



Deliberate erring improves scientific reasoning: Learning experimentation skills by intentionally designing invalid experiments

Sarah Shi Hui Wong^{1,2}

Received: 31 March 2025 / Accepted: 16 July 2026
© The Author(s) 2026

Abstract

Mastering skills in designing valid, controlled experiments is foundational for scientific reasoning and inquiry. Whereas prevailing interventions to develop experimentation skills often induce cognitive conflict through comparing others' invalid experiments with valid ones, this study tested the hypothesis that deliberately committing and correcting one's own errors in invalid designs yields better learning and transfer than studying others' errors. University students ($N=276$) were instructed on experimental design principles (e.g., control of variables), then practiced applying them using one of three learning methods: *derr-explain-fix* (intentionally designing an invalid experiment then explaining and correcting one's deliberate errors) or *spot-explain-fix* (spotting, explaining, and correcting others' deliberate errors in invalid designs) or *errorless practice* (designing a valid experiment and explaining it). All students were then tested on their near and far transfer in designing an experiment that was structurally analogous to or more complex than the one during practice. They were also assessed on critiquing a given design's validity, and their negative knowledge in anticipating problem-specific design errors. Both errorful learning methods outperformed errorless practice on the near transfer, design critique, and error anticipation tests. Crucially, the *derr-explain-fix* method improved near and far transfer more than the *spot-explain-fix* method, which did not yield better far transfer than errorless practice. Yet, students lacked metacognitive awareness of the benefits of deliberate erring and, more broadly, errorful learning. Overall, these findings demonstrate that intentionally designing and correcting one's own invalid experiments during practice is an effective way to enhance learning and transfer of experimental design principles.

Keywords Learning from errors · Deliberate errors · Incorrect worked examples · Control-of-variables strategy · Scientific reasoning · Negative knowledge

✉ Sarah Shi Hui Wong
sarahwong@smu.edu.sg

¹ School of Social Sciences, Singapore Management University, 10 Canning Rise, #05-01, Singapore 179873, Singapore

² Division of Social Sciences, Yale-NUS College, Singapore, Singapore

Introduction

Experimentation skills are crucial in scientific reasoning and inquiry when students construct new knowledge. Broadly, scientific reasoning involves a dual search of two problem spaces: a hypothesis space (e.g., generating hypotheses) and an experiment space (e.g., constructing experiments to test one's hypotheses; Klahr & Dunbar, 1988). Although scientific discovery often begins by asking good research questions that drive and catalyze the inquiry process (Tawfik et al., 2020; Wong et al., 2023), students must further develop skills in designing valid experiments to bring their research ideas and questions into fruition (e.g., Chen & Klahr, 1999; van Riesen et al., 2018). Indeed, developing students' competence in designing fair investigations has been emphasized in national science curricula such as the *Next Generation Science Standards* (National Research Council, 2013) in the U.S., England's *National Curriculum* (Department for Education, 2015), and the *Primary Science Syllabus* in Singapore (Ministry of Education, 2022).

To enhance students' experimentation skills, the present study tested a novel deliberate erring intervention. Specifically, this research compared the relative efficacy of designing valid versus deliberately invalid experiments during practice, and further tested the hypothesis that students in fact learn and transfer better from producing their own intentionally invalid designs than the prevailing approach of receiving others' invalid designs.

Experimentation skills and design principles

When designing experiments, one may rely on domain-general heuristics or principles (Klahr et al., 1993). A cardinal experimental design principle is the *control-of-variables strategy* (CVS), which involves designing unconfounded experiments (i.e., varying only the key variable of interest while holding constant all other variables) so that valid causal inferences can be made (Chen & Klahr, 1999). CVS has been viewed as comprising four distinct subskills, with the last two listed subskills being the most difficult: (a) *identification* of a valid or controlled design from multiple alternative designs, (b) *interpretation* when drawing correct inferences from valid designs, (c) *planning* valid experiments with controlled comparisons, and (d) *understanding* the indeterminacy of invalid experiments (Chen & Klahr, 1999; Peteranderl et al., 2023; Schwichow, Christoph, et al., 2016). The logic of controlling variables relates to the domain-general idea of causality across the natural and social sciences (Woodward, 2003), such that CVS has been widely considered as essential—and even the hallmark—in the development of scientific reasoning (Schwichow et al., 2020).

Beyond control of variables, though, experimental design also encompasses a suite of other important principles such as including an appropriate control group, measuring relevant covariates, random assignment, and using a pretest-posttest design. Such design principles are commonly included in undergraduate research methods course syllabi and textbooks, particularly since students may not spontaneously use such principles or may hold misconceptions about them (Nachtigall et al., 2020). Hence, this study examined holistic learning of these principles, in addition to CVS that has been the centerpiece of much research (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lorch et al., 2010, 2014).

Learning experimentation skills

Learning the sophisticated use of experimentation skills can be challenging. Although a rudimentary grasp of experimentation skills develops in elementary school (Sandoval et al., 2014; Zimmerman, 2007), even adults struggle with evaluating and designing valid experiments for complex research problems (Bullock et al., 2009; Lin & Lehman, 1999). Instruction, even if brief, can benefit students' acquisition and use of experimental design principles (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lazonder & Egberink, 2014; Lorch et al., 2017; Schalk et al., 2019). For instance, young children in Chen and Klahr's (1999) seminal study were able to learn and apply CVS to design unconfounded experiments after explicit training with probe questions, but not when they only received probe questions.

However, not all kinds of instruction are equally effective for improving experimentation skills (Martella et al., 2020). Meta-analytic evidence suggests that *how* instruction is implemented matters—CVS interventions that use strategies to induce cognitive conflict have significantly larger effects ($g=0.80$) than those that do not ($g=0.53$; Schwichow, Croker, et al., 2016). Rooted in Piagetian theory, *cognitive conflict* involves explicitly challenging students' expectations and conceptions with contradictions to create disequilibrium, which is resolved when students refine or restructure their knowledge schemas for conceptual change (Limón, 2001). When learning experimentation skills, cognitive conflict occurs when students are confronted with problematic features of a design (e.g., confounds) that prompt them to think about why such designs are inadequate or invalid (Schwichow et al., 2020). Such problematic features can also be considered design errors, in that they deviate from the canonical solutions for designing fair experiments to enable valid causal inferences. In particular, cognitive conflict is often induced through demonstrations or comparisons of valid versus invalid (i.e., erroneous) designs.

Learning from valid versus invalid experimental designs

In interventions for teaching experimentation skills, students are typically presented with examples of invalid designs and asked to evaluate whether they enable a valid comparison to test a particular hypothesis (Schwichow, Croker, et al., 2016). Upon recognizing that the design is invalid, students are then asked to fix or correct it so that it becomes valid. This approach parallels the use of incorrect worked examples in mathematics, whereby learners are presented with examples of solutions containing common errors, which they must then spot and explain (e.g., Durkin & Rittle-Johnson, 2012; Große & Renkl, 2007; Heemsoth & Heinze, 2014). Indeed, besides the benefits of incorrect worked examples for mathematical learning and transfer (Booth et al., 2015), studying invalid designs promotes transfer of experimental principles (Corral & Carpenter, 2020; Lorch et al., 2014). For instance, Lorch et al. (2014) compared learning from invalid designs (i.e., explaining why a given confounded experiment is bad, then correcting it) versus valid designs (i.e., explaining why a given valid experiment is good, then generating another valid comparison). On transfer tests that required designing experiments in a new domain and medium or evaluating the validity of specific comparisons, students performed better when they had learned from invalid than valid designs.

Indeed, impasse-driven learning theory posits that successful learning requires encountering impasses (e.g., errors or obstacles) that motivate students to actively construct a better understanding and overcome the impasse (VanLehn et al., 2003). Presumably, studying others' errors in invalid designs and comparing them with valid ones induces cognitive conflict that prompts students to think deeply about the logic of design principles, recognize their diagnostic features, and revise one's faulty knowledge structures for a deeper understanding that aids transfer (Corral & Carpenter, 2020; Loibl et al., 2017; Lorch et al., 2014). For instance, when reflecting on others' errors, students may develop *negative knowledge*—knowledge of incorrect responses or what *not* to do (Gartmeier et al., 2008)—that yields richer mental models for better learning than studying correct solutions only (Barbieri et al., 2021; Heemsoth & Heinze, 2014). Interestingly, though, recent evidence suggests that deliberately committing and correcting one's own errors can yield superior learning and transfer than studying others' errors (Wong, 2023; Yap & Wong, 2024).

Learning from deliberate errors

According to Wong and Lim's (2019b) Prevention–Permission–Promotion (3P) framework, errors in learning can be observed, allowed, induced, or guided when not avoided. Whereas learning from others' invalid designs involves observing errors, a less studied approach is to guide students into deliberately making and correcting errors in low-stakes contexts, even after they have been instructed on the correct responses. Of note, deliberate erring has been found to enhance learning in a counterintuitive phenomenon termed the *derring effect*, as coined from a portmanteau of the words “deliberate” and “erring” (Wong & Lim, 2022a, 2022b). In the first studies of the derring effect, deliberate erring was implemented during open-book study by having students write each sentence in to-be-learned scientific term-definition concept lists (Wong & Lim, 2022b) or expository texts (Wong & Lim, 2022a) such that it contained a plausible conceptual error that they had intentionally generated, then striking out their error and correcting it. The key finding was that deliberate erring enhanced recall and application performance more than errorless methods such as copying, concept-mapping, and generating synonyms or examples.

To probe the generalizability of the derring effect, further work has begun to test whether it holds over other effective errorful learning methods. For instance, Wong (2023) compared deliberate erring against not only generating synonyms, but also a spot-and-fix condition in which students additionally spotted and corrected their peers' deliberate errors when studying scientific texts. On a far transfer test that required applying the learned concepts in remote knowledge domains, deliberate erring outperformed both the synonym generation and spot-and-fix methods. This advantage of deliberate erring has also been extended to procedural transfer in mathematics. Yap and Wong (2024) had students receive instruction on mathematical algorithms then solve practice problems either via errorless repeated practice, deliberately committing and correcting procedural errors, or studying incorrect worked examples using a spot-explain-fix method (i.e., finding, explaining, and correcting their peers' deliberate errors). On transfer test questions that were structurally more complex or that required students to construct a coherent mental representation of the problem situation, deliberate erring outperformed both errorless practice and the spot-explain-fix method. Taken together, these findings suggest that deliberate erring is a promising approach to boost not only recall, but also higher order outcomes such as transfer. Crucially, these benefits are

specific to guiding students to commit and correct their own errors—observing and correcting others’ errors does not yield the same extent of gains (Metcalf & Xu, 2018).

Several candidate mechanisms have been proposed to explain the derring effect. Based on theories of failure-driven learning that encountering errors triggers inquiry processes that may not otherwise occur during errorless learning (Tawfik et al., 2015), it is possible that deliberate erring evokes mental processes that are not typically invited by the material but that are useful for learning it (Wong, 2023; Wong & Lim, 2022a). For instance, deliberate erring may afford opportunities for students to develop negative knowledge, thus promoting a richer understanding that includes not only correct responses but also incorrect ones for better learning. Exploring incorrect responses may also ironically weaken and suppress those pathways, thereby strengthening pathways to the correct responses (Kornell et al., 2009). Alternatively, intentionally generating errors may direct attention toward their correction for enhanced processing (Wong & Lim, 2022b). For instance, 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 incorrect responses may stimulate greater depth of search (Siegler & Chen, 2008), whereby students must also reflect on the correct response and its scope of application (Heemsoth & Heinze, 2014), toward developing more sophisticated strategies.

Such deeper processing could enhance learning and transfer to new problems. Failure to transfer often occurs when initial learning is shallow (Chi & VanLehn, 2012); students are more likely to successfully transfer their knowledge with deeper initial learning (Kimball & Holyoak, 2000), particularly when they have mindfully abstracted a more general representation of the to-be-learned information that can bridge multiple contexts (Salomon & Perkins, 1989). For instance, comparing correct versus incorrect solutions during deliberate erring could induce structural alignment processes (i.e., aligning relational structures) that promote the noticing of both structural commonalities and differences, which can then be applied more broadly (Gentner & Gunn, 2001; Gentner & Markman, 1997). Indeed, that deliberate erring facilitates transfer to problems that demand constructing a mental representation of the problem situation (Yap & Wong, 2024) suggests that this technique may foster recognition, encoding, and/or mapping of the target knowledge’s deep structural features for better transfer.

Further, students may engage in more or increased levels of such processes when generating their own errors than studying others’ errors. Some evidence suggests that constructing one’s own representations of knowledge directs students’ attention more strongly toward the underlying structure of the problem for better learning and transfer, relative to simply receiving those representations (Terwel et al., 2009). Hence, although learning from one’s own intentionally invalid designs versus others’ invalid designs both involve comparing incorrect versus correct responses, deliberately generating one’s own errors may promote deeper learning of the target knowledge, thereby facilitating transfer more than observing others’ errors (Wong, 2023; Yap & Wong, 2024).

The present study

The present study tested whether deliberately committing and correcting one’s own errors—intentionally designing invalid experiments—enhances learning and transfer more than the prevailing approach of observing and correcting others’ errors when learning experimenta-

tion skills. Here, university students were first instructed on experimental design principles, then practiced applying them to design an experiment that tested a given hypothesis using one of three learning methods: *derr-explain-fix* (i.e., deliberately designing an invalid experiment then explaining and correcting one's errors) or *spot-explain-fix* (i.e., spotting, explaining, and correcting others' errors in invalid designs) or *errorless practice* (i.e., designing a valid experiment and explaining one's solution; control condition).

Using a multidimensional approach, students' learning was assessed across a range of four outcomes: near transfer, far transfer, design critique, and negative knowledge. This approach allowed for systematic testing and differentiation of any variation in the intervention's effects across distinct experimentation subskills (Peteranderl et al., 2023; Schwichow, Christoph, et al., 2016). Specifically, all students were first tested on their near and far transfer of the principles in designing a valid experiment to test a novel given hypothesis. The near transfer test involved a design that was structurally similar to the one that students had encountered during practice, albeit with different surface features (Gentner et al., 1993). Conversely, the far transfer test involved a structurally more complex design that students had not encountered during practice, and that took on a novel problem representation.¹ In addition, a design critique test measured students' ability to analyze and evaluate the validity of a given design based on the principles they had learned.

Respectively, the transfer and design critique tests thus assessed the CVS subskills of *planning* valid experiments and *understanding* the indeterminacy of invalid ones (Chen & Klahr, 1999), which have been considered the most difficult CVS subskills (Peteranderl et al., 2023; Schwichow, Christoph, et al., 2016). In line with prior research (Schwichow, Croker, et al., 2016), inducing cognitive conflict via learning from invalid designs was expected to enhance transfer and design critique performance more than errorless practice with valid designs. Importantly, to the extent that deliberately committing and correcting one's own errors produces better transfer than spotting and correcting others' errors (Wong, 2023; Yap & Wong, 2024), the *derr-explain-fix* method would outperform the *spot-explain-fix* method.

Finally, as a novel measure of their negative knowledge of the design principles (i.e., how *not* to design an experiment), students completed an error anticipation test in which they described as many problem-specific design errors as possible that a researcher might make in a given context. Based on extant findings that studying incorrect rather than correct examples promotes negative knowledge (Heemsoth & Heinze, 2014; Joung et al., 2006), both errorful learning methods were expected to yield better error anticipation performance than the errorless practice control.

At the same time, students have often been found to be unaware of the benefits of errorful learning (e.g., Huelser & Metcalfe, 2012; Yang et al., 2017), including deliberate erring (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024). Yet, accurate metacognitive knowledge is vital for effective self-regulated learning (Bjork et al., 2013). If students are inaccurate in estimating their learning and recognizing effective learning methods, they may fail to make optimal study choices (Metcalfe & Finn, 2008; Yang et al., 2017). Hence, in this study, students' metacognitive judgments of their learning were collected before and after the tests to examine their subjective assessments of the learning methods' effects on their performance.

¹ As in extant research (e.g., Chen & Klahr, 1999), the metric for transfer "distance" per se is undefined, although the relative ordering of the "near" versus "far" transfer tests is unambiguous.

Method

Participants and design

The participants were 276 students (201 women; 75 men) between the ages of 18 and 45 ($M=21.35$, $SD=2.91$) recruited from both the Psychology Research Participation Program and university-wide Research Participation Scheme at a large public university in Singapore, who received either course credit or monetary reimbursement for their participation. The students came from a range of majors and years in college. Extant deliberate erring studies have reported medium to large effect sizes for transfer ranging from $d=0.46$ to 0.77 (Wong, 2023; Wong & Lim, 2022a; Yap & Wong, 2024). Based on the most conservative effect size ($d=0.46$), a power analysis (G*Power) indicated that at least 76 participants per condition were required for two-tailed between-subjects pairwise comparisons (i.e., t -tests) at 80% power and $\alpha = .05$. Participants were randomly assigned ($n=92$ per condition) to use one of three learning methods: *derr-explain-fix*, *spot-explain-fix*, or *errorless practice*. The four learning outcomes were their near transfer, far transfer, design critique, and error anticipation test performance.

All participants provided their written informed consent. This study was approved by and conducted in accordance with the ethical standards of the university's Institutional Review Board.

Materials and procedure

Prior knowledge questionnaire

Before attending the experiment, participants completed a 9-item online questionnaire that measured their self-assessed prior knowledge of the to-be-learned experimental design principles (adapted from Fiorella & Mayer, 2013). Participants rated their knowledge of designing an experiment that compares three or more groups (1 = *very low*; 7 = *very high*), and reported whether they possessed prior knowledge of eight content items (*yes/no*): (a) "I have taken a university course in research methods and/or statistics", (b) "I know what a confound means", (c) "I know what a covariate is", (d) "I know how random assignment works", (e) "I know what a factorial design is", (f) "I know what an interaction effect is", (g) "I have used an ANOVA before", and (h) "I know the basic characteristics of experimental control to determine causality". As opposed to presenting actual test questions, this relatively more indirect measure of prior knowledge was used to avoid inducing any pretesting effects (Richland et al., 2009) when participants later studied the lesson materials. A prior knowledge score was computed for each participant by summing their prior knowledge rating (out of 7) and the number of content items that they reported having prior knowledge of (out of 8); the maximum possible score was 15.

Experimental design lesson and conceptual understanding check questions

Participants were run in groups of up to six per session, and completed the experiment individually. Upon arriving for their experimental session, participants were told that they would be studying research design principles that they would later be tested on. The nature

of the tests was not revealed. All participants were then instructed on the principles via a 25-min video-recorded lesson (transcript available in Online Supplementary Materials). The lesson was presented as an automated slideshow with audio narration that sequentially introduced five key principles in experimental design: (a) “control group”, (b) “measuring relevant covariates”, (c) “systematically varying key variables”, (d) “random assignment”, and (e) “pretest-posttest design”. For each principle, the lesson explained its definition, described an example of how it could be applied to test a given hypothesis, and explained with examples of valid and invalid designs why the principle was important for experimental rigor. So that participants could focus on understanding the principles without excessive cognitive load from having to simultaneously process multiple research scenarios, all five principles were illustrated using the same scenario involving a three-level single-factor design (note-taking: *no note-taking* vs. *writing notes* vs. *typing notes*; see Appendix A).

Similar to how textbooks often include checkpoint questions at the end of each topic or chapter, a conceptual understanding check question was presented onscreen after each principle had been taught (available in Online Supplementary Materials). A sample check question was: “Explain what a control group is, and how it can help researchers to determine whether an experimental treatment has produced an effect”. Participants had 1.5 min to write down their response, before the solution was presented as feedback onscreen. After all five principles had been taught, the lesson concluded with a brief summary.

Post-lesson questionnaire

At the end of the lesson, participants completed a post-lesson questionnaire (adapted from Wong & Lim, 2022a) in which they rated how interesting and understandable the lesson was, respectively (1 = *not at all*; 7 = *extremely*). As a measure of their situational emotions during the lesson, participants also completed the 9-item Situational Emotion Short Scale (Randler et al., 2011). The scale includes three dimensions with three items each: (a) interest (e.g., “I found the topic important”), (b) well-being (e.g., “I enjoyed the lesson”), and (c) boredom (e.g., “I felt bored”). All ratings were made on a 7-point scale (1 = *not at all*; 7 = *very much*). Participants’ mean rating on each of the three dimensions was computed. The scale demonstrated good reliability in the present study, Cronbach’s $\alpha = .83, .88$, and $.78$ for the interest, well-being, and boredom dimensions, respectively.

Intervention

All participants were then given 15 min to practice applying the five principles to design an experiment that tested a given novel hypothesis involving a single factor with three levels (coffee-drinking: 0 cups vs. 1 cup vs. 2 cups; see Appendix B). In the *errorless practice* condition, participants applied the principles to design an optimal study to test the given hypothesis, then explained their solutions. Conversely, participants in the *derr-explain-fix* condition applied the principles to design a non-optimal study to test the given hypothesis, explained their deliberate errors, then corrected them. In the *spot-explain-fix* condition, each participant was presented with a unique set of deliberate errors that a peer had made in the *derr-explain-fix* condition. This yoked procedure controlled for the quality of errors that par-

ticipants were exposed to in both errorful conditions.² Using the spot-explain-fix method, participants applied the principles to spot their peers' errors, explained those errors, then corrected them. Specific instructions for all learning methods are available in Appendix C.

Across all conditions, participants were prompted to explicitly explain their responses since reason justification improves transfer of design principles (Lin & Lehman, 1999). To ensure that participants understood what was required of them, they were shown a sample response for their learning method using the same research scenario from the lesson (see Appendix D) before they independently completed the intervention without feedback. All participants engaged in open-book practice whereby they were allowed access to a hand-out detailing the principles' definitions (available in Online Supplementary Materials). This simulated naturalistic learning contexts and also ensured that participants' errors in the derr-explain-fix condition were deliberately made with knowledge of the correct principles. All participants were told that if they completed the activity before the 15 min were up, they should use the remaining time to check their responses and ensure they had written in as much specific detail as possible. Thus, the total intervention time was equated across all conditions.

Post-intervention questionnaire

A post-intervention questionnaire measured participants' perceived learning after the intervention. First, they completed a 3-item perceived competence scale (Loibl & Rummel, 2015; Nachtigall et al., 2020): (a) "I am satisfied with my performance during the learning activity", (b) "I was skilled in the activities during the learning activity", and (c) "I think I was pretty good at the learning activity". All ratings were made on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Participants' mean perceived competence score was computed. The scale demonstrated high reliability in this study, Cronbach's $\alpha = .92$.

Next, participants made six metacognitive higher order judgments of learning (JOL+; Wong & Lim, 2019a) on a 11-point scale from 0% (*definitely not well*) to 100% (*definitely well*): (a) "How well do you think you understood the experimental design principles?", (b) "How likely are you to remember the experimental design principles on a future test?", (c) "How much do you think you have learned about the experimental design principles?", (d) "How well do you think you will be able to design a study applying the experimental design principles?", (e) "How likely will you be able to analyze whether a study design has applied the experimental design principles?", and (f) "How well do you think you will be able to evaluate a study design?". Participants' mean JOL+ rating was computed.

Tests

Then, participants completed the near transfer, far transfer, design critique, and error anticipation tests. Participants were given 10 min to complete each test without access to the lesson materials, and were allowed to take brief self-paced breaks in-between the tests.

² To collect an initial sample of deliberate errors for presentation in the spot-explain-fix condition, 10 participants were first run in the derr-explain-fix condition, thereby contributing 10 unique sets of deliberate errors to be presented to 10 participants in the spot-explain-fix condition. Thereafter, both the derr-explain-fix and spot-explain-fix conditions were run simultaneously with yoking done on a rolling basis, such that each derr-explain-fix participant contributed a unique set of deliberate errors that were presented to a randomly yoked participant in the spot-explain-fix condition.

Near transfer test In the near transfer test, participants were presented with a novel research scenario that was structurally analogous to the one that they had encountered in the intervention, albeit with different surface features (Gentner et al., 1993). Specifically, the near transfer scenario involved a three-level single-factor design (photo-editing: *no photo-editing* vs. *photo-cropping* vs. *photo-filtering*) on the effects of photo-editing on memory of artworks (see Appendix E). The test required that participants apply the five principles they had learned to design a study to test the given hypothesis.

Far transfer test In the far transfer test, a novel research scenario was presented as a transcript of a dialogue among three science teachers (adapted from Nachtigall et al., 2020; see Appendix F). Thus, the far transfer scenario took on a problem representation that differed from that in the intervention and near transfer test scenarios. Importantly, the far transfer scenario involved a structurally more complex design that participants had not previously encountered but had to infer from the dialogue. Specifically, the scenario required a 2×2 factorial design on the effects of problem-solving context (*working alone* vs. *working in groups*) and computer support (*present* vs. *absent*) on science problem-solving performance, with English language proficiency as a covariate. Participants were asked to apply the five principles they had learned to design a study to test the given hypotheses.

Design critique test The design critique test required participants to analyze and evaluate the validity of a given design. All participants were presented with a scenario of a design that tested whether people believe they are at lower risk of catching the flu from others in the dark or with the lights on (see Appendix G). Participants then analyzed the extent that the design was sound based on the five principles they had learned. The design was adapted from an actual retracted psychological study. Thus, before responding to the test question, participants were asked to indicate whether they knew about the given scenario prior to the experiment. No participant reported having prior knowledge of the scenario.

Error anticipation test The error anticipation test assessed participants' ability to anticipate problem-specific errors that others might make when designing experiments (Barbieri et al., 2021). Participants were presented with a scenario where a researcher wanted to test whether students would learn Korean vocabulary better if they spread out their revision over multiple sessions instead of cramming in a single session (adapted from Goodwin & Goodwin, 2014; see Appendix H). Participants were asked to describe as many potential design errors as possible that the researcher might make within the specific study context.

Method effectiveness rating

Finally, after completing all four tests, participants rated how effective their learning method had been in helping their test performance (1 = *not at all*; 7 = *extremely*).

Scoring procedure

Two raters independently scored all 276 (100%) scripts. Discrepancies were reviewed and resolved to reach 100% agreement. Appendix I presents sample responses for the near transfer, far transfer, and design critique tests, whereas Appendix J presents sample responses for the error anticipation test.

Conceptual understanding check questions

Participants received 1 point for each of the five conceptual understanding check questions that they answered correctly during the lesson. Interrater reliability was high, absolute agreement intraclass correlation (ICC) = .90, 95% CI [.88, .92], based on a two-way random-effects model.

Near transfer test

Following extant scoring procedures (Nachtigall et al., 2020), participants' near transfer test responses were scored by awarding points for two sets of principles (see Table 1): *implicit design principles* that had not been explicitly explained during the lesson as they do not represent typical student misconceptions (e.g., specifying the correct conditions), and *explicit design principles* that had been explicitly taught during the lesson as they build on typical student misconceptions (e.g., including an appropriate control group). Under the implicit design principles, participants were awarded 1 point for proposing a design with the correct number of conditions (i.e., three conditions), and additional points for correctly naming each condition (i.e., up to 3 points). Under the explicit design principles, participants could earn 2 points for correctly applying and elaborating on each of the five taught principles, for up to 10 points. Thus, the maximum possible score was 14.

Importantly, participants' responses had to effectively address the specific hypothesis in the test question to earn points. For example, responses proposing a control group in which participants took photos of the artworks but did not edit them would earn 1 point for the "control group" principle, with an additional point for explaining how exactly other variables would be held constant except for the act of editing one's photos (e.g., having the control group use the same smartphone as the experimental groups to photograph the same artworks). In contrast, responses that did not effectively address the specific hypothesis (e.g., "There should be a control group.") or incorrectly applied the principles (e.g., "The control group will be people who don't take photos of the artworks.") did not earn any points. Interrater reliability was high, ICC = .96, 95% CI [.96, .97].

Far transfer test

The far transfer test was scored using similar criteria as the near transfer test (see Table 1), except that an additional mark could be earned for the implicit design principle of correctly naming each condition since the far transfer scenario had four conditions. Thus, the maximum possible score was 15. Interrater reliability was high, ICC = .96, 95% CI [.94, .98].

Table 1 Scoring criteria for near transfer test, far transfer test, and design critique test

Principle	Criterion	Near Transfer Test	Far Transfer Test	Design Critique Test
Implicit design principle (i.e., not focused on during instruction)	Specifying correct number of conditions	✓	✓	
	Correctly naming each condition	✓✓✓	✓✓✓✓	
Control group	Naming an appropriate control group in the specific study context (i.e., specifying that a particular condition is the control)	✓	✓	✓
	Explaining how the control group would experience the same conditions as the experimental groups in all ways, except for the treatment	✓	✓	✓
Measuring relevant covariates	Naming an appropriate covariate in the specific study context	✓	✓	✓
	Explaining how the covariate would be measured	✓	✓	✓
Systematically varying key variables	Naming a potential confound in the specific study context	✓	✓	✓
	Explaining how the confound can be avoided in the specific study context	✓	✓	✓
Random assignment	Randomly assigning participants to conditions	✓	✓	✓
	Explaining how random assignment would specifically be done	✓	✓	✓
Pretest-posttest design	Including a pretest and posttest (e.g., specifying that they occur before vs. after the treatment)	✓	✓	✓
	Explaining what the pretest and posttest would specifically involve or measure	✓	✓	✓

Note. Each tick (✓) represents 1 point that could be earned in each test. For the implicit design principle, the near transfer test scenario involved three experimental conditions (i.e., 3 points for correctly naming each condition), whereas the far transfer test scenario involved a factorial design with four experimental conditions (i.e., 4 points for correctly naming each condition); the implicit design principle was not applicable to the design critique test. The maximum possible score was 14 for the near transfer test, 15 for the far transfer test, and 10 for the design critique test

Design critique test

The design critique test was scored using similar criteria as the transfer tests (see Table 1), except that the implicit design principle was not applicable since the design critique test involved a scenario with given conditions. Thus, the maximum possible score was 10—participants could earn 2 points for each of the five explicit design principles in correctly applying and elaborating on them to analyze whether the given research scenario was valid. For example, a response would earn 1 point for the “systematically varying key variables” principle for identifying season of testing as a confound in the scenario (e.g., “Dr. Lee should not have run one group in winter and the other group in spring as the season could be a confound should participants believe, for example, that they are more vulnerable to catching the flu from others during winter.”), and an additional point for explaining how this confound could have been avoided (e.g., “Both groups should have been tested in the same season instead.”). Conversely, responses that incorrectly applied the principles (e.g., “We

cannot test the randomness of her groupings as we do not know how she obtained her sample.”) did not earn any points. Interrater reliability was high, ICC = .88, 95% CI [.85, .91].

Error anticipation test

Following extant scoring procedures (Barbieri et al., 2021), one point was awarded for each problem-specific error that participants proposed on the error anticipation test; problem-general errors did not receive points. Problem-specific errors demonstrated a deep understanding of the potential design errors that could plausibly arise in the specific study context, whereas problem-general errors could more generally apply to other study contexts too. Interrater reliability was high, ICC = .87, 95% CI [.84, .90].

Results

Preliminary analyses

One-way between-subjects analyses of variance (ANOVAs) were run to ascertain that the three learning groups did not differ in their self-assessed prior knowledge, performance on the conceptual understanding check questions, perceptions of the lesson, and situational emotions during the lesson. Means and standard deviations are presented in Table 2.

Prior knowledge questionnaire and conceptual understanding check questions

Before the lesson, participants reported relatively low prior knowledge of the design principles, with no significant difference across conditions, $F(2, 273)=0.48$, $p=.62$, $\eta^2=.004$. After receiving instruction, participants displayed high basic understanding of the taught principles on the five conceptual understanding check questions during the lesson, with no significant difference across conditions, $F(2, 273)=0.54$, $p=.59$, $\eta^2=.004$.

Table 2 Means and standard deviations for prior knowledge questionnaire, conceptual understanding check questions, post-lesson questionnaire, post-intervention questionnaire, and method effectiveness ratings

Variables	Errorless Practice		Spot-Explain-Fix		Derr-Explain-Fix	
	M	SD	M	SD	M	SD
Prior knowledge	4.57	2.67	4.82	2.76	4.93	2.36
Conceptual understanding check questions	4.33	0.76	4.43	0.73	4.36	0.70
Post-lesson questionnaire						
Lesson interestingness	4.68	1.26	4.58	1.16	4.75	1.17
Lesson understandability	5.96	0.93	6.05	0.82	6.10	0.87
Situational interest	4.81	1.11	4.79	1.18	4.90	1.01
Situational well-being	4.80	1.10	4.81	0.98	5.00	0.98
Situational boredom	3.18	1.28	2.94	1.22	2.97	1.17
Post-intervention questionnaire						
Perceived competence	4.32	1.07	4.24	1.08	4.08	1.06
Judgments of higher order learning (JOL+)	64.47	14.04	64.40	14.23	63.30	13.59
Method effectiveness	4.91	1.16	4.92	1.15	4.95	1.11

Note. $N=276$

Post-lesson questionnaire

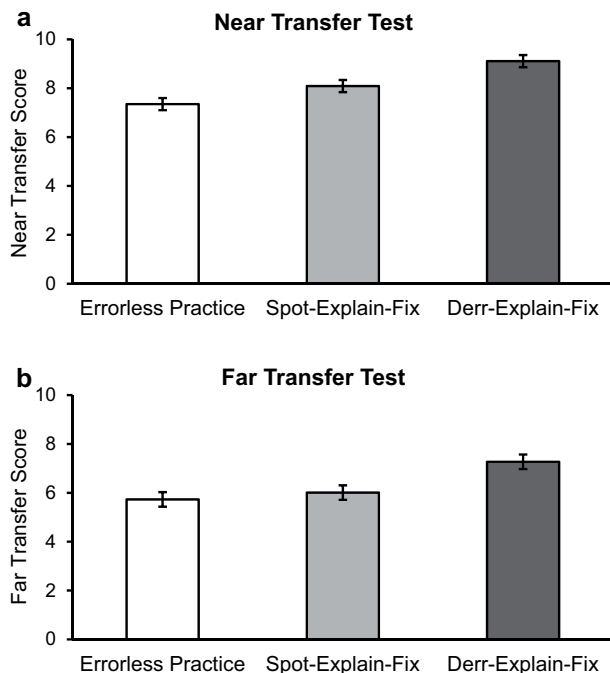
Across conditions, participants did not significantly differ in how interesting they perceived the lesson to be, $F(2, 273)=0.50$, $p=.61$, $\eta^2=.004$, or how understandable they perceived the lesson to be, $F(2, 273)=0.64$, $p=.53$, $\eta^2=.01$. In addition, the three learning groups did not differ in their self-reported situational emotions during the lesson across the interest dimension, $F(2, 273)=0.25$, $p=.78$, $\eta^2=.002$, wellbeing dimension, $F(2, 273)=1.10$, $p=.33$, $\eta^2=.01$, and boredom dimension, $F(2, 273)=1.10$, $p=.33$, $\eta^2=.01$.

Main analyses

Near transfer test performance

A one-way between-subjects ANOVA revealed that the learning groups significantly differed in their near transfer test performance, $F(2, 273)=12.66$, $p<.001$, $\eta^2=.09$. As expected, the spot-explain-fix ($M=8.09$, $SD=2.51$) and derr-explain-fix ($M=9.11$, $SD=2.23$) methods produced better near transfer than errorless practice ($M=7.35$, $SD=2.40$), $p=.036$ and $p<.001$, $d=0.30$ and 0.76 , respectively. Importantly, the derr-explain-fix group outperformed the spot-explain-fix group, $p=.004$, $d=0.43$. Thus, students' near transfer benefited more from errorful than errorless learning, with additional gains from generating their own deliberately invalid designs than learning from others' invalid designs (Fig. 1a).

Fig. 1 Near transfer and far transfer test performance. *Note.* (a) and (b) show the mean near transfer and far transfer test scores, respectively. Error bars indicate standard errors



Far transfer test performance

The three learning groups also significantly differed in their far transfer test performance, $F(2, 273)=7.58$, $p=.001$, $\eta^2=.05$, albeit with a different pattern. The derr-explain-fix group ($M=7.27$, $SD=2.56$) outperformed both the errorless practice ($M=5.73$, $SD=2.93$) and spot-explain-fix ($M=6.01$, $SD=3.07$) groups, $p<.001$ and $p=.003$, $d=0.56$ and 0.45 , respectively, whereas the latter two groups did not differ, $p=.50$, $d=0.09$. Hence, the advantage of deliberate erring, but not that of observing and correcting others' invalid designs, persisted even on the structurally more complex far transfer test (Fig. 1b).

Design critique test performance

An errorful learning advantage emerged on the design critique test, $F(2, 273)=4.07$, $p=.018$, $\eta^2=.03$, whereby the spot-explain-fix ($M=3.66$, $SD=1.76$) and derr-explain-fix ($M=3.61$, $SD=1.49$) groups outperformed the errorless practice control ($M=3.08$, $SD=1.35$), $p=.01$ and $.02$, both $ds=0.37$, respectively. Both errorful learning groups did not significantly differ, $p=.81$, $d=-0.03$. Thus, learning from invalid designs—whether self- or other-generated—enhanced students' ability to understand and critique the indeterminacy of invalid experiments, relative to learning from valid designs only (Fig. 2).

Error anticipation test performance

As with the design critique test, there was an errorful learning advantage on the error anticipation test, $F(2, 273)=13.62$, $p<.001$, $\eta^2=.09$. The spot-explain-fix ($M=1.48$, $SD=1.11$) and derr-explain-fix ($M=1.46$, $SD=0.94$) groups anticipated more problem-specific errors than the errorless practice control ($M=0.82$, $SD=0.86$), both $ps<.001$, $d=0.66$ and 0.71 , respectively. Both errorful learning groups did not significantly differ, $p=.88$, $d=-0.02$. Hence, learning from invalid designs—whether self- or other-generated—produced better negative knowledge than learning from valid designs only (Fig. 3).

Metacognitive ratings

Yet, participants' metacognitive judgments revealed that they were largely unaware of the benefits of deliberate erring and, more broadly, errorful learning. After the intervention, the three learning groups neither differed in their mean perceived competence ratings, $F(2, 273)=1.19$, $p=.31$, $\eta^2=.01$, nor mean JOL+ ratings, $F(2, 273)=0.21$, $p=.81$, $\eta^2=.002$. Even after the tests when participants had experienced the effects of their learning method,

Fig. 2 Design critique test performance.
Note. Error bars indicate standard errors

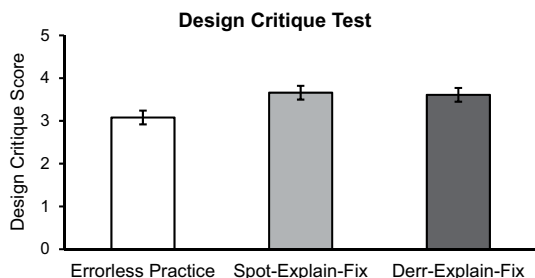
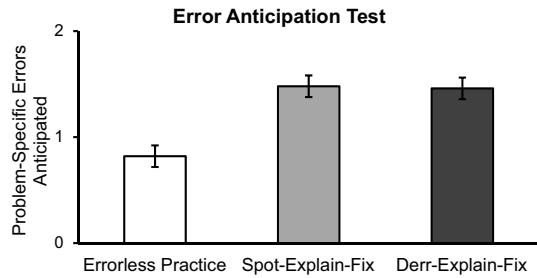


Fig. 3 Error anticipation test performance.
 Note. Error bars indicate standard errors



the three groups did not differ in how effective they perceived their learning method to be, $F(2, 273)=0.02$, $p=.98$, $\eta^2<.001$. Means and standard deviations are presented in Table 2.

Discussion

To foster experimentation skills, prevailing interventions often induce cognitive conflict by having students compare given examples of invalid versus valid designs (Schwichow, Croker, et al., 2016). Drawing on the derring effect (Wong & Lim, 2022a, 2022b), though, this study found that deliberately committing and correcting one's own errors in invalid designs yields better learning and transfer of experimental design principles, relative to observing and correcting others' errors, as well as errorless practice.

First, consistent with prior research on the benefits of using invalid designs in CVS interventions (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lorch et al., 2010, 2014), an errorful learning advantage emerged. As compared to designing valid experiments during errorless practice, learning from invalid designs—whether self- or other-generated—produced: (a) better near transfer of the learned principles when designing a valid experiment for a structurally analogous problem, (b) better understanding of the indeterminacy of invalid experiments on a design critique test, and (c) better negative knowledge in anticipating problem-specific design errors that others might make in a given research scenario.

But despite boosting near transfer more than errorless practice, learning from others' invalid designs did not yield better far transfer when designing a valid experiment for a structurally more complex problem. Rather, designing and correcting *one's own* deliberately invalid experiments produced greater benefits for students' transfer—both near and far—of the learned principles, relative to spotting, explaining, and correcting the deliberate errors in their peers' invalid designs. This advantage of self- over other-generated errors resonates with extant findings in errorful learning research (e.g., Kapur, 2014; Metcalfe & Xu, 2018), while extending the derring effect to the transfer of more general principles beyond specific concepts or procedures (Wong, 2023; Yap & Wong, 2024). These findings also illuminate the nuanced differences in the effects of observing versus guiding errors (Wong & Lim, 2019b), going beyond typical dichotomous comparisons of errorless versus errorful learning.

Theoretical implications

The differential effects across the four tests highlight the importance of distinguishing among various experimentation subskills (Peteranderl et al., 2023; Schwichow et al., 2020). Whereas both errorful learning methods similarly improved students' skills in understanding the indeterminacy of invalid experiments (i.e., design critique test) and anticipating problem-specific errors (i.e., error anticipation test), the spot-explain-fix method fell short against deliberate erring in boosting students' skills in planning valid experiments for both structurally analogous (i.e., near transfer test) and dissimilar (i.e., far transfer test) problems.

On one hand, these data reveal that engaging with errors rather than avoiding them during practice enriches students' mental models with negative knowledge of incorrect or inadequate solutions, as borne out on the design critique and error anticipation tests. These findings replicate and extend prior research that has observed similar benefits of learning from incorrect worked examples in mathematics (Barbieri et al., 2021; Heemsoth & Heinze, 2014) and from case studies of others' errors in firefighting practices (Joung et al., 2006). Importantly, the present results demonstrate that this learning advantage occurs not only when observing others' errors, but also when deliberately erring oneself. In sum, learning from others' and one's own errors in invalid designs boosts negative knowledge of what *not* to do, relative to errorless practice with valid designs only.

Yet, better negative knowledge of what constitute invalid experiments did not insure better far transfer in designing valid experiments for more complex problems. Despite improving students' design critique and error anticipation performance, the spot-explain-fix method fared no better than errorless practice on the far transfer test. Moreover, despite rivaling the derr-explain-fix method on the design critique and error anticipation tests, the spot-explain-fix method produced weaker near and far transfer. Thus, a key implication is that knowing what not to do does not necessarily suffice for knowing what to do (Gartmeier et al., 2008; Heemsoth & Heinze, 2014), especially in novel and more challenging situations. Rather, deliberately committing and correcting one's own errors is vital for deep learning of the correct responses and strategies, besides incorrect ones.

Given that the spot-explain-fix group received the same deliberate errors that their peers had generated in the derr-explain-fix condition, both groups' varying test performance cannot be due to differences in the quality of errors that they had been exposed to. Rather, failure to transfer often stems from a lack of deep initial learning—when students have deeply learned the source information, they are more likely to successfully transfer it to new problems (Chi & VanLehn, 2012; Gick & Holyoak, 1983). For instance, generating a correct response (e.g., a valid design) and considering why it is correct may evoke a more superficial search of one's existing knowledge, whereas cognitive conflict when confronting the inadequacies of an error (e.g., an invalid design) may evoke a deeper search of one's knowledge (Siegler & Chen, 2008) in which one also reflects on the correct response and the conditions that would undermine it (Heemsoth & Heinze, 2014), toward developing more sophisticated strategies. Thus, both errorful learning methods improved near transfer more than errorless practice. However, generating one's own errors in the derr-explain-fix condition may have promoted deeper comparisons between these errors and the correct solutions, relative to receiving others' ready-made errors in the spot-explain-fix condition (Terwel et al., 2009). Consequently, the derr-explain-fix group may have been more likely to attend to

and abstract the critical structural features of the design principles, gaining a richer understanding that enhanced not only their near transfer but also far transfer.

Educational implications

Whereas extant deliberate erring research has only examined the transfer of specific scientific concepts and routinized mathematical procedures (Wong, 2023; Yap & Wong, 2024), the present study reveals that the benefits of this approach extend to the transfer of more general principles such as those in experimental design (Barnett & Ceci, 2002). Moreover, beyond extant studies' predominant focus on CVS mastery when students learn experimentation skills (Schwichow, Croker, et al., 2016), this research adopted a broader approach in also developing and assessing students' learning of other important design principles such as including an appropriate control group, measuring relevant covariates, random assignment, and using a pretest-posttest design. The data show that deliberate erring is an effective technique to improve holistic learning and transfer of a suite of experimental design principles including CVS. Given that real-world phenomena often comprise multiple interacting variables to be coordinated in skilled scientific thinking (Kuhn et al., 2008), it is noteworthy that the derring effect persisted on the far transfer test that involved planning a structurally more complex factorial design that students had not previously encountered. In educational settings, students could thus be instructed on the fundamentals of research design, then practice designing deliberately invalid experiments and explaining and correcting their errors for deep learning.

Although this study found that first-hand deliberate erring yields better transfer than spotting and correcting others' deliberate errors (see also Wong, 2023; Yap & Wong, 2024), it should be noted that the latter method is not harmful for learning. Indeed, the spot-explain-fix method outperformed errorless practice on the near transfer, design critique, and error anticipation tests. Hence, after designing their own deliberately invalid experiments, students could potentially exchange their responses with their peers to spot and discuss each other's deliberate errors. Collaborative discourse and problem-solving can foster meaningful learning by enabling students to co-construct knowledge (Hmelo-Silver & Barrows, 2008; Nussbaum, 2008). For instance, exploring a diverse variety of suboptimal or incorrect representations and solution methods could provide opportunities for students to build on each other's ideas, thus further enriching their understanding of the to-be-learned principles (Kapur & Bielaczyc, 2012).

However, students tend to be unaware of the benefits of errorful learning, even after profiting from it. In this study, for instance, the three learning groups neither differed in their perceived competence and JOL+ ratings after the intervention, nor in how effective they perceived their learning method to be after the tests. Such lack of metacognitive awareness has been consistently observed in studies on deliberate errors (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024) and spontaneous errors or incorrect guesses (e.g., Huelser & Metcalfe, 2012; Yang et al., 2017). Because metacognitive knowledge influences students' choices in deciding what and how to study (Metcalfe & Finn, 2008), it would be vital for teachers to develop and support their students' metacognitive skills to become more sophisticated self-regulated learners (Bjork et al., 2013).

Limitations and future directions

Beyond the present focus on students' experimentation skills in designing valid experiments and understanding the indeterminacy of invalid ones, scientific thinking also includes skills in formulating hypotheses (Klahr & Dunbar, 1988) and evaluating evidence when making inductive causal inferences from data during hypothesis-testing (Zimmerman, 2007; see also Wong et al., 2019). Thus, whereas deliberate erring in this study involved violating experimental design principles to create invalid designs, future work could explore how deliberate errors can be adapted in instructional design to target diverse experimentation skills. For instance, recent evidence suggests that students' argumentative reasoning is enhanced when they teach others about a controversial issue while generating deliberately weak arguments about it for their audience to spot, relative to teaching with good arguments they have generated (Wong, 2025). By extension, to improve students' skills in evaluating scientific claims and evidence, they could potentially be asked to make intentionally flawed arguments or inferences from given data, then correct them while explaining the justificatory standards they have applied. Likewise, given that teaching others a topic enhances students' ability to formulate research questions that create new knowledge about that topic (Lim et al., 2026; Wong et al., 2023; Wong & Lim, 2026), it could be profitable to examine whether teaching with deliberate errors yields additional gains for generating creative research questions and hypotheses.

It would also be worthwhile to test the effectiveness of deliberate erring with students of different age groups and levels of prior knowledge. In this and earlier studies (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024), the derring effect has been established with college-aged students who possess relatively low self-reported prior knowledge about the learning material. Although these university students were able to deliberately commit and correct errors when learning unfamiliar material, it is possible that younger students may require more guidance to effectively use this learning method. Likewise, the benefits of errorful learning may be augmented when students have sufficient prior knowledge to productively self-explain and correct errors (Große & Renkl, 2007; Heemsoth & Heinze, 2014). If so, scaffolding could be vital for students with low prior knowledge (Durkin & Rittle-Johnson, 2012). In this study, for instance, providing the prerequisite knowledge during the experimental design lesson and implementing the intervention as open-book practice could have enabled students' fruitful engagement with deliberate errors without imposing excessive cognitive load, particularly since they reported relatively low prior knowledge about the design principles. Examining whether and how learner characteristics influence the efficacy of deliberate erring would inform how this technique can be optimally implemented or tailored in the classroom. Whereas this study measured students' self-assessed prior knowledge of the material, future work could directly test students' prior knowledge given that they may not always hold accurate perceptions of their learning.

Relatedly, students in this study first received explicit instruction on the to-be-learned principles, then practiced applying them. In contrast, proponents of problem-solving-before-instruction (PS-I) approaches suggest that problem-solving should precede explicit instruction to better prepare students for future learning (e.g., Kapur & Bielaczyc, 2012; Schwartz & Martin, 2004). As an example of a PS-I approach, "productive failure" involves delaying instructional structure by first having learners generate and explore multiple representations and solution methods to a complex problem, then providing teacher-led consoli-

dation to compare learners' (inevitably) suboptimal or failed solutions with the canonical ones (Kapur & Bielaczyc, 2012). Despite improving conceptual knowledge and transfer, PS-I approaches have been found to yield limited benefits for domain-general skills such as applying CVS (Nachtigall et al., 2020; Sinha & Kapur, 2021). Whereas PS-I induces errors in learning, deliberate erring involves guiding errors and the spot-explain-fix method involves observing errors (Wong & Lim, 2019b), with all three errorful approaches similarly involving the comparison of correct versus incorrect solutions. Thus, to understand when each errorful approach should optimally be applied to enhance learning, an interesting future prospect is to compare their nuanced effects across diverse educational outcomes, and to explore potential synergies among them.

Finally, another future direction based on current developments in education can be envisioned. The rapid proliferation of generative artificial intelligence (AI) tools begs answers to such questions as how AI can be used to deliver personalized instructional support at scale and to engage students in deep, rather than shallow, learning processes. In response, future work could further test how, when applying deliberate erring, students can effectively collaborate with AI (e.g., using a "think first, ChatGPT later" approach) for yet greater learning benefits (Wong & Qiu, 2026).

Conclusion

Intuitively, one might assume that considering wrong ways to design experiments could create interference that confuses or misleads students. But contrary to this view, prevailing effective interventions have shown that learning from errors in examples of invalid designs enhances scientific reasoning (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lorch et al., 2010, 2014). This research asked whether students in fact learn better from producing their own intentionally invalid designs than receiving others' invalid designs. The present findings showcase the differential effects of observing versus guiding errors (Wong & Lim, 2019b), and demonstrate the critical importance of deliberately committing and correcting one's own errors. When students practiced intentionally designing and correcting invalid experiments after instruction, they learned and transferred experimental design principles better than errorless practice with valid designs or spotting and correcting the deliberate errors in their peers' invalid designs. Counterintuitively, intentionally designing bad experiments enables the design of good ones.

Appendix A

Research scenario in experimental design lesson

John would like to find out whether taking notes during lectures improves students' test performance. He has also noticed that some students write their notes during lectures, whereas other students type their notes. John wonders how these different note-taking methods may affect students' test performance.

Appendix B

Research scenario in intervention

Mel is interested in testing whether drinking caffeine improves students' memory performance. She has also noticed that some of her classmates drink one cup of coffee each day, whereas others drink two cups. She wonders whether students who consume more caffeine will be able to remember more biology concepts that their lecturer assigns each week.

Appendix C

Instructions for learning methods

Errorless practice

Applying the five principles you have learned, design an *optimal* study to test Mel's ideas. For each principle, propose at least one specific solution in your study design. Explain your design in as much detail as possible, and provide your own specific examples for each solution that you have proposed in the context of Mel's study.

Derr-explain-fix

Applying the five principles you have learned, design a *non-optimal* study to test Mel's ideas. For each principle, propose at least one specific error in your study design. Explain your design in as much detail as possible, and provide your own specific examples for each error that you have proposed in the context of Mel's study. Then, correct each error by proposing a solution for it.

Spot-explain-fix

The following is a study design that has been proposed to test Mel's ideas: [yoked response from the derr-explain-fix condition].

Applying the five principles you have learned, explain why this proposed design is a *non-optimal* study to test Mel's ideas. For each principle, identify at least one specific error in the proposed study design. Explain your answer in as much detail as possible, and provide your own specific solutions for each error that you have identified in the context of Mel's study.

Appendix D

Sample responses for learning methods

Learning Method	Sample Response
Errorless practice	<i>“Solution:</i> All students should take the test at the same venue. <i>Explanation:</i> This prevents venue from being a confound, which would be problematic as we wouldn’t be able to determine whether the students’ better test performance was truly because they had taken notes or because they had completed the test in a quiet and more comfortable air-conditioned lecture hall, rather than a noisy and hot canteen.”
Derr-explain-fix	<i>“Deliberate error:</i> Students who had written or typed their notes stay in the lecture hall to complete the test, whereas the students who had not taken any notes complete the test in the canteen. <i>Explanation:</i> The venue is a confound, which is problematic as we can’t determine whether the students’ better test performance was truly because they had taken notes or because they had completed the test in a quiet and more comfortable air-conditioned lecture hall, rather than a noisy and hot canteen. <i>Solution:</i> All students should take the test at the same venue.”
Spot-explain-fix	<i>“Spotted error:</i> In the proposed design, the test venue was not kept the same for all three groups of students. Students who had written or typed their notes stayed in the lecture hall to complete the test, whereas the students who had not taken any notes completed the test in the canteen. <i>Explanation:</i> The venue is a confound, which is problematic as we can’t determine whether the students’ better test performance was truly because they had taken notes or because they had completed the test in a quiet and more comfortable air-conditioned lecture hall, rather than a noisy and hot canteen. <i>Solution:</i> All students should take the test at the same venue.”

Note. All participants were shown the sample response for their respective learning method so that they understood what was required of them before they began the intervention

Appendix E

Near transfer test

Sam works at the National Gallery. He has noticed that visitors often photograph the artworks on display along with the artworks’ titles and artists’ names, then edit their photos on their smartphone. Whereas some visitors crop their photos, others apply a color-changing filter. Sam is interested in how the act of editing one’s photos affects people’s memory for the content of the original artwork. He wonders whether visitors who edit their photos in a certain way will be able to remember the artwork’s content more accurately.

Applying the five principles you have learned earlier, propose a study design to test Sam’s ideas. For each principle, elaborate and provide your own specific examples in the context of Sam’s ideas. Describe and explain your study design in as much specific detail as possible.

Appendix F

Far transfer test

At the Ministry of Education’s teacher conference, the science teachers discuss how to improve teaching and learning:

Ms. A: “I think we should include more group work in our classes. If students work in groups on their tasks, they might be more motivated than when they work alone.”

Ms. B: “Maybe students will be even more motivated if they have computer support on their tasks? I think using technology to improve our students’ learning is important. But I don’t know if working in groups with a computer is better than working alone.”

Mr. C: “But what about our students who have just joined our school from overseas? These students may not have English as their first language, and they may be disadvantaged from group and/or computer work if they have language difficulties.”

Minister X thinks that all three teachers’ suggestions are important, but isn’t sure whether any of their suggestions is better than the current teaching method of having students solve science tasks by working alone.

Applying the five principles you have learned earlier, propose a study design to test all three teachers’ suggestions. For each principle, elaborate and provide your own specific examples in the context of the teachers’ suggestions. Describe and explain your study design in as much specific detail as possible.

Appendix G

Design critique test

Dr. Lee wanted to find out whether people believe that they are at lower risk of catching the flu from others in the dark. To test this idea, Dr. Lee had participants watch a documentary film either with the lights on or off while seated next to a trained confederate, who intentionally coughed and sneezed precisely the same number of times and in equal loudness during the session. At the end of the session, participants rated their likelihood of contracting the flu. Dr. Lee collected data over two semesters in a year with participants at her university. During the spring semester, she tested 50 participants who watched the film in a dark room with the lights off. She also tested 50 participants during the winter semester who watched the film with the lights on.

Based on the five principles you have learned earlier, analyze and evaluate the extent that Dr. Lee’s study design allows her to soundly test her ideas. Explain your answer in as much specific detail as possible.

Appendix H

Error anticipation test

Dr. Kim is teaching a Korean language module with three tutorial classes this semester. She is interested in finding out whether students will be able to learn Korean vocabulary better if they spread out their revision over two or three sessions, instead of cramming their revision in a single session before their final exam.

What design errors do you think Dr. Kim may make when designing her study? Describe as many potential errors as you can, and in as much specific detail as possible within the context of Dr. Kim's study.

Appendix I

Sample responses on near transfer, far transfer, and design critique tests

Test	Sample Responses	Scoring Remarks
Near transfer	"All groups will include visitors who use a smartphone to capture a picture of the same artwork. The study will have a control group of visitors who do not edit the photo at all. The first experimental group will have visitors crop the photo only. The second experimental group will have visitors apply a color-changing filter only. There will also be a questionnaire prior to the experiment to gauge the visitors' levels of interest in art, to account for covariates. In terms of systematically varying variables, to avoid differences in time of day that the visitors take the picture affecting the outcome (e.g. during peak hours, the gallery is noisier and this may reduce one's ability to recall the artwork), all groups will take a picture at the same time of day. There will be random assignment of visitors to each group, to prevent people who already have prior knowledge of the artwork or are familiar with the artwork before the experiment being in the same group and to equalize all experimental groups. Finally there will be a pretest to test each group's level of prior knowledge about the artwork and to measure any changes in memory of the artwork before and after the experiment."	This response scored 13 points (out of a maximum of 14 points for the near transfer test). It earned 4 points for the implicit design principle (proposing a design with three conditions and correctly naming each condition), and 9 points for the explicit design principles (applying and elaborating on the principles of "control group", "measuring relevant covariates", "systematically varying key variables", and "pretest-posttest design", and applying but not elaborating on the "random assignment" principle).
Far transfer	"The teacher should choose 200 students, and assign a number to each of the students and use a random number generator to assign each student into one of the 4 groups: Working on a science task in groups, working on a science task alone on a computer, working on a science task in groups on a computer, and working on a science task alone (control group). Before the entire experiment, each student should go through 2 tests, one on their English proficiency and one on their Science proficiency. A questionnaire on the students' interest in Science should also be conducted as students with a lack of interest may be less motivated to do work and perform poorer on the Science task. Afterwards, the students will be assigned to work on their tasks based on their groupings, and then they will do a posttest Science proficiency test. The result of the posttest Science proficiency test will be compared to the pretest Science proficiency test to see whether the respective groupings helped in the learning of Science."	This response scored 12 points (out of a maximum of 15 points for the far transfer test). It earned 5 points for the implicit design principle (proposing a design with four conditions and correctly naming each condition), and 7 points for the explicit design principles (applying and elaborating on the principles of "measuring relevant covariates", "random assignment", and "pretest-posttest design", and applying but not elaborating on the "control group" principle).

Test	Sample Responses	Scoring Remarks
Design critique	“Covariates such as health status should be controlled for by giving all participants a questionnaire about their health status before the study, because the presence of any pre-existing health condition can affect one’s likelihood of contracting the flu. The confound, which is the semester when the data were collected, is a variable that is not of interest but was not kept the same across the groups. The time at which the study was conducted (whether it was during spring or winter) could have influenced the likelihood of participants contracting the flu. She should have randomly assigned the participants to both groups by using a random number generator to ensure all groups are equivalent. Dr Lee should have a pretest to allow participants to rate their likelihood of contracting the flu to check that participants have been properly randomized and to measure how their likelihood of contracting the flu has changed as a result of the treatment (lights on or off).”	This response scored 7 points (out of a maximum of 10 points for the design critique test) for applying and elaborating on the principles of “measuring relevant covariates”, “random assignment”, and “pretest-posttest design”, and for applying (but not elaborating on) the principle of “systematically varying key variables”.

Appendix J

Sample problem-specific versus problem-general errors in error anticipation test

Problem-Specific Errors	Problem-General Errors
“Dr. Kim may conveniently assign each of her three tutorial classes to one experimental group, instead of randomly assigning her students.”	“Dr. Kim may fail to randomly assign the different groups.”
“Different total durations of revision could introduce confounds, whereby the groups that spread out their revision get more total time to revise compared to the group that crammed their revision in a single session. For example, a single session could be 6 hr, but the multiple sessions of three could be 3 hr each, totaling to 9 hr.”	“She may not keep all variables constant besides number of revision sessions.”
“Dr. Kim may not have taken into account some students have learned another language that has similar vocabulary as Korean. Some Chinese terms and Japanese terms have similar pronunciation for some words, which may make it easier for those students to learn.”	“Dr. Kim may not control for participants’ IQ and motivation.”

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11251-026-09819-x>.

Acknowledgements The author thanks Jenell Ong, Sophia Xuefei Qiu, and Zheng’En Liao for their assistance with data collection and scoring.

Author contributions 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.

Funding This research was supported by a Singapore Management University grant, a Yale-NUS College Educational 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.

Declarations

Competing interests The author has no competing interests to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Barbieri, C. A., Booth, J. L., Begolli, K. N., & McCann, N. (2021). The effect of worked examples on student learning and error anticipation in algebra. *Instructional Science*, 49(4), 419–439. <https://doi.org/10.1007/s11251-021-09545-6>
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>
- Booth, J. L., McGinn, K. M., Young, L. K., & Barbieri, C. (2015). Simple practice doesn't always make perfect: Evidence from the worked example effect. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 24–32. <https://doi.org/10.1177/2372732215601691>
- Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp. 173–197). Psychology Press. <https://doi.org/10.4324/9780203888544-13>
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098–1120. <https://doi.org/10.1111/1467-8624.00081>
- Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist*, 47(3), 177–188. <https://doi.org/10.1080/00461520.2012.695709>
- Corral, D., & Carpenter, S. K. (2020). Facilitating transfer through incorrect examples and explanatory feedback. *Quarterly Journal of Experimental Psychology*, 73(9), 1340–1359. <https://doi.org/10.1177/1747021820909454>
- Department for Education (2015, May 6). *National curriculum in England: Science programmes of study*. <https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study>
- 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>
- 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>
- 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>
- Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory & Cognition*, 29(4), 565–577. <https://doi.org/10.3758/BF03200458>
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1), 45–56. <https://doi.org/10.1037/0003-066X.52.1.45>
- Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25(4), 524–575. <https://doi.org/10.1006/cogp.1993.1013>
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15(1), 1–38. [https://doi.org/10.1016/0010-0285\(83\)90002-6](https://doi.org/10.1016/0010-0285(83)90002-6)

- Goodwin, C. J., & Goodwin, K. A. (2014). *Research in psychology: Methods and design* (7th ed.). John Wiley.
- 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>
- 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>
- Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, 26(1), 48–94. <https://doi.org/10.1080/07370000701798495>
- Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, 40(4), 514–527. <https://doi.org/10.3758/s13421-011-0167-z>
- Joung, W., Hesketh, B., & Neal, A. (2006). Using “war stories” to train for adaptive performance: Is it better to learn from error or success? *Applied Psychology: An International Review*, 55(2), 282–302. <https://doi.org/10.1111/j.1464-0597.2006.00244.x>
- Kapur, M. (2014). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences*, 23(4), 651–677. <https://doi.org/10.1080/10508406.2013.819000>
- Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, 21(1), 45–83. <https://doi.org/10.1080/10508406.2011.591717>
- Kimball, D. R., & Holyoak, K. J. (2000). Transfer and expertise. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 109–122). Oxford University Press.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1
- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661–667. <https://doi.org/10.1111/j.0956-7976.2004.00737.x>
- Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, 25(1), 111–146. <https://doi.org/10.1006/cogp.1993.1003>
- 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>
- Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, 23(4), 435–451. <https://doi.org/10.1016/j.cogdev.2008.09.006>
- Lazonder, A. W., & Egberink, A. (2014). Children’s acquisition and use of the control-of-variables strategy: Effects of explicit and implicit instructional guidance. *Instructional Science*, 42(2), 291–304. <https://doi.org/10.1007/s11251-013-9284-3>
- Lim, S. W. H., Wong, S. S. H., & Visessuvanapoom, P. (2026). Durable benefits of learning-by-teaching for research question generation performance: A field experiment. *Journal of Experimental Education*, 94(3), 399–419. <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](https://doi.org/10.1016/S0959-4752(00)00037-2)
- Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, 36(7), 837–858. [https://doi.org/10.1002/\(SICI\)1098-2736\(199909\)36:7%3C837::AID-TEA6%3E3.0.CO;2-U](https://doi.org/10.1002/(SICI)1098-2736(199909)36:7%3C837::AID-TEA6%3E3.0.CO;2-U)
- Loibl, K., & Rummel, N. (2015). Productive failure as strategy against the double curse of incompetence. *Learning: Research and Practice*, 1(2), 113–121. <https://doi.org/10.1080/23735082.2015.1071231>
- 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>
- Lorch, R. F., Jr., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, 102(1), 90–101. <https://doi.org/10.1037/a0017972>
- Lorch, R. F., Jr., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology*, 106(1), 18–35. <https://doi.org/10.1037/a0034375>
- Lorch, R. F., Jr., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H.-T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, 51, 391–403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>

- Martella, A. M., Klahr, D., & Li, W. (2020). The relative effectiveness of different active learning implementations in teaching elementary school students how to design simple experiments. *Journal of Educational Psychology*, 112(8), 1582–1596. <https://doi.org/10.1037/edu0000449>
- 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>
- Metcalfe, J., & Xu, J. (2018). Learning from one's own errors and those of others. *Psychonomic Bulletin & Review*, 25(1), 402–408. <https://doi.org/10.3758/s13423-017-1287-7>
- Ministry of Education, Singapore (2022). *Science teaching & learning syllabus: Primary three to six standard / foundation*. <https://www.moe.gov.sg/api/media/ba3562d3-5b31-4459-8693-45cde7b97273/Primary-Science-Syllabus-2023.pdf>
- Nachtigall, V., Serova, K., & Rummel, N. (2020). When failure fails to be productive: Probing the effectiveness of productive failure for learning beyond STEM domains. *Instructional Science*, 48(6), 651–697. <https://doi.org/10.1007/s11251-020-09525-2>
- National Research Council. (2013). *Next generation science standards: For states, by states*. The National Academies Press. <https://doi.org/10.17226/18290>
- Nussbaum, E. M. (2008). 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>
- Peteranderl, S., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2023). What skills related to the control-of-variables strategy need to be taught, and who gains most? Differential effects of a training intervention. *Journal of Educational Psychology*, 115(6), 813–835. <https://doi.org/10.1037/edu0000799>
- Randler, C., Hummel, E., Gläser-Zikuda, M., Vollmer, C., Bogner, F. X., & Mayring, P. (2011). Reliability and validation of a short scale to measure situational emotions in science education. *International Journal of Environmental and Science Education*, 6(4), 359–370.
- Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15(3), 243–257. <https://doi.org/10.1037/a0016496>
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, 24(2), 113–142. https://doi.org/10.1207/s15326985ep2402_1
- Sandoval, W. A., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. *Educational Psychologist*, 49(2), 139–152. <https://doi.org/10.1080/00461520.2014.917589>
- Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction*, 59, 34–45. <https://doi.org/10.1016/j.learninstruc.2018.09.006>
- Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129–184. https://doi.org/10.1207/s1532690xc2202_1
- Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education*, 38(2), 216–237. <https://doi.org/10.1080/09500693.2015.1137651>
- Schwichow, M., Croker, S., Zimmerman, C., Höfler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37–63. <https://doi.org/10.1016/j.dr.2015.12.001>
- Schwichow, M., Osterhaus, C., & Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Contemporary Educational Psychology*, 63, Article 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- 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>
- Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, 91(5), 761–798. <https://doi.org/10.3102/00346543211019105>
- Tawfik, A. A., Graesser, A., Gatewood, J., & Gishbaugher, J. (2020). Role of questions in inquiry-based instruction: Towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, 68(2), 653–678. <https://doi.org/10.1007/s11423-020-09738-9>
- 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>
- Terwel, J., van Oers, B., van Dijk, I., & van den Eeden, P. (2009). Are representations to be provided or generated in primary mathematics education? Effects on transfer. *Educational Research and Evaluation*, 15(1), 25–44. <https://doi.org/10.1080/13803610802481265>

- 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
- van Riesen, S., Gijlers, H., Anjewierden, A., & de Jong, T. (2018). Supporting learners' experiment design. *Educational Technology Research and Development*, 66(2), 475–491. <https://doi.org/10.1007/s11423-017-9568-4>
- 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. (2025). Learning by teaching with deliberate errors promotes argumentative reasoning. *Journal of Educational Psychology*, 117(7), 1013–1038. <https://doi.org/10.1037/edu0000934>
- 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. (2026). The powerful teacher: A power hypothesis for the benefits of learning-by-teaching. *Educational Psychology Review*, 38(1), Article 26. <https://doi.org/10.1007/s10648-025-10097-1>
- 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>
- Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2019). Retrieval practice enhances analogical problem solving. *Journal of Experimental Education*, 87(1), 128–138. <https://doi.org/10.1080/00220973.2017.1409185>
- Wong, S. S. H., & Qiu, S. X. (2026). Think first, ChatGPT later: Guiding human–AI collaboration for learning gains in independent human creativity. *Educational Psychology Review*, 38(1), Article 45. <https://doi.org/10.1007/s10648-026-10118-7>
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. <https://doi.org/10.1093/0195155270.001.0001>
- 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>
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172–223. <https://doi.org/10.1016/j.dr.2006.12.001>