AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

© 2025 American Psychological Association
ISSN: 0022-0663

2025, Vol. 117, No. 7, 1013–1038
https://doi.org/10.1037/edu0000934

Learning by Teaching With Deliberate Errors Promotes Argumentative Reasoning

Sarah Shi Hui Wong[1, 2]
[1] Division of Social Sciences, Yale-NUS College
[2] Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore

Reasoning and arguing well lies at the core of thinking and constructing knowledge about complex, controversial issues. Leveraging the techniques of learning by teaching and deliberate erring, the present study developed and tested a novel intervention—*learning by misteaching*—to boost argumentative reasoning. University students ($N = 208$) were trained on argumentation strategies and studied a dual-position argumentative text on a controversial topic using one of three learning methods: notetaking, correct teaching, or misteaching. The notetaking group prepared to be tested and wrote study notes while generating good arguments about the topic, whereas both teaching groups prepared to teach and wrote a verbatim teaching script about the topic exactly as how they would orate a lecture while generating good arguments (correct teaching) or deliberately weak arguments (misteaching) for their intended audience to spot. All students were then tested on their basic recall of the text and higher order argumentative reasoning in integrating opposing views to form conclusions about the topic (e.g., weighing arguments and counterarguments, developing new alternative solutions or compromises). On both tests, students who had taught outperformed their peers who had written study notes. Importantly, misteaching produced additional gains for argumentative reasoning over correct teaching, even after controlling for recall performance. Yet, students' metacognitive judgments revealed that they were largely unaware of these benefits even after the tests. Overall, these findings demonstrate how learning by teaching and deliberate erring can be strategically combined to improve higher order outcomes such as argumentative reasoning, while highlighting the counterintuitive benefits of intentionally making errors in low-stakes contexts.

Educational Impact and Implications Statement

The skill to reason and argue well is vital for 21st-century education and democratic participation but is challenging to master. This study demonstrates how argumentative reasoning (e.g., weighing and integrating arguments, designing new solutions) can be enhanced via *learning by misteaching*—a novel combination of the techniques of learning by teaching and deliberate erring. Students displayed not only superior recall but also argumentative reasoning when they had taught a controversial issue by writing a verbatim teaching script than study notes. Crucially, students' argumentative reasoning further benefited from teaching incorrectly with deliberately weak arguments they had generated for their intended audience to spot, relative to teaching correctly with good arguments. This advantage of misteaching held even when controlling for students' recall of the material, suggesting that it was not merely driven by better memory per se. Learning by teaching with deliberate errors is a promising way to boost argumentative reasoning.

Keywords: learning by teaching, explaining, learning from errors, argumentation, reasoning

Supplemental materials: https://doi.org/10.1037/edu0000934.supp

In our increasingly interconnected world where information abounds, one must be able to think and reason well about complex, controversial issues. Argumentation—the construction and evaluation of arguments—has been viewed as the main function of human reasoning (Mercier, 2016; Mercier & Sperber, 2011) and the core of thinking since the time of the ancient Greek philosophers

This article was published Online First January 9, 2025.
Alexandra List served as action editor.
Sarah Shi Hui Wong https://orcid.org/0000-0003-4243-212X
This research was supported by a Yale-NUS College Research Grant (H-607-02-000047) and a National University of Singapore Educational Research Grant (C-581-000-222-091) awarded to Sarah Shi Hui Wong. The author thanks Huiyi Liu and Sophia Xuefei Qiu for their assistance with data collection and scoring. Materials and data for this study are available in the online supplemental materials.
Sarah Shi Hui Wong served as lead for conceptualization, data curation, formal analysis, funding acquisition, methodology, resources, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Sarah Shi Hui Wong, Division of Social Sciences, Yale-NUS College, 28 College Avenue West, #01-501, Singapore 138533, Singapore. Email: psywshs@nus.edu.sg

(Kuhn, 1991). Today, argumentation remains central to meaningful participation in democratic societies (Asterhan & Schwarz, 2016). Moreover, argumentation has been associated with better learning (Andriessen & Baker, 2014; Asterhan & Schwarz, 2007; Chinn, 2006; Iordanou et al., 2019), scientific literacy (Driver et al., 2000; Kuhn, 1993; J. Osborne, 2010), and epistemic cognition when students acquire, construct, and use knowledge (Greene & Yu, 2016; Greene et al., 2018; Iordanou & Constantinou, 2015; Iordanou et al., 2016; Kuhn et al., 2013; Ryu & Sandoval, 2012).

Developing students' argumentative reasoning is thus a crucial educational goal, as outlined in curricular and policy initiatives such as the Next Generation Science Standards (NGSS Lead States, 2013) and Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). The present research leverages two potent techniques—learning by teaching and deliberate erring—to promote argumentative reasoning.

Argumentative Reasoning

Argumentation involves not only producing arguments but also responding to them with counterarguments (van Eemeren et al., 1996, 2014; D. Walton, 2007). Good arguments are able to resist and defeat opposing arguments to emerge ultimately undefeated (Pollock, 1987). Hence, whereas students often display a myside bias in considering only one side of the issue and ignoring counterarguments (Perkins, 1985; Perkins et al., 1991; Wolfe & Britt, 2008), stronger and more cogent argumentation in fact demands integrating arguments and counterarguments to reach an overall conclusion or consensus (i.e., *integrative argumentation*; Nussbaum, 2008b, 2021; Nussbaum & Schraw, 2007). Unlike adversarial discourse that aims to win through persuasion, integrative argumentation aligns with deliberative or collaborative discourse that aims to decide on optimal actions by reconciling different views to achieve a common goal (Felton et al., 2022; Nussbaum, 2008a, 2021; Rapanta & Felton, 2022; D. Walton, 2010).

To construct an integrative argument, two kinds of stratagems can be used: refutational and integrative stratagems (Nussbaum & Edwards, 2011). Refutational stratagems include *refutation* when rebutting a counterargument to explain why it is false or flawed. By weakening counterarguments instead of ignoring or dismissing them, refutation yields stronger arguments and better impressions of the author (Wolfe et al., 2009). However, refutational stratagems are also considered the least integrative because they mainly emphasize the counterargument (Nussbaum & Edwards, 2011). Conversely, integrative stratagems include weighing and design claims. *Weighing* is a sophisticated strategy that involves contrasting arguments on both sides and showing that one side surpasses the other in strength of evidence and/or importance of values when weighing benefits and costs (Nussbaum, 2021; Nussbaum & Putney, 2020). *Design claims* develop a compromise or new alternative solution by preserving the benefits of an argument while reducing the costs cited in a counterargument. Thus, weighing and design claims explicitly attend to both sides of an issue. Given their particular pertinence for integrative argumentation, this study focused on students' use of integrative stratagems. Indeed, balancing multiple views and searching for ways to integrate them have more broadly been regarded as forms of wise reasoning when navigating conflicts in life (Grossmann et al., 2020).

Yet, teaching argumentative reasoning is challenging (Newell et al., 2011). High school and undergraduate students and even teachers have been observed to engage in weak argumentation, as characterized by failures to generate two-sided arguments or justify their positions with evidence (Iordanou et al., 2020; Jiménez-Aleixandre et al., 2000; Lytzerinou & Iordanou, 2020; Sadler, 2004). Although even young children possess some rudimentary skills in argumentation (Köymen et al., 2014; Mercier, 2011; Mercier et al., 2014), developing more complex integrative argumentation skills is not spontaneous (Rapanta et al., 2013) and requires substantial scaffolding (Andriessen & Baker, 2014). For instance, explicit instruction and tools such as graphic organizers, critical questions or prompts, and computer-based systems have been used to support students' argumentation (for reviews, see Ferretti & Graham, 2019; Newell et al., 2011; Noroozi et al., 2012; Scheuer et al., 2010).

Crucially, for students to productively argue even when equipped with instruction and tools, they must acquire a deep understanding of the topic at hand and relate it to their prior knowledge, such that this new knowledge can be used meaningfully in discourse (Baytelman et al., 2020; Christodoulou & Diakidoy, 2020; Iordanou et al., 2019; Murphy et al., 2018). For instance, without understanding agricultural biotechnology and its implications, one would be hard-pressed to write a high-quality argumentative response on whether it should be introduced in Africa, even if one were well-versed in integrative stratagems. Accordingly, to promote deep learning for better argumentative reasoning, generative strategies that encourage students to make sense of new information by actively organizing and integrating it with their prior knowledge (Fiorella, 2023; Fiorella & Mayer, 2016; Wittrock, 1974, 1989) could be helpful. In particular, one such generative strategy is learning by teaching.

Learning by Teaching

Taking on the role of a "tutor" and teaching others benefits students' own learning of the material (Bargh & Schul, 1980; Duran & Topping, 2017; for recent reviews, see Kobayashi, 2024; Lachner et al., 2022; Ribosa & Duran, 2022), with unique gains from expecting to teach, actually teaching, and responding to audience or "tutee" questions (Benware & Deci, 1984; Fiorella & Mayer, 2013, 2014; Guerrero & Wiley, 2021; Kobayashi, 2022a; Nestojko et al., 2014; Roscoe & Chi, 2008). Whereas learning by teaching has often been implemented via peer tutoring (e.g., Roscoe & Chi, 2007, 2008) or teachable agents in computer-based environments (e.g., Biswas et al., 2005, 2016; Chin et al., 2010), tutors can benefit from teaching even without interacting with a real, remote, or fictitious audience (Lachner et al., 2022). For instance, learning-by-teaching effects have emerged when orally delivering video-recorded lectures to fictitious others (e.g., Fiorella & Mayer, 2013, 2014; Hoogerheide et al., 2014, 2016; Koh et al., 2018; Lachner et al., 2020; Wong et al., 2023) or even when writing verbatim teaching scripts as exact transcripts of how one would orate a lecture (i.e., *silent teaching*; K. Y. L. Lim et al., 2021; S. W. H. Lim et al., 2024).

Three main nonmutually exclusive accounts have been proposed to explain learning-by-teaching effects: (a) the retrieval hypothesis, (b) the generative hypothesis, and (c) the social presence hypothesis (see Lachner et al., 2022 for a review). First, the *retrieval hypothesis* suggests that tutors engage in retrieval practice when teaching from memory (Koh et al., 2018; see also Kobayashi, 2022b), thereby improving their durable learning of the material (Karpicke, 2017). Second, the *generative hypothesis* posits that teaching induces generative processes such as selecting relevant information, organizing it into a coherent mental representation, and integrating it with one's prior knowledge (Fiorella & Mayer, 2016). As tutors monitor their comprehension, construct inferences, and integrate ideas during their teaching explanations, this reflective knowledge-building process boosts their learning (Roscoe & Chi, 2007). Third and relatedly, the *social presence hypothesis* proposes that teaching triggers greater physiological arousal and generative processing when tutors perceive their audience—whether actual or imagined—as "real" and "present" (Hoogerheide et al., 2016, 2019a, 2019b; Jacob et al., 2020; Lachner et al., 2021; see also Kreijns et al., 2022). For instance, in anticipating what their audience knows or does not know (Nickerson, 1999), tutors may adapt their teaching such as generating more elaborations for less knowledgeable tutees (H. H. Clark & Brennan, 1991; Wittwer et al., 2010). However, it should also be noted that simply increasing social presence does not necessarily improve learning (Jacob et al., 2021), especially when it evokes excessive negative emotional arousal and distraction (Wang et al., 2023) or when tutors engage in limited knowledge-building even with heightened feelings of social presence (Ribosa & Duran, 2023).

Together, these accounts underpin the idea that teaching favors the tutor's building of rich mental models of the material for better learning. According to Kintsch's (1988, 1994) construction–integration model, a text can be encoded at the *textbase* level (i.e., mentally representing its propositional content) and *situation model* level (i.e., a global representation of the text that is elaborated from and integrated with one's prior knowledge). Whereas a text can be recalled or (superficially) summarized even if only processed at the textbase level, forming a situation model is crucial for deeper understanding (Kintsch, 1994). By inducing elaborate processing, teaching could support the tutor's construction of a situation model for better knowledge generalization and learning (Coleman et al., 1997; Fiorella, 2023). Indeed, teaching has been found to improve not only the tutor's basic recall, comprehension, and transfer to new problems (Lachner et al., 2022) but also more complex outcomes such as research question generation (S. W. H. Lim et al., 2024; Wong et al., 2023).

The view is that the deep learning that teaching affords could aid the tutor's argumentative reasoning too. Whether students can reason well depends on their knowledge of the argumentation topic (Chinn & Duncan, 2018; Stein & Bernas, 1999; von Aufschnaiter et al., 2008; Voss & Van Dyke, 2001), besides their *need for cognition* (i.e., tendencies to engage in and enjoy effortful cognitive activities; Cacioppo et al., 1983, 1996), *epistemological understanding* (i.e., beliefs about knowing and knowledge; Greene et al., 2018; Mason & Boscolo, 2004; Mason & Scirica, 2006; Wu & Tsai, 2011), and competence in reasoning strategies (J. F. Osborne et al., 2016). More robust topic knowledge positively predicts the quantity, quality, and diversity of arguments that students construct for better argumentative reasoning (Baytelman et al., 2020), while better recall and understanding of an argumentative text's content has been associated with better critical evaluation of its arguments (Christodoulou & Diakidoy, 2020; Neuman & Weizman, 2003). Moreover, to the extent that tutors experience heightened social presence (e.g., Hoogerheide et al., 2016), they may be poised for stronger integrative argumentation if they take their audience's perspective into account. Perspective taking—reasoning from other viewpoints—prompts people to distance themselves from their own perspectives, which could reduce the myside bias during argumentative reasoning (Beatty & Thompson, 2012; Thompson et al., 2005), particularly for weak arguments (McCrudden et al., 2017).

Teaching With Good Versus Poor Arguments

Skilled arguers are further able to evaluate argument strength when mustering their best response on the topic. For instance, when weighing arguments and counterarguments, one must recognize and reject weaker arguments while accepting better ones. Likewise, when forming design claims, one must synthesize the strengths of both sides while mitigating their weaknesses to propose an in-between position or new alternative solution. To these ends, understanding what a good argument is (not) could be paramount. In defeasible reasoning, good arguments are rationally compelling in that their premises provide support for the conclusion, even if this conclusion could later be falsified by additional information (Koons, 2022; Pollock, 1987). Conversely, poor arguments may appear to be rationally compelling in that their premises are plausible, but are not actually strong in justifying their conclusion (see Hahn & Oaksford, 2007 for a discussion of argument strength).

One could posit that teaching with only good arguments enables tutors' understanding of how such responses can be applied for strong argumentation. Indeed, poor arguments can be considered errors, which have traditionally been viewed as aversive events to be avoided in learning so that they are not ingrained and repeated in the future (Ausubel, 1968; Bandura, 1986; Skinner, 1958). But against this view, errors have been vindicated by burgeoning evidence that they can in fact benefit learning (Metcalfe, 2017; Wong & Lim, 2019b). Thus, a counterintuitive proposition is that tutors may actually gain more from deliberately generating weak arguments as they teach.

Learning From Deliberate Errors

According to Wong and Lim's (2019b) Prevention–Permission–Promotion (3P) framework, errors can be observed, allowed, induced, or guided when not avoided in learning. Of particular interest, recent studies have shown that guiding students to deliberately err in low-stakes contexts enhances their learning, even when they already know the correct answers. This phenomenon has been termed the *derring effect*, as coined from a portmanteau of the words "deliberate" and "erring" (Wong & Lim, 2022a, 2022b). For instance, when studying scientific expository texts or term-definition concepts, students display better recall when they have deliberately committed and corrected errors (e.g., generated conceptually wrong answers) during open-book study than avoided errors by copying, creating concept maps, or generating alternative correct responses (Wong & Lim, 2022a, 2022b). Besides boosting recall, deliberate erring improves higher order knowledge application (Wong & Lim, 2022a) and even far transfer of learning to different knowledge domains (Wong, 2023). Moreover, the derring effect has been observed not only with conceptual errors but also with procedural errors when deliberately executing incorrect problem-solving procedures in mathematics (Yap & Wong, 2024).

Although the mechanisms underlying the derring effect have yet to be fully specified, several theoretical accounts are conceivable (Wong, 2023; Wong & Lim, 2022a, 2022b). Broadly, theories of failure-driven learning propose that encountering errors challenges one's existing mental models and triggers an inquiry process that may not otherwise occur during errorless learning (A. A. Tawfik et al., 2015). Likewise, deliberate erring may evoke mental processes that are not typically invited by the material but that are useful for learning it (Wong, 2023; Wong & Lim, 2022a; for discussions of material-appropriate processing, see McDaniel & Butler, 2011; McDaniel & Einstein, 1989). For instance, considering what *not* to do or what something is *not* when generating errors could enhance learning of a correct response more than considering what *else* it is (Gartmeier et al., 2008; Oser & Spychiger, 2005). It may be that error generation potentiates encoding of the target response by more strongly directing attention to subsequent correction for better learning (Wong & Lim, 2022a, 2022b). Alternatively, exploring incorrect solutions may ironically weaken and cull those pathways, thereby increasing the relative retrieval strength of the correct solutions (Kornell et al., 2009). Moreover, contrasting incorrect versus correct responses may increase students' awareness of their knowledge gaps and recognition of the problem's deep features (Loibl et al., 2017), such that they are more likely to seek to repair or refine their mental models for better future performance (Chi, 2000; VanLehn, 1999; VanLehn et al., 2003).

Applied in argumentation, deliberately generating errors in the form of weak arguments may preempt and expose faulty reasoning, paradoxically enabling richer mental models of what constitutes good arguments and why. For instance, when searching one's knowledge to produce weak arguments, one must determine the conditions and parameters for success and failure. Whereas considering why a correct response is correct may involve a more superficial search that simply cites one's existing knowledge, confronting the inadequacies of an error may stimulate a deeper search of one's knowledge (Siegler & Chen, 2008) whereby one must also reflect on the correct response and the conditions that would undermine it (Heemsoth & Heinze, 2014), toward developing more sophisticated strategies (Siegler, 2002).

In this way, deliberately generating weak arguments may promote a richer understanding that includes not only correct (or better) responses and strategies but also incorrect (or worse) ones for better argumentative reasoning. For instance, identifying weak arguments during deliberate erring may prepare students to recognize stronger ones when weighing the relative merits of the arguments for and against a topic. Engaging with weak arguments may also provide opportunities for discovering new solutions or compromises when creating design claims, which may not be afforded by learning from good arguments alone (Parviainen & Eriksson, 2006). Indeed, some evidence suggests that studying cases of failure when solving ill-structured problems, as opposed to cases of success, improves students' argumentation in considering alternative perspectives (A. Tawfik & Jonassen, 2013). Thus, tutors' argumentative reasoning may further benefit from generating weak, rather than good, arguments when teaching an argumentation topic, besides the potential learning gains from teaching-induced social presence (e.g., Hoogerheide et al., 2016) and reflective knowledge-building (e.g., Fiorella & Mayer, 2016; Roscoe & Chi, 2007).

The Present Study: Learning by Misteaching

To promote argumentative reasoning, the present study tested a novel technique—*learning by misteaching*—that combined learning by teaching and deliberate erring. Participants were first instructed on the qualities of good versus poor arguments, then presented with a dual-position argumentative text on a controversial topic, either "Will biotech solve Africa's food problems?" or "Should we continue to study sex differences?" Participants engaged in open-book study of the text using one of three learning methods: notetaking, correct teaching, or misteaching. The notetaking control group prepared to be tested and wrote study notes about the text, while generating as many good and plausible arguments of their own about the topic. Conversely, both teaching groups prepared to teach by writing teaching notes about the topic, then taught "silently" by writing a verbatim (i.e., word-for-word) teaching script. Crucially, whereas the correct teaching group generated as many good and plausible arguments of their own in their teaching script for their intended audience's learning, the misteaching group deliberately generated as many poor yet plausible arguments of their own for their audience to spot. Subsequently, all participants were trained on argumentation stratagems and then tested on their recall and argumentative reasoning of the text they had studied. Finally, all participants made a metacognitive judgment of their learning method's effectiveness.

The main hypotheses were: To the extent that teaching others yields deeper learning than generating egocentric content (e.g., Fiorella & Mayer, 2016; Lachner et al., 2022), both teaching groups would outperform the notetaking group on the recall and argumentative reasoning tests. Furthermore, in line with the derring effect (Wong & Lim, 2022a, 2022b), misteaching was expected to confer additional gains over correct teaching. Deliberately generating poor yet plausible arguments may enable the tutor to better understand the conditions for successful arguments when weighing different positions or creating new solutions. Thus, misteaching should be more helpful than correct teaching for the tutor's argumentative reasoning, although not necessarily their basic recall of the text's content.

Because prior research has found that need for cognition predicts students' recall of arguments (Cacioppo et al., 1983), and that epistemological understanding predicts argumentation performance (Greene et al., 2018; Mason & Boscolo, 2004; Mason & Scirica, 2006), both variables were measured alongside participants' English language proficiency at the start of the study to ascertain baseline equivalence of the three learning groups.

In addition, to explore participants' learning processes and the characteristics of their study notes and teaching scripts, these were scored on the number of (a) self-generated arguments, (b) self-other referential terms such as "me" and "you," which served as a proxy for perceived social presence (e.g., Hoogerheide et al., 2016; Jacob et al., 2020; Lachner et al., 2018; see also Chafe, 1982; Sindoni, 2013), (c) elaborations (e.g., Jacob et al., 2020; Lachner et al., 2018), and (d) monitoring statements (e.g., Roscoe, 2014). The last three variables served as measures of teaching quality based on the social presence hypothesis (e.g., Hoogerheide et al., 2016) and generative hypothesis (e.g., Fiorella & Mayer, 2016; Roscoe & Chi, 2007) of learning by teaching. As per the social presence hypothesis, teaching with an audience in mind would evoke heightened social presence, such that both teaching groups should use more self-other referential terms than the notetaking group, in turn mediating any advantage of learning by teaching on test performance. Additionally, the generative hypothesis predicts that teaching others stimulates generative and metacognitive processing (i.e., greater knowledge-building), such that both teaching groups should produce more elaborations and monitoring statements than the notetaking group, in turn mediating any advantage of learning by teaching on test performance. To examine these hypotheses, exploratory analyses tested: (a) whether the learning groups significantly differed on the three teaching quality measures, (b) whether these measures correlated with test performance, and if so, (c) whether these measures mediated the effects of learning method.

Method

Transparency and Openness

This study reports how the sample size was determined, all data exclusions, all manipulations, and all measures, and it follows the American Psychological Association Journal Article Reporting Standards. Materials and data for this study are available in the online supplemental materials. Data were analyzed using SPSS Version 26. This study's design and analyses were not preregistered.

Participants and Design

The participants were 216 students (153 were female) between the ages of 18 and 36 ($M = 21.58$, $SD = 2.48$) recruited from both the Psychology Research Participation Program and university-wide Research Participation Scheme at the National University of Singapore, who received either course credit or monetary reimbursement for their participation. The students came from a range of majors and years in college. The outcomes reported below are based on data from 208 participants; eight participants who failed to adhere to the experimental instructions were excluded from analyses. A power analysis (G*Power; Faul et al., 2007) indicated that the present sample size afforded sufficient sensitivity to detect effects of $d \geq 0.48$ for two-tailed between-subjects pairwise comparisons (i.e., t tests) at 80% power and $\alpha = .05$, similar to the effect size of noninteractive teaching with preparing-to-teach over a nonteaching control ($g = 0.48$) reported in Kobayashi's (2019b) meta-analysis, and the moderate effect sizes of deliberate erring over errorless control methods for higher order learning outcomes ($d = 0.46$–0.77) reported in extant research (Wong, 2023; Wong & Lim, 2022a; Yap & Wong, 2024).

The key independent variable was learning method, whereby participants were randomly assigned to one of three learning groups: notetaking (control condition, $n = 70$), correct teaching ($n = 70$), or misteaching ($n = 68$). To ascertain that any effects of the learning methods generalized across argumentation topics, argumentative text was included as a second between-subjects factor for control purposes, whereby participants were randomly assigned to study a text on either "biotech" or "sex differences." The main learning outcomes of interest were: (a) participants' recall performance, as assessed by the number of idea units from the texts that they correctly recalled, and (b) participants' argumentative reasoning performance, as assessed by the number of integrative stratagems that they used at the microlevel, and their holistic argumentation quality at the macrolevel. All participants provided their written informed consent. This study was conducted with ethics approval from the university's institutional review board.

Materials

Preexperiment Questionnaire

To assess baseline equivalence of the three learning groups, a preexperiment questionnaire measured their need for cognition, epistemological understanding, and English language proficiency (e.g., Wong & Lim, 2019a). Participants completed the questionnaire online before attending the experiment.

Need for Cognition Scale. The 18-item short form of the Need for Cognition Scale (Cacioppo et al., 1984) measured participants' need for cognition on a 5-point scale (1 = *extremely uncharacteristic* to 5 = *extremely characteristic*). A sample item was: "I would prefer complex to simple problems." Participants' mean need for cognition score was computed. The scale had high internal consistency in this study, Cronbach's $\alpha = .86$.

Epistemological Understanding Scale. The 15-item Epistemological Understanding Scale (Kuhn et al., 2000) was administered as a measure of participants' levels of epistemological understanding (absolutist, multiplist, or evaluativist). A sample item was: "Robin believes one book's explanation of how the brain works versus Chris believes another book's explanation of how the brain works." After being shown each pair of contrasting statements, participants were asked to indicate if "only one could be right" (absolutist answers, scored as one point) or if "both could have some rightness." If participants selected the latter, they were then asked to indicate whether "one could not be more right than the other" (multiplist answers, scored as two points) or "one could be more right" (evaluativist answers, scored as three points). The maximum possible score was 45. The scale had acceptable internal consistency in this study, Cronbach's $\alpha = .76$.

English Proficiency Test. Participants' English language proficiency was assessed through 10 questions adapted from the Verbal Reasoning section of the Graduate Record Examinations. The maximum possible score was 10.

Argumentation Training

To introduce all participants to argumentation, they were presented with a training handout on arguments and counterarguments (available in the online supplemental materials). Specifically, the handout explained the qualities of three types of arguments: (a) *good and plausible* arguments that are plausible and strong in supporting their conclusion, (b) *poor yet plausible* arguments that are plausible but not strong in supporting their conclusion, and (c) *poor and implausible* arguments that are neither plausible nor strong in supporting their conclusion. Each type of argument was illustrated with examples using the training topic of "Should candy be banned from school?" (see Table 1), which was not related to either of the argumentative texts on "biotech" and "sex differences."

For instance, the argument that "candy should be banned from school because it has high calories" is poor yet plausible—whereas it is plausible that candy does have high calories, this argument is not strong in supporting the conclusion that candy should be banned from school just on this basis. Indeed, not all high-calorie foods are unhealthy; some high-calorie foods such as meat and milk can be healthy for children. Conversely, the argument that "candy should be banned from school because children don't have the right to eat what they want" is poor and implausible—it is neither plausible nor strong in supporting its conclusion but is based solely on a personal opinion without any relevant or sound evidence.[1]

[1] It should be noted that the conceptual difference between "poor" and "implausible" could have been more clearly defined during training, but differentiating between them was not essential because the main goal was for participants to generate either good arguments in the notetaking and correct teaching conditions or flawed arguments in the misteaching condition.

Table 1

Training Examples of Good Versus Poor Arguments on the Topic of "Should Candy Be Banned From School?"

Argument	"For"	"Against"
Good and plausible	Yes, candy should be banned from school because it makes children overly active and hurts concentration. For example, children who eat sweets have been shown to be easily distracted and get out of their seats more.	No, candy should not be banned from school because it is an effective incentive to motivate children to study. For example, children are shown to be more satisfied and ready to learn after immediate incentives are given.
Poor yet plausible	Yes, candy should be banned from school because it has high calories. Food with high calories is unhealthy.	No, candy should not be banned from school because adults eat candy themselves. It is unfair to treat children and adults differently.
Poor and implausible	Yes, candy should be banned from school because children don't have the right to eat what they want.	No, candy should not be banned from school because I like candy.

Prestudying Questionnaire

A prestudying questionnaire assessed participants' prior attitude and familiarity toward their randomly assigned argumentation topic, as well as the personal importance of the topic to them (adapted from Kobayashi, 2010; Wong & Lim, 2019a). Participants were presented with a 55-word introductory paragraph on the topic of their argumentative text, either "biotech" or "sex differences" (available in the online supplemental materials). After reading the introductory paragraph, participants rated their prior attitude toward the topic on four items: (a) "Agricultural biotechnology should be introduced in Africa"/"Sex differences should continue to be studied," (b) "African society will greatly benefit from agricultural biotechnology"/"Society will greatly benefit from the study of sex differences," (c) "The introduction of agricultural biotechnology will bring a lot of problems to African society"/ "The study of sex differences will bring a lot of problems to society," and (d) "Agricultural biotechnology is unsuitable for Africa"/ "The study of sex differences is unsuitable for society." All ratings were made on a 7-point scale (1 = *strongly disagree* to 7 = *strongly agree*). Participants' prior attitude scores were computed as their mean rating across all four items; negative items were reverse-scored such that higher scores indicated more positive attitudes. The prior attitude measure had acceptable internal consistency, Cronbach's $\alpha = .72$.

Participants indicated their familiarity with their argumentation topic on two items: (a) "How familiar are you with (agricultural biotechnology/the study of sex differences)?," and (b) "How familiar are you with the controversy over the (introduction of agricultural biotechnology in Africa/study of sex differences in society)?" Both items were rated on a 7-point scale (1 = *not familiar at all* to 7 = *extremely familiar*). Participants' familiarity scores were computed as their mean rating across both items. The prior familiarity measure had acceptable internal consistency, Cronbach's $\alpha = .76$.

In addition, participants rated the personal importance of their argumentation topic on a 7-point scale (1 = *not important at all* to 7 = *extremely important*): "How personally important is the (introduction of agricultural biotechnology in Africa/study of sex differences in society) to you?"

Argumentative Texts

The argumentative texts were two dual-position passages on the topics of "biotech" and "sex differences" (available in the online supplemental materials). Both texts were adapted from the "Taking Sides" McGraw-Hill Contemporary Learning Series (Moseley, 2007; Paul, 2002), which has been used in higher education and educational psychology research (e.g., Agarwal, 2019). Each text was organized in two columns that presented arguments for versus against the topic, with four paragraphs describing each opposing side. Both texts contained arguments of varying qualities, thus affording room for participants to engage in integrative argumentation at test (e.g., weighing the relative merits of the arguments vs. counterarguments). The "biotech" text (612 words) presented the debate on the introduction of agricultural biotechnology in Africa and had a Flesch-Kincaid grade level of 14; the "sex differences" text (606 words) presented the debate on the study of sex differences in society and had a Flesch-Kincaid grade level of 17.

Judgments of Higher Order Learning (JOL+)

As a metacomprehension monitoring intervention to guide all participants toward the higher order learning outcome of integrative argumentation, a series of JOL+ questions was presented (adapted from Wong & Lim, 2019a). The JOL+ questions directed students' attention toward the critical processes required for integrative argumentation (e.g., addressing counterarguments, weighing arguments and counterarguments, creating compromises or novel alternative solutions to reconcile opposing sides of an issue). Specifically, all participants were asked to respond to these JOL+ questions when learning their argumentative text: (a) "How difficult do you think it is to argue the case for/against (introducing agricultural biotechnology in Africa/ continuing to study sex differences)?" (0 = *definitely not difficult* to 100 = *definitely difficult*), (b) "How confident do you think you are in arguing the case for/against (introducing agricultural biotechnology in Africa/continuing to study sex differences)?" (0 = *definitely not confident* to 100 = *definitely confident*), and (c) "If you are asked to argue for/against (introducing agricultural biotechnology in Africa/ continuing to study sex differences), how well do you think you can (i) address the counterclaims to your claims, (ii) argue that one claim is weaker than another claim, (iii) comment on the benefits and costs of a claim, or the benefits and costs of one claim compared to another, and (iv) create novel solutions to resolve the arguments on both sides of the issue?" (each item rated 0 = *definitely not well* to 100 = *definitely well*). All JOL+ ratings were made on an 11-point scale (i.e., 0, 10, 20, …, 100). For each participant, their mean rating across all JOL+ questions was computed.

Poststudying Questionnaire

A four-item poststudying questionnaire was administered after participants had studied the argumentative text. Specifically, participants rated how interesting and understandable the text was on a 7-point scale (1 = *not at all* to 7 = *extremely*). They also reported how much information in the text they knew prior to reading it (i.e., prior knowledge quantity; 1 = *not very much* to 7 = *very much*), and how well they knew the subject matter in the text prior to reading it (i.e., prior knowledge quality; 1 = *not very well* to 7 = *very well*).

Integrative Argumentation Training

To ensure that all participants understood what was required of them in integrative argumentation, they were trained on this critical learning outcome. The integrative argumentation training materials (available in the online supplemental materials) were adapted from training protocols by Nussbaum (2008b), Nussbaum and Schraw (2007), and Wong and Lim (2019a). First, participants were presented with a handout that introduced and explained the strategies of refutation, weighing, and design claims. Each strategy was illustrated using the training topic of "Should candy be banned from school?" Participants were also shown a written sample of how all three strategies could be used in combination to form an integrative conclusion.

Next, participants practiced using the three strategies to write an integrative argumentation response on the topic of "Should the university lease personal parking spaces?," which was not related to either of the argumentative texts on "biotech" and "sex differences." To scaffold participants' response formulation, they were provided with an argumentation vee diagram (AVD; adapted from Novak & Gowin, 1984; Nussbaum, 2008b) that contained arguments for and against the topic. Shaped like a "V," the AVD is a graphic organizer that is intended to reduce cognitive load by facilitating the organization of arguments and counterarguments. Using the AVD with instruction on argumentation strategies has been found to enhance students' integrative argumentation (Nussbaum, 2008b). Figure 1 shows a sample completed AVD on the topic of "Should candy be banned from school?," with arguments for and against banning candy organized side by side to facilitate comparisons between both sides.

To further guide participants' use of the argumentation strategies, five "critical questions" (Nussbaum & Edwards, 2011; D. N. Walton, 1996) were included below the AVD: (a) "Are any of the arguments not as important as others?," (b) "Are any of the arguments unlikely?," (c) "Is there a creative solution to any problem raised?," (d) "Is the creative solution practical? (Consider costs.)," and (e) "For any argument, can you think of any examples to the contrary or other likely explanations?" Supplementing the AVD with critical questions has been found to support students' production of integrated arguments (Nussbaum & Edwards, 2011; Nussbaum & Putney, 2020; Nussbaum et al., 2019).

After participants had practiced formulating an integrative argumentation response, they were shown a handout with three sample answers ranging from low to high quality. Each sample answer was accompanied by an explanation of its strengths and/or weaknesses. Participants also received a "Criteria for a Good Argument" handout (adapted from Nussbaum & Schraw, 2007) that explained the qualities of good argumentation such as stating a clear position, providing supporting reasons, using counterargumentation, integrating arguments and counterarguments, and organizing one's answer.

Recall and Argumentative Reasoning Tests

During the recall and argumentative reasoning tests, all participants were provided with a blank AVD as a planning aid (available in the online supplemental materials).

Procedure

Figure 2 depicts a flowchart of the procedure. Before attending the experiment, all participants completed the online preexperiment questionnaire. Upon arriving for the experiment, participants underwent three experimental phases: practice, studying, and test. Participants were run in groups of up to eight per session in a lab setting and completed the experiment individually. To ensure that all participants understood what was required of them, the experimenter verbally reinforced all training and task instructions during the experiment. The total experimental duration was approximately 90 min.

Practice Phase

Participants were told that they would be studying an argumentative text and generating their own arguments on the topic. All participants were then trained on argumentation—they received a handout that introduced arguments and counterarguments using the sample topic of "Should candy be banned from school?," and that explained and provided examples of arguments that are good and plausible, or poor yet plausible, or poor and implausible. Participants were then given 5 min to practice using their randomly assigned learning method to generate "for" and "against" arguments on the topic of "Should candy be banned from school?" Specifically, the notetaking and correct teaching groups practiced generating good and plausible arguments (i.e., errorless learning), whereas the misteaching group practiced generating poor yet plausible arguments (i.e., deliberate erring). After the 5-min period, all participants received concise verbal feedback within a couple of sentences on whether their generated arguments fulfilled the given criteria and, if not, how they could modify their arguments appropriately with reference to the training examples (see Table 1).

Studying Phase

After the practice phase, participants were presented with an introductory paragraph on their randomly assigned argumentative text (either "biotech" or "sex differences"). They then completed the prestudying questionnaire.

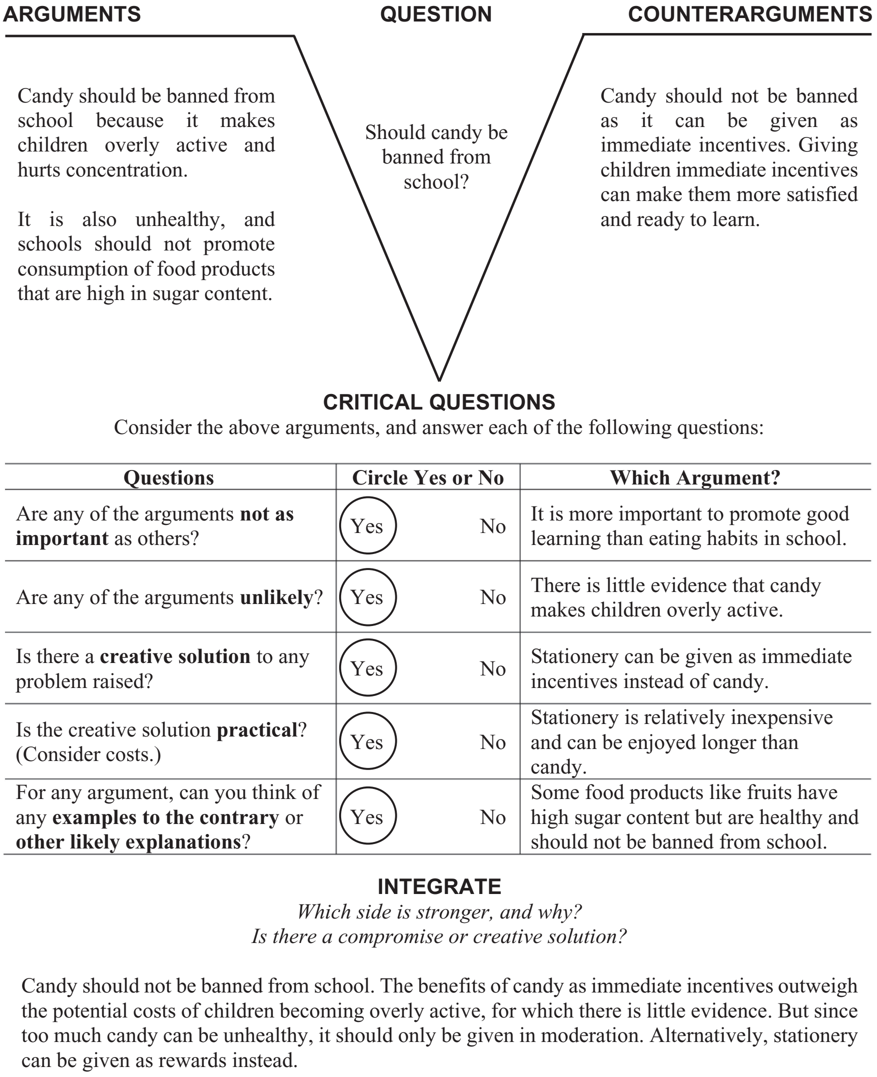
Next, participants began a 25-min studying period—all participants were first given 2 min to read their respective argumentative text, then studied the text using their randomly assigned learning method for 23 min. Thus, the total studying duration was exactly matched across all three learning conditions. The modality of studying (written format) was held constant across all methods to distill the unique effects of teaching and deliberate erring.

Notetaking participants were told that they would later be tested on the text content, and were given 15 min to write study notes about the topic in preparation for the test, while generating as many good and plausible "for" and "against" arguments of their own about the topic. The specific nature of the test was not divulged. After the 15-min period, the text was collected back and participants answered the JOL+ questions as a metacomprehension monitoring intervention. Then, they were given 8 min to refine and revise their study notes, as well as their self-generated arguments, to their best abilities.

In contrast, *correct teaching* participants were told that they would later be asked to teach the text content to their peers, and were given 15 min to write teaching notes in preparation to teach while generating as many good and plausible "for" and "against" arguments of

Figure 1

Sample Completed AVD With Critical Questions on "Should Candy Be Banned From School?"



ARGUMENTS

Candy should be banned from school because it makes children overly active and hurts concentration.

It is also unhealthy, and schools should not promote consumption of food products that are high in sugar content.

QUESTION

Should candy be banned from school?

COUNTERARGUMENTS

Candy should not be banned as it can be given as immediate incentives. Giving children immediate incentives can make them more satisfied and ready to learn.

CRITICAL QUESTIONS

Consider the above arguments, and answer each of the following questions:

Questions	Circle Yes or No	Which Argument?
Are any of the arguments **not as important** as others?	[✓ Yes] No	It is more important to promote good learning than eating habits in school.
Are any of the arguments **unlikely**?	[✓ Yes] No	There is little evidence that candy makes children overly active.
Is there a **creative solution** to any problem raised?	[✓ Yes] No	Stationery can be given as immediate incentives instead of candy.
Is the creative solution **practical**? (Consider costs.)	[✓ Yes] No	Stationery is relatively inexpensive and can be enjoyed longer than candy.
For any argument, can you think of any **examples to the contrary** or **other likely explanations**?	[✓ Yes] No	Some food products like fruits have high sugar content but are healthy and should not be banned from school.

INTEGRATE

Which side is stronger, and why?
Is there a compromise or creative solution?

Candy should not be banned from school. The benefits of candy as immediate incentives outweigh the potential costs of children becoming overly active, for which there is little evidence. But since too much candy can be unhealthy, it should only be given in moderation. Alternatively, stationery can be given as rewards instead.

Note. AVD = argumentation vee diagram. The sample completed AVD was created based on materials used by Nussbaum and Edwards (2011) and Wong and Lim (2019a).

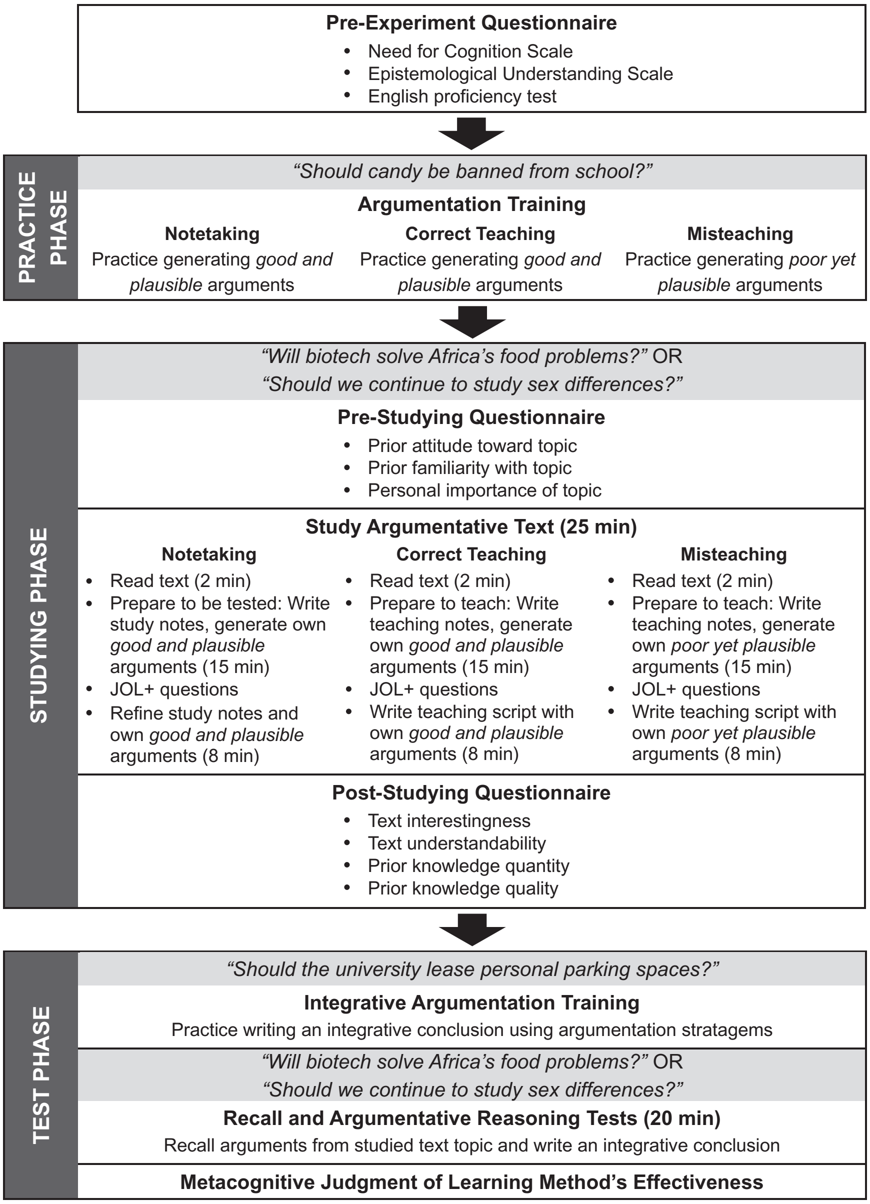
their own about the topic that their peers would be asked to learn. The specific nature of the teaching was not divulged. After the 15-min period, the text was collected back and participants answered the JOL+ questions. Then, they were given 8 min to teach "silently" with reference to their teaching notes. Specifically, they wrote a verbatim (i.e., word-for-word) teaching script on the text as how they would exactly deliver a lecture about the topic when orally teaching their peers (K. Y. L. Lim et al., 2021; S. W. H. Lim et al., 2024), while including as many good and plausible "for" and "against" arguments of their own about the topic. They were also provided with a brief lecture opening as a sample of a verbatim teaching script:

> Good day, everyone! In this lecture, we will learn about the controversy over whether [agricultural biotechnology should be introduced in Africa/scientists should continue to study sex differences in society]. I will be sharing some good and plausible arguments on this topic. As I teach, please learn these arguments. Let's begin ….

Likewise, *misteaching* participants were told that they would later be asked to teach the text content to their peers, and were given 15 min to write teaching notes in preparation to teach. However, participants were asked to deliberately err by generating as many poor yet plausible "for" and "against" arguments of their own about the topic that their peers would be asked to spot. As in the correct teaching condition, the exact nature of the teaching was not disclosed. After the 15-min period, the text was collected back and participants answered the JOL+ questions. They were then given 8 min to write a verbatim teaching script with reference to their teaching notes, while intentionally including as many poor yet plausible "for" and

Figure 2

Flowchart of Experimental Procedure


Pre-Experiment Questionnaire
• Need for Cognition Scale
• Epistemological Understanding Scale
• English proficiency test
PRACTICE PHASE
"Should candy be banned from school?"
Argumentation Training
Notetaking
Practice generating good and plausible arguments
Correct Teaching
Practice generating good and plausible arguments
Misteaching
Practice generating poor yet plausible arguments
STUDYING PHASE
"Will biotech solve Africa's food problems?" OR
"Should we continue to study sex differences?"
Pre-Studying Questionnaire
• Prior attitude toward topic
• Prior familiarity with topic
• Personal importance of topic
Study Argumentative Text (25 min)
Notetaking
• Read text (2 min)
• Prepare to be tested: Write study notes, generate own good and plausible arguments (15 min)
• JOL+ questions
• Refine study notes and own good and plausible arguments (8 min)
Correct Teaching
• Read text (2 min)
• Prepare to teach: Write teaching notes, generate own good and plausible arguments (15 min)
• JOL+ questions
• Write teaching script with own good and plausible arguments (8 min)
Misteaching
• Read text (2 min)
• Prepare to teach: Write teaching notes, generate own poor yet plausible arguments (15 min)
• JOL+ questions
• Write teaching script with own poor yet plausible arguments (8 min)
Post-Studying Questionnaire
• Text interestingness
• Text understandability
• Prior knowledge quantity
• Prior knowledge quality
TEST PHASE
"Should the university lease personal parking spaces?"
Integrative Argumentation Training
Practice writing an integrative conclusion using argumentation stratagems
"Will biotech solve Africa's food problems?" OR
"Should we continue to study sex differences?"
Recall and Argumentative Reasoning Tests (20 min)
Recall arguments from studied text topic and write an integrative conclusion
Metacognitive Judgment of Learning Method's Effectiveness

Note. JOL+ = judgments of higher order learning.

"against" arguments of their own about the topic. Participants were not required to refute their own poor arguments. A brief lecture opening was provided as a sample of a verbatim teaching script:

> Good day, everyone! In this lecture, we will learn about the controversy over whether [agricultural biotechnology should be introduced in Africa/scientists should continue to study sex differences in society]. I will be intentionally sharing some poor yet plausible arguments on this topic. As I teach, please spot these flawed arguments. Let's begin ….

Because the learning benefits of teaching have been attributed, at least in part, to retrieval practice when teaching unaided from memory (Koh et al., 2018), participants in both teaching groups were allowed to refer to their teaching notes when writing their teaching script, just as how notetaking participants could access their study notes when refining them. This procedure dissociated teaching from retrieval-based learning, ensuring that any observed learning benefits of teaching in the present study could not be due to retrieval practice. Across all three conditions, participants completed the studying phase independently without feedback on their responses. At the end of the 25-min studying period, all participants completed the poststudying questionnaire.

Test Phase

After the studying phase, participants were told that they would be writing an integrative response on the argumentative topic they had studied and that they would first be trained on using argumentation stratagems.[2] Participants received a training handout that explained and illustrated how refutation, weighing, and design claims could be used to formulate an integrative conclusion on the topic of "Should candy be banned from school?" Next, participants were given 5 min to practice using the argumentation stratagems to write an integrative conclusion on the topic of "Should the university lease personal parking spaces?" To guide participants' responses, they were provided with an AVD with critical questions and "for" and "against" arguments on the topic. After completing their practice responses, participants were shown the "Criteria for a Good Argument" handout, as well as a handout with three sample integrative conclusions and accompanying explanations of their strengths and/or weaknesses.

After the training, participants began the recall and argumentative reasoning tests. They were first asked to recall the "for" and "against" arguments on the topic that they had studied (either "Will biotech solve Africa's food problems?" or "Should we continue to study sex differences?"), and to write them down on a blank AVD that was provided as a planning device. Then, participants applied the argumentation stratagems they had learned during training to write an integrative conclusion on the topic (see sample responses in Appendix A). They were reminded to include as much information as possible and to incorporate all stratagems taught for strong argumentation (full instructions are available in the online supplemental materials). Participants were given 20 min to complete both tests without reference to the argumentative text, although they were allowed to access the "Criteria for a Good Argument" handout. They were also told that if they finished before the time was up, they should spend the remaining time reviewing their response.

After completing the tests, participants rated the effectiveness of their learning method for their test performance on a 7-point scale (1 = *not at all* to 7 = *extremely*). Finally, participants provided their demographic information and were debriefed and thanked.

Scoring Procedure

Interrater Reliability

Two raters independently scored 70 of the 208 (33%) scripts. Interrater reliability was high for all scored variables, all absolute agreement intraclass correlations (ICCs) ≥ .90 based on a two-way random-effects model (see Table 2). Discrepancies between both raters were reviewed and resolved to yield 100% agreement. Given the high interrater agreement, the remaining responses were scored by one rater.

Recall Performance

Participants' recall test performance was scored by awarding one point for each idea unit from the argumentative text that they correctly recalled. For scoring purposes, each text was decomposed into 64 idea units (available in the online supplemental materials). A sample idea unit in the "biotech" text was "A biotech fix would be costly for the farmer"; a sample idea unit in the "sex differences" text was "Genders need not be understood through dichotomous opposition." Both verbatim restatements and paraphrases that preserved the meaning of the idea units were considered correct. As an example, for the idea unit, "A biotech fix would be costly for the farmer," an acceptable paraphrased response was "Biotech would be expensive for farmers," whereas an inadequate response was "Biotech is costly."

Table 2
Interrater Reliability

Variable	ICC	95% CI
Recall performance	.96	[0.83, 0.99]
Argumentative reasoning		
Integrative argumentation performance	.90	[0.84, 0.94]
Holistic argumentation quality	.92	[0.87, 0.95]
Self-generated arguments	.96	[0.94, 0.98]
Self-other referential terms	.999	[0.998, 0.999]
Elaborations	.90	[0.81, 0.94]
Monitoring statements	1.00	

Note. There was perfect interrater absolute agreement in scoring the number of monitoring statements. ICC = intraclass correlation; CI = confidence interval.

Argumentative Reasoning

Integrative Argumentation Performance. Based on extant scoring procedures (e.g., Nussbaum & Edwards, 2011; Nussbaum et al., 2019; Wong & Lim, 2019a), participants' integrative argumentation performance was scored by awarding one point for each instance that they used an integrative stratagem (i.e., weighing or design claim) in their test responses. Table 3 provides descriptions and examples of integrative stratagems.

Holistic Argumentation Quality. In addition, the overall quality of participants' argumentative reasoning on the test was scored using a 7-point holistic scoring rubric (available in Appendix B). The rubric was based on those used by Anmarkrud et al. (2014), Reznitskaya et al. (2009), and Nussbaum and Schraw (2007). Aligning with the "Criteria for a Good Argument" handout that participants received, the holistic scoring scale reflected how developed participants' integrative conclusions were based on the clarity of position, elaboration and support of arguments with reasons, discussion of alternative perspectives and counterarguments, and overall essay organization.

Responses that received lower holistic scores between 1 and 4 tended to be poorly organized and were differentiated based on whether they contained a clear position on the issue, the number of reasons presented to support the position, and whether they mentioned alternative perspectives. For example, responses that did not contain a position received a score of 1, whereas those that contained a position supported by at least four distinct reasons and that mentioned (but did not discuss) alternative perspectives could qualify for a score of 4. Responses that received higher holistic scores between 5 and 7 were further differentiated based on the extent that they were well-organized and developed in discussing alternative perspectives for strong integrative argumentation.

Process Measures

To explore participants' learning processes and the characteristics of their studying phase responses, their study notes and teaching

[2] Although it was technically possible to combine the argumentation and integrative argumentation trainings during the practice phase, they were conducted as separate segments to avoid imposing excessive cognitive load in this relatively complex task (e.g., if participants had to simultaneously learn how to generate arguments and integrate them). Whether training outcomes are optimized when students learn integrative argumentation stratagems as a follow-up to or prerequisite for generating good versus poor arguments is an open question that future work can address by directly manipulating the sequence of both trainings.

Table 3
Descriptions and Examples of Integrative Stratagems

		Sample responses	
Integrative stratagem	Description	“Biotech”	“Sex differences”
Weighing	Contrasting the relative merits of an argument or counterargument	“Although biotech is costly, the initial investment outlay will be outweighed by the long-term economic benefits. By increasing yield, the economy as a whole has an opportunity to strengthen and diversify itself. The increase in agricultural productive capacity could open up secondary markets in value-added production, creating more profitable livelihoods for locals.”	“The advantages of continuing the study of sex differences are greater than the social and political repercussions, which can be mitigated by anti-discriminatory policies and regulation. The study of sex differences will benefit us more in the long-run because life-saving medical knowledge from such studies can help produce effective healthcare treatments for both sexes.”
Design claims	Developing an in-between position that combines the merits of both sides	“Biotech can be implemented to boost yield, while tackling the structural issue of converting the crops into profits. The idea of traditional farm cooperatives can be synergized with modern biotech methods. By pooling their resources, farmers can benefit from economies of scale to spread out the cost of biotech, gain marketing and distribution advantages, and reap greater profits.”	“We can come to a compromise by acknowledging the presence of sex differences, while adopting a triangulation approach in future research that considers biological, sociocultural, and political factors to study sex differences in a more nuanced and holistic way.”
	Suggesting alternative solutions	“Biotech companies can be encouraged to start their own farms instead. It would be in the interest of such profit-driven firms to ensure that the crops produced reach the market, and to leverage on their intellectual property to increase food production. They will then need to enlist the help of local farmers to work on their plantations and pay them a fair wage. This will increase the overall food production in Africa and alleviate the poverty of resource-poor farmers.”	“To address gender inequality and create real social change, we should focus our efforts on how scientific knowledge is used. Instead of trying to stop all research on sex differences, we should call out the distortion or misuse of such data to advance harmful causes, oppress others, or propagate any forms of hate speech towards any groups in society.”

scripts were scored on the number of (a) self-generated arguments, (b) self-other referential terms, (c) elaborations, and (d) monitoring statements. The last three variables served as measures of participants’ teaching quality.

Self-generated arguments were participants’ own arguments expressing a main idea that had not been put forth in the text. One point was awarded for each novel argument, which could be expressed either in a single or multiple sentences. A sample self-generated argument for the “biotech” text was:

> Some genetically modified food may cause allergic reactions that are not in the present literature that we know about. In the worst-case scenario, some of these adverse effects may be deadly and hence biotech agricultural products may have serious ethical implications.

A sample self-generated argument for the “sex differences” text was: “We should continue to study sex differences as such differences may give rise to different medical conditions that may also require different forms of treatment.”

Self-other referential terms included words such as “I,” “me,” “you,” “us,” “let’s,” “our,” “ourselves,” “we,” “your,” and “yourself.” The number of instances that participants used such terms served as a proxy for perceived social presence (e.g., Hoogerheide et al., 2016; Jacob et al., 2020, 2021; Lachner et al., 2018; K. Y. L. Lim et al., 2021).

Elaborations were statements on the text’s arguments and counterarguments that participants related to their prior knowledge, such as generating personal examples, analogies, and inferences that were not explicitly stated in the text (e.g., Fiorella & Kuhlmann, 2020; Jacob et al., 2020, 2021; Lachner et al., 2018, 2020; K. Y. L. Lim et al., 2021). *Monitoring statements* were instances where participants monitored understanding, evaluated correctness, or indicated content that was worth paying attention to based on importance or interest (e.g., Fiorella & Kuhlmann, 2020; Lachner et al., 2020; Roscoe, 2014). Together, elaborations and monitoring statements have been considered elements of reflective knowledge-building that promote tutors’ learning (Roscoe & Chi, 2007). Sample elaborations and monitoring statements are presented in Table 4.

Results

Preliminary Analyses

One-way between-subjects analyses of variance (ANOVAs) were conducted to ascertain that the three learning groups did not differ in their responses on the preexperiment questionnaire, prestudying questionnaire, JOL+ questions, and poststudying questionnaire. Table 5 shows the means and standard deviations.

Preexperiment Questionnaire

At baseline, the three learning groups did not significantly differ in their mean need for cognition scores, $F(2, 205) = 0.06$, $p = .94$, $\eta^2 = .001$, levels of epistemological understanding, $F(2, 205) = 1.09$, $p = .34$, $\eta^2 = .01$, and English proficiency scores, $F(2, 205) = 1.70$, $p = .19$, $\eta^2 = .02$. Thus, these variables were not considered in subsequent analyses.

Prestudying Questionnaire

Likewise, across learning conditions, participants did not significantly differ in their prior attitude toward the argumentation topic,

Table 4
Sample Elaborations and Monitoring Statements

Statement type	Sample responses
Elaborations	
Examples	"For instance, crops can be genetically modified to provide them with resistance against certain kinds of diseases." "For example, there are more male political leaders than female political leaders in the world today."
Analogies	"Contrary to the common belief that agricultural biotechnology is akin to introducing a foreign concept to an indigenous group…" "It's just like studying the difference between young people and the aged, the difference between two animals, the difference between two fruits."
Inferences	"Agricultural biotechnology can solve challenges to crop production such as diseases, pests, and weeds, *so this can save costs buying pesticides and costs of removing the weeds or disease-ridden crops* [emphasis added]." "Research has shown that the stereotype of women is generally more positive compared to men, *which would therefore not be harmful to the feminist movement* [emphasis added]."
Monitoring statements	
Monitoring understanding	"As you know…" "I don't really understand the logic."
Evaluating correctness	"Is this true?" "… is scientifically proven, I think."
Directing attention	"It is important to know this." "This is interesting."

$F(2, 205) = 0.53$, $p = .59$, $\eta^2 = .01$, prior familiarity with the topic, $F(2, 205) = 0.06$, $p = .95$, $\eta^2 = .001$, and perceived personal importance of the topic, $F(2, 205) = 0.02$, $p = .98$, $\eta^2 < .001$.

JOL+ Questions

The JOL+ questions were administered as a metacomprehension monitoring intervention to guide all participants toward the higher order learning outcome of argumentative reasoning. Nevertheless, for completeness, it was ascertained that the learning groups did not significantly differ in their mean JOL+ ratings, $F(2, 205) = 1.45$, $p = .24$, $\eta^2 = .01$.

Poststudying Questionnaire

Participants across all conditions reported relatively low prior knowledge of the argumentative texts' content, with no significant differences in their prior knowledge quantity, $F(2, 205) = 2.09$, $p = .13$, $\eta^2 = .02$, and quality, $F(2, 205) = 1.48$, $p = .23$, $\eta^2 = .01$. An unexpected finding was that after the studying phase, the learning groups differed in their perceptions of the text's interestingness, $F(2, 205) = 5.84$, $p = .003$, $\eta^2 = .05$, and understandability, $F(2, 205) = 3.79$, $p = .02$, $\eta^2 = .04$. Specifically, the misteaching group rated the text as less interesting and understandable than the notetaking group, $p = .002$ and $.03$, $d = -0.52$ and -0.38, and the correct teaching group, $p = .005$ and $.01$, $d = -0.46$ and -0.41, respectively. The notetaking and correct teaching groups did not significantly differ in how interesting or understandable they perceived the text to be, $p = .77$ and $.76$, $d = -0.05$ and 0.05, respectively.

Table 5
Means and Standard Deviations for Preexperiment, Prestudying, and Poststudying Questionnaires and JOL+ Questions

Variable	Notetaking		Correct teaching		Misteaching	
	M	*SD*	*M*	*SD*	*M*	*SD*
Preexperiment questionnaire						
Need for cognition	3.05	0.63	3.06	0.52	3.08	0.56
Epistemological understanding	33.44	4.44	34.41	4.61	33.43	4.54
English proficiency	3.56	1.82	3.59	1.94	3.09	1.54
Prestudying questionnaire						
Prior attitude toward topic	5.11	0.92	4.96	0.75	5.01	0.97
Prior familiarity with topic	2.48	1.11	2.41	1.22	2.44	1.08
Personal importance of topic	3.43	1.63	3.37	1.57	3.38	1.68
JOL+	46.41	15.57	44.83	13.51	42.38	12.67
Poststudying questionnaire						
Text interestingness	4.37	1.30	4.30	1.41	3.62	1.57
Text understandability	4.50	1.27	4.57	1.38	3.99	1.43
Prior knowledge quantity	2.63	1.48	2.71	1.48	2.25	1.29
Prior knowledge quality	2.60	1.46	2.36	1.39	2.21	1.20

Note. $N = 208$. JOL+ = judgments of higher order learning.

Main Analyses

Recall Performance

To analyze participants' recall performance, a 3 (learning method) × 2 (argumentation topic) between-subjects ANOVA was conducted with the total idea units that participants recalled from the argumentative text as the dependent variable. As predicted, there was a learning-by-teaching effect on recall performance, $F(2, 202) = 3.97$, $p = .02$, $\eta_p^2 = .04$. Specifically, the correct teaching ($M = 5.64$, $SD = 4.98$) and misteaching ($M = 5.82$, $SD = 4.07$) groups outperformed the notetaking group ($M = 4.17$, $SD = 2.95$), $p = .026$ and $.01$, $d = 0.36$ and 0.46, respectively. Both teaching groups did not significantly differ in their recall performance, $p = .72$, $d = 0.04$. Thus, teaching the argumentative text—whether "correctly" with good and plausible arguments or deliberately "incorrectly" with poor yet plausible arguments—improved recall of the material, relative to writing study notes while generating good and plausible arguments (Figure 3A).

Overall, participants recalled more idea units from the "biotech" ($M = 6.32$, $SD = 4.26$) than "sex differences" ($M = 4.11$, $SD = 3.71$) text, $F(1, 202) = 16.85$, $p < .001$, $\eta_p^2 = .08$. Nevertheless, the Learning Method × Argumentation Topic interaction was nonsignificant, $F(2, 202) = 0.08$, $p = .93$, $\eta_p^2 = .001$, indicating that the learning-by-teaching advantage held reliably across topics.

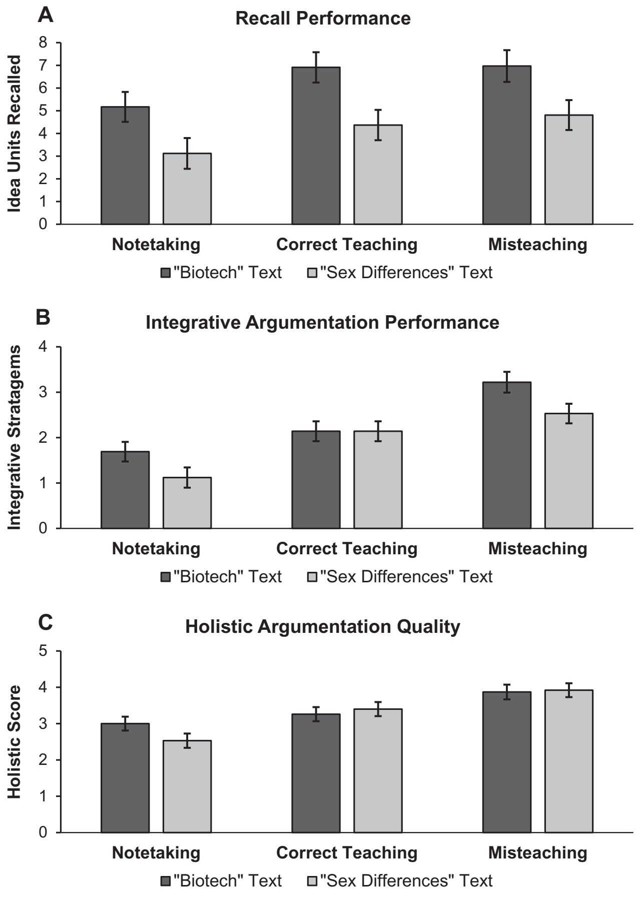
Argumentative Reasoning

Participants' integrative conclusions in their test responses contained an average of 238.68 words ($SD = 96.79$). There was no significant difference in the word count of participants' integrative conclusions across the notetaking ($M = 238.13$, $SD = 111.15$), correct teaching ($M = 245.43$, $SD = 87.32$), and misteaching ($M = 232.31$, $SD = 90.99$) conditions, $F(2, 205) = 0.32$, $p = .73$, $\eta^2 = .003$.

Integrative Argumentation Performance. A 3 (learning method) × 2 (argumentation topic) between-subjects ANOVA was conducted to analyze the number of integrative stratagems that participants used at test. As with their recall performance, there was a significant learning-by-teaching effect on their integrative argumentation performance, $F(2, 202) = 22.00$, $p < .001$, $\eta_p^2 = .18$. The correct teaching ($M = 2.14$, $SD = 1.03$) and misteaching ($M = 2.85$, $SD = 1.68$) groups successfully used more integrative stratagems than the notetaking group ($M = 1.41$, $SD = 1.16$), $p = .001$ and $p < .001$, $d = 0.67$ and 1.00, respectively. Importantly, the misteaching group outperformed the correct teaching group, $p = .001$, $d = 0.51$. Thus, teaching others benefited the tutor's integrative argumentation performance more than writing study notes for their own learning, with an additional benefit from deliberately incorrect teaching than correct teaching (Figure 3B).

Figure 3

Performance on Recall and Argumentative Reasoning Tests



Note. (A) The mean number of idea units recalled on the recall test. (B) and (C) The mean argumentative reasoning test scores, as assessed by the number of integrative stratagems used at the microlevel and holistic argumentation quality at the macrolevel, respectively. Error bars indicate standard errors.

Overall, participants used more integrative stratagems for the "biotech" ($M = 2.32$, $SD = 1.59$) than "sex differences" ($M = 1.94$, $SD = 1.24$) text, $F(1, 202) = 5.51$, $p = .02$, $\eta_p^2 = .03$. However, the Learning Method × Argumentation Topic interaction was nonsignificant, $F(2, 202) = 1.41$, $p = .25$, $\eta_p^2 = .01$, indicating that the advantage of misteaching persisted across both topics.

At the same time, participants' integrative argumentation performance significantly and positively correlated with their recall performance, $r(206) = .15$, $p = .03$. Hence, an arising question was whether the teaching groups' better argumentative reasoning was simply driven by their better recall of the material. To address this question, participants' recall performance was added as a covariate in a 3 (learning method) × 2 (argumentation topic) analysis of covariance (ANCOVA) with their integrative argumentation performance as the dependent variable. To first check the homogeneity of slopes assumption, interactions between the covariate and independent variables were entered into the model, alongside the main effects. None of the interactions were significant, all $ps > .05$, indicating that the homogeneity of slopes assumption had been met. The interaction terms were then removed from the model, which was reestimated. The ANCOVA revealed that the benefit of misteaching persisted even after controlling for the total idea units that participants recalled, $F(2, 201) = 20.41$, $p < .001$, $\eta_p^2 = .17$. Controlling for participants' recall performance, the misteaching group ($M_{\text{adjusted}} = 2.86$) used significantly more integrative stratagems than the correct teaching ($M_{\text{adjusted}} = 2.14$) and notetaking ($M_{\text{adjusted}} = 1.42$) groups, $p = .001$ and $p < .001$, respectively. In addition, the correct teaching group outperformed the notetaking group, $p = .002$. This suggests that the teaching groups' better integrative argumentation performance was not merely due to better recall per se.

Holistic Argumentation Quality. Similarly, a 3 (learning method) × 2 (argumentation topic) between-subjects ANOVA was conducted to analyze participants' holistic argumentation quality scores. Once again, there was a significant learning-by-teaching effect, $F(2, 202) = 16.73$, $p < .001$, $\eta_p^2 = .14$. Both the correct teaching ($M = 3.33$, $SD = 0.97$) and misteaching ($M = 3.90$, $SD = 1.44$) groups displayed higher overall argumentation quality than the notetaking group ($M = 2.77$, $SD = 0.98$), $p = .004$ and $p < .001$, $d = 0.57$ and 0.92, respectively. Moreover, echoing participants' integrative argumentation performance, the misteaching group outperformed the correct teaching group, $p = .004$, $d = 0.47$. Thus, students' holistic argumentation quality benefited more from writing a verbatim teaching script than study notes about an argumentative text, with an additional benefit from deliberately incorrect than correct teaching (Figure 3C).

Overall, participants' holistic argumentation quality did not differ across the "biotech" ($M = 3.36$, $SD = 1.19$) and "sex differences" ($M = 3.30$, $SD = 1.28$) texts, $F(1, 202) = 0.36$, $p = .55$, $\eta_p^2 = .002$. Neither was there a Learning Method × Argumentation Topic interaction, $F(2, 202) = 1.43$, $p = .24$, $\eta_p^2 = .01$.

Participants' holistic argumentation quality was significantly and positively associated with their integrative argumentation performance, $r(206) = .59$, $p < .001$, but not their recall performance, $r(206) = .10$, $p = .15$. Hence, at the microlevel, using more integrative stratagems, but not recalling more idea units from the argumentative text, was linked to higher overall quality of participants' argumentative reasoning at the macrolevel in their clarity of position, elaboration and support of arguments with reasons, discussion of alternative perspectives and counterarguments, and overall essay organization.

Metacognitive Judgments

In contrast to their test performance, participants' metacognitive ratings after the tests revealed that they inaccurately judged the effectiveness of the learning methods. Participants' effectiveness ratings did not significantly differ across the notetaking ($M = 4.16$, $SD = 1.34$), correct teaching ($M = 4.23$, $SD = 1.22$), and misteaching ($M = 3.84$, $SD = 1.31$) conditions, $F(2, 205) = 1.79$, $p = .17$, $\eta^2 = .02$. Thus, participants failed to recognize that (mis)teaching had helped their learning, even after having just experienced its benefits for their test performance.

Process Measures

To examine the characteristics and quality of participants' studying phase responses, the number of self-generated arguments, self-other referential terms, elaborations, and monitoring statements in their study notes and teaching scripts were analyzed. Table 6 displays the means and standard deviations. Overall, participants produced few self-other referential terms, elaborations, and monitoring statements.

Self-Generated Arguments

Across learning groups, there was no significant difference in the number of arguments that participants generated during the studying phase, $F(2, 205) = 1.02$, $p = .36$, $\eta^2 = .01$. This suggests that the learning-by-teaching and derring effects observed are not due to participants generating more or fewer arguments of their own.

Teaching Quality

Social Presence. As a proxy for perceived social presence, the number of self-other referential terms in participants' study notes and teaching scripts was analyzed, revealing a significant difference across learning groups, $F(2, 205) = 6.14$, $p = .003$, $\eta^2 = .06$. The correct teaching and misteaching groups used more self-other referential terms than the notetaking group, $p = .002$ and $.004$, $d = 0.56$ and 0.49, respectively. Both teaching groups did not significantly differ, $p = .85$, $d = -0.03$. Thus, writing a correct or an incorrect verbatim teaching script triggered greater perceived social presence than writing study notes.

Elaborations. Likewise, the learning groups differed in the number of elaborations in their studying phase responses, $F(2, 205) = 6.25$, $p = .002$, $\eta^2 = .06$. The correct teaching and misteaching groups generated more elaborations than the notetaking group, $p = .002$ and $.004$, $d = 0.60$ and 0.47, respectively. Both teaching groups did not significantly differ in their number of elaborations, $p = .76$, $d = -0.05$. Hence, writing a correct or an incorrect verbatim teaching script induced more generative processing than writing study notes.

Monitoring Statements. The notetaking group did not produce any metacognitive monitoring statements in their study notes; monitoring statement scores were nonnormally distributed with skewness of 3.44 ($SE = 0.17$) and kurtosis of 12.89 ($SE = 0.34$). Thus, nonparametric bootstrapping with 10,000 samples was applied to robustly estimate the standard errors and 95% bias-corrected and accelerated confidence intervals (BCa CI) for the mean differences between the learning groups. The correct teaching and misteaching groups generated more monitoring statements than the notetaking group, $M_{\text{difference}} = 0.21$ and 0.24, bootstrap $SE = 0.06$ and 0.07, 95% BCa CI [0.10, 0.34] and [0.12, 0.37], respectively. Both teaching groups did not differ in their number of monitoring statements, $M_{\text{difference}} = 0.02$, bootstrap $SE = 0.09$, 95% BCa CI [−0.16, 0.21].

Teaching Quality and Test Performance

Correlations. To explore whether the teaching quality measures were associated with test performance, correlational analyses were run. The number of elaborations and monitoring statements in participants' study notes and teaching scripts positively correlated with their holistic argumentation quality at test, $r(206) = .21$ and $.14$, $p = .002$ and $.048$, respectively. However, social presence was not significantly associated with holistic argumentation quality, $r(206) = .03$, $p = .67$. Neither were there any significant correlations between the three teaching quality measures—social presence, elaborations, and monitoring statements—and recall performance, $r(206) = -.12$, $.05$, and $-.04$, $p = .09$, $.49$, and $.57$, respectively, nor integrative argumentation performance, $r(206) = -.01$, $.02$, and $.10$, $p = .85$, $.73$, and $.14$, respectively.

Mediation Analyses. Accordingly, regression analyses were conducted to test whether elaborations and monitoring statements (i.e., reflective knowledge-building; Roscoe & Chi, 2007) operating in parallel mediated the teaching groups' superior holistic argumentation quality over the notetaking group, as would be predicted by the generative hypothesis of learning by teaching (Fiorella &

Table 6

Means and Standard Deviations for Number of Self-Generated Arguments, Self-Other Referential Terms, Elaborations, and Monitoring Statements

	Notetaking		Correct teaching		Misteaching	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*
Self-generated arguments	3.40	1.39	3.13	1.65	3.53	1.97
Teaching quality						
Self-other referential terms (perceived social presence)	1.77	3.25	3.61	3.36	3.50	3.82
Elaborations	0.70	1.05	1.36	1.13	1.29	1.43
Monitoring statements	0.00	0.00	0.21	0.54	0.24	0.55

Mayer, 2016). Following guidelines from Hayes and Preacher (2014), a percentile bootstrap estimation approach with 10,000 samples was used in Model 4 of Hayes' (2013) PROCESS macro. The multicategorical predictor learning method was dummy coded with notetaking as the reference group. In mediation analyses with a multicategorical predictor, evidence that at least one relative indirect effect differs from zero supports the conclusion that a mediator variable mediates the effect of the predictor on the outcome (Hayes & Preacher, 2014).

Figure 4 depicts the results of the parallel mediation analyses. With elaborations as a mediator, the relative indirect effects of learning method on holistic argumentation quality were significant for both correct teaching versus notetaking, 0.09, 95% CI [0.003, 0.22], and misteaching versus notetaking, 0.08, 95% CI [0.001, 0.21]. However, with monitoring statements as a mediator, there were no significant relative indirect effects of learning method on holistic argumentation quality for both correct teaching versus notetaking, 0.03, 95% CI [−0.07, 0.11], and misteaching versus notetaking, 0.03, 95% CI [−0.06, 0.13]. Thus, although writing verbatim teaching scripts induced more reflective knowledge-building—elaborations and monitoring statements—than writing study notes, only the number of elaborations mediated the benefit of learning by teaching for students' holistic argumentation quality.

Discussion

Without argument and critique, the construction of reliable knowledge that survives scrutiny would be impossible (J. Osborne, 2010). To promote students' argumentative reasoning, the present study tested a novel intervention—learning by misteaching—that combined the potent techniques of teaching others and deliberate erring. The results provided evidence for both the learning-by-teaching and derring effects.

Relative to writing study notes in preparation for a test, preparing to teach others an argumentative text and then actually teaching by writing a verbatim teaching script improved not only students' recall of the material but also their argumentative reasoning. Specifically, students who had taught were later more successful in integrating arguments and counterarguments about the topic, such as weighing arguments on both sides and forming design claims that developed a compromise or new alternative solution. Besides their use of integrative stratagems at the microlevel, students' holistic argumentation quality at the macrolevel benefited more from writing teaching scripts than study notes. These benefits occurred whether students taught correctly with good arguments or incorrectly with deliberately weak arguments they had generated. Together, these findings extend learning-by-teaching effects to the complex, higher order outcome of argumentative reasoning, beyond extant studies' predominant focus on tutors' basic recall and comprehension (e.g., Fiorella & Mayer, 2013, 2014; Hoogerheide et al., 2019b; Jacob et al., 2020; K. Y. L. Lim et al., 2021) or transfer (e.g., Hoogerheide et al., 2014, 2016, 2019a; Lachner et al., 2018).

Crucially, deliberately incorrect teaching produced further gains over correct teaching for the tutor's argumentative reasoning. This advantage was not driven by better memory of the text content per se, persisting even when the tutor's recall performance was controlled for. Indeed, both correct teaching and misteaching groups did not significantly differ in their superior recall over the notetaking group. At the same time, this finding should not be taken to mean that deliberate erring cannot enhance memory. Extant studies on the derring effect have consistently shown that deliberately generating conceptual errors improves recall of scientific texts and concepts more than errorless learning (Wong, 2023; Wong & Lim, 2022a, 2022b). Rather, based on transfer-appropriate processing, performance is enhanced when the processing stimulated during initial acquisition is appropriate for the processing demands of the criterial test (Morris et al., 1977). Hence, the nature of deliberate erring in this study—intentionally generating poor yet plausible arguments—likely oriented students toward the cognitive processes needed specifically for successful argumentative reasoning, without necessarily conferring additional recall gains.

Despite the benefits of teaching and deliberate erring, these went largely unappreciated in students' metacognitive judgments. Even after having just experienced the techniques' effects on their test performance, the three learning groups did not differ in their perceptions of how effective the techniques had been. This suggests that test experience alone is not necessarily sufficient to promote accurate knowledge about learning techniques, as observed in extant studies on learning by teaching (e.g., Fiorella & Mayer, 2013, 2014) and deliberate erring (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024).

Theoretical Explanations for Learning by Misteaching

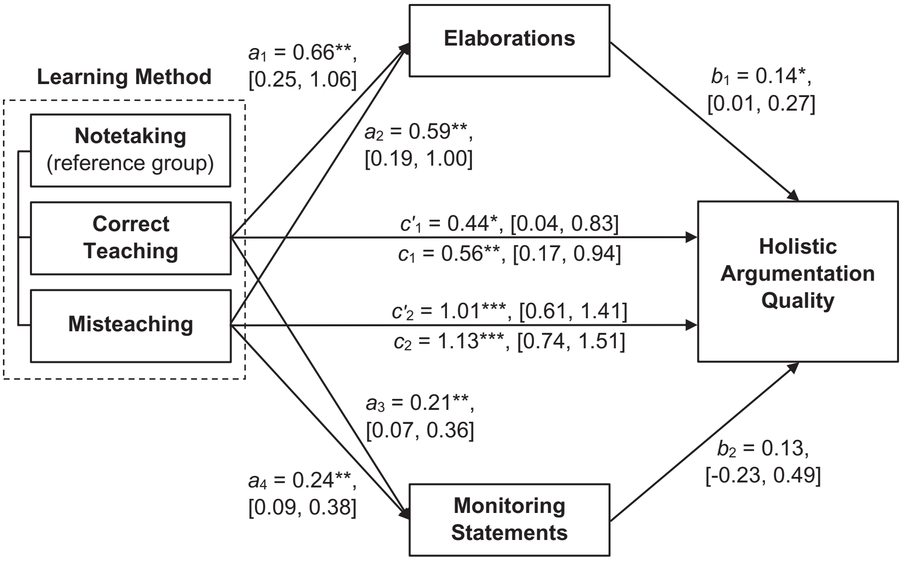
It is likely that teaching and deliberate erring each evoke distinct processes that uniquely contribute to the benefits of learning by misteaching, although this does not negate the possibility that they may also interact synergistically such that misteaching as a whole is greater than the sum of its parts. The present paradigm renders some candidate mechanisms less plausible while offering supporting evidence for others. First, the learning-by-teaching benefit observed here cannot be attributed to retrieval practice when tutors teach from memory (Koh et al., 2018), since all students accessed the argumentative text when writing their notes, and then accessed their notes when refining them or writing a teaching script. Moreover, whereas students may choose to invest more time studying when they expect to teach than take a test (Tauber et al., 2022), studying time was equated across all conditions, as was the studying modality (written format).

Rather, students are more likely to engage in beneficial generative processes when preparing to teach and actually teaching (Fiorella & Mayer, 2016). In taking on the role of a teacher, students may enact behaviors that they perceive as defined by this role. For instance, when bearing an intended audience in mind and anticipating their audience's learning needs, tutors may select relevant information, organize it, and integrate it with their prior knowledge when teaching (Bargh & Schul, 1980; Fiorella & Mayer, 2016), while monitoring their own understanding to repair knowledge gaps (Roscoe & Chi, 2007). Conversely, egocentric generative activities such as writing study notes for one's own learning may not trigger such processes or do so to lesser degrees, particularly when students adopt a knowledge-telling bias in summarizing the material with little elaboration (Roscoe & Chi, 2007).

Indeed, both teaching groups in this study used more self-other referential terms than the notetaking group, implicating higher social presence in perceiving their audience as "real" (e.g., Hoogerheide et al., 2016). However, higher social presence was not associated with better recall and argumentative reasoning test performance,

Figure 4

Results of Parallel Mediation Analyses


Learning Method
Notetaking (reference group)
Correct Teaching
Misteaching
Elaborations
Monitoring Statements
Holistic Argumentation Quality
a₁ = 0.66**, [0.25, 1.06]
a₂ = 0.59**, [0.19, 1.00]
a₃ = 0.21**, [0.07, 0.36]
a₄ = 0.24**, [0.09, 0.38]
b₁ = 0.14*, [0.01, 0.27]
b₂ = 0.13, [-0.23, 0.49]
c'₁ = 0.44*, [0.04, 0.83]
c₁ = 0.56**, [0.17, 0.94]
c'₂ = 1.01***, [0.61, 1.41]
c₂ = 1.13***, [0.74, 1.51]

Note. Parallel mediation model with two mediators (elaborations and monitoring statements), learning method as the multicategorical predictor (dummy coded with notetaking as the reference group), and holistic argumentation quality as the outcome. Unstandardized regression coefficients are presented for each path. The *a* path coefficients represent differences on the mediators between each teaching group and the notetaking group. The *b* path coefficients represent effects of each mediator on holistic argumentation quality, controlling for learning method and the other mediator. The *c*′ path coefficients represent the relative direct effects of correct teaching and misteaching on holistic argumentation quality, relative to notetaking, controlling for both mediators. The *c* path coefficients represent the relative total effects of correct teaching and misteaching on holistic argumentation quality, relative to notetaking. Values in brackets represent the 95% confidence intervals for the regression coefficients using a percentile bootstrap estimation approach with 10,000 samples.

* $p < .05$. ** $p < .01$. *** $p < .001$.

thus offering limited evidence for the social presence hypothesis of learning by teaching (e.g., Jacob et al., 2021; Lachner et al., 2018).

On the other hand, the present data lend some support to the generative hypothesis of learning by teaching (Fiorella & Mayer, 2016). Both teaching groups produced more elaborations and monitoring statements than the notetaking group, indicating that teaching stimulates generative and metacognitive processes that have collectively been considered reflective knowledge-building (Roscoe & Chi, 2007). In turn, these knowledge-building processes may support tutors' construction of a more elaborate situation model of the topic (Kintsch, 1988) that aids their argumentative reasoning. Indeed, the mediation analyses in this study found that the teaching groups' greater number of elaborations mediated their superior holistic argumentation quality over the notetaking group (e.g., Fiorella & Kuhlmann, 2020; Lachner et al., 2018). Although metacognitive monitoring is also an important element of knowledge-building, it is possible that it was not a significant mediator here because participants produced rather few monitoring statements on overall (Fiorella & Kuhlmann, 2020).

Besides the knowledge-building processes that teaching induces, deliberately erring by producing weak arguments may better prepare tutors to use integrative stratagems for a further boost in their argumentative reasoning. As the present data show, the advantage of misteaching over correct teaching cannot be attributed to increased levels of the same processes associated with learning by teaching—social presence, elaboration, and metacognitive monitoring—since both teaching groups did not significantly differ on these counts. Rather, deliberately generating poor yet plausible arguments, as opposed to good arguments, may prompt tutors to consider why some arguments fail to carry the day. When intentionally seeking out and preempting such flawed reasoning, tutors may build a richer understanding of what a good argument is (not) and avoid pitfalls for better argumentation (e.g., Gartmeier et al., 2008; Oser & Spychiger, 2005), as when rejecting weaker arguments in favor of stronger ones when weighing both sides of an issue, and developing design claims that mitigate the weaknesses of both sides while synthesizing their merits. This view also aligns with findings from refutation text research that students revise their misconceptions and learn concepts better by studying texts that explain why a wrong idea is wrong, as opposed to texts that explain why a right idea is right (e.g., Broughton et al., 2010; Hynd & Alvermann, 1986; for reviews, see Schroeder & Kucera, 2022; Tippett, 2010; Zengilowski et al., 2021). Presumably, when incorrect and correct information are coactivated and integrated into a mental network or representation (Kendeou, 2024; Kendeou & O'Brien, 2014), students are more likely to experience cognitive conflict that could trigger additional processing for deeper learning and conceptual change (Kendeou et al., 2019; Limón, 2001).

More broadly, the misteaching advantage converges with theories of failure-driven learning that highlight how encountering and reflecting on errors may catalyze learning (A. A. Tawfik et al., 2015). For instance, Piaget's (1952) theory of cognitive development suggests that learning is driven by a state of cognitive disequilibrium when students are confronted with obstacles or contradictions that push them to restore equilibrium by assimilating new knowledge or accommodating it by modifying their existing schemas. Likewise, VanLehn's (1999) theory of impasse-driven learning posits that encountering an impasse (e.g., making an error) motivates students to resolve it by constructing a better understanding of the material. By extension, such principles could apply to deliberate erring too during misteaching. Future research is needed to probe these processes more deeply.

Educational Implications

Adding to the nascent evidence for the benefits of writing verbatim teaching scripts (K. Y. L. Lim et al., 2021; S. W. H. Lim et al., 2024), this work demonstrates how learning by teaching can be implemented efficiently and accessibly to enhance not only basic recall but also higher order argumentative reasoning. Unlike delivering video-recorded lectures, writing teaching scripts eliminates practical barriers such as the need for technical equipment. Thus, teachers could viably use this activity as a type of assignment to boost their students' learning. For instance, a recent field experiment in a research and statistical methods course had undergraduates study statistical concepts by writing verbatim teaching scripts or study notes in a take-home open-book assignment (S. W. H. Lim et al., 2024). On a high-stakes final exam 1 month later, the students were more successful in generating higher order research questions that created new knowledge about the concepts for which they had written teaching scripts than study notes and in applying those concepts to design a study that tested a given hypothesis. These findings illustrate how writing verbatim teaching scripts can be feasibly applied in real-world classrooms to improve meaningful learning.

By extension, students could be encouraged to write verbatim teaching scripts for better argumentative reasoning, along with receiving instruction on argumentation strategies. For instance, when learning about a topic, students could be asked to teach the topic to their peers by writing a teaching script exactly as how they would orate a lecture, then form an integrative conclusion after being trained on argumentation strategies.

Importantly, students reap additional gains from teaching *incorrectly* with deliberately weak arguments they have generated. That misteaching yields better argumentative reasoning than correct teaching points to the counterintuitive benefits of actively promoting errors in learning (Wong & Lim, 2019b). Understandably, when the stakes are high, unintentionally erring by producing poor arguments could backfire by incurring reputational costs that deter people from engaging fully in argumentation or even from venturing to put forth arguments at all (Mercier et al., 2017). But when the stakes are low during practice and the goal is precisely to make errors for others to spot, then deliberately producing flawed arguments in fact enhances learning.

In particular, this research provides the first evidence for the utility of fusing learning by teaching and deliberate erring via the novel technique of learning by misteaching. Students need not be compelled to use effective learning techniques one at a time in isolation, but can productively combine them to optimize learning gains. This approach resonates with growing interest and calls to examine how different learning techniques can fruitfully complement each other (Roelle et al., 2023), beyond pitting them against each other in "horse race studies" (Renkl, 2015; Salomon, 2002).

Moreover, learning by misteaching illuminates new prospects for exploring how deliberate erring can be viably implemented in various ways to achieve diverse learning goals. In extant deliberate erring studies, students typically write each sentence in a given text such that it contains a deliberate conceptual error they have generated. This form of deliberate conceptual erring has been found to improve recall, application, and far transfer of the text content (Wong, 2023; Wong & Lim, 2022a, 2022b). Alternatively, for better procedural transfer and problem-solving in domains such as mathematics, students could deliberately err by intentionally executing incorrect procedures when solving practice problems (Yap & Wong, 2024). In comparison, this study shows how deliberate errors can be integrated in a written teaching script through generating flawed arguments to ultimately achieve better argumentative reasoning.

The key implication is that the nature of students' deliberate errors can be flexibly adapted to strategically target their learning goals. Whereas this study focused on argumentative reasoning, learning by misteaching can plausibly be applied to boost other valued educational outcomes too. For instance, since writing correct teaching scripts about to-be-learned concepts improves students' ability to ask good research questions about those concepts (S. W. H. Lim et al., 2024), it could be fruitful to test whether writing deliberately incorrect teaching scripts yields additional gains. By probing the generalizability of learning-by-misteaching benefits across diverse domains and outcomes, we would achieve a deeper understanding of this technique's utility and how it can be effectively applied.

Yet, the students in this study were unaware that (mis)teaching had benefited them, even after experiencing its effects on their test performance. Such inaccurate metacognitive knowledge could prevent students from choosing to use techniques that actually help them more (Metcalfe & Finn, 2008). To address this problem, teachers could guide their students to update their metacognitive knowledge with instruction on how learning and memory work (McCabe, 2011), while supporting their students' commitment, planning, and monitoring to use effective learning techniques (Bernacki et al., 2020; McDaniel & Einstein, 2020).

Limitations and Future Directions

Here, students engaged in solitary reasoning when studying and formulating a response about an argumentation topic. But argumentation can also occur collaboratively as a social process or dialogue when students work together to construct and critique arguments (e.g., Casado-Ledesma et al., 2021; A.-M. Clark et al., 2003; Mateos et al., 2018; for a review, see Nussbaum, 2008a). Hence, an interesting future prospect is to test the dynamic effects of learning by misteaching in collaborative argumentation. Teaching-based activities inherently bear potential for interactions with one's audience that could enhance the tutor and tutee's learning (Kobayashi, 2019a; Roscoe, 2014; Roscoe & Chi, 2008). For instance, after writing a verbatim teaching script with deliberately weak arguments, students could exchange their scripts with their peers to spot and discuss each other's deliberate errors. Although first-hand deliberate erring yields better learning than spotting and correcting others' deliberate errors (Wong, 2023; Yap & Wong, 2024), both approaches could have synergistic effects when implemented together. When interacting with friendly critics, students gain opportunities to exchange ideas, coconstruct new arguments that integrate different views, reflect on gaps in their understanding, and revise or refine their mental models for deeper learning and conceptual change (Amigues, 1988; Keil, 2006; Leitão, 2000). Thus, group dialogues could create shared discourse norms (Kuhn et al., 2013) while boosting students' argumentation skills (Larrain et al., 2021) and content knowledge-building (Chinn, 2006; Felton et al., 2015), such that they become even better individual reasoners (Mercier et al., 2017).

Future work could also probe the scope of learning-by-misteaching benefits for argumentative reasoning more broadly, beyond this study's focus on integrative stratagems and holistic argumentation quality. For instance, students could be asked to evaluate individual arguments by rating their strength (e.g., McCrudden et al., 2017) and (re)appraising their plausibility (e.g., Lombardi et al., 2013, 2018). In addition, whereas participants in this study engaged with a single dual-position text, more complex tasks could involve navigating and integrating multiple texts that contain corroborating and conflicting information (e.g., Anmarkrud et al., 2014; for reviews, see Barzilai et al., 2018; Britt & Rouet, 2020). Students' skillful and accurate use of multiple sources and evidence could then be assessed (e.g., Brante & Strømsø, 2018; Du & List, 2021; Iordanou & Constantinou, 2015), while examining their justificatory standards when evaluating knowledge claims and evidence (e.g., List, 2024; List et al., 2022; Lombardi et al., 2016).

Furthermore, there is merit in considering how the implementation of misteaching can be optimized for potentially greater gains. The misteaching group was not required to refute their deliberately poor arguments when teaching since these were intended for their audience to spot. However, generating errors may enhance encoding of their subsequent correction (Kornell et al., 2009; Potts & Shanks, 2014; Potts et al., 2019), such that correcting one's deliberate errors yields additional learning benefits over leaving them uncorrected (Wong & Lim, 2022b). Hence, it may be fruitful to test whether tutors profit further from refuting their poor arguments after teaching (e.g., via individual reflection and/or group dialogues).

Relatedly, participants' teaching quality was relatively low; their teaching scripts contained few self-other referential terms, elaborations, and monitoring statements on overall. This likely occurred because the present study examined students' spontaneous teaching explanations—whereas participants were trained on argumentation, they were not explicitly guided or prompted to generate higher quality teaching explanations (e.g., Fiorella & Kuhlmann, 2020). Indeed, some learning-by-teaching studies have found that tutors may not spontaneously produce high-quality explanations (Jacob et al., 2021; Lachner et al., 2021; Roscoe & Chi, 2007). For instance, tutors' feelings of social presence can fluctuate during their teaching, and they may offer few elaborations even when they have their intended audience in mind (Ribosa & Duran, 2023). Consequently, lower teaching quality may constrain the tutor's learning (Roscoe, 2014; Roscoe & Chi, 2007). It would thus be worth exploring how students' teaching quality can be boosted through training or scaffolding for more robust learning.

In addition, an unexpected but intriguing finding was that after the intervention, the misteaching group rated the argumentative text as less interesting and understandable than the notetaking and correct teaching groups. On one hand, participants' ratings may in part reflect their subjective emotional and/or cognitive experience during learning. Generating errors can be experienced as less fluent than errorless learning, which could lead people to perceive the material as more difficult to learn (Potts & Shanks, 2014; Yang et al., 2017). On the other hand, if the misteaching group implicitly adopted an evaluation goal while reading the argumentative text to generate flawed arguments, they could have developed more critical or conservative evaluations of the text's argumentative content and quality (Diakidoy et al., 2017). Future work ought to disentangle such possibilities.

Although this study focused on student tutors' learning, the findings also raise questions for teachers and what it means to teach effectively. Teachers typically strive to teach correctly with accurate content—imparting wrong information that goes uncorrected could be costly for students' learning and can be perceived unfavorably as an indicator of poor content knowledge (Kearney et al., 1991). Unsurprisingly, teachers may thus experience anxiety about appearing incompetent when they make mistakes, becoming trapped in trying to maintain a "persona of perfectionism" that could lead to burnout and exit from the profession (Hargreaves & Tucker, 1991). In tackling such issues, new possibilities arise from the approach of deliberately making errors in one's teaching for students to spot as part of the intentional learning design. Research on incorrect worked examples has shown that having students identify and explain errors in incorrect examples can yield better learning than studying correct examples only (e.g., Booth et al., 2013; Durkin & Rittle-Johnson, 2012; Große & Renkl, 2007). Hence, teachers may be reassured that having students spot and correct deliberate errors does not harm their learning (Wong, 2023).

Still, much less is known about the emotional effects of deliberate erring on the teacher's part. Besides benefiting teachers' learning, might deliberate erring also enable them to reframe errors as meaningful teachable moments rather than debilitating events? If so, teachers may be empowered to embrace errors and take risks to enhance their professional practice (Phelps, 2000).

Conclusion

The skill to reason and argue well is vital for 21st-century education and democratic participation. Leveraging the techniques of learning by teaching and deliberate erring, this study unveiled their joint advantage for argumentative reasoning. Students displayed not only superior recall but also argumentative reasoning when they had taught the material by writing a verbatim teaching script, relative to writing study notes for their own learning. Moreover, teaching incorrectly with deliberately weak arguments than correctly with good arguments further enhanced argumentative reasoning. Whereas learning techniques have often been studied parallel to each other and errors have traditionally been regarded as events to be avoided in learning, the present data attest that it is worth rethinking these approaches. When potent learning techniques join forces and errors are strategically sought out than avoided, students and teachers have much more to gain than previously imagined.

References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Amigues, R. (1988). Peer interaction in solving physics problems: Sociocognitive confrontation and metacognitive aspects. *Journal of Experimental Child Psychology*, *45*(1), 141–158. https://doi.org/10.1016/0022-0965(88)90054-9

Andriessen, J., & Baker, M. (2014). Arguing to learn. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 439–460). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526

Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, *30*, 64–76. https://doi.org/10.1016/j.lindif.2013.01.007

Asterhan, C. S. C., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology*, *99*(3), 626–639. https://doi.org/10.1037/0022-0663.99.3.626

Asterhan, C. S. C., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, *51*(2), 164–187. https://doi.org/10.1080/00461520.2016.1155458

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart and Winston.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, *72*(5), 593–604. https://doi.org/10.1037/0022-0663.72.5.593

Barzilai, S., Zohar, A. R., & Mor-Hagani, S. (2018). Promoting integration of multiple texts: A review of instructional approaches and practices. *Educational Psychology Review*, *30*(3), 973–999. https://doi.org/10.1007/s10648-018-9436-8

Baytelman, A., Iordanou, K., & Constantinou, C. P. (2020). Epistemic beliefs and prior knowledge as predictors of the construction of different types of arguments on socioscientific issues. *Journal of Research in Science Teaching*, *57*(8), 1199–1227. https://doi.org/10.1002/tea.21627

Beatty, E. L., & Thompson, V. A. (2012). Effects of perspective and belief on analytic reasoning in a scientific reasoning task. *Thinking & Reasoning*, *18*(4), 441–460. https://doi.org/10.1080/13546783.2012.687892

Benware, C. A., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*(4), 755–765. https://doi.org/10.3102/00028312021004755

Bernacki, M. L., Vosicka, L., & Utz, J. C. (2020). Can a brief, digital skill training intervention help undergraduates “learn to learn” and improve their STEM achievement? *Journal of Educational Psychology*, *112*(4), 765–781. https://doi.org/10.1037/edu0000405

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, *19*(3–4), 363–392. https://doi.org/10.1080/08839510590910200

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty’s brain. *International Journal of Artificial Intelligence in Education*, *26*(1), 350–364. https://doi.org/10.1007/s40593-015-0057-9

Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*, *25*, 24–34. https://doi.org/10.1016/j.learninstruc.2012.11.002

Brante, E. W., & Strømsø, H. I. (2018). Sourcing in text comprehension: A review of interventions targeting sourcing skills. *Educational Psychology Review*, *30*(3), 773–799. https://doi.org/10.1007/s10648-017-9421-7

Britt, M. A., & Rouet, J. (2020). Multiple document comprehension. In L.-F. Zhang (Ed.), *The Oxford encyclopedia of educational psychology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.867

Broughton, S. H., Sinatra, G. M., & Reynolds, R. E. (2010). The nature of the refutation text effect: An investigation of attention allocation. *The Journal of Educational Research*, *103*(6), 407–423. https://doi.org/10.1080/00220670903383101

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818. https://doi.org/10.1037/0022-3514.45.4.805

Casado-Ledesma, L., Cuevas, I., Van den Bergh, H., Rijlaarsdam, G., Mateos, M., Granado-Peinado, M., & Martín, E. (2021). Teaching argumentative synthesis writing through deliberative dialogues: Instructional practices in secondary education. *Instructional Science*, *49*(4), 515–559. https://doi.org/10.1007/s11251-021-09548-3

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Ablex.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Lawrence Erlbaum Associates.

Chin, D. B., Dohmen, I. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L. (2010). Preparing students for future learning with teachable agents. *Educational Technology Research and Development*, *58*(6), 649–669. https://doi.org/10.1007/s11423-010-9154-5

Chinn, C. A. (2006). Learning to argue. In A. M. O’Donnell, C. E. Hmelo-Silver, & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 355–383). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203826843

Chinn, C. A., & Duncan, R. G. (2018). What is the value of general knowledge of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge* (pp. 77–101). Routledge. https://doi.org/10.4324/9780203731826

Christodoulou, S. A., & Diakidoy, I.-A. N. (2020). The contribution of argument knowledge to the comprehension and critical evaluation of argumentative text. *Contemporary Educational Psychology*, *63*, Article 101903. https://doi.org/10.1016/j.cedpsych.2020.101903

Clark, A.-M., Anderson, R. C., Kuo, L.-J., Kim, I.-H., Archodidou, A., & Nguyen-Jahiel, K. (2003). Collaborative reasoning: Expanding ways for children to talk and think in school. *Educational Psychology Review*, *15*(2), 181–198. https://doi.org/10.1023/A:1023429215151

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Coleman, E. B., Brown, A. L., & Rivkin, I. D. (1997). The effect of instructional explanations on learning from scientific texts. *The Journal of the Learning Sciences*, *6*(4), 347–365. https://doi.org/10.1207/s15327809jls0604_1

Diakidoy, I.-A. N., Ioannou, M. C., & Christodoulou, S. A. (2017). Reading argumentative texts: Comprehension and evaluation goals and outcomes. *Reading and Writing*, *30*(9), 1869–1890. https://doi.org/10.1007/s11145-017-9757-x

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

Du, H., & List, A. (2021). Evidence use in argument writing based on multiple texts. *Reading Research Quarterly*, *56*(4), 715–735. https://doi.org/10.1002/rrq.366

Duran, D., & Topping, K. J. (2017). *Learning by teaching: Evidence-based strategies to enhance learning in the classroom*. Routledge. https://doi.org/10.4324/9781315649047

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, *22*(3), 206–214. https://doi.org/10.1016/j.learninstruc.2011.11.001

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Felton, M., Crowell, A., Garcia-Mila, M., & Villarroel, C. (2022). Capturing deliberative argument: An analytic coding scheme for studying argumentative dialogue and its benefits for learning. *Learning, Culture and Social Interaction*, *36*, Article 100350. https://doi.org/10.1016/j.lcsi.2019.100350

Felton, M., Garcia-Mila, M., Villarroel, C., & Gilabert, S. (2015). Arguing collaboratively: Argumentative discourse types and their potential for knowledge building. *British Journal of Educational Psychology*, *85*(3), 372–386. https://doi.org/10.1111/bjep.12078

Ferretti, R. P., & Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Reading and Writing*, *32*(6), 1345–1357. https://doi.org/10.1007/s11145-019-09950-x

Fiorella, L. (2023). Making sense of generative learning. *Educational Psychology Review*, *35*(2), Article 50. https://doi.org/10.1007/s10648-023-09769-7

Fiorella, L., & Kuhlmann, S. (2020). Creating drawings enhances learning by teaching. *Journal of Educational Psychology*, *112*(4), 811–822. https://doi.org/10.1037/edu0000392

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, *38*(4), 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75–85. https://doi.org/10.1016/j.cedpsych.2014.01.001

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning*, *1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Greene, J. A., Cartiff, B. M., & Duke, R. F. (2018). A meta-analytic review of the relationship between epistemic cognition and academic achievement. *Journal of Educational Psychology*, *110*(8), 1084–1111. https://doi.org/10.1037/edu0000263

Greene, J. A., & Yu, S. B. (2016). Educating critical thinkers: The role of epistemic cognition. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 45–53. https://doi.org/10.1177/2372732215622223

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., Fournier, M. A., Hu, C. S., Nusbaum, H. C., & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. *Psychological Inquiry*, *31*(2), 103–133. https://doi.org/10.1080/1047840X.2020.1750917

Guerrero, T. A., & Wiley, J. (2021). Expecting to teach affects learning during study of expository texts. *Journal of Educational Psychology*, *113*(7), 1281–1303. https://doi.org/10.1037/edu0000657

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. https://doi.org/10.1037/0033-295X.114.3.704

Hargreaves, A., & Tucker, E. (1991). Teaching and guilt: Exploring the feelings of teaching. *Teaching and Teacher Education*, *7*(5-6), 491–505. https://doi.org/10.1016/0742-051X(91)90044-P

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 451–470. https://doi.org/10.1111/bmsp.12028

Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science*, *42*(4), 639–657. https://doi.org/10.1007/s11251-013-9302-5

Hoogerheide, V., Deijkers, L., Loyens, S. M. M., Heijltjes, A., & van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology*, *44–45*, 95–106. https://doi.org/10.1016/j.cedpsych.2016.02.005

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108–119. https://doi.org/10.1016/j.learninstruc.2014.04.005

Hoogerheide, V., Renkl, A., Fiorella, L., Paas, F., & van Gog, T. (2019a). Enhancing example-based learning: Teaching on video increases arousal and improves problem-solving performance. *Journal of Educational Psychology*, *111*(1), 45–56. https://doi.org/10.1037/edu0000272

Hoogerheide, V., Visee, J., Lachner, A., & van Gog, T. (2019b). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, Article 101226. https://doi.org/10.1016/j.learninstruc.2019.101226

Hynd, C., & Alvermann, D. E. (1986). The role of refutation text in overcoming difficulty with science concepts. *Journal of Reading*, *29*(5), 440–446.

Iordanou, K., & Constantinou, C. P. (2015). Supporting use of evidence in argumentation through practice in argumentation and reflection in the context of SOCRATES learning environment. *Science Education*, *99*(2), 282–311. https://doi.org/10.1002/sce.21152

Iordanou, K., Kendeou, P., & Beker, K. (2016). Argumentative reasoning. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 39–53). Routledge. https://doi.org/10.4324/9781315795225

Iordanou, K., Kendeou, P., & Zembylas, M. (2020). Examining my-side bias during and after reading controversial historical accounts. *Metacognition and Learning*, *15*(3), 319–342. https://doi.org/10.1007/s11409-020-09240-w

Iordanou, K., Kuhn, D., Matos, F., Shi, Y., & Hemberger, L. (2019). Learning by arguing. *Learning and Instruction*, *63*, Article 101207. https://doi.org/10.1016/j.learninstruc.2019.05.004

Jacob, L., Lachner, A., & Scheiter, K. (2020). Learning by explaining orally or in written form? Text complexity matters. *Learning and Instruction*, *68*, Article 101344. https://doi.org/10.1016/j.learninstruc.2020.101344

Jacob, L., Lachner, A., & Scheiter, K. (2021). Does increasing social presence enhance the effectiveness of writing explanations? *PLOS ONE*, *16*(4), Article e0250406. https://doi.org/10.1371/journal.pone.0250406

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, *84*(6), 757–792. https://doi.org/10.1002/1098-237X(200011)84:6<757::AID-SCE5>3.0.CO;2-F

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, volume 2 of learning and memory: A comprehensive reference* (J. H. Byrne, Series ed., pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Kearney, P., Plax, T. G., Hays, E. R., & Ivey, M. J. (1991). College teacher misbehaviors: What students don't like about what teachers say and do. *Communication Quarterly*, *39*(4), 309–324. https://doi.org/10.1080/01463379109369808

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kendeou, P. (2024). A theory of knowledge revision: The development of the KReC framework. *Educational Psychology Review*, *36*(2), Article 44. https://doi.org/10.1007/s10648-024-09885-y

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition*, *47*(1), 33–46. https://doi.org/10.3758/s13421-018-0848-y

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) framework: Processes and mechanisms. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353–378). MIT Press. https://doi.org/10.7551/mitpress/9737.003.0022

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Kobayashi, K. (2010). Strategic use of multiple texts for the evaluation of arguments. *Reading Psychology*, *31*(2), 121–149. https://doi.org/10.1080/02702710902754192

Kobayashi, K. (2019a). Interactivity: A potential determinant of learning by preparing to teach and teaching. *Frontiers in Psychology*, *9*, Article 2755. https://doi.org/10.3389/fpsyg.2018.02755

Kobayashi, K. (2019b). Learning by preparing-to-teach and teaching: A meta-analysis. *Japanese Psychological Research*, *61*(3), 192–203. https://doi.org/10.1111/jpr.12221

Kobayashi, K. (2022a). Learning by teaching face-to-face: The contributions of preparing-to-teach, initial explanation, and interaction phases. *European Journal of Psychology of Education*, *37*(2), 551–566. https://doi.org/10.1007/s10212-021-00547-z

Kobayashi, K. (2022b). The retrieval practice hypothesis in research on learning by teaching: Current status and challenges. *Frontiers in Psychology*, *13*, Article 842668. https://doi.org/10.3389/fpsyg.2022.842668

Kobayashi, K. (2024). Interactive learning effects of preparing to teach and teaching: A meta-analytic approach. *Educational Psychology Review*, *36*(1), Article 26. https://doi.org/10.1007/s10648-024-09871-4

Koh, A. W. L., Lee, S. C., & Lim, S. W. H. (2018). The learning benefits of teaching: A retrieval practice hypothesis. *Applied Cognitive Psychology*, *32*(3), 401–410. https://doi.org/10.1002/acp.3410

Koons, R. (2022). Defeasible reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 ed.). The Metaphysics Research Lab. https://plato.stanford.edu/archives/sum2022/entries/reasoning-defeasible/

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Köymen, B., Rosenbaum, L., & Tomasello, M. (2014). Reasoning during joint decision-making by preschool peers. *Cognitive Development*, *32*, 74–85. https://doi.org/10.1016/j.cogdev.2014.09.001

Kreijns, K., Xu, K., & Weidlich, J. (2022). Social presence: Conceptualization and measurement. *Educational Psychology Review*, *34*(1), 139–170. https://doi.org/10.1007/s10648-021-09623-8

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*(3), 319–337. https://doi.org/10.1002/sce.3730770306

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, *15*(3), 309–328. https://doi.org/10.1016/S0885-2014(00)00030-7

Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cognition and Instruction*, *31*(4), 456–496. https://doi.org/10.1080/07370008.2013.830618

Lachner, A., Backfisch, I., Hoogerheide, V., van Gog, T., & Renkl, A. (2020). Timing matters! Explaining between study phases enhances students' learning. *Journal of Educational Psychology*, *112*(4), 841–853. https://doi.org/10.1037/edu0000396

Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2022). Learning-by-teaching without audience presence or interaction: When and why does it work? *Educational Psychology Review*, *34*(2), 575–607. https://doi.org/10.1007/s10648-021-09643-4

Lachner, A., Jacob, L., & Hoogerheide, V. (2021). Learning by writing explanations: Is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction*, *74*, Article 101438. https://doi.org/10.1016/j.learninstruc.2020.101438

Lachner, A., Ly, K.-T., & Nückles, M. (2018). Providing written or oral explanations? Differential effects of the modality of explaining on students' conceptual learning and transfer. *The Journal of Experimental Education*, *86*(3), 344–361. https://doi.org/10.1080/00220973.2017.1363691

Larrain, A., Singer, V., Strasser, K., Howe, C., López, P., Pinochet, J., Moran, C., Sánchez, Á, Silva, M., & Villavicencio, C. (2021). Argumentation skills mediate the effect of peer argumentation on content knowledge in middle-school students. *Journal of Educational Psychology*, *113*(4), 736–753. https://doi.org/10.1037/edu0000619

Leitão, S. (2000). The potential of argument in knowledge building. *Human Development*, *43*(6), 332–360. https://doi.org/10.1159/000022695

Lim, K. Y. L., Wong, S. S. H., & Lim, S. W. H. (2021). The "silent teacher": Learning by teaching via writing a verbatim teaching script. *Applied Cognitive Psychology*, *35*(6), 1492–1501. https://doi.org/10.1002/acp.3881

Lim, S. W. H., Wong, S. S. H., & Visessuvanapoom, P. (2024). Durable benefits of learning-by-teaching for research question generation performance: A field experiment. *The Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2024.2364625

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction*, *11*(4-5), 357–380. https://doi.org/10.1016/S0959-4752(00)00037-2

List, A. (2024). The limits of reasoning: Students' evaluations of anecdotal, descriptive, correlational, and causal evidence. *The Journal of Experimental Education*, *92*(1), 1–31. https://doi.org/10.1080/00220973.2023.2174487

List, A., Du, H., & Lyu, B. (2022). Examining undergraduates' text-based evidence identification, evaluation, and use. *Reading and Writing*, *35*(5), 1059–1089. https://doi.org/10.1007/s11145-021-10219-5

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Lombardi, D., Bickel, E. S., Bailey, J. M., & Burrell, S. (2018). High school students' evaluations, plausibility (re) appraisals, and knowledge about topics in Earth science. *Science Education*, *102*(1), 153–177. https://doi.org/10.1002/sce.21315

Lombardi, D., Brandt, C. B., Bickel, E. S., & Burg, C. (2016). Students' evaluations about climate change. *International Journal of Science Education*, *38*(8), 1392–1414. https://doi.org/10.1080/09500693.2016.1193912

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction*, *27*, 50–62. https://doi.org/10.1016/j.learninstruc.2013.03.001

Lytzerinou, E., & Iordanou, K. (2020). Teachers' ability to construct arguments, but not their perceived self-efficacy of teaching, predicts their ability to evaluate arguments. *International Journal of Science Education*, *42*(4), 617–634. https://doi.org/10.1080/09500693.2020.1722864

Mason, L., & Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology*, *29*(2), 103–128. https://doi.org/10.1016/j.cedpsych.2004.01.001

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction*, *16*(5), 492–509. https://doi.org/10.1016/j.learninstruc.2006.09.007

Mateos, M., Martín, E., Cuevas, I., Villalón, R., Martínez, I., & González-Lamas, J. (2018). Improving written argumentative synthesis by teaching the integration of conflicting information from multiple sources. *Cognition and Instruction*, *36*(2), 119–138. https://doi.org/10.1080/07370008.2018.1425300

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

McCrudden, M. T., Barnes, A., McTigue, E. M., Welch, C., & MacDonald, E. (2017). The effect of perspective-taking on reasoning about strong and weak belief-relevant arguments. *Thinking & Reasoning*, *23*(2), 115–133. https://doi.org/10.1080/13546783.2016.1234411

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*(2), 113–145. https://doi.org/10.1007/BF01326639

McDaniel, M. A., & Einstein, G. O. (2020). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science*, *15*(6), 1363–1381. https://doi.org/10.1177/1745691620920723

Mercier, H. (2011). Reasoning serves argumentation in children. *Cognitive Development*, *26*(3), 177–191. https://doi.org/10.1016/j.cogdev.2010.12.001

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Mercier, H., Bernard, S., & Clément, F. (2014). Early sensitivity to arguments: How preschoolers weight circular arguments. *Journal of Experimental Child Psychology*, *125*, 102–109. https://doi.org/10.1016/j.jecp.2013.11.011

Mercier, H., Boudry, M., Paglieri, F., & Trouche, E. (2017). Natural-born arguers: Teaching how to make the best of our reasoning abilities. *Educational Psychologist*, *52*(1), 1–16. https://doi.org/10.1080/00461520.2016.1207537

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Moseley, W. G. (2007). *Taking sides: Clashing views on African issues* (2nd ed.). McGraw-Hill.

Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B. D., Li, M., Montalbano, C., & Wei, L. (2018). Quality talk: Developing students' discourse to promote high-level comprehension. *American Educational Research Journal*, *55*(5), 1113–1160. https://doi.org/10.3102/0002831218771303

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards*. https://corestandards.org/

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition*, *42*(7), 1038–1048. https://doi.org/10.3758/s13421-014-0416-z

Neuman, Y., & Weizman, E. (2003). The role of text representation in students' ability to identify fallacious arguments. *The Quarterly Journal of Experimental Psychology*, *56*(5), 849–864. https://doi.org/10.1080/02724980244000666

Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly*, *46*(3), 273–304. https://doi.org/10.1598/RRQ.46.3.4

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press. https://doi.org/10.17226/18290

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Noroozi, O., Weinberger, A., Biemans, H. J. A., Mulder, M., & Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCL): A synthesis of 15 years of research. *Educational Research Review*, *7*(2), 79–106. https://doi.org/10.1016/j.edurev.2011.11.006

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press.

Nussbaum, E. M. (2008a). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology*, *33*(3), 345–359. https://doi.org/10.1016/j.cedpsych.2008.06.001

Nussbaum, E. M. (2008b). Using argumentation vee diagrams (AVDs) for promoting argument-counterargument integration in reflective writing. *Journal of Educational Psychology*, *100*(3), 549–565. https://doi.org/10.1037/0022-0663.100.3.549

Nussbaum, E. M. (2021). Critical integrative argumentation: Toward complexity in students' thinking. *Educational Psychologist*, *56*(1), 1–17. https://doi.org/10.1080/00461520.2020.1845173

Nussbaum, E. M., Dove, I. J., Slife, N., Kardash, C. M., Turgut, R., & Vallett, D. (2019). Using critical questions to evaluate written and oral arguments in an undergraduate general education seminar: A quasi-experimental study. *Reading and Writing*, *32*(6), 1531–1552. https://doi.org/10.1007/s11145-018-9848-3

Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences*, *20*(3), 443–488. https://doi.org/10.1080/10508406.2011.564567

Nussbaum, E. M., & Putney, L. G. (2020). Learning to use benefit-cost arguments: A microgenetic study of argument-counterargument integration in an undergraduate seminar course. *Journal of Educational Psychology*, *112*(3), 444–465. https://doi.org/10.1037/edu0000412

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *Journal of Experimental Education*, *76*(1), 59–92. https://doi.org/10.3200/JEXE.76.1.59-92

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846. https://doi.org/10.1002/tea.21316

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft. Zur Theorie des Negativen Wissens und zur Praxis der Fehlerkultur* [Learning is painful. On the theory of negative knowledge and the practice of error culture]. Beltz.

Parviainen, J., & Eriksson, M. (2006). Negative knowledge, expertise and organisations. *International Journal of Management Concepts and Philosophy*, *2*(2), 140–153. https://doi.org/10.1504/IJMCP.2006.010265

Paul, E. L. (2002). *Taking sides: Clashing views on controversial issues in sex and gender* (2nd ed.). McGraw-Hill/Dushkin.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77*(5), 562–571. https://doi.org/10.1037/0022-0663.77.5.562

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83–105). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203052228

Phelps, P. H. (2000). Mistakes as vehicles for educating teachers. *Action in Teacher Education*, *21*(4), 41–49. https://doi.org/10.1080/01626620.2000.10462979

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). WW Norton & Co. https://doi.org/10.1037/11494-000

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, *11*(4), 481–518. https://doi.org/10.1207/s15516709cog1104_4

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1023–1041. https://doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Rapanta, C., & Felton, M. K. (2022). Learning to argue through dialogue: A review of instructional approaches. *Educational Psychology Review*, *34*(2), 477–509. https://doi.org/10.1007/s10648-021-09637-2

Rapanta, C., Garcia-Mila, M., & Gilabert, S. (2013). What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education. *Review of Educational Research*, *83*(4), 483–520. https://doi.org/10.3102/0034654313487606

Renkl, A. (2015). Different roads lead to Rome: The case of principle-based cognitive skills. *Learning: Research and Practice*, *1*(1), 79–90. https://doi.org/10.1080/23735082.2015.994255

Reznitskaya, A., Kuo, L.-J., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers? *Learning and Individual Differences*, *19*(2), 219–224. https://doi.org/10.1016/j.lindif.2008.11.001

Ribosa, J., & Duran, D. (2022). Do students learn what they teach when generating teaching materials for others? A meta-analysis through the lens of learning by teaching. *Educational Research Review*, *37*, Article 100475. https://doi.org/10.1016/j.edurev.2022.100475

Ribosa, J., & Duran, D. (2023). Students' feelings of social presence when creating learning-by-teaching educational videos for a potential audience. *International Journal of Educational Research*, *117*, Article 102128. https://doi.org/10.1016/j.ijer.2022.102128

Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together? On the relationship between research on retrieval practice and generative learning using the case of follow-up learning tasks. *Educational Psychology Review*, *35*(4), Article 102. https://doi.org/10.1007/s10648-023-09810-9

Roscoe, R. D. (2014). Self-monitoring and knowledge-building in learning by teaching. *Instructional Science*, *42*(3), 327–351. https://doi.org/10.1007/s11251-013-9283-4

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*(4), 534–574. https://doi.org/10.3102/0034654307309920

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/10.1007/s11251-007-9034-5

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, *96*(3), 488–526. https://doi.org/10.1002/sce.21006

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536. https://doi.org/10.1002/tea.20009

Salomon, G. (2002). Technology and pedagogy: Why don't we see the promised revolution? *Educational Technology*, *42*(2), 71–75.

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, *5*(1), 43–102. https://doi.org/10.1007/s11412-009-9080-x

Schroeder, N. L., & Kucera, A. C. (2022). Refutation text facilitates learning: A meta-analysis of between-subjects experiments. *Educational Psychology Review*, *34*(2), 957–987. https://doi.org/10.1007/s10648-021-09656-z

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). Cambridge University Press. https://doi.org/10.1017/CBO9780511489709.002

Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science*, *11*(4), 433–448. https://doi.org/10.1111/j.1467-7687.2008.00689.x

Sindoni, M. G. (2013). *Spoken and written discourse in online interactions: A multimodal approach*. Routledge. https://doi.org/10.4324/9780203587935

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. *Science*, *128*(3330), 969–977. https://doi.org/10.1126/science.128.3330.969

Stein, N. L., & Bernas, R. (1999). The early emergence of argumentative knowledge and skill. In G. Rijlaarsdam & E. Espéret (Series Eds.) & J. Andriessen & P. Coirier (Vol. Eds.), Studies in writing: Volume 5. *Foundations of argumentative text processing* (pp. 97–116). Amsterdam University Press.

Tauber, S. K., Thakkar, V. J., & Pleshek, M. A. (2022). How does the type of expected evaluation impact students' self-regulated learning? *Journal of Applied Research in Memory and Cognition*, *11*(1), 106–119. https://doi.org/10.1016/j.jarmac.2021.08.002

Tawfik, A., & Jonassen, D. (2013). The effects of successful versus failure-based cases on argumentation while solving decision-making problems. *Educational Technology Research and Development*, *61*(3), 385–406. https://doi.org/10.1007/s11423-013-9294-5

Tawfik, A. A., Rong, H., & Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development*, *63*(6), 975–994. https://doi.org/10.1007/s11423-015-9399-0

Thompson, V. A., Evans, J. S. B. T., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language*, *53*(2), 238–257. https://doi.org/10.1016/j.jml.2005.03.001

Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. *International Journal of Science and Mathematics Education*, *8*(6), 951–970. https://doi.org/10.1007/s10763-010-9203-x

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Henkemans, A. F. S., Verheij, B., & Wagemans, J. H. M. (2014). *Handbook of argumentation theory*. Springer. https://doi.org/10.1007/978-90-481-9473-5

van Eemeren, F. H., Grootendorst, R., Henkemans, F. S., Blair, J. A., Johnson, R. H., Krabbe, E. C. W., Plantin, C., Walton, D. N., Willard, C. A., Woods, J., & Zarefsky, D. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Lawrence Erlbaum Associates.

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: An evaluation of cascade. *Journal of the Learning Sciences*, *8*(1), 71–125. https://doi.org/10.1207/s15327809jls0801_3

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*(3), 209–249. https://doi.org/10.1207/S1532690XCI2103_01

von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, *45*(1), 101–131. https://doi.org/10.1002/tea.20213

Voss, J. F., & Van Dyke, J. A. (2001). Argumentation in psychology: Background comments. *Discourse Processes*, *32*(2–3), 89–111. https://doi.org/10.1080/0163853X.2001.9651593

Walton, D. (2007). *Dialogue theory for critical argumentation*. John Benjamins.

Walton, D. (2010). Types of dialogue and burdens of proof. In P. Baroni, F. Cerutti, M. Giacomin, & G. R. Simari (Eds.), *Computational models of argument: Proceedings of COMMA 2010* (pp. 13–24). IOS Press. https://doi.org/10.3233/978-1-60750-619-5-13

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Erlbaum.

Wang, F., Cheng, M., & Mayer, R. E. (2023). Improving learning-by-teaching without audience interaction as a generative learning activity by minimizing the social presence of the audience. *Journal of Educational Psychology*, *115*(6), 783–797. https://doi.org/10.1037/edu0000801

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist*, *11*(2), 87–95. https://doi.org/10.1080/00461527409529129

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*(4), 345–376. https://doi.org/10.1207/s15326985ep2404_2

Wittwer, J., Nückles, M., Landmann, N., & Renkl, A. (2010). Can tutors be supported in giving effective explanations? *Journal of Educational Psychology*, *102*(1), 74–89. https://doi.org/10.1037/a0016727

Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, *14*(1), 1–27. https://doi.org/10.1080/13546780701527674

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication*, *26*(2), 183–209. https://doi.org/10.1177/0741088309333019

Wong, S. S. H. (2023). Deliberate erring improves far transfer of learning more than errorless elaboration and spotting and correcting others' errors. *Educational Psychology Review*, *35*(1), Article 16. https://doi.org/10.1007/s10648-023-09739-z

Wong, S. S. H., Lim, K. Y. L., & Lim, S. W. H. (2023). To ask better questions, teach: Learning-by-teaching enhances research question generation more than retrieval practice and concept-mapping. *Journal of Educational Psychology*, *115*(6), 798–812. https://doi.org/10.1037/edu0000802

Wong, S. S. H., & Lim, S. W. H. (2019a). From JOLs to JOLs+: Directing learners' attention in retrieval practice to boost integrative argumentation. *Journal of Experimental Psychology: Applied*, *25*(4), 543–557. https://doi.org/10.1037/xap0000225

Wong, S. S. H., & Lim, S. W. H. (2019b). Prevention–permission–promotion: A review of approaches to errors in learning. *Educational Psychologist*, *54*(1), 1–19. https://doi.org/10.1080/00461520.2018.1501693

Wong, S. S. H., & Lim, S. W. H. (2022a). Deliberate errors promote meaningful learning. *Journal of Educational Psychology*, *114*(8), 1817–1831. https://doi.org/10.1037/edu0000720

Wong, S. S. H., & Lim, S. W. H. (2022b). The derring effect: Deliberate errors enhance learning. *Journal of Experimental Psychology: General*, *151*(1), 25–40. https://doi.org/10.1037/xge0001072

Wu, Y.-T., & Tsai, C.-C. (2011). High school students' informal reasoning regarding a socio-scientific issue, with relation to scientific epistemological beliefs and cognitive structures. *International Journal of Science Education*, *33*(3), 371–400. https://doi.org/10.1080/09500690903505661

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073–1092. https://doi.org/10.1037/xlm0000363

Yap, J. B. K., & Wong, S. S. H. (2024). Deliberately making and correcting errors in mathematical problem-solving practice improves procedural transfer to more complex problems. *Journal of Educational Psychology*, *116*(7), 1112–1128. https://doi.org/10.1037/edu0000850

Zengilowski, A., Schuetze, B. A., Nash, B. L., & Schallert, D. L. (2021). A critical review of the refutation text literature: Methodological confounds, theoretical problems, and possible solutions. *Educational Psychologist*, *56*(3), 175–195. https://doi.org/10.1080/00461520.2020.1861948

(*Appendices follow*)

Appendix A

Sample Argumentative Reasoning Test Responses

"Will Biotech Solve Africa's Food Problems?"

"The Against arguments are stronger. While using biotechnology promises potentially huge returns that will benefit 75% of the African population, the start-up costs of innovating, acquiring and adapting these technologies are costly. Considering how Africa relies mostly on foreign aid and has preexisting debts, it is unlikely that Africa is able to finance biotechnology sustainably.

Indeed, biotechnology could be implemented using a bottom-up approach where farmers, women and people living in rural areas are involved in the decision-making process. However, as their education level tends to be low and it is likely that their knowledge in technology is not high, the additional financial resources needed to educate them will put an additional strain on limited financial resources. In contrast, the natural method of fallowing does not require start-up costs nor education costs. While the returns are much lower and slower as compared to what biotechnology promises to deliver, keeping costs low is preferred given Africa's financial circumstances.

Lastly, in the event that Africa is able to finance biotechnology, it will reap high yield productivity that can elevate many out of the poverty cycle. However, despite reaping high yields, most of the profits are not distributed within Africa itself. Instead, foreign marketing and distributing companies are the ones pocketing most of the profits. This is because they have the bargaining power to suppress supply costs that are paid to African farmers while having the market power to overcharge prices that consumers have to pay for the eventual product in supermarkets.

In conclusion, the For arguments are too optimistic without providing actual solutions or avenues for which biotechnology can be financed. Without any sustainable means of financing biotechnology, the mass implementation of biotechnology will only further entrench Africa in debt and poverty before any returns could be realized. However, the biotechnology program could still be progressively implemented, albeit on a smaller scale. In collaboration with international banks or tech companies, Africa could look to pilot the biotechnology program in a few farms to test out the results. As such, profits could be used to fund the roll-out of the biotechnology program in other farms and accumulatively, Africa would have adequate self-financing capability to finance the entire biotechnology program few years down the road without incurring any debts."

"Should We Continue to Study Sex Differences?"

"The arguments supporting sex difference research are more relevant than those opposing its continuation. As the arguments show, most of the opponents of such research use sensationalized evidence that completely misrepresents empirical data that have been found through scientific inquiries into the biological differences between sexes. Although the opponents do have a relevant point that much of current gender discrimination against women stems from a focus on differences between sexes, this does not mean that research into this field should not be conducted at all. Instead, it should be supplemented with sociological research as well as understandings of how researchers and participants also likely possess some form of bias ingrained through societal norms and expectations of different genders. Hence, it is a weak argument on the opponents' side to argue for a complete stop in sex difference research on the grounds that its only purpose is division of genders, as this is a very reductive argument of the actual research conducted.

In addition, sex difference research does not only perpetuate the understanding purely of differences between sexes. Instead, it can also find overlaps and similarities between male and female sexes biologically, and can even center around how societal norms and generational traditions have impacted the development and division between sexes. It can be a useful tool in research into gender identity and how it interplays with the traditional understanding of biological sex, which could help to improve the process of gender reassignment surgery and hormone therapy, as well as advance research into the field of gender dysmorphia and gender identity. Although many opponents argue that such research would further demean women and perpetuate an understanding that women are biologically inferior to men, this is a result of oversimplification and misrepresentation of empirical data, and should be rectified through the avoidance of headline sensationalization of research findings in this field as is often done by both sides of the argument.

One point of contestation on the side of proponents of the argument is that there is no negative conception of women, and women are viewed even more positively than men. I would argue that this evidence is shaky as the study was only conducted on college students in the United States and Canada, which is an extremely small sample size not indicative of a general population and hence cannot be generalized. However, although the claim made should not have been drawn from the provided evidence, I believe that further analysis into the topic can provide the proponents of sex difference research better findings to come to that conclusion.

Scientific research should not be completely halted in this field. Although there are many possible negative implications of the findings of such research, they stem from controllable and avoidable situations such as accurate representation of findings and appropriate usage of empirical data. In conclusion, there needs to be an integration of social science into this field of research to avoid some of the valid points made by the opponents of such research that gender and sex differences are not a purely biological issue as it has been informed by much of our societal understandings of sex and gender. However, the argument cannot be extrapolated to advocate a complete cessation of such research. There needs to be a good middle ground found to understand how society's definitions of gender and sex have informed gender roles and stereotypes, and how they factor into research bias."

(*Appendices continue*)

Appendix B

Holistic Scoring Rubric for Argumentative Reasoning Test

Score	Description
7	Highly developed position • The essay states a clear position on the issue, supporting reasons, opposing reasons, elaborations, and rebuttals • There is consistent discussion of opposing perspective(s) • The essay is well organized and focused, no irrelevant information is included, repetition is low
6	Well-developed position • The essay states a clear position on the issue supported by elaborated reasons • There is consistent discussion of opposing perspective(s) • The essay is generally well organized and focused
5	Between the standards for 4 and 6 • The essay states a clear position on the issue supported by elaborated reasons • There is some consideration of alternative perspective(s) but they are not well developed—there is little or no attempt at reconciling the alternative perspective(s) in own argumentation • The essay may contain irrelevant and/or repetitive information
4	Partially developed position • The essay contains a position on the issue supported by four or more distinct or elaborated reasons, which are often presented in a list-like fashion • Alternative perspective(s) may be mentioned but are not discussed • The essay may contain inconsistencies, irrelevant information, and/or problems with organization and clarity
3	Between the standards for 2 and 4 • The essay contains a position on the issue supported by four or more distinct or elaborated reasons, which are often presented in a list-like fashion; alternative perspective(s) are not mentioned or discussed OR • The essay contains a position on the issue supported by fewer than four reasons; alternative perspective(s) may be mentioned but are not discussed • The essay contains a lot of irrelevant and/or repetitive and/or inconsistent information
2	Minimally developed position • The essay contains a position on the issue supported by fewer than four reasons • Alternative perspective(s) are not mentioned or discussed • The reasons are not elaborated, or are unrelated to or inconsistent with the position, or are incoherent
1	Undeveloped position • The essay responds to the topic in some way but does not contain a position on the issue • Alternative perspective(s) are not mentioned or discussed • The essay may contain irrelevant information

Note. The holistic scoring rubric was developed based on rubrics used by Anmarkrud et al. (2014), Reznitskaya et al. (2009), and Nussbaum and Schraw (2007).

Received March 1, 2024
Revision received October 16, 2024
Accepted October 25, 2024 ■