Journal of Applied Research in Memory and Cognition

Learning to Call Bullsh*t via Induction: Categorization Training Improves Critical Thinking Performance

Benjamin A. Motz, Emily R. Fyfe, and Taylor P. Guba

Online First Publication, August 4, 2022. http://dx.doi.org/10.1037/mac0000053

Journal of Applied Research in Memory and Cognition

ISSN: 2211-3681

https://doi.org/10.1037/mac0000053

EMPIRICAL ARTICLE

Learning to Call Bullsh*t via Induction: Categorization Training Improves Critical Thinking Performance

Benjamin A. Motz[1, 2], Emily R. Fyfe[1], and Taylor P. Guba[1]

[1] Department of Psychological and Brain Sciences and Cognitive Science Program, Indiana University Bloomington, United States

[2] eLearning Research and Practice Lab, Indiana University Bloomington, United States

Participants performed a categorization training task, where each trial presented a claim based on an observation, and participants marked which fallacy or bias, if any, applied to the claim. In two studies, we measure the effect of this training task on critical thinking, measured using an open-ended critical thinking assessment. In Study 1, we pilot these materials in an online college course and observe credible improvements in critical thinking performance. In Study 2, we conduct a randomized controlled experiment and observe credible improvements in critical thinking relative to no training, and relative to comparable learning activities focused on conventional curricular content. We infer that the categorization training task facilitated inductive learning of patterns of biased and flawed reasoning, which improved participants' ability to identify such patterns in a delayed critical thinking assessment. Such categorization training shows promise as an effective and practical method for improving learners' resistance to online misinformation.

General Audience Summary

Helping students learn to think critically is an important priority across education levels, but there is currently no clear consensus on how to teach critical thinking. Many proposed teaching strategies are very elaborate and involve extensive discussion and analysis, which makes it difficult for teachers to incorporate critical thinking instruction into their existing curriculum. In our current research, we

---

Benjamin A. Motz https://orcid.org/0000-0002-0379-2184

Emily R. Fyfe https://orcid.org/0000-0002-5420-360X

Taylor P. Guba https://orcid.org/0000-0002-6131-2417

Taylor P. Guba is now at University of Delaware.

Study materials, data, and analysis scripts are publicly available at https://osf.io/vcgzf/. This research was made possible by support from the Reboot Foundation, and portions of this research have been reported on the Reboot Foundation's blog, but the funder had no role in research design nor analysis and there are no conflicts of interest related to the conduct of this study. The authors appreciate the contributions of Alexandria Hartburg, Ulrich Boser, Michelle Sublette, Emily Cheshire Brown, Helen Lee Bouygues, and Steve Sloman.

The current research was conducted under human subjects research protocols approved by the Indiana University institutional review board (IU IRB).

Study 1 was conducted under protocol number 1505778051, which the IU IRB deemed to be Exempt. The authors did not administer a formal informed consent process. This is because the present study did not deviate from routine instruction, all students received the same learning activities, and the responses and performance data were all deidentified prior to analysis.

Study 2 was conducted under protocol number 2011544331, which the IU IRB also deemed to be Exempt. The authors did not administer a formal informed consent process for this study, which used volunteer research participants recruited on MTurk. In the study's listing on MTurk, we did include an IRB-approved study information sheet (SIS), which explained participants' rights as research subjects, provided contact information for questions or concerns, and explained the study. The SIS concluded with the statement, "Thank you for agreeing to participate in our research. Before you begin, please note that the data you provide may be collected and used by Amazon as per its privacy agreement. Additionally, this research is for participants over the age of 18*; if you are under the age of 18, please do not complete this survey."

Benjamin A. Motz played a lead role in conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, validation, visualization, and writing of original draft; a supporting role in supervision; and an equal role in data curation and project administration. Emily R. Fyfe played a lead role in supervision and manuscript review and editing; a supporting role in conceptualization, investigation, and validation; and equal role in data curation, resources, methodology and project administration. Taylor P. Guba played a lead role in measurement and coding; a supporting role in manuscript review and editing; and an equal role in data curation, formal analysis and methodology.

The data are available at https://osf.io/vcgzf.

The experimental materials are available at https://osf.io/vcgzf.

The preregistered design and analysis plan are accessible at https://osf.io/9j583.

Correspondence concerning this article should be addressed to Benjamin A. Motz, Department of Psychological and Brain Sciences and Cognitive Science Program, Indiana University Bloomington, 1101 East 10th Street, Bloomington, IN 47405, United States. Email: bmotz@indiana.edu

explore whether a more focused approach might improve students' critical thinking performance. Specifically, we examine an inductive approach to critical thinking instruction. Inductive learning, commonly known as learning by example, happens when a person learns a general pattern from multiple examples of the pattern. In this case, we ask whether participants can learn to identify illogical or biased claims, a common measure of critical thinking ability, by presenting them with multiple examples of different kinds of fallacies and biases. We presented participants with sets of example scenarios in which an individual makes a claim based on an observation, and participants categorized which fallacy or bias, if any, the individual in the scenario was committing. In two studies, we find that this critical thinking categorization practice improves performance on a delayed open-ended critical thinking assessment. These exciting results show that critical thinking skills can be taught using a well established, specific, psychologically grounded method, and these results come at a critical juncture in our society. In a time of fake news, misinformation and a lack of critical thinking skills among the populace can have dire consequences. We believe our current results are of broad interest to educators and applied cognitive psychologists, and show promise for improving peoples' defenses against misinformation.

Keywords: critical thinking, inductive learning, causal reasoning, categorization, misinformation

Supplemental materials: https://doi.org/10.1037/mac0000053.supp

The importance of critical thinking is a consensus view across education. Both postsecondary institutions (Association of American Colleges and Universities, 2005) and now the current generation of K-12 common standards place heavy emphasis, explicitly, on developing higher order thinking and analytical skills (Porter et al., 2011). But underneath this consensus, there are mixed views on how to improve critical thinking, with similarly mixed outcomes (Abrami et al., 2015; Niu et al., 2013; Tiruneh et al., 2014). In this study, we examine a specific operationalization of critical thinking, and evaluate a corresponding training regimen, using an inductive learning paradigm, to improve critical thinking performance.

What is critical thinking? Although controversy persists (Johnson & Hamby, 2015), an increasingly common view defines critical thinking as the ability to detect and identify fallacies (such as logical errors or biases) in peoples' claims (Bergstrom & West, 2021; Lawson et al., 2015; Schmaltz et al., 2017; Schmaltz & Lilienfeld, 2014; Sternberg, 1999). Essentially, when presented with a statement about the world, someone who is skilled at critical thinking should be able to discriminate whether the claim is justified based on the information provided. Further, when claims are unjustified, they should be able to describe the specific issues that call into question their validity. This view of critical thinking lacks the scope and grandeur of broader definitions (which may include elements of creativity, self-efficacy, problem-solving, and attitudinal dispositions, such as skepticism; Ennis, 1987; Facione, 1990; Halpern, 2013), but has the converse benefits of discriminant validity and specificity. Moreover, this proposed view of critical thinking has a clear operationalization in a real-world task of increasing importance.

Individuals urgently need to be able to counteract growing volumes of misinformation (false and misleading information) and disinformation (*intentionally* misleading information; Lazer et al., 2018; Machete & Turpin, 2020). The prevalence of misleading online news and information presents real threats to democracy, public discourse, and public health (Lewandowsky et al., 2017). For example, online misinformation (that has been thoroughly debunked) about the side effects from vaccines has resulted in large-scale vaccine hesitancy, creating secondary public health crises caused not by illness, but by a contagion of fallacious claims (Horton, 2020; Poland & Spier, 2010). These claims, and our susceptibility to them, follow well-worn psychological paths, such as our human tendency to view correlational findings, or even just rare coincidences, as providing evidence of causal mechanisms. When exposed to these types of claims, one who is skilled at critical thinking should be able to recognize that mere associations do not warrant claims about one factor causing another, and to challenge the validity of claims.

Individuals also urgently need to detect and correct biased claims. Another example of our susceptibility is *confirmation bias*, the tendency to privilege evidence that supports one's views while ignoring evidence that contradicts them, a common feature of ideological extremism (Lilienfeld et al., 2009; Tavris & Aronson, 2020). In social media, the construction of polarized echo chambers is attributed to confirmation bias, as users promote posts and articles that only reflect one side of a debate (Garimella et al., 2018). Users who share one-sided posts do not believe themselves to be spreading false information, they simply neglect evidence to the contrary (Pennycook et al., 2021; Pennycook & Rand, 2019, 2021). Again, when exposed to such a scenario, one who is skilled at critical thinking should be able to recognize the single sidedness of the arguments, and to challenge the validity of claims.

Increasingly, many people look to our education system to provide students with the critical thinking skills necessary to mitigate their susceptibility to biased and misleading claims. However, developing a learner's ability to think critically—to identify fallacies such as confirmation bias or a causal inference from rare coincidences—requires dedicated effort. Many students simply show no measurable improvement in general critical thinking ability during college (Arum & Roksa, 2011). Moreover, improving a student's understanding of domain knowledge (e.g., neuroscience) does not necessarily improve a student's ability to identify fallacies within that domain (e.g., neuromyths; Im et al., 2018). Incidentally, most college-age learners maintain views that have no evidentiary basis whatsoever, such as the existence of people who can move objects with their minds (70% endorsement), or that certain objects such as rabbits' feet genuinely bring good luck (69% endorsement; Lobato et al., 2014), providing indirect but complementary evidence of low-critical thinking ability (Lantian et al., 2021). If education systems are to improve students' critical thinking skills, a more focused educational strategy is necessary.

Reviews and meta-analyses of instructional strategies for improving critical thinking have examined a wide range of published approaches, with varying degrees of efficacy (Abrami et al., 2015; Niu et al., 2013; Tiruneh et al., 2014). Across these reviews, three general patterns emerge: effective instructional strategies (a) tend to be embedded within the discipline, (b) tend to expose students to authentic critical thinking problems in real contexts, and (c) tend to occur over extended periods of time. Situating or contextualizing instruction within a discipline, rather than presenting instruction in abstraction, is a well-known feature of effective educational programs in critical thinking (Willingham, 2008). And the importance of exposure to authentic problems over time mirrors recent findings that emphasize the importance of deliberate practice (Heijltjes et al., 2014; van Gelder, 2005) for improving critical thinking skills.

Improving Critical Thinking via Category Induction

These features of effective critical thinking instruction (contextualized materials, deliberate practice, and repeated exposure) bear striking resemblance to principles of inductive learning. Inductive learning, broadly defined as learning by example, involves extracting a general rule, pattern, or schema based on key structural features (and not the variable surface features) present across multiple examples. Like critical thinking instruction, inductive learning is improved by exposure to more concrete examples (Fong et al., 1986; Fyfe et al., 2014), by spacing exposure over time (Carvalho & Goldstone, 2015; Wahlheim et al., 2011), and by practice testing (Izawa, 1967; Tulving, 1967). And while inductive learning might commonly be associated with contrived laboratory-based experiments, the general framework has constructively been applied to more naturalistic learning tasks, such as learning to discriminate artists' painting styles (Kang & Pashler, 2012; Kornell & Bjork, 2008), learning to medically diagnose skin disorders (Brooks et al., 1991), learning to sort natural rock specimens with complex category structures (Nosofsky et al., 2018), and learning to classify concepts in real education settings (Carvalho et al., 2016; Day et al., 2015; Dunlosky et al., 2013; Quilici & Mayer, 2002). For all these tasks, the inductive approach is to present an example of a problem, have the learner respond to the problem, provide feedback, then repeat.

But by contrast, learning activities currently proposed in the critical thinking literature are generally much more elaborate. For example, Renaud and Murray's (2008) critical thinking intervention involved essay questions that required multiple pages of written text response. Zohar and Tamir (1993) infused critical thinking scenarios into a biology curriculum, where activities extended for hours, sometimes covering multiple class periods with lively student debates. Self and Self (2017), who also propose exposing students to regular practical critical thinking scenarios, advocate that this exposure should methodically cover each of the 10 steps of Kolb's experiential learning process (Kolb, 2015). While such training regimens are impressive in their detail and scope, they also create practical barriers for implementation, such as heavy incursion into existing curricular activities and steep grading demands. It is also unclear if these laborious efforts cause better critical thinking outcomes than a more basic inductive approach.

If we define critical thinking as the ability to detect and identify fallacies, an inductive approach should simply present examples of potentially fallacious claims, situated in a course's content domain and distributed across the semester, and invite students to identify which fallacy (if any, among a set of fallacies) are evident in the claim. These might be presented as multiple-choice problems, where students categorize claims according to basic fallacies and receive prompt feedback on their responses.

This approach is strong because it relies on categorization, which is fundamental to human cognition (Harnad, 2017), and category learning builds structure in the organization of semantic knowledge (Goldstone et al., 2017). Showing real-world examples of a category improves one's ability to correctly identify new (unstudied) examples of the category, even when the category is very high level, such as decision-making principles (Rawson et al., 2015), intuitions about statistics (Fong et al., 1986), or the deep meaning of proverbs (Tullis & Goldstone, 2016). Along these lines, one might expect improvement in our ability to identify critical thinking fallacies when people are shown examples and tasked to categorize the fallacy present. In effect, learners are inducing mental models of the category structure of these fallacies (Cheng et al., 1986; Goldstone et al., 2013; Nosofsky, 1986; Smith et al., 1992). Research suggests that knowledge of these categories, in particular, has general utility: The specific ability to detect and reject fallacious claims has been shown to predict academic performance, resistance to misconceptions, and the ability to weigh and interpret evidence (Bensley et al., 2014; Bensley & Lilienfeld, 2017; Kowalski & Taylor, 2004; McCutcheon et al., 1992).

Preliminary evidence of these kinds of improvements has already been observed in a critical thinking study using a multiple-choice format for induction. Hitchcock (2004) saw better critical thinking performance among students who performed computer-guided practice with multiple-choice questions than students who engaged in small group discussion, but this study lacked random assignment. As Hitchcock concluded, "more work needs to be done," to build confidence that critical thinking gains can be achieved by this approach.

Researchers have recently begun to build this evidence base, using a tactic presently referred to as "prebunking" (Basol et al., 2021; Lewandowsky & van der Linden, 2021). The prebunking approach involves inductive exposure to small innocuous doses of misinformation, which progressively build an individual's ability to respond appropriately to misleading claims in the wild. Like the idea that one might become inoculated from persuasion (Papageorgis & McGuire, 1961), prebunking has shown promise in building an individual's cognitive defenses against common forms of misinformation.

The Current Research

Here, we explore whether an inductive training regimen, where learners categorize example claims according to different types of fallacies and biases, improves general critical thinking performance. Considering that critical thinking instruction should be situated within disciplinary content, we developed materials for implementation in a college Introductory Psychology course (although this general approach could easily be adapted to other courses). Introductory Psychology is one of the most popular college courses, taken by roughly 1.5 million students each year in the United States (Gurung et al., 2016), with programs placing heavy emphasis on the importance of critical thinking (Homa et al., 2013), and where critical thinking outcomes similarly include the ability to recognize flaws in explanations (Halonen et al., 2013).

In the first study, we pilot a set of critical thinking practice problems in a live Introductory Psychology class. In the second study, we randomly assign online research participants to receive the same critical thinking practice problems, to receive non-critical thinking Introductory Psychology practice questions, or to a no-training control condition. In both studies, we measure critical thinking performance pre- and post-training using the Psychology Critical Thinking Exam (PCTE, updated version; Lawson et al., 2015). The PCTE is an open-ended assessment where each item describes a fallacious claim drawn from some evidence, and examinees are asked to describe the problem (if one exists) with the conclusion. The open-ended nature of the PCTE provides a reasonably authentic measure of whether participants can accurately identify and describe these fallacies from memory, in the absence of any cues.

Study 1 (Classroom Pilot)

Study 1 was an exploratory pilot study, designed to test for possible benefits of inductive learning using multiple-choice questions about critical thinking scenarios in a live education setting. All enrolled students are exposed to critical thinking categorization practice, and we measure improvements in critical thinking performance by comparing scores on pretest (at the start of the semester) and posttest (at the end of the semester). All materials are publicly available at https://osf.io/vcgzf/ (Motz et al., 2022).

Method

Participants

Students enrolled in online sections of Introductory Psychology at Indiana University Bloomington during the Spring 2017 semester constituted a convenience sample for examining the effect of critical thinking training. Students who withdrew from the course are not included, resulting in 504 students enrolled. Aggregate institutional records from these sections indicate that 50.2% are female, with an average age of 19.9 years at the end of the semester. However, 122 students did not access either the pretest or posttest assessment (which were ungraded; any response earned a small amount of completion credit toward the course score), and another 22 students accessed neither of them, so these students are also excluded, leaving 360 students who all submitted a pretest and a posttest.

Materials

The PCTE (updated version; Lawson et al., 2015) was used to assess critical thinking. Each item in the PCTE is a short description of a fictional scenario where someone makes a fallacious or biased inference based on some observation, and examinees are instructed to “state whether or not there is a problem with the person’s conclusions and explain the problem.” There are seven such problems in the full PCTE, which we label here as follows: Random Chance, Lack of Control, Correlation is not Causation, Overgeneralization, Experimenter Bias, Confirmation Bias, and Oversimplification. However, in the present study, we excluded items related to Oversimplification because these were in direct conflict with the students’ textbook (Schacter et al., 2016), which explicitly promoted the importance of parsimony, stating that the “simplest theory that explains all the evidence is the best one” (p. 41). Moreover, the original authors of the PCTE found group differences specifically on the Oversimplification items and suggested that specialized training may be necessary for students to understand this principle in particular. So in the present study, we use a reduced form of the PCTE, including only the first six fallacies, assessed across 12 items, with two items per fallacy (see Table 1).

We developed four problem sets (training materials) for students to practice categorizing scenarios according to these fallacies. Each problem set contained seven multiple-choice questions, where each question included a scenario where someone drew a potentially fallacious conclusion from their own observations, and the options corresponded to the six fallacies in the PCTE, plus a seventh, where there was no problem with the claim. The scenarios in these multiple-choice practice tests were designed to be related to the content areas of the four Introductory Psychology units in which they were embedded: Neuroscience, Sensation and Perception, Memory, and Learning. In these practice exercises, we carefully controlled the properties of the individuals described in each scenario. Half the items described a male and the other half described a female (randomly assigned across scenarios), and the names of these individuals were designed to be representative of the U.S. population (forenames drawn from U.S. Social Security Administration records, and surnames drawn from U.S. Census records). Because it was not known whether the individuals in these training exercises should be presented as authoritative researchers or as lay public (the PCTE includes a mix of both), we randomly assigned students to receive either of two versions of the training scenarios throughout the semester, one describing inferences seemingly drawn by experts (e.g., Professor Denise Crane), and the other where the person drawing inferences was described by a first name only, with no implication of expertise (e.g., Denise). All training materials are available at https://osf.io/cg7k6/.

Procedure

Pretests and posttests were created by splitting the PCTE into two forms with six items each (one item per fallacy), and students were randomly assigned either form for pretest, and the other was used for posttest (as in Lawson et al., 2015). The pretest was administered in

Table 1
Fallacies and Biases Included in Training and Tests

Fallacy	Description
Confirmation Bias	Inferences based only on positive evidence and ignoring disconfirming evidence
Correlation is not Causation	Inferring a causal relationship on the basis of correlational data
Experimenter Bias	Inferences based on questions that were biased, loaded, or leading
Lack of Control	Inferring that improvement was due to an experimental intervention without comparison to a control group
Overgeneralization	Inferring that findings generalize to a larger group based on a biased sample
Random Chance	Inferring systematic relationships from chance occurrences

the first week of the semester, and the posttest was administered in the final (16th) week of the semester. These were ungraded, but students received a small amount of course credit for completion.

Training. Students received direct instruction on critical thinking during the third week of the semester (in a content unit about research methods). This included a statement on the importance of critical thinking, having students read definitions and descriptions of each of the six fallacies included in the PCTE, and a short graded multiple-choice practice test (which was presented on the same page as the definitions) that students could attempt as many times as they wanted prior to the deadline. The practice test included six scenarios (similar to the PCTE items) illustrating each of the six fallacies, as well as a seventh item where there was nothing wrong with the inference in the scenario, and students were instructed to identify the problem (if any) in multiple-choice format (with 7 labeled options: Random Chance, Lack of Control, Correlation is not Causation, Overgeneralization, Experimenter bias, Confirmation bias, and there is no fallacy in this conclusion).

Then, in the subsequent four units of the course (Neuroscience, Sensation & Perception, Memory, and Learning), training activities were embedded amidst typical online Introductory Psychology lessons and content. Each training activity involved the completion of a graded multiple-choice problem set, as described above (see the Materials section). The order of items within each problem set, and the order of response options, was randomized for each individual. There were roughly 2 weeks between the administration of each problem set, which could be completed multiple times prior to the due date, allowing students to maximize their scores.

Data Analysis

Response Scoring. Items on the PCTE (Lawson et al., 2015), which were used for the pretest and posttest, are open-ended questions which ask participants to explain the problem in a scenario. We manually scored participants' responses using the original PCTE scoring rubric, which was modeled on a scale used by Williams et al. (2003): 0 (*no problem identified*), 1 (*a problem recognized but misidentified*), 2 (*identified main problem, but also mentioned less relevant problems*), and 3 (*identified only the main problem*). We did not require participants to respond with specific words, terms, or labels, rather to describe the problem in their own words—for example, the response "although the weight loss supplement may have actually caused their weight-loss, there are other factors that could have," was scored as 3, the maximum score, for a "Lack of Control" item (PCTE item #3). Blank responses were scored as 0. In two cases when non-English responses were provided, we used Google Translate to translate the students' responses to English and then scored the responses accordingly. Responses were scored by a trained coder who was not able to see whether the responses were from a pretest or posttest.

Statistical Analysis. The goal of statistical analysis in this study is to examine differences in PCTE scores between pretest and posttest. Toward this goal, we carried out two different analytical approaches. Because PCTE scores are ordinal, we first analyzed the data using a hierarchical ordered probit model, adapted from the model described by Liddell and Kruschke (2018); but due to the complexity of this approach, we have opted to make it available as an online supplement (see Supplemental Materials). Additionally, we also implemented a more intuitive and coarse analysis, simply measuring individual test performance as the percent of items that participants got precisely correct (i.e., scored 3 out of 3 on a particular item), and subtracting pretest performance from posttest performance, creating a metric improvement score for each individual. Only this second analysis is described in the current article's main text, however, both yielded identical inferences.

We use Bayesian estimation for our statistical analyses. Bayesian estimation provides a statistical framework for directly estimating parameter values (such as the typical PCTE improvement score) and quantifying the uncertainty in these estimates. We estimated the tendency of participants' improvements pretest–posttest using a robust, hierarchical, Bayesian version of the *t* test, described in Kruschke (2013). The prior for group-level estimates of this improvement score had a mean of 0, and a standard deviation that was 100 times the empirical standard deviation of the sample, so that priors were vague and uninformed. Considering that our goal is to examine differences between pretest and posttest, and not to test the improbability of our results under a null hypothesis that such differences do not exist, Bayesian methods are particularly well suited to the present study.

We estimated parameter values in JAGS (Plummer, 2003), using the runjags package (Denwood, 2016) for R. The full model specifications are available at https://osf.io/vcgzf/. Our estimation runs used 500 steps of adaptation, 1,000 steps of burn-in, and then 250,000 Markov chain Monte Carlo (MCMC) steps across four chains, thinned to every fifth step for a total of 50,000 saved steps total. The effective sample size (ESS) was at least 20,000 for all model estimates, well above the 10,000 recommended by Kruschke (2014).

Results and Discussion

Prior to any critical thinking training, participants received a score of 3 (the maximum score) on 30.3% of the pretest questions, while 3.4% of responses received a score of 2, 40.0% received a score of 1, and 26.7% received a score of 0.

Following critical thinking training (at the end of the semester), participants received a score of 3 on 40.3% of the posttest questions, while 4.2% of responses received a score of 2, 32.2% received a score of 1, and 23.3% received a score of 0. The rarity of instances of a score of 2 (identified main problem, but also mentioned less relevant problems) is likely because PCTE instructions explicitly referred to "the problem" in singular form ("state whether or not there is a problem ... and explain the problem").

Lawson et al. (2015) did not report details of the distribution of PCTE scores, and instead reported the average cumulative score across all items for different groups of students in their study, which ranged from 35.5% of points possible (senior art majors) to 77.5% (senior psychology majors who recently received critical thinking instruction). If we similarly aggregate the current results (maximum score of 18; students can earn up to 3 points on each of the 6 items), we observe cumulative percent scores of 45.8% at pretest and 53.8% at posttest for students in the training group, roughly in line with Lawson's observations.

At pretest, participants were precisely correct (receiving a score of 3 out of 3) on 1.82 items (out of 6 items total, 30.3%), and at posttest, were precisely correct on 2.42 items (40.3%). We estimated the mean tendency of this improvement to be 0.596 more items (95% highest density interval [HDI]: 0.410–0.782), a credibly nonzero improvement in the number of correct responses on the PCTE following training (see Figure 1). In standardized terms, this corresponds to a Cohen's *d* of 0.38; however, this standard effect size is insensitive to the present study's within-subject design.

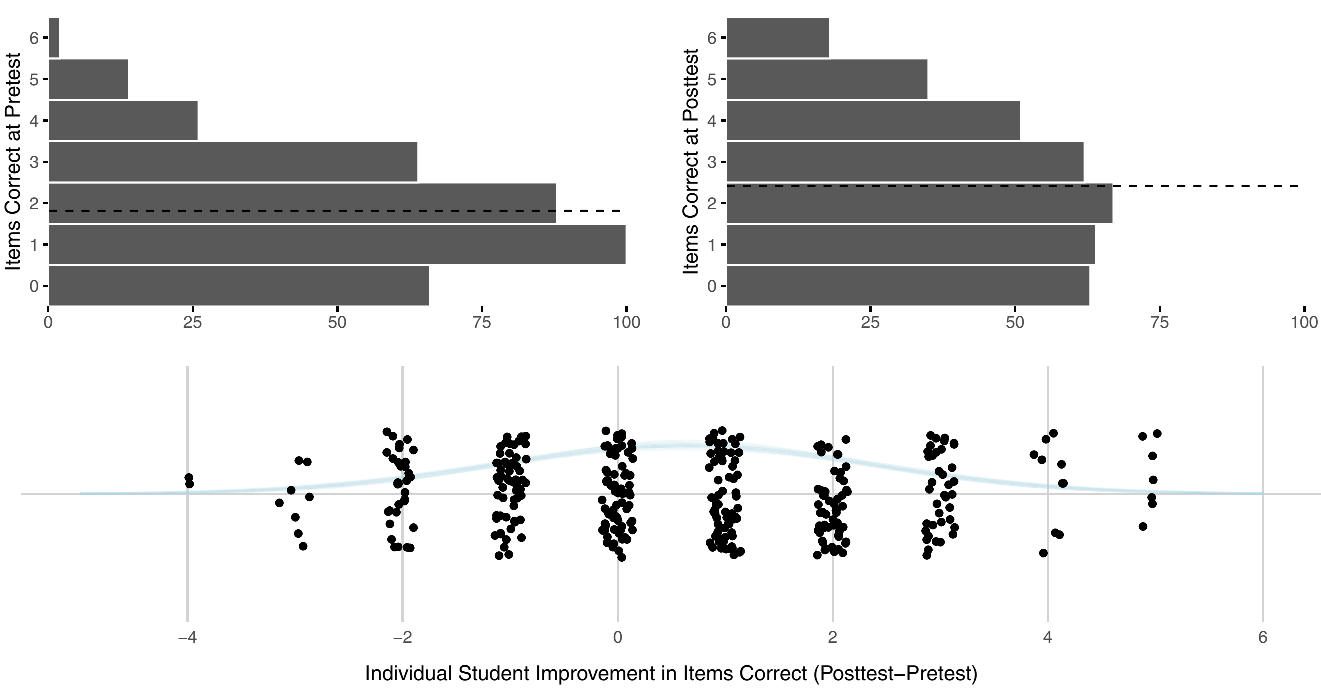
For participants whose scenarios included titled full names (e.g., Professor Denise Crane), these metric improvement scores were slightly smaller than scores from participants whose scenarios only included first names (e.g., Denise), but the difference between these groups was not credibly different from zero (difference mode = −0.172; 95% HDI: −0.56 to 0.17). Nevertheless, for our subsequent work (see Study 2 Procedure section, below) we opted to only use first names in the training materials.

There were modest differences in critical thinking performance across the different fallacies present in the PCTE, as shown in Figure 2. Students were least likely to respond correctly on items presenting the "Correlation is not Causation" fallacy, both at pretest and at posttest. This is not surprising, considering that people routinely consider patterns of correlation to provide evidence of causality (Sloman, 2005). But paradoxically, items presenting causal claims from uncontrolled experiments (labeled "Lack of Control") were most likely to be correct, both at pretest and at posttest, suggesting that respondents may have a reasonable grasp of some features of causality.

We investigated further, examining student performance on the training task, which had multiple-choice rather than open-ended items. Looking only at students' first attempt at each quiz, we calculated a confusion matrix, as shown in Figure 3. Again, "Correlation is not Causation" had the lowest percent correct. We find that the most common incorrect response, when presented with a "Correlation is not Causation" claim, was "Overgeneralization" (15% of responses). There is an intuitive explanation of this pattern, that participants may be skeptical of whether a correlational pattern would *generalize* to a new sample or setting, which is not inconsistent with concerns of whether a general causal relationship is present (S. Sloman, personal communication). This may also explain the proportionately high frequency of "Correlation is not Causation" responses on "Random Chance" items, also possibly stemming from concerns about a finding's generalizability. On the other side of the spectrum, items about "Experimenter Bias" had the highest percent correct, consistent with the observation that identification of this fallacy had the largest improvement from pretest to posttest (as shown in Figure 2).

Taking stock, we observe credible improvements in critical thinking scores from pretest to posttest among those students who were exposed to explicit, direct, critical thinking instruction and practice during their semester of enrollment in Introductory Psychology. However, these students were *also* exposed to routine instruction in Psychology (and in their other courses) so the improvement pre to post might also be attributable to other curricular content. Another confound is that the students received both direct instruction on critical thinking fallacies *and* received opportunities to practice categorizing claims according to these fallacies.

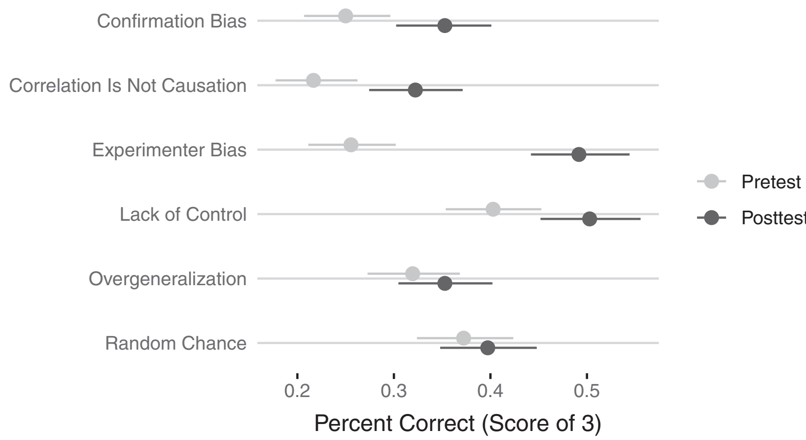
Figure 1
Performance Outcomes in Study 1



Note. Histograms on the top panels illustrate the number of items participants got precisely correct (score of 3 out of 3) on pretest (left panel) and posttest (right panel). Dashed lines indicate group means. The plot on the bottom panel shows individual student-level improvement from pretest to posttest. Each dot illustrates a single student's improvement, with minor horizontal jitter added for visualization (improvement scores are integers). Semitransparent blue lines show posterior estimates of the mean and standard deviation, modeled as a normal distribution at the group level, which is credibly centered above zero. See the online article for the color version of this figure.

Figure 2

Pretest and Posttest Performance for Each Fallacy in Study 1



Note. Horizontal bars indicate the percent of all participants in the training group who scored 3 out of 3 for that fallacy, on pretest and on posttest. Error bars show the 95% highest density interval (HDI) of each point estimate, as modeled with a simple binomial distribution.

Thus, this present study conflates the possible benefits of training with the benefits of categorization practice (this was also the case in Bensley & Spero, 2014), limiting our ability to explain the students' improvements by category induction. If we were to draw causal inferences from this group alone, we would certainly be guilty of the "Lack of Control" fallacy.

While the present results provide modest correlational evidence in support of the benefits of critical thinking training in practice, they lack the controls that would be characteristic of a randomized controlled experiment. Ideally, one would administer a pretest to a full cohort of students, then randomize these students to receive critical thinking training using categorization practice or no practice,

Figure 3

Confusion Matrix for Training Task in Study 1

Note. Horizontal rows indicate the correct response for each training item, and columns indicate participants' actual responses. Values (and color codings) indicate row percentages. Only includes data from the participant's first attempt at each training set. Light blue squares along the diagonal indicate a high proportion of correct responses. Proportionately frequent errors include "Correlation is not Causation" responses to "Random Chance" items, and "Overgeneralization" responses to "Correlation is not Causation" items, as discussed in the main text. See the online article for the color version of this figure.

and finally administer a posttest. However, given the present suggestive evidence of the benefits of training, it would be potentially unethical to withhold this training from enrolled students. Instead, we conduct a controlled experiment with online research volunteers using MTurk.

Study 2 (Randomized Controlled Experiment)

Study 2 was confirmatory, designed to examine the benefits of categorization practice with critical thinking scenarios in a randomized controlled experiment. Prior to data collection, we publicly registered all study methods at https://osf.io/9j583, which included fully specified analysis scripts based on simulated data. This study, as described below, contained no deviations from this preregistration, and all data are available at https://osf.io/vcgzf/.

Method

Participants

Participants were recruited on MTurk (https://www.mturk.com), with the requirements that they were based in the United States, were at least 18 years old, and had an MTurk approval rating above 50%. The study's sample size was determined by a financial constraint—we continued to recruit participants in batches until funding was exhausted. Initially, 601 potential participants accessed the screening session, and of these, 497 participants completed the screener and correctly entered their MTurk ID in a validation question. Sixty of those who completed the screener were excluded because they did not write coherent responses to simple questions, leaving 437 participants who passed the screener and were invited to complete the full study. Following these preregistered exclusions and attrition (either due to not completing a session or not initiating an invited session over the multiday online study), a total of 253 participants completed the full sequence. All results described hereafter are limited to these 253 participants with complete data sets. According to responses provided in the initial screening session, 68.4% of these participants had completed a bachelor's degree or higher, and 85.0% agreed or strongly agreed with the statement "I am skilled at critical thinking." We collected no other demographic data about participants.

Procedure

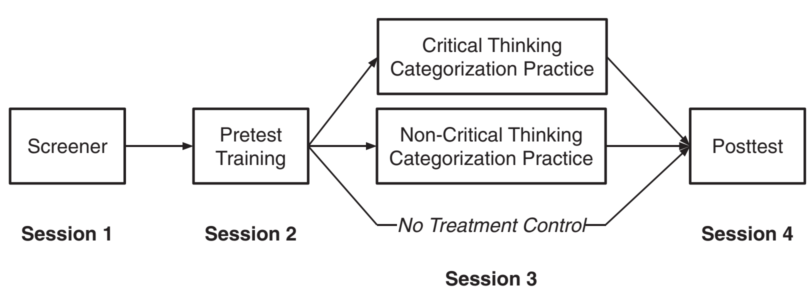
The present study had four sessions: (a) Screener, (2) Pretest and Training, (c) Intervention, and (4) Posttest (see Figure 4), all conducted online using Qualtrics (Provo, Utah). Each day we monitored who had satisfactorily completed each session and sent invitations for subsequent sessions accordingly (using MTurk's messaging system). In effect, the average time between participants' completion of each session was about 1.4 days.

Session 1: Screener. The purpose of the Screener was to identify human participants (not automated bots) who could follow instructions and provide coherent responses to open-ended questions. Participants were paid $0.20 this session, and the median completion time for participants in the final sample was 3.2 min.

Session 2: Pretest and Training. This session included a pretest assessment of participants' critical thinking, immediately followed by training about critical thinking fallacies. Just like the training group in Study 1, the pretest was half the items on the PCTE, which was counterbalanced: Half the participants got odd-numbered items, and the other half got even-numbered items for the pretest. We used the same reduced version of the PCTE (omitting items related to Oversimplification), with 12 items total (6-item pretest, 6-item posttest).

After the pretest, participants completed a training round that included two phases: a study phase and a test phase. During the study phase, participants were asked to read and study instructional text about the six fallacies described in Table 1. During the subsequent test phase, participants were asked to respond to seven multiple-choice items, corresponding to each of the six fallacies plus an additional item where there was no fallacy, all available at https://osf.io/n2hu8. Each item included a scenario similar to the PCTE items, and participants were asked to mark which fallacy (if any) the person committed. The training text from the study phase was included on the same page as each test item, so participants could refer back to definitions of these fallacies when selecting their

Figure 4
Overview of Study 2 Procedure



Note. Participants who satisfactorily completed each session were invited to the subsequent session. Participants were randomly assigned to one of three conditions for Session 3; those who were assigned to the no-treatment control condition were invited directly to Session 4 after completing Session 2. The average time between each session was 1.4 days. All participants received some critical thinking training in Session 2; the key manipulation is the form of practice (or lack thereof) in Session 3. See main text for details of each session.

responses. Participants received feedback indicating the correct answer after each selection.

Participants were paid $2 for completing Session 2, and the median completion time for participants in the final sample was 20.5 min.

Session 3: Intervention. Participants were randomly assigned to one of three intervention conditions: (a) critical thinking categorization practice ($n = 81$); (b) non-critical thinking categorization practice ($n = 91$); or (c) no-intervention control ($n = 81$). Participants in the no-intervention control group were not invited to complete Session 3; after completing Session 2 they were invited directly to Session 4. Participants in the other two conditions were paid $3 for completing Session 3.

For the critical thinking practice group, we administered the same four practice sets as in Study 1, each containing seven multiple-choice critical thinking scenarios. However, whereas practice sets in Study 1 differentiated the individual's name in each scenario (some participants saw "Professor Denise Crane" while others simply saw "Denise"), in Study 2 all practice sets only used the individual's first name ("Denise") for simplicity, and because we observed directionally better performance when participants were shown first name only. The sets were still organized around typical content units in an Introductory Psychology course (neuroscience, sensation and perception, memory, and learning). The order of items was randomized within each of these sets; however, all participants received items in the same order. Items were shown one at a time, and multiple-choice response options were randomized separately for each participant. After participants selected their answer, verification feedback (correct/incorrect) was shown, and participants had to get at least five of seven questions correct in each set; if they got fewer than five items correct, they had to repeat the set a maximum of three times. The median completion time was 36.6 min for the final sample of participants in the critical thinking practice group.

For the non-critical thinking practice group, we modified existing multiple-choice Introductory Psychology exam questions corresponding to each content unit (neuroscience, sensation and perception, memory, and learning), so that they were length matched with the critical thinking practice items, and so that there were seven response options, available at https://osf.io/bq5a8 (also see Table 2). Each item asked participants to categorize a scenario according to a structure or concept in psychology, with no mention of logical fallacies. Like the critical thinking questions, these non-critical thinking questions were shown alongside full definitions of all the relevant structures and concepts, verification feedback was shown after each response, and participants had to get five of seven correct to proceed to the next set. The median completion time was 32.6 min for the final sample of participants in the non-critical thinking practice group.

Session 4: Posttest. In this session, participants completed the remaining half of the PCTE items. As with the pretest, no information about the fallacies was displayed during the posttest. Participants were paid $5 for completing Session 4, and the median completion time was 11.1 min.

Data Analysis

Response Scoring. Similar to Study 1, responses to PCTE items were scored by a trained coder who was not able to see participants' condition assignments, nor whether the responses were from a pretest or posttest. Scoring again used the PCTE's original coding scheme: 0 (*no problem identified*), 1 (*a problem recognized but misidentified*), 2 (*identified main problem, but also mentioned less relevant problems*), and 3 (*identified only the main problem*). Our primary analytical goal is to assess whether improvement from pretest to posttest differs between the three intervention conditions:

Table 2
Examples of Items From Session 3 (Intervention) in Study 2

Critical-thinking categorization practice	Non-critical thinking categorization practice
Directions: Read the scenario below and select which fallacy (if any) the person made. You can use the descriptions of the fallacies in the table to help. After you have made a selection, click the arrow to proceed. You will get feedback on your answer, which you can look over.	Directions: Read the scenario below and select which concept applies to that scenario. You can use the definitions of the concepts in the table to help. After you have made a selection, click the arrow to proceed. You will get feedback on your answer, which you can look over.
Many people have suggested that some of the symptoms of ADHD might be caused by hyperactivity of the mirror neuron system. In order to treat these symptoms, Douglas has designed a therapy that should improve cognitive control of mirror neurons. This therapy, which involves watching cartoons of animals while keeping perfectly still, was administered to one hundred 12-year-olds who had previously been diagnosed with ADHD. After 6 months of therapy, these patients showed an overall decline in symptoms of ADHD. Douglas concluded that the therapy reduces symptoms of ADHD.	A group of people suffering from depression recently began a new experimental drug therapy trial. Like most pharmacological treatments for depression, this experimental drug affects the action of a certain neurotransmitter. However, this new drug is designed to avoid some of the adverse effects of previous drug therapies affecting this neurotransmitter, such as insomnia, drowsiness, and food cravings. Therapy for people suffering from depression often involves drugs that affect the action of which neurotransmitter?
○ Random chance ● Lack of control ○ Correlation is not causation ○ Overgeneralization ○ Confirmation bias ○ Experimenter bias ○ There is no fallacy in this conclusion	○ Cerebellum ○ Substantia nigra ● Serotonin ○ Prefrontal cortex ○ Temporal lobe ○ Amygdala ○ Hippocampus

Note. ADHD = attention deficit hyperactivity disorder.

(a) critical thinking categorization practice; (b) non-critical thinking categorization practice; or (c) no-intervention control.

Statistical Analysis. We registered two separate statistical analyses prior to data collection, and both were implemented without deviation from their registered forms. These analyses were identical to the analyses described in Study 1 (ordered probit model and metric improvement model), but extended to include pairwise contrasts in improvement pre to post between groups (Study 1 only had pre-to-post contrasts for one group). Additionally, the prior for group-level estimates of improvement scores for each treatment group used the mean improvement for the full study sample. As with Study 1, the ordered probit analysis is described in Supplemental Material, and the metric analysis is described in the main text below; both yielded identical findings.

We estimated model parameters with 100 adaptation steps, 500 burn-in steps, and 100,000 samples thinned to every fifth step (20,000 saved samples) across four MCMC chains, again using JAGS and the runjags package for R, and the full registered model specification is available at https://osf.io/vju8w/. The ESS was at least 20,000 for all model estimates, well above the 10,000 recommended by Kruschke (2014).

Results and Discussion

Averaging across all treatment conditions, prior to any training, participants got 2.4 items precisely correct (scored 3 out of 3; 40.3%), received a score of 2 on 0.4 items (6%), received a score of 1 on 2.0 items (33%), and received a score of 0 on 1.2 items (21%), out of six items total on the pretest. These pretest scores are modestly higher than pretest scores observed in our previous college sample (Study 1), which may be due to differences in age or participants' education level at time of the study; for example, 68.4% of participants in Study 2 reported having completed a bachelor's degree or higher, compared with 0% in Study 1, who were early in college enrollment. Moreover, attrition during Study 2's four sessions may have produced a more selective final sample than Study 1.

On the posttest, still averaging across all treatment conditions, participants got a score of 3 on 3.1 items (52.3%), a score of 2 on 0.4 items (7.2%), a score of 1 on 1.4 items (23.7%), and a score of 0 on 1 item (16.8%), out of six items total.

According to Lawson et al.'s (2015) point tallying method, these results correspond to an average pretest score of 55.3% (9.96 points out of 18 possible) and an average posttest score of 65% (11.7 points).

The number of items participants got precisely correct (3 out of 3), again aggregating across all treatment conditions, was 2.4 at pretest and 3.1 at posttest. Posterior estimates of average improvement (posttest–pretest; full sample) had a mode of 0.72, and a 95% HDI of 0.55–0.93. It makes sense that we see credibly nonzero improvement across the full study sample, considering that all participants received some training on critical thinking fallacies immediately after taking the pretest.

In our hierarchical model, we also estimated improvement from pretest to posttest for each treatment condition separately at the group level. The modal improvement estimate in the critical thinking practice condition was 1.33 items (95% HDI: 0.96–1.67), in the non-critical thinking practice condition was 0.46 items (95% HDI: 0.13–0.81), and in the no-treatment control was 0.41 items (95% HDI: 0.13–0.70). These improvement scores, with group-level distribution estimates, are shown in Figure 5, visibly demonstrating that scores are larger in the critical thinking practice condition than in the other two conditions. Preregistered pairwise contrasts find that improvement in the critical thinking practice condition was credibly larger than the non-critical thinking practice condition (difference mode = 0.82; 95% HDI: 0.33–1.32), and larger than the no-treatment control (difference mode = 0.89; 95% HDI: 0.44–1.35). There was no evidence of any credible difference between the non-critical thinking practice condition and the no-treatment control condition (difference mode = 0.035; 95% HDI: −0.395 to 0.491). For participants in the critical thinking practice condition, improvement from pretest to posttest had Cohen's d of 0.84; however, this standardized effect size measure is not sensitive to the within-subject nature of these improvement scores.

The magnitude of these differences is notably larger than observed in Study 1, which might again be attributable to sample differences between these studies, or to the relative recency of practice and training at posttest in Study 2. On this second point, the posttest was administered 1.4 days after practice in Study 2, but roughly 14 days after the most recent practice session in Study 1.

Considering that all three treatment groups (critical thinking practice, non-critical thinking practice, and no practice) took both pretest and posttest, Study 2 affords the opportunity to investigate differences in the patterns of improvement for the six different fallacies, for the different treatment groups. In an exploratory (not preregistered) analysis we estimated parameters for a metric improvement model similar to the one described above, but with an additional hierarchical level for the different fallacies (see https://osf.io/bfvg3/). Consistent with Study 1, we observe the smallest improvement for PCTE items related to "Correlation is not Causation," and largest improvement for "Experimenter Bias." Unlike Study 1, we observe robust improvements for items related to "Random Chance" in the group that received critical thinking practice (Figure 6). However, interpreting random chance coincidences as evidence of a causal mechanism is, in a sense, misattributing a correlational pattern for causation, in the special case involving few observations.

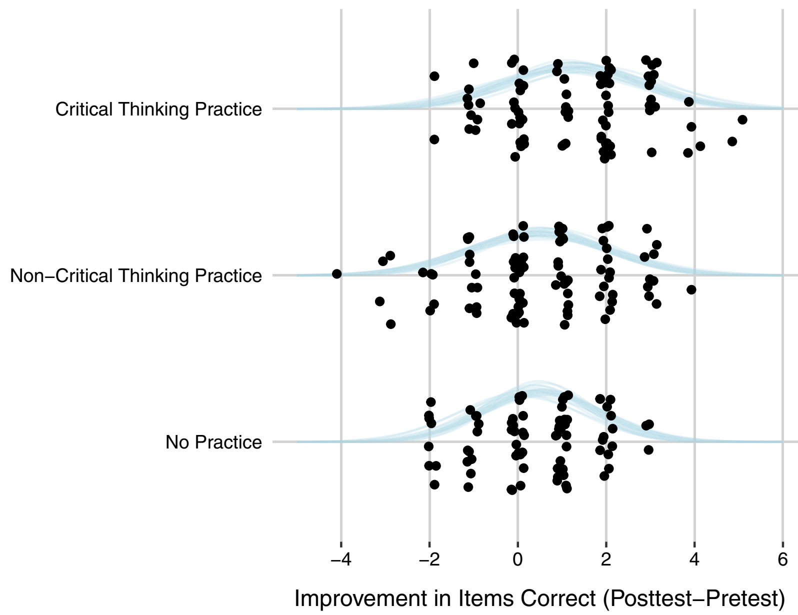
In Study 2, all participants received training about logical fallacies, and in all three conditions, participants demonstrated improved performance on an open-ended critical thinking assessment. However, participants who had a separate session practicing categorizing scenarios according to these logical fallacies had significantly higher gains. Participants assigned to this condition had roughly three times larger improvement pre to post (1.33 more items correct) than participants who practiced categorizing basic concepts (0.46 items) and participants who had no practice (0.41 items).

General Discussion

The current research extends a robust research lineage demonstrating that the strategic use of abstract "rules for reasoning" is a skill, and that this skill can be trained through induction (reviewed in Holland et al., 1989; Nisbett, 2013). In the present study, we show that such an approach can be constructively applied to train participants to categorize claims according to different logical fallacies and biases, and that this training

Figure 5

Improvement in the Number of Items Correct Between Pretest and Posttest in Study 2

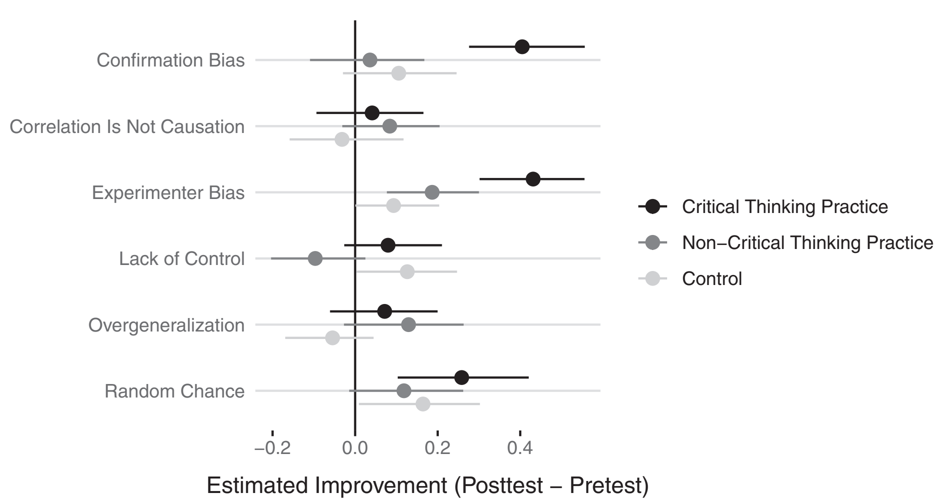


Note. Each dot illustrates a single participant's improvement, with minor horizontal jitter added for visualization (improvement scores are integers). Semitransparent blue lines show credible estimates of the mean and standard deviation, modeled as a normal distribution at the group level. Results indicate larger improvements in the critical thinking practice condition (top row) compared with the other two conditions. See the online article for the color version of this figure.

improves performance on a delayed, open-ended critical thinking assessment. In Study 1, we observed this improvement over an academic semester in an authentic college education setting, and in Study 2, we conducted a preregistered randomized controlled experiment, demonstrating appreciable improvements caused by induction with critical thinking practice problems, beyond those observed by a single training session, and by practice with traditional learning activities.

Figure 6

Improvement From Pretest and Posttest in Study 2, Separated by Fallacy



Note. Horizontal location indicates the magnitude of improvement from pretest to posttest estimated on a latent metric scale. Points show the most credible estimate, error bars show the 95% highest density interval (HDI).

By practicing the categorization of fallacies and biases, we believe that participants induce mental models of the abstract patterns exemplified in the training materials. Armed with such mental models, we find that participants later exposed to a logical fallacy or a biased claim are better equipped to recognize and label the claim as fallacious, a skill commonly associated with critical thinking ability, and one that is urgently in demand, considering the volume of misinformation online.

The central feature of the current work is our inductive approach to critical thinking training, presenting multiple sets of contextualized example claims that participants categorized using a simple multiple-choice response format. In this regard, our research was opposite of Renaud and Murray's (2008) approach; rather than designing a critical thinking intervention involving open-form essay questions, and evaluating critical thinking improvement using a multiple-choice assessment instrument (Watson–Glaser Critical Thinking Appraisal), we created an intervention using multiple-choice practice questions and assessed improvement on an open-form critical thinking assessment. The current results add to existing evidence that multiple-choice practice questions can develop higher order understanding (Hitchcock, 2004; Jensen et al., 2014; Scully, 2017), and we believe that our proposed intervention is more portable and easy to integrate in education settings, addressing calls for incorporating modular critical thinking practice throughout a course (Stevens et al., 2016).

Meta-analyses of critical thinking instruction in education settings have inferred that contextualized materials, deliberate practice, and repeated exposure are common elements of successful strategies for improving critical thinking (Abrami et al., 2015; Niu et al., 2013; Tiruneh et al., 2014; Willingham, 2008). However, with rare exceptions (Heijltjes et al., 2014; Hitchcock, 2004), we have found no scholarly review that directly considered, let alone examined, the use of an inductive strategy for building learners' critical thinking skills. Instead, education scholars tend to advocate for deep analysis of open-ended cases and in-depth group discussion, strategies that are laudable, but time-intensive and threaten incursion into other curricular instruction. Cáceres et al. (2020) observed that many teachers do make ad hoc efforts to infuse critical thinking into subject matter instruction, albeit implemented informally, without dedicated modules or lessons. These efforts suggest that many educators are willing to incorporate critical thinking instruction into their curriculum, and the instructional strategy we propose herein offers a flexible and effective method for doing so.

Our current implementation takes the form of an online quiz, a routine learning device in contemporary education settings. However, alternative formats might also confer similar benefits, such as a recent web browser-based game for detecting fake news (Basol et al., 2020; Roozenbeek & van der Linden, 2019). This game's prebunking approach extends a metaphor where people become inoculated against persuasive techniques (McGuire, 1964; Papageorgis & McGuire, 1961), by exposing them to weakened forms of misinformation, progressively training their resistance to fake news. If resistance to misinformation is analogous to being able to accurately identify fallacies in claims, we see the current results as providing converging evidence of the benefits of induction to building such cognitive defenses.

In the present study, we found that inductive training, and these resulting defenses, were not equally resistant to all fallacies. The single most difficult fallacy to identify, as evidenced by it having the lowest scores during pretest, training, posttest, and improvement from pretest to posttest, involved causal claims drawn from correlational data (labeled "Correlation is not Causation"). Indeed, in the course of natural human experience, correlations typically provide evidence that a causal mechanism is at play, and humans are generally disposed to infer causal structure from correlational patterns (Sloman, 2005). Nevertheless, special cases of fallacious correlational inferences, where change over time is attributed to an antecedent action (labeled "Lack of Control") and where rare coincidences are attributed to a causal relationship (labeled "Random Chance"), showed improvements during training in comparison to "Correlation is not Causation" items. Evidently, the current training develops sensitivities to specific features of correlational inferences, but general causal claims from correlational findings are more resistant to induction, a topic that is ripe for future research. In contrast, the items that did not involve correlational patterns ("Confirmation Bias" and "Experimenter Bias") showed consistent and credible improvements attributable to training. We speculate that this may reflect the relative ease with which people attribute the causes of others' behaviors to individual predispositions (e.g., being biased) rather than to situations (e.g., being misled by correlational data), a classic observation in social psychology (Jones & Nisbett, 1987; Malle, 2006).

Still, we acknowledge limitations of the present findings. These issues include the size of the training effect, the generalizability of training materials, and the transfer to real-world situations. Regarding the size of the observed effect, the difference in the magnitude of improvement (from pretest to posttest) between Studies 1 and 2, which differed substantially in the delay between training and posttest, raises questions about the persistence of these effects. Even so, it may yet be possible to strengthen the inductive training regimen, and perhaps to see more enduring effects over longer delays, by presenting more examples and/or repeating the examples to increase the number of training items (in the present study, critical thinking training only involved 4 sets of 7 questions, 28 items total). Expanding the set of practice problems, perhaps further to align with other disciplinary content (beyond Introductory Psychology) would also test the generalizability of this approach. We see no reason to doubt that similar results would be observed when students are shown critical thinking scenarios related to history, health, economics, or other domains, but we refrain from making claims about the breadth of possible benefits on the basis of the current results alone. Further, the question of whether inductive training has isolated benefits for our narrow operationalization of critical thinking ability, or whether it facilitates broader benefits (such as a tendency to approach claims with skepticism, or other attitudinal dispositions) also begs for further examination. We adopted the PCTE as our critical thinking assessment for its face validity to situations where people need to judge the validity of claims that they might encounter online. Additional research might directly examine whether improvement in PCTE scores corresponds to an improved ability to identify fallacies encountered in one's day-to-day experience.

Democracy depends on a well-informed society. Considering the volumes of misinformation that presently circulate online, it is critically important to equip individuals to identify and reject fallacious and biased claims, or more bluntly, to be able to call bullsh*t. The current work provides evidence in support of a specific and flexible method for doing so in common education settings.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275–314. https://doi.org/10.3102/0034654314551063

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college*.

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 1–18. https://doi.org/10.1177/20539517211013868

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), Article 2. https://doi.org/10.5334/joc.91

Bensley, D. A., & Lilienfeld, S. O. (2017). Psychological misconceptions: Recent scientific advances and unresolved issues. *Current Directions in Psychological Science*, *26*(4), 377–382. https://doi.org/10.1177/0963721417699026

Bensley, D. A., Lilienfeld, S. O., & Powell, L. A. (2014). A new measure of psychological misconceptions: Relations with academic background, critical thinking, and acceptance of paranormal and pseudoscientific claims. *Learning and Individual Differences*, *36*, 9–18. https://doi.org/10.1016/j.lindif.2014.07.009

Bensley, D. A., & Spero, R. A. (2014). Improving critical thinking skills and metacognitive monitoring through direct infusion. *Thinking Skills and Creativity*, *12*, 55–68. https://doi.org/10.1016/j.tsc.2014.02.001

Bergstrom, C. T., & West, J. D. (2021). *Calling bullshit: The art of skepticism in a data-driven world*. Random House.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278–287. https://doi.org/10.1037/0096-3445.120.3.278

Cáceres, M., Nussbaum, M., & Ortiz, J. (2020). Integrating critical thinking into the classroom: A teacher's perspective. *Thinking Skills and Creativity*, *37*, Article 100674. https://doi.org/10.1016/j.tsc.2020.100674

Carvalho, P. F., Braithwaite, D. W., de Leeuw, J. R., Motz, B. A., & Goldstone, R. L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLOS ONE*, *11*(3), Article e0152115. https://doi.org/10.1371/journal.pone.0152115

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, Article 505. https://doi.org/10.3389/fpsyg.2015.00505

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*(3), 293–328. https://doi.org/10.1016/0010-0285(86)90002-2

Day, S. B., Motz, B. A., & Goldstone, R. L. (2015). The cognitive costs of context: The effects of concreteness and immersiveness in instructional examples. *Frontiers in Psychology*, *6*, Article 1876. https://doi.org/10.3389/fpsyg.2015.01876

Denwood, M. (2016). Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (Vol. 275, pp. 9–26). W. H. Freeman/Times Books/Henry Holt.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations* (ERIC document ED315423). American Philosophical Association.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292. https://doi.org/10.1016/0010-0285(86)90001-0

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, *26*(1), 9–25. https://doi.org/10.1007/s10648-014-9249-3

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference* (pp. 913–922). Association for Computing Machinery. https://doi.org/10.1145/3178876.3186139

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2013). Concepts and categorization. In A. F. Healy (Ed.), *Handbook of psychology: Experimental psychology* (Vol. 4, pp. 607–630). Wiley.

Goldstone, R. L., Rogosky, B. J., Pevtzow, R., & Blair, M. (2017). The construction of perceptual and semantic features during category learning. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 851–882). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00034-8

Gurung, R. A. R., Hackathorn, J., Enns, C., Frantz, S., Cacioppo, J. T., Loop, T., & Freeman, J. E. (2016). Strengthening introductory psychology: A new model for teaching the introductory course. *American Psychologist*, *71*(2), 112–124. https://doi.org/10.1037/a0040012

Halonen, J. S., Buskist, W., Dunn, D. S., Freeman, J., Hill, G. W., Enns, C., Landrum, R. E., McCarthy, M., Rudmann, J., Stoloff, M., & Kaslow, N. (2013). *APA guidelines for the undergraduate psychology major* (Version 2.0). American Psychological Association.

Halpern, D. F. (2013). *Thought and knowledge: An introduction to critical thinking*. Psychology Press. https://doi.org/10.4324/9781315885278

Harnad, S. (2017). To cognize is to categorize: Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 21–54). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00002-6

Heijltjes, A., van Gog, T., & Paas, F. (2014). Improving students' critical thinking: Empirical support for explicit instructions combined with practice. *Applied Cognitive Psychology*, *28*(4), 518–530. https://doi.org/10.1002/acp.3025

Hitchcock, D. (2004). The effectiveness of computer assisted instruction in critical thinking. *Informal Logic*, *24*(3), 183–217. https://doi.org/10.22329/il.v24i3.2145

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT Press. https://doi.org/10.7551/mitpress/3729.001.0001

Homa, N., Hackathorn, J., Brown, C. M., Garczynski, A., Solomon, E. D., Tennial, R., Sanborn, U. A., & Gurung, R. A. R. (2013). An analysis of learning objectives and content coverage in introductory psychology syllabi. *Teaching of Psychology*, *40*(3), 169–174. https://doi.org/10.1177/0098628313487456

Horton, R. (2020). Offline: Managing the COVID-19 vaccine infodemic. *Lancet*, *396*(10261), Article 1474. https://doi.org/10.1016/S0140-6736(20)32315-1

Im, S.-H., Cho, J.-Y., Dubinsky, J. M., & Varma, S. (2018). Taking an educational psychology course improves neuroscience literacy but does not reduce belief in neuromyths. *PLOS ONE*, *13*(2), Article e0192163. https://doi.org/10.1371/journal.pone.0192163

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, *75*(2), 194–209. https://doi.org/10.1037/h0024971

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test … or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Johnson, R. H., & Hamby, B. (2015). A meta-level approach to the problem of defining “critical thinking”. *Argumentation*, *29*(4), 417–430. https://doi.org/10.1007/s10503-015-9356-4

Jones, E. E., & Nisbett, R. E. (1987). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Lawrence Erlbaum.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kolb, D. A. (2015). *Experiential learning: Experience as the source of learning and development*. Pearson Education.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kowalski, P., & Taylor, A. K. (2004). Ability and critical thinking as predictors of change in students’ psychological misconceptions. *Journal of Instructional Psychology*, *31*(4), 297–303.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lantian, A., Bagneux, V., Delouvée, S., & Gauvrit, N. (2021). Maybe a free thinker but not a critical one: High conspiracy belief is associated with low critical thinking ability. *Applied Cognitive Psychology*, *35*(3), 674–684. https://doi.org/10.1002/acp.3790

Lawson, T. J., Jordan-Fleming, M. K., & Bodle, J. H. (2015). Measuring psychological critical thinking: An update. *Teaching of Psychology*, *42*(3), 248–253. https://doi.org/10.1177/0098628315587624

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 1–38. https://doi.org/10.1080/10463283.2021.1876983

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*(4), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x

Lobato, E., Mendoza, J., Sims, V., & Chin, M. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance among a university population. *Applied Cognitive Psychology*, *28*(5), 617–625. https://doi.org/10.1002/acp.3042

Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology* (Vol. 12067, pp. 235–246). Springer. https://doi.org/10.1007/978-3-030-45002-1_20

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*(6), 895–919. https://doi.org/10.1037/0033-2909.132.6.895

McCutcheon, L. E., Apperson, J. M., Hanson, E., & Wynn, V. (1992). Relationships among critical thinking skills, academic achievement, and misconceptions about psychology. *Psychological Reports*, *71*(2), 635–639. https://doi.org/10.2466/pr0.1992.71.2.635

McGuire, W. J. (1964). Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0

Motz, B. A., Fyfe, E. R., & Guba, T. P. (2022). *Do you think you can tell?* Open Science Framework. https://doi.org/10.17605/OSF.IO/VCGZF

Nisbett, R. E. (2013). *Rules for reasoning*. Psychology Press. https://doi.org/10.4324/9780203763230

Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students’ critical thinking skills? A meta-analysis. *Educational Research Review*, *9*, 114–128. https://doi.org/10.1016/j.edurev.2012.12.002

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*(3), 328–353. https://doi.org/10.1037/xge0000369

Papageorgis, D., & McGuire, W. J. (1961). The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments. *Journal of Abnormal and Social Psychology*, *62*(3), 475–481. https://doi.org/10.1037/h0048430

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.3406

Poland, G. A., & Spier, R. (2010). Fear, misinformation, and innumerates: How the Wakefield paper, the press, and advocacy groups damaged the public health. *Vaccine*, *28*(12), 2361–2362. https://doi.org/10.1016/j.vaccine.2010.02.052

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards. *Educational Researcher*, *40*(3), 103–116. https://doi.org/10.3102/0013189X11405038

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*(3), 325–342. https://doi.org/10.1002/acp.796

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review*, *27*(3), 483–504. https://doi.org/10.1007/s10648-014-9273-3

Renaud, R. D., & Murray, H. G. (2008). A comparison of a subject-specific and a general measure of critical thinking. *Thinking Skills and Creativity*, *3*(2), 85–93. https://doi.org/10.1016/j.tsc.2008.03.005

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10. https://doi.org/10.1057/s41599-019-0279-9

Schacter, D., Gilbert, D., Wegner, D., & Hood, B. (2016). *Psychology*. Worth Publishers.

Schmaltz, R., & Lilienfeld, S. O. (2014). Hauntings, homeopathy, and the Hopkinsville Goblins: Using pseudoscience to teach scientific thinking. *Frontiers in Psychology*, *5*, Article 336. https://doi.org/10.3389/fpsyg.2014.00336

Schmaltz, R. M., Jansen, E., & Wenckowski, N. (2017). Redefining critical thinking: Teaching students to think like scientists. *Frontiers in Psychology*, *8*, Article 459. https://doi.org/10.3389/fpsyg.2017.00459

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, *22*(4), 1–13. https://doi.org/10.7275/SWGT-RJ52

Self, T. T., & Self, J. T. (2017). Reality in the classroom: Teaching critical thinking with scenarios. *Journal of Hospitality & Tourism Education*, *29*(4), 187–193. https://doi.org/10.1080/10963758.2017.1382367

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195183115.001.0001

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, *16*(1), 1–40. https://doi.org/10.1207/s15516709cog1601_1

Sternberg, R. J. (1999). Teaching psychology students to be savvy consumers and producers of research questions. *Teaching of Psychology*, *26*(3), 211–213. https://doi.org/10.1207/S15328023TOP260311

Stevens, C., Witkow, M. R., & Smelt, B. (2016). Strengthening scientific reasoning skills in introductory psychology: Evidence from community college and liberal arts classrooms. *Scholarship of Teaching and Learning in Psychology*, *2*(4), 245–260. https://doi.org/10.1037/stl0000070

Tavris, C., & Aronson, E. (2020). *Mistakes were made (but not by me) third edition: Why we justify foolish beliefs, bad decisions, and hurtful acts*. Houghton Mifflin Harcourt.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, *4*(1). https://doi.org/10.5539/hes.v4n1p1

Tullis, J. G., & Goldstone, R. L. (2016). Comparison versus reminding. *Cognitive Research: Principles and Implications*, *1*(1), Article 20. https://doi.org/10.1186/s41235-016-0028-1

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175–184. https://doi.org/10.1016/S0022-5371(67)80092-6

van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, *53*(1), 41–48. https://doi.org/10.3200/CTCH.53.1.41-48

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750–763. https://doi.org/10.3758/s13421-010-0063-y

Williams, R. L., Oliver, R., Allin, J. L., Winn, B., & Booher, C. S. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology*, *30*(3), 220–223. https://doi.org/10.1207/S15328023TOP3003_04

Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21–32. https://doi.org/10.3200/AEPR.109.4.21-32

Zohar, A., & Tamir, P. (1993). Incorporating critical thinking into a regular high school biology curriculum. *School Science and Mathematics*, *93*(3), 136–140. https://doi.org/10.1111/j.1949-8594.1993.tb12211.x

Received January 27, 2022
Revision received June 6, 2022
Accepted June 27, 2022 ■