of the Association for Computational Linguistics (ACL). 2017 (July): 7–12. DOI: 10.5281/zenodo.6855718 Veritas Health Innovation. Covidence Systematic Review Software. (Melbourne, Australia, 2023). Available at www.covidence.org.
Data analysis	We used R (version 4.4.1) on RStudio (version 2024.04.2+764) and the following R packages to run the analysis: tidyverse (ver. 2.0.0), janitor (ver. 2.2.0), googlesheets4 (ver. 1.1.1), dplyr (ver. 1.1.4), compute.es (ver. 0.2-5), metaSEM (ver. 1.4.0), lavaan (ver. 0.6-18), gt (ver. 0.11.0), metadat (ver. 1.2-0), meta (ver. 7.0-0), metafor (ver. 4.6-0), ggplot2 (ver. 3.5.1), renv (ver. 1.0.7), esc (ver. 0.5.1), stringr (ver. 1.5.1), metacart (ver. 2.0-3), readxl (ver. 1.4.3), robvis (ver. 0.3.0), PRISMA2020 (ver.1.1.1), PublicationBias (ver. 2.4.0), weightr (ver. 2.0.2), svglite (ver.2.1.3), MetaUtility (ver. 2.1.2), rsvg (ver. 2.6.0), knitr (ver. 1.48) The data, R codes, and instructions (on a README file) are available on https://osf.io/7x5z6.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We collected our data from the following databases: Scopus, PsycINFO, ERIC, Web of Science, and Proquest Dissertations and Theses. The complete datasets are available in the Data folder (clean_data.RDS, moderators.RDS, models_one_rob_at_a_time.RDS, prisma_data.csv, and Risk_of_Bias.RDS) at https://osf.io/7x5z6/files/osfstorage

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We have moderated by gender where available, using the original study author's description of sex versus gender. We have presented the results for this moderation analysis in Supplementary Note 1 as per recommendations from a peer reviewer.
Reporting on race, ethnicity, or other socially relevant groupings	Information on race, ethnicity, or other socially relevant grouping was seldom available so was not assessed via moderation analyses.
Population characteristics	Population characteristics were described in Supplementary Table 2 Characteristics of Included Studies
Recruitment	Participants were not recruited, as this was a systematic review. Studies were identified by the search strategy and selection criteria identified in the Methods section.
Ethics oversight	Ethics approval not required by our institution for meta-analyses

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Systematic review and meta-analysis with majority quantitative data in the meta-analysis, whereas qualitative data was used to synthesise findings of the studies where data was insufficient to run a meta-analyses (i.e., models with less than three studies to compare).
Research sample	Studies of randomised controlled trials in educational settings. Studies were sought from the following databases: Scopus, PsycINFO, ERIC, Web of Science, and Proquest Dissertations and Theses. Studies were included if they fit the predetermined inclusion criteria.
Sampling strategy	Our study is a systematic review and meta-analysis that analysed existing published data rather than collecting primary data. Hence, we followed PRISMA guidelines for systematic reviews, conducting a comprehensive literature search using databases with predefined inclusion and exclusion criteria (available on the OSF pre-registration page https://osf.io/xrm4g). Sample size was determined by the available studies meeting our eligibility criteria after screening. While there is no strict minimum sample size for a meta-analysis, generally aiming for five to ten studies minimum is sufficient (Turner, R. M., Bird, S. M., & Higgins, J. P. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PloS one, 8(3), e59202. https://doi.org/10.1371/journal.pone.0059202).
Data collection	After conducting comprehensive data search from the databases, we used EndNote X9 to remove duplicate studies and used RobotSearch to first filter the non-RCT studies. We then used Covidence to screen the title and abstract of the remaining studies (done independently and in duplicate across two reviewers). We sought full-text articles that passed title and abstract screening. The full-text screen was also done independently and in duplicate between two reviewers. In both screening stages, any conflicts were resolved through discussion or a consultation with a third reviewer. One reviewer then developed a data extraction form on Google Sheets, which was piloted and revised alongside another reviewer. The form extracted details regarding participants, outcome information, intervention, comparison, and any metric that could be used to calculate an effect size (e.g., means, standard

deviations, confidence intervals, p-values). When the included studies did not present exact values but presented figured, we extracted data from figures using WebPlotDigitizer (v4, now available at www.automeris.io). As the study was a systematic review, the researchers were not blinded to the experimental condition or the study hypothesis to be able to conduct the screening sufficiently.

Timing

The initial data search was finished on 4th August 2022. Additional searches using the references from included studies were conducted on 6th April 2023.

Data exclusions

Full reasons for data exclusions are outlined in the PRISMA flow chart. Our initial search yielded 83,277 studies in total. After removing duplicate records (49,807) and records marked as ineligible by RobotSearch (13,834), we started screening 19,636 studies. Based on a predetermined exclusion criteria (available on the OSF pre-registration page https://osf.io/xrm4g), we excluded 18,704 studies from title/abstract screening. From 928 studies in the full-text screening, we could not find the full-text papers for 38 studies. From 890 studies in full-text screening, 263 had no debiasing outcome, 250 were not RCTs, 108 had clinical outcomes, 83 were not done in educational setting, 66 studies had participants that were not students, 46 had no original data, 19 studies had no explicit discussion of cognitive biases in the intervention, and 16 studies had interventions that were less than 10 minutes long. From the remaining included studies, we ran citation search that yielded 2,818 findings. Of these, 158 were eligible for full-text screening, but 1 study could not be obtained. From 157 studies that were screened for full-text, 20 had no debiasing outcome, 51 were not RCT, 1 had clinical outcome, 19 were not done in educational settings, 15 studies did not have students as participants, 23 studies had no original data, 9 studies had no explicit discussion of cognitive biases, and 9 studies had interventions that were less than 10 minutes. This further selection yielded 53 studies across 49 papers and a pooled sample size of 10,941 participants.

Non-participation

N/A - No participants dropped out or declined participation as this was a systematic review. We had no direct involvement with participants of the included studies. We assessed risk of bias from non-participation that was reported in each included study.

Randomization

N/A - Randomisation was not relevant to our study as this was a systematic review. We assessed risk of bias from the randomisation procedure stated in the included studies.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

N/A

Novel plant genotypes

N/A

Authentication

N/A