Contemporary Educational Psychology 77 (2024) 102277

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

"My drawing is quite different!" Drawbacks of comparing generative drawings to instructional visuals

Logan Fiorella [a,*], Allison J. Jaeger [b], Alexis Capobianco [a], Anna Burnett [a]

[a] University of Georgia, United States
[b] Mississippi State University, United States

ARTICLE INFO

Keywords:
Generative learning
Drawing
Comparison
Metacognition
Comprehension

ABSTRACT

This study tested how prompting learners to compare their drawings to instructional visuals affects their perceived and actual performance. Undergraduates (n = 116) created two drawings while studying a text on the human circulatory system. Then they made a series of retrospective and prospective judgments of their drawing performance and prospective judgments of their comprehension. In a subsequent restudy phase, students were randomly assigned to either compare their drawings to instructional visuals (compare group; n = 56) or to restudy the text and review their drawings without receiving instructional visuals (control group; n = 60), followed by a series of new judgments of drawing and comprehension. All students then completed drawing and comprehension post-tests. Results indicated that comparing one's drawings to instructional visuals caused students to become underconfident in the quality of their drawings (lower retrospective accuracy) and overconfident in their future drawing performance (lower prospective accuracy). Exploratory analyses indicated that the compare group tended to make surface-level (rather than conceptual) comparisons when processing the provided visuals, such as attending to the aesthetic style or conventions used in the instructional visuals. Furthermore, despite a strong link between drawing and comprehension performance, comparing drawings to instructional visuals did not significantly affect students' judgments of comprehension. These findings highlight potential drawbacks of comparing generative drawings to instructional visuals in learning by drawing.

1. Introduction

Drawing can be an effective strategy for learning from science texts, particularly among undergraduates (Ainsworth & Scheiter, 2021; Brod, 2021; Fiorella & Zhang, 2018). Indeed, several studies have found that prompting students to generate drawings results in higher comprehension than only restudying or summarizing the text (e.g., Bobek & Tversky, 2016; Leopold & Leutner, 2012). There is also emerging evidence that drawing may provide *metacognitive* benefits (e.g., Fiorella & Jaeger, 2023; Thiede et al., 2022; van de Pol et al., 2020)—that is, the experience of drawing may help students become aware of what they do or do not understand. However, drawing is not always effective (e.g., Leopold et al., 2009; see Fiorella and Zhang, 2018); students often struggle to generate high-quality drawings and might not accurately assess the quality of their understanding without instructional support. One common form of drawing support is to prompt learners to compare their drawings to 'expert' instructor-provided visuals (e.g., Van Meter, 2001; Schmeck et al., 2014), such as the types of instructional visuals that students view in textbooks. On one hand, comparing generated and provided instructional visuals may help students correct errors and form more accurate judgments of their own drawing and comprehension performance—what we refer to as the *comparison facilitation hypothesis*. On the other hand, instructional visuals may interfere with the accuracy of students' metacognitive judgments, particularly if students focus on irrelevant or surface-level comparisons that are not diagnostic of their actual learning—what we refer to as the *comparison interference hypothesis*. In the present study, we tested the potential facilitative or detrimental effects of comparing learner-generated and provided instructional visuals on students' perceived and actual performance.

1.1. *Using Instructor-Provided visuals to support Learner-Generated drawings*

According to generative learning theory (Fiorella, 2023; Fiorella & Mayer, 2016; Wittrock, 1989), drawing supports comprehension by encouraging learners to *select* relevant information from a lesson,

* Corresponding author at: Department of Educational Psychology, University of Georgia, United States.
E-mail address: lfiorella@uga.edu (L. Fiorella).

https://doi.org/10.1016/j.cedpsych.2024.102277

Available online 1 May 2024

organize it into a coherent structure, and *integrate* it with their existing knowledge (see also Van Meter & Firetto, 2013). For example, when creating a drawing from a text on the human circulatory system, learners attempt to visually depict and label the key spatial relationships described in the text—such as relationships among the four chambers of the heart, the lungs, and the body—which in turn may help them generate inferences about how the system works—such as the mechanisms underlying systemic and pulmonary circulation. However, the effectiveness of learning by drawing depends on managing several potential boundary conditions (Brod, 2021; Fiorella & Zhang, 2018; Leutner & Schmeck, 2022). Most notably, drawing can be cognitively demanding and time consuming (e.g., Leutner et al., 2009; Van Meter, 2001; Zhang & Fiorella, 2021), and without appropriate background knowledge or instructional support, students might struggle to (a) generate high-quality drawings and/or (b) accurately self-evaluate the quality of their drawings and comprehension. This is important because the quality of students' drawings during learning is generally predictive of their performance on subsequent drawing and comprehension tests (e.g., Schwamborn et al., 2010).

One common way to support learning by drawing is to provide learners an opportunity to actively *compare* their own drawings to a provided instructional visualization, which should help them detect and correct potential knowledge gaps or misconceptions (Leutner & Schmeck, 2022; Nicol, 2021; Zhang & Fiorella, 2023). For example, upon studying a provided visual of the circulatory system, a student might notice that they incorrectly represented the functions of the left and right sides of the heart in their own drawing. Research suggests that, in some cases, comparing one's drawings to provided instructional visuals can support comprehension better than either drawing without comparing or only studying provided visuals. For example, early work by Van Meter (2001) found that students prompted to compare their drawings of the nervous system to expert instructor-provided visuals exhibited better subsequent comprehension than students who only created drawings or students who only studied provided visuals. Furthermore, think aloud data indicated that students who compared their drawings to the instructional visuals engaged in more self-monitoring events, such as looking back to a previously read part of the text or generating questions to oneself.

However, other work has produced null or even negative effects of combining instructor-provided and learner-generated visuals. In a study by Schwamborn and colleagues (2011), high school students learned about the chemical processes of washing with soap and water from either text only, provided instructional visuals, generated visuals, or generated and provided visuals. Results indicated that instructional visuals supported comprehension and drawing test performance, whereas generated visuals only supported drawing test performance. There was no evidence for an added benefit of pairing generated visuals with instructional visuals, particularly for comprehension. Following a similar design, Schmeck and colleagues (2014) found that generating one's own visuals enhanced comprehension test performance compared to students who studied instructional visuals or who generated their own visuals *and* received instructional visuals. Thus, in this case, instructional visuals appeared to *interfere* with comprehension performance compared to only generating one's own visuals.

Recent work by Zhang and Fiorella (2019, 2021) also highlights the mixed effects of combining generated and instructional visuals. For example, one study found that studying an instructional visual of the circulatory system after creating one's own drawing resulted in better transfer compared to only studying instructional visuals (Zhang & Fiorella, 2019). However, comparing to instructional visuals was not significantly more effective for transfer than only generating visuals or studying an instructional visual *before* drawing. In slight contrast, a related subsequent study suggested that comparing to instructional visuals enhanced the effectiveness of drawing versus either drawing without comparing or drawing with only verbal forms of support (Zhang & Fiorella, 2021).

Taken together, the existing evidence suggests comparing one's drawings to instructional visuals *can* be an effective way to support learning by drawing; however, in some cases, instructional visuals may not boost and may even interfere with learning. One issue that has not been examined in this prior research is the potential *metacognitive* effect of comparing one's drawings to instructional visuals. The metacognition literature suggests the presence of instructional visuals can sometimes interfere with students' ability to make accurate judgments of their own learning (e.g., Jaeger & Fiorella, 2023; Wiley, 2019), such as when predicting how well one will perform on an upcoming test. Provided instructional visuals might cause students to focus on superficial metacognitive cues that are not diagnostic of their actual understanding, such as the feeling of fluency or familiarity associated with viewing a professional illustration. Indeed, studies suggest the presence of decorative images can cause students to overestimate their learning (e.g., Cardwell et al., 2017; Ikeda et al., 2013), and other work suggests even instructionally-relevant visuals can interfere with the accuracy of students' judgments (e.g., Jaeger & Fiorella, 2023; Serra & Dunloksy, 2010). Students make more accurate judgments when they focus on conceptual cues present in the provided visuals, such as by using the visual to self-explain and reflect on one's understanding (e.g., Jaeger & Fiorella, 2023; Jaeger & Wiley, 2014). However, comprehending instructional visuals, particularly in science, can be cognitively demanding for students (Cromley et al., 2010), as it requires mentally representing complex spatiotemporal relationships depicted via discipline-specific conventions (Gilbert, 2008; Hegarty et al., 1991; Kozma, 2003; Novick, 2006).

In the context of learning by drawing, the effectiveness of comparing drawings to instructional visuals depends on the extent to which students productively compare the conceptual similarities and differences between instructional visuals and their own drawings and revise their knowledge accordingly. This is challenging because students' drawings will likely contain conceptual errors, as well as other differences from conventional instructional visuals that are unrelated to the conceptual fidelity of their drawing. For example, instructional visuals of the circulatory system commonly follow conventions such as 'flipping' the orientation of left and right sides of the heart, representing the heart's structures is three dimensions, or using red and blue to represent oxygenated or deoxygenated blood, respectively. Other conventions are more idiosyncratic across visuals, such as the precise relative size and location of the heart, lungs, body, and blood vessels. Importantly, these features are orthogonal to the *conceptual fidelity* of one's drawing (e.g., Ainsworth, 2006; Butcher, 2006)—whether one's drawing accurately depicts the structures and relationships among the circulatory system (e.g., showing that blood flows from the right ventricle to the lungs). Given the ubiquity of instructional visuals in science education (e.g., in textbooks, lectures slides, instructional videos, simulations, etc.; Ainsworth, 2006), it is important to understand how these representations interact with learner-generated visualizations, particularly how they affect students' metacognitive and learning outcomes.

If students focus primarily on differences between their drawings and the provided visuals that are superficial and irrelevant to their understanding, they are unlikely to accurately assess and correct their understanding (e.g., Jaeger, Marzano, & Shipley, 2020). This is consistent with the broader literature on feedback and learning from errors (Metcalfe, 2017; Nicol, 2021; Shute, 2008). For example, Zhang and Fiorella (2023) distinguish between *surface* errors and *deep* errors in their recent model of learning from errors. According to their model, detecting and correcting deep errors requires not only noticing that one's response is conceptually different from provided feedback but also actively self-explaining *why* one's response was wrong, which many learners may not do spontaneously. The metacognition and instructional visuals literatures suggests learning from complex visuals is cognitively demanding (Ainsworth, 2006; Cromley et al., 2010) and students might be particularly inclined to focus on surface-level differences between their drawings and provided visuals (i.e., surface errors), such as salient perceptual or aesthetic differences (e.g., differences in style or conventions), which may lead to less accurate judgments of learning, and potentially worse learning outcomes (e.g., see Bjork, Dunlosky, & Kornell, 2013). Similarly, students might fixate on the perceived fluency associated with aesthetic features of a provided visual, which may cause students to be overconfident about their learning. In the present study, we systematically address this issue by examining how comparing generative drawings to instructional visuals affects students' metacognitive judgments and performance during learning by drawing. The primary goal was to understand whether students use the instructional visuals to accurately diagnose the conceptual quality of their drawings and their understanding, or whether they fixate on superficial differences (such as style and conventions) between their drawings and the instructional visuals.

2. The present study

The present study tested how comparing one's drawings to instructional visuals affects students' perceived and actual drawing and comprehension performance. In an initial learning phase, undergraduates created two drawings while learning from a text about the human circulatory system. Then they made a series of initial judgments of learning: retrospective and prospective judgments of their drawing performance, as well as prospective judgments of their comprehension.[1] In a subsequent restudy phase, students were randomly assigned to restudy the text and compare their drawings to two instructional visuals (compare group), or to only restudy the text and review their own drawings (control group). The instructional visuals were designed to match the content of the text and be representative of the types of visuals commonly found in textbook chapters about the human circulatory system. After the restudy phase, all students made a new series of judgments of drawing and comprehension performance. In the test phase, students completed post-tests assessing their drawing and comprehension performance.

Our primary focus was to examine whether comparing one's drawings to representative instructional visuals affects the *accuracy* of students' judgments of drawing and comprehension. In line with prior research, we determined judgment accuracy by calculating the difference between students' perceived drawing or comprehension performance and their actual drawing or comprehension performance (e.g., Griffin, Mielicki, & Wiley, 2019). Specifically, for each type of judgment, we computed *absolute accuracy* (the absolute difference between perceived and actual performance) and *confidence bias* (the signed difference between perceived and actual performance). Thus, absolute accuracy reflects the degree of error in one's judgments, whereas confidence bias reflects the degree of overconfidence or underconfidence in one's judgments.

2.1. Hypotheses

We tested two competing hypotheses about the effects of comparing drawings to provided visuals. According to the *comparison facilitation hypothesis*, instructional visuals should help learners better detect and correct conceptual errors in their drawings—such as incorrectly depicting connections among the heart and lungs—thereby supporting metacognitive and learning outcomes. That is, students who compare to instructional visuals should exhibit more accurate retrospective and prospective judgments of their drawing performance, and more accurate prospective judgments of their comprehension. Comparing to instructional visuals should also support better performance on the final drawing and comprehension post-tests.

Alternatively, according to the *comparison interference hypothesis*, instructional visuals may cause learners to focus on superficial differences in their drawings—such as differences in the aesthetics, style, or conventions of the instructional visuals—thereby inhibiting metacognitive and learning outcomes. That is, students who compare their drawings to instructional visuals should exhibit less accurate retrospective and prospective judgments of their drawing performance, and potentially, less accurate prospective judgments of comprehension. Consequently, comparison to instructional visuals may also interfere with performance on the final drawing and comprehension post-tests.

In line with prior research, we expected drawing performance during learning to be positively associated with subsequent comprehension post-test performance (e.g., Schwamborn et al., 2010). However, an open question is whether comparing drawings to instructional visuals has similar or unique effects on students' drawing and comprehension judgments. One possibility is that comparing to instructional visuals primarily affects students' retrospective and prospective drawing judgments without affecting their prospective comprehension judgments. Another possibility is that comparing one's drawings to provided visuals might inform students' judgments of drawing *and* judgments of comprehension. Addressing this issue will provide insight into whether students use the perceived quality of their drawings as an indicator of their level of comprehension.

3. Method

3.1. Participants and design

Participants were 120 undergraduates recruited from introductory education or biology courses at a large southeastern university who received course credit or a $20 gift card for completing the study. The mean age of participants was 19.5 years ($SD = 1.6$), and there were 91 females, 25 males, and 4 reported other. The participants represented a wide range of (intended) majors, but most were interested in fields related to either education (30 %), STEM (22 %), or communications sciences (17 %). None of the participants had completed a college-level introductory biology course. Four participants were removed from the analyses because of an experimenter error during data collection, resulting in a final sample of 116. A post-hoc power analysis using G*Power indicates that the study is highly powered (power = .89) for detecting small effect sizes ($f = .15$) at alpha = .05.

Participants were randomly assigned to the compare group ($n = 56$) or the control group ($n = 60$).[2] The groups did not significantly differ in mean age, $t(114) = 1.24$, $p = .218$, or gender distribution, $\chi^2(2) = 2.37$, $p = .307$. This study was approved by and conducted in accordance with the ethical standards of the Institutional Review Board (IRB).

3.2. Transparency and open science

The materials, measures, and data for this study are publicly available via the Open Science Framework: https://osf.io/w3sqc/. The hypotheses for this study were not preregistered, though the study tests

[1] Most prior metacognition research includes only prospective judgments, such as predicting how well one will perform on a later test (e.g., Griffin et al., 2019). We included retrospective judgments to additionally examine how comparing one's drawings to provided visuals affects the accuracy of students' self-assessments of drawing quality.

[2] This study also included an additional between-subjects manipulation related to the instructions students received about the drawing activity at the beginning of the experiment. Some students were prompted to use drawing as a learning activity (i.e., the goal of drawing is to support comprehension), whereas other students were prompted to use drawing as a metacognitive activity (i.e., the goal of drawing is to make better judgments of one's learning). Analyses indicated that this subtle manipulation did not have statistically significant effects on any of the judgment accuracy or performance measures. Thus, to improve clarity in presenting the key findings from this study, we do not include this factor in our analyses reported in the Results section.

two competing hypotheses derived from prior research, as described above.

3.3. Materials and measures

3.3.1. Text and drawing prompts

The learning materials consisted of a text about the human circulatory system and two drawing prompts. The text was adapted from Zhang & Fiorella (2021) and consisted of 726 words broken into two main parts. Part 1 described the path of blood flow through the circulatory.

system, and Part 2 described the path of blood flow through the valves of the heart. The full text is presented in the Appendix A. For each part, students were prompted to create a corresponding drawing on paper: Drawing 1:"Use the paper in front of you to draw the path of blood flow throughout the circulatory system. Label each of the key parts," Drawing 2: "Use the paper in front of you to draw the path of blood flow through the valves of the heart. Label each of the key parts.".

3.3.2. Instructional visuals

During the restudy phase of the experiment, participants assigned to the compare group received the text again along with two instructional visuals corresponding to each part of the lesson (Figs. 1 and 2, respectively). The visuals were created specifically for this project by a professional medical illustrator. The illustrator designed the visuals to complement the content presented in the text and to follow common instructional conventions for depicting the circulatory system: e.g., the use of arrows to depict the path of blood flow, the use of red and blue to represent oxygenated and deoxygenated blood, respectively, the 'flipped' perspective of the left and right sides of the heart, and the relative size of various structures, such as the location and size of the lungs. The visuals were also designed to minimize extraneous features irrelevant to the function of the circulatory system, such as excessive detail and realism (e.g., Ainsworth, 2006; Butcher, 2006). Thus, the instructional visuals are representative of the types of visuals that students would study when learning about the circulatory system, such as in textbooks, lecture slides, or instructional videos. It is important to note that the visuals were *not* designed to necessarily match the specific style and conventions students might use to create drawings from the text. Based on prior work, we anticipated the appearance of students' drawings would vary considerably across individuals (e.g., Zhang & Fiorella, 2021), as well as in comparison to the instructional visuals. Our primary research question was whether students would effectively evaluate the conceptual quality of their drawings compared to the instructional visuals, despite other inevitable stylistic or perceptual differences.

3.4. Comparison to instructional visuals

Participants in the compare group were prompted to describe how their drawings compared to the instructional visuals by typing into a text box. Participants assigned to the control group restudied the text and reviewed their drawings without receiving the instructional visuals.

As an exploratory measure, we used participants' responses to the comparison prompts to determine the types of comparisons students made between their drawings and the instructional visuals. The explicit goal of the study and the drawing task was for students to understand how the circulatory system works. Thus, we first distinguished among comparisons that reflected information that was relevant or irrelevant to one's conceptual understanding of the structures and functions of the system. Relevant or *conceptual* comparisons included mentioning structures or processes in one's drawing that were correctly or incorrectly depicted or labeled, such as "I did not draw the lungs and the various connections between the lung and heart such as the vena cavae and pulmonary vein," "My semilunar valves are in the wrong place," or "I did not draw the direction of the valves correctly." Irrelevant or *surface-level* comparisons included making vague statements that did not specify a difference (e.g., "My drawing is quite different"), or mentioning superficial differences in the aesthetic appearance, style, or conventions of the instructional visual. For example, students might mention that their drawing do not follow the same conventions for representing the left and right side of the heart ("I also mixed the left and

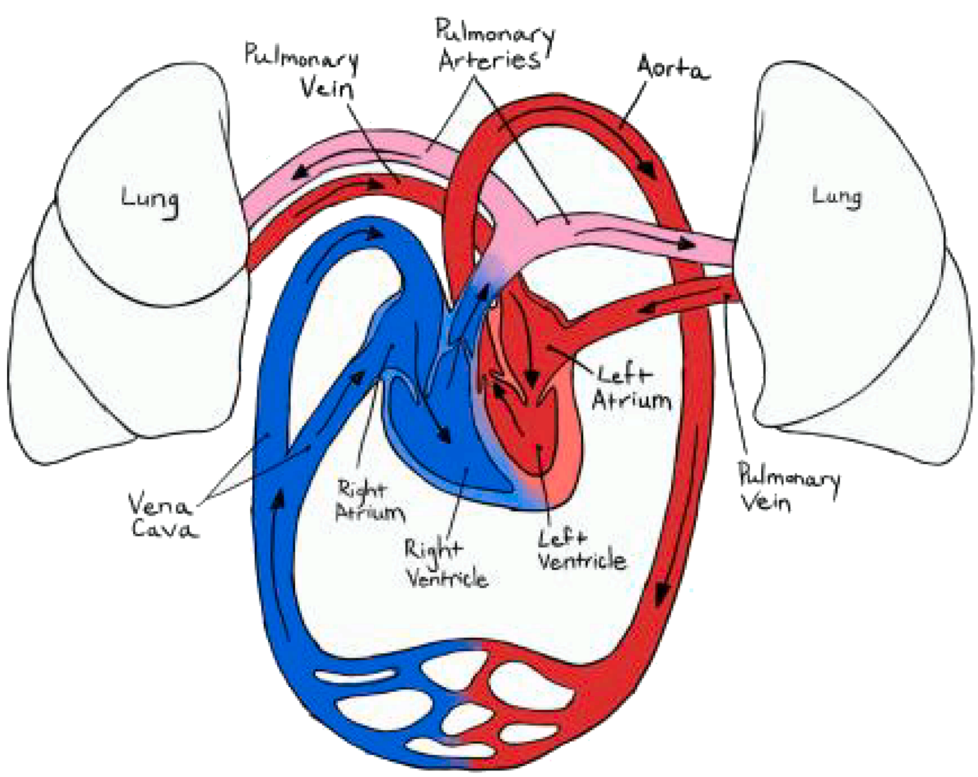

Fig. 1. Instructor-Provided Visualization of Circulatory System for the Compare Group.

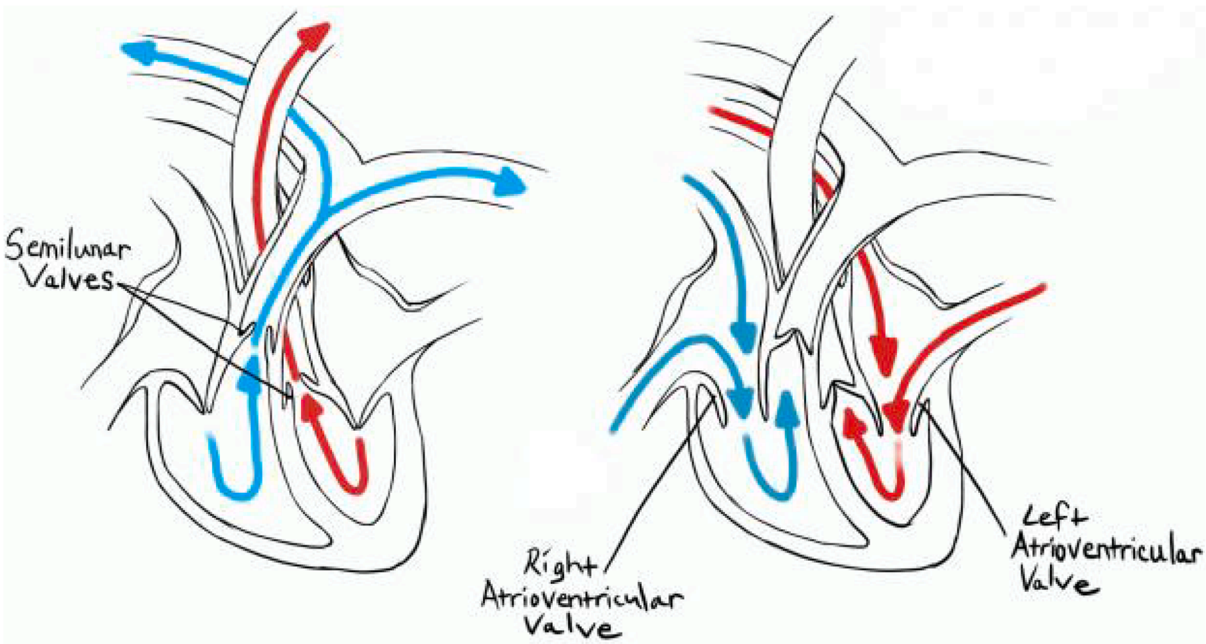

Fig. 2. Instructor-Provided Visualization of Heart's Valves for the Compare Group.

right side up and switched them in my drawing"), that the relative size or location of certain structures is different in their drawings ("I drew the heart way too big"), that their drawings differ in their level of detail ("My drawing is a lot simpler") or that their drawings differ in their aesthetic quality ("The sample drawing is a lot clearer than mine and much easier to follow"). These differences in style or conventions are not necessary for understanding the structure and function of the circulatory system. That is, a student could 'flip' the right and left sides of the heart in their drawing but still correctly depict their respective role in pulmonary and systemic circulation. Similarly, students might depict the lungs or the body in various shapes, sizes, or locations relative to the heart, but these features are orthogonal to whether students accurately depicted the connections among the heart, body, and lungs. Importantly, our method of scoring the quality of participants' drawings (described below) focused only on whether students accurately depicted and labeled the correct structures and connections of the circulatory system described in the provided text—not whether they followed the same stylistic conventions as the instructional visuals.

Two research assistants coded a subset ($n = 33$) of participants' responses to the comparison prompts for conceptual and surface-level comparisons. Inter-rater reliability between was high (ICCs $> .84$), and so one rater scored all remaining responses. To provide further insight into the comparison process, we also calculated the frequency of specific types of surface-level differences that students might have mentioned: (a) the use of color, (b) the left/right orientation of the heart, (c) the dimensionality of the visuals (2D vs. 3D), (d) the relative size or location of structures depicted in the visuals (e.g., the size or location of the lungs), or (e) the level of detail presented in the visuals, and (f) the number of visuals generated vs. provided.

3.4.1. *Judgments of drawing*

Participants made two types of judgments of drawing: retrospective judgments and prospective judgments. For *retrospective* judgments, participants responded to one prompt for each of their drawings: e.g., "For Drawing 1, you created a drawing of the path of blood flow through the circulatory system. Your drawing will be scored based on how complete and accurate it is. On a scale from 1 to 5, how complete and accurate is your Drawing 1?" Participants responded on a scale from 1, "very low completeness and accuracy," to 5, "very high completeness and accuracy." For *prospective* judgments, participants responded to the following prompt: "Later you will be asked to create new drawings of the circulatory system. Your drawings will be scored based on how complete and accurate they are. On a scale from 1 to 5, how complete and accurate do you think your drawings will be?" Participants responded on a scale from 1, "very low completeness and accuracy," to 5, "very high completeness and accuracy."

3.4.2. *Judgments of comprehension*

Participants also made one additional prospective judgment regarding their comprehension of the learning material. The prompt asked students to predict their ability to *explain* the learning material: "Later you will be asked to write an explanation of how the circulatory system works. Your explanation will be scored based on how complete and accurate it is. On a scale from 1 to 5, how complete and accurate do you think your explanation will be?" Participants responded on a scale from 1, "very low completeness and accuracy," to 5, "very high completeness and accuracy."

3.4.3. *Drawing performance*

We used a lab-developed rubric to score the quality of students' two drawings during the learning phase and the final drawing test. The final drawing test required participants to reproduce drawings 1 and 2 without access to the text or their original drawings. Participants received one point for each component accurately situated and labeled in their drawings, such as depicting the aorta coming out of the left ventricle. All components included in the rubric were explicitly described in the text. As mentioned above, participants were not graded on whether they followed the specific style or conventions depicted in the instructional visuals, such as the relative size, shape or location of structures or the orientation of the left and right sides of the heart. Drawing 1 was worth 16 possible points; drawing 2 was worth 5 possible points. Two research assistants scored all drawings for 32 participants. Inter-rater reliability was high (ICC's $> .80$), and so the remaining drawings were distributed between the two raters. Figs. 3 and 4 present example participant drawings for drawings 1 and 2, respectively.

3.4.4. *Comprehension performance*

To assess comprehension, participants completed an explanation

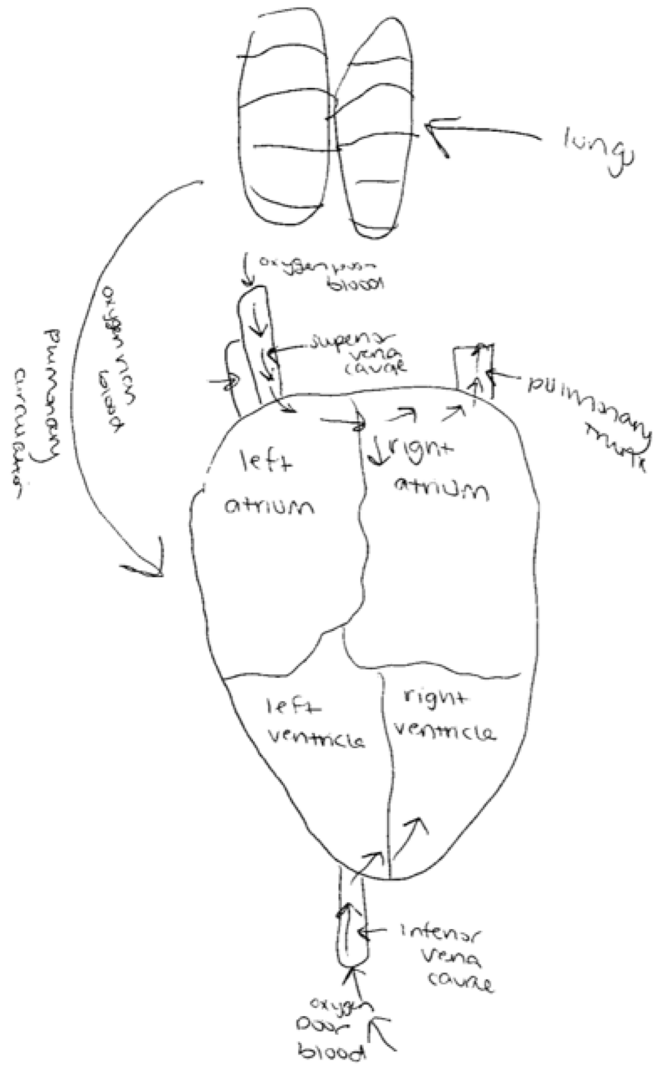

Fig. 3. Example Participant Drawing for Drawing 1.

test. The explanation test consisted of one free-response item: "Please write an explanation of how the human circulatory system works. Your explanation will be scored based on how complete and accurate it is." We scored students' explanations using a lab-developed rubric. Participants received one point for each component they included in their explanation for a maximum of 33 possible points. For example, participants received one point each for stating that systemic circulation supplies oxygen-rich blood to the body or that blood is received through the vena cavae. Two research assistants scored all explanation tests for 32 participants. Inter-rater reliability was high (ICC = .95), and so the remaining explanation tests were distributed between the two raters. It is important to note that to maximize sensitivity of the measure, our rubric included all possible components from the text that students could theoretically include in their explanations. We did not expect participants to include all or even most of these components in their response. Thus, this measure is designed to detect relative differences in performance across groups rather than yield absolute percentage scores that reflect their level of understanding.

3.4.5. Judgment accuracy

To determine the accuracy of participants' drawing and comprehension judgments, we calculated the difference between their subjective judgments and their actual performance. For example, to calculate the accuracy of students' retrospective drawing judgments we computed the difference between their retrospective judgments (on 1 to 5 scale) and their actual drawing performance during the learning phase (with raw scores converted to a 1 to 5 scale[3]). We followed the same procedure for prospective drawing judgment accuracy and prospective comprehension judgment accuracy. For all judgment types, we computed absolute accuracy (i.e., the absolute difference between judgments and performance) and confidence bias (i.e., the signed difference between judgments and performance). Absolute accuracy indicates the degree of accuracy, whereas confidence bias indicates the extent to which one is overconfident or underconfident. Table 1 summarizes each of the judgment accuracy measures for drawing and comprehension performance.

3.4.6. Additional measures

We also included a brief demographics survey (e.g., age, gender), a prior knowledge test, and a measure of spatial ability. The prior knowledge test consisted of 6 short-answer items, each worth one point, assessing background knowledge related to the circulatory system, e.g., "What is the major difference between arteries and veins?" Two raters scored prior knowledge tests for 31 participants. Inter-rather reliability was high (ICC = .93), and so one rater scored all remaining responses. The measure of spatial ability was the Paper Folding Test, which consists of 10 items requiring participants to imagine a paper being folded, punched with a hole, and then reopened. Prior research indicates performance on the Paper Folding Test is associated with one's ability to draw and comprehend science texts (Zhang and Fiorella, 2019; Jaeger et al., 2018).

3.5. Procedure

Fig. 5 provides an overview of the experimental procedure. Upon providing informed consent, participants completed the prior knowledge test and received general instructions for the learning phase. All participants were informed that the goal of the study was for them to understand how the human circulatory system works and that the drawing activity was intended to improve their understanding. During the learning phase, all participants read the text on the circulatory system, which contained prompts for them to create two drawings using black pen on provided blank sheets of paper. Participants studied the text and created drawings 1 and 2 at their own pace, and learning time was recorded. The experimenter then collected participants' drawings, and participants made their initial judgments of learning: retrospective and prospective judgments of drawing and prospective judgments of comprehension.

Next, the experimenter handed back the participants' drawings, and participants completed the restudy phase, during which they were randomly assigned to the compare group or the control group. The compare group compared their drawings to the provided visuals; the control group restudied the text and reviewed their drawings without provided visuals. Both groups completed the restudy phase at their own pace, and restudy time was recorded. Then the experimenter collected participants' drawings, and all participants made new retrospective and prospective judgments of drawing, as well as new prospective judgments of comprehension. After their new judgments, participants completed the final drawing test (during which they recreated drawings 1 and 2 without the text), and the explanation test. Finally, participants completed the Paper Folding Test and the brief demographics survey. The entire experiment lasted approximately 60 min.

[3] We used the following formula to rescale the raw scores: NewValue = (((OldValue – OldMin) x (NewMax –NewMin)) / (OldMax – OldMin) + NewMin. For example, a score of 6 out of 16 on drawing 1 converted to a 1 to 5 scale would equal 2.5.

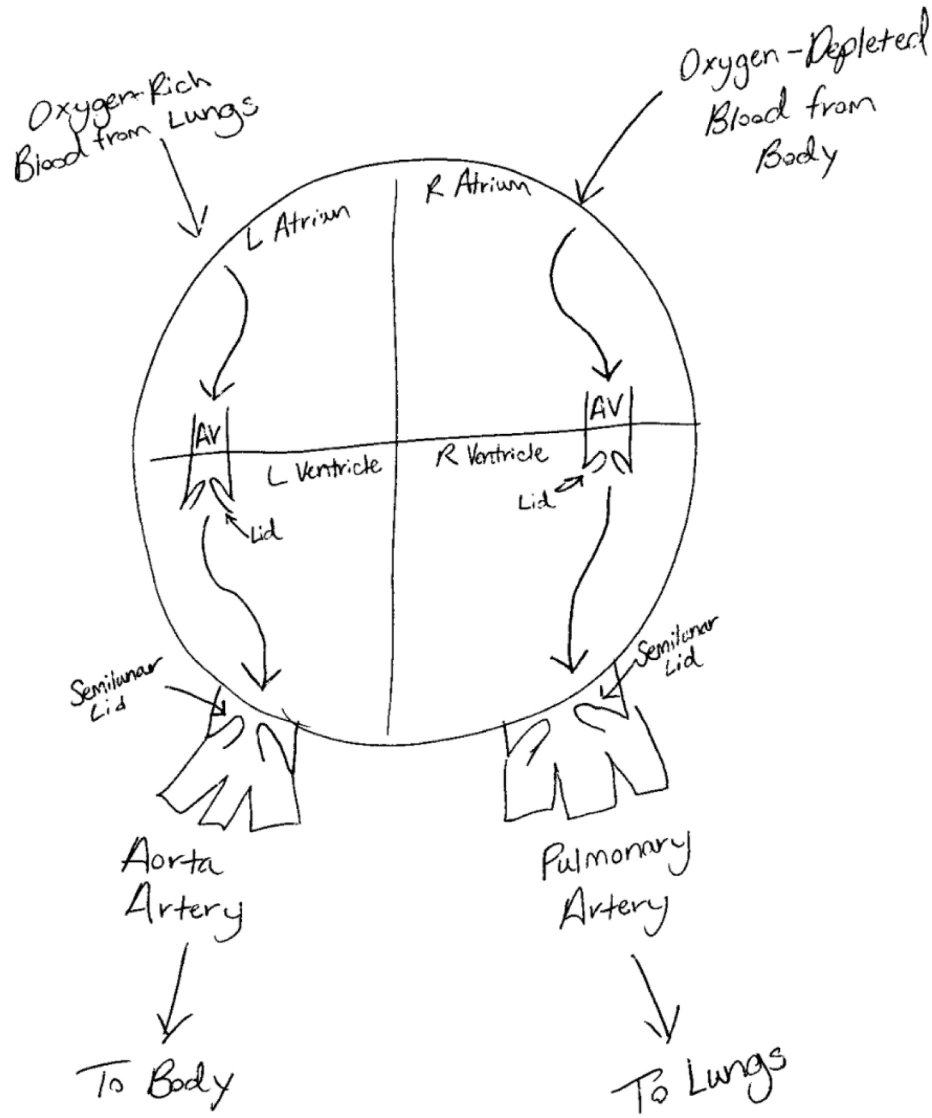

Fig. 4. Example Participant Drawing for Drawing 2.

Table 1
Summary of Judgment Accuracy Measures.

Type	Judgment	Metrics
Retrospective	Drawing 1 performance	Absolute accuracy Confidence bias
	Drawing 2 performance	Absolute accuracy Confidence bias
Prospective	Drawing test performance	Absolute accuracy Confidence bias
	Explanation test performance	Absolute accuracy Confidence bias

4. Results

4.1. Preliminary analyses

First, we tested whether the two groups significantly differed in prior knowledge, spatial ability, time spent during the learning phase, or time spent during the restudy phase. Independent samples *t*-tests revealed no significant differences between groups for performance on the prior knowledge test, $t(114) = 0.29$, $p = .773$, the Paper Folding Test, $t(114) = 1.64$, $p = .105$, or for time spent during the learning phase, $t(114) = 0.07$, $p = .948$. The compare group spent significantly more time than the control group in the restudy phase, $t(114) = 3.92$, $p < .001$, $d = .73$; however, this was expected because the compare group required additional time to process and compare the provided visuals to their own drawings.

4.2. Drawing and comprehension performance

Next, we examined whether comparing one's drawings to provided visuals affected students' actual level of drawing and comprehension performance. Drawing performance was assessed at two time points: the learning phase and the final drawing post-test. Comprehension performance was assessed via the explanation test.

4.2.1. Drawing performance

Table 2 presents descriptive statistics for students' performance on drawings 1 and 2 during the learning phase and the test phase. To test whether comparing to provided visuals caused changes in actual drawing performance, we used repeated measures ANOVA with group (compare or control) as the between-subjects factor, time (learning phase or test phase) as the within-subjects factor, and performance on drawing 1 and 2 as dependent measures. Regarding change in performance on drawing 1, the group by time interaction was not statistically

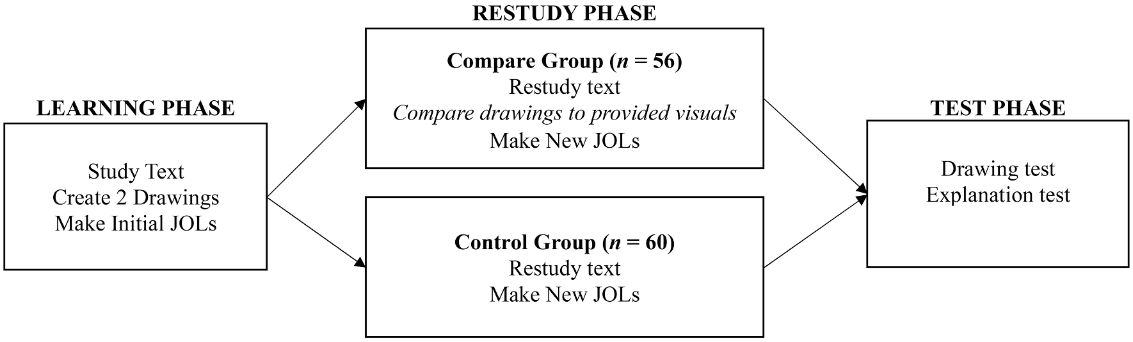

Fig. 5. Overview of Experimental Procedure. Note: JOLs = Judgments of Learning: retrospective judgments of drawing and prospective judgments of drawing and comprehension.

Table 2
Drawing Performance During Learning Phase and Test Phase.

Group	Drawing 1 Performance				Drawing 2 Performance			
	Learning Phase		Test Phase		Learning Phase		Test Phase	
	M	SD	M	SD	M	SD	M	SD
Compare	6.7	3.6	3.9	3.1	2.1	1.8	1.2	1.7
Control	7.6	4.0	5.3	4.0	2.2	1.9	1.9	1.8

Note. Drawing 1: maximum 16 points; Drawing 2: maximum 5 points.

significant, $F(1, 114) = 0.41$, $p = .524$, suggesting that both groups showed a similar pattern of performance. Regarding change in performance on drawing 2, there was a significant group by time interaction, $F(1,114) = 4.90$, $p = .029$, partial $\eta^2 = .04$, such that the compare group showed a stronger *decline* in performance for drawing 2. There were also significant main effects of time for both drawings (drawing 1: $F(1,114) = 61.59$, $p < .001$, partial $\eta^2 = .35$; drawing 2: $F(1,114) = 18.60$, $p < .001$, partial $\eta^2 = .14$), such that students generally performed worse on the final test (when they did not have access to the text) than on their initial drawings during the learning phase (when they did have access to the text). There were not significant main effects of group for either drawing (drawing 1: $F(1,114) = 3.76$, $p = .055$; drawing 2: $F(1,114) = 1.65$, $p = .202$). Overall, comparing one's drawing to provided visuals did not affect drawing performance on drawing 1, and it hindered performance on drawing 2.

4.2.2. Comprehension performance

Levene's test indicated unequal variances between groups on the explanation test; thus, we conducted an independent-samples *t*-test without assuming equal variances. The results indicated no significant difference between the compare group ($M = 5.5$; SD = 4.4) and the control group ($M = 6.6$; SD = 5.6) on the explanation test, $t(104.25) = 1.18$, $p = .243$.

4.3. Accuracy of drawing and comprehension judgments

Our primary hypotheses concerned whether comparing one's drawings to provided instructional visuals changes the accuracy of students' judgments of comprehension and drawing.[4] The comparison facilitation hypothesis posits that comparing generative drawings to instructional visuals will increase judgment accuracy, whereas the comparison interference hypothesis posits comparing to instructional visuals will decrease judgment accuracy. To test these competing hypotheses, we used repeated measures analysis of variance (ANOVA) with group (compare or restudy) as the between-subjects factor, time (learning phase or restudy phase) as the within-subjects factor, and the accuracy of students' retrospective drawing judgments, prospective drawing judgments, and prospective comprehension judgments as dependent variables. We tested for statistically significant *group by time interactions*,[5] which would indicate a change in judgment accuracy after comparing to provided visuals.

4.3.1. Accuracy of retrospective drawing judgments

Table 3 presents descriptive statistics for the accuracy of students' retrospective judgments for drawings 1 and 2. Regarding absolute accuracy, there were no significant group by time interactions for drawing 1, $F(1,114) = 1.60$, $p = .209$, or drawing 2, $F(1,114) = 0.03$, $p = .865$. However, regarding confidence bias, there were large significant interaction effects for drawing 1, $F(1,114) = 23.23$, $p < .001$, partial $\eta^2 = .17$, and drawing 2, $F(1,114) = 20.78$, $p < .001$, partial $\eta^2 = .15$, such that the compare group became more *underconfident* in their retrospective judgments of drawing 1 and 2 (see Fig. 6). This pattern supports the comparison interference hypothesis: participants who compared their drawings to provided instructional visuals tended to rate their drawings as lower quality than they actually were.

Table 3
Accuracy of Students' Retrospective Judgments for Drawings 1 and 2.

Group	Drawing 1 Absolute Accuracy				Drawing 1 Confidence Bias			
	Learning Phase		Restudy Phase		Learning Phase		Restudy Phase	
	M	SD	M	SD	M	SD	M	SD
Compare	0.76	0.5	0.90	0.7	0.14	0.9	−0.37	1.0
Control	1.02	0.7	0.99	0.7	−0.14	0.9	0.01	1.2
	Drawing 2 Absolute Accuracy				Drawing 2 Confidence Bias			
	Learning Phase		Restudy Phase		Learning Phase		Restudy Phase	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Compare	1.06	0.7	1.13	0.9	−0.10	1.3	−0.47	1.4
Control	1.02	0.7	1.07	0.7	−0.21	1.2	−0.04	1.3

Note. Absolute accuracy: values reflect absolute difference between perceived score and actual score. Confidence bias: values reflect the signed difference between perceived score and actual score (positive values reflect overconfidence; negative values reflect underconfidence).

[4] We focus here on the *accuracy* of students' judgments (i.e., the difference between students' judgments and their actual performance) rather than the overall magnitude of students' judgments. For full reporting of the magnitude of students' judgments, please see the Supplemental Materials (Supplemental Analysis #1).

[5] Given the design of the experiment, main effects of group and time are not particularly informative. Thus, for clarity, we focus here on results concerning group by time interaction effects. For full reporting of main effects of group and time, please see the Supplemental Materials (Supplemental Analysis #2).

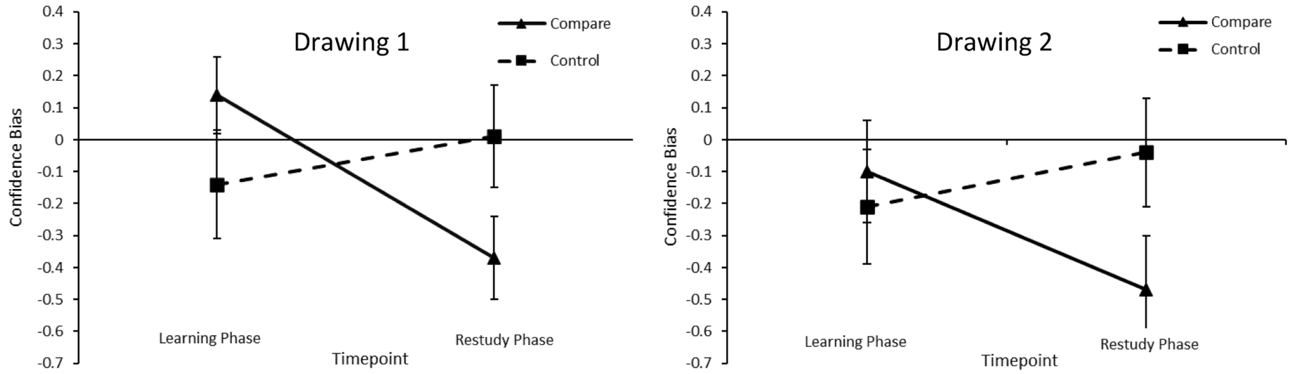

Fig. 6. Accuracy of Students' Retrospective Confidence Bias for Drawing 1 (left) and Drawing 2 (right). Note. Error bars represent standard error of the mean. Values above 0 reflect overconfidence; values below 0 reflect underconfidence.

4.3.2. *Accuracy of prospective drawing judgments*

Table 4 presents descriptive statistics for the accuracy of students' prospective judgments of drawing performance. Regarding absolute accuracy, there was a significant group by time interaction, $F(1,114) = 8.85$, $p = .004$, partial $\eta^2 = .07$, such that the compare group exhibited significantly *reduced* absolute accuracy than the control group (i.e., they showed a greater absolute degree of error in their judgments). Regarding confidence bias, the group by time interaction was not statistically significant, $F(1,114) = 3.30$, $p = .072$, partial $\eta^2 = .03$; however, the pattern suggests the compare group became relatively more overconfident (i.e., they thought they would perform better than they did). Overall, the results support the comparison interference hypothesis: participants who compared their drawings to provided instructional visuals were less accurate in their judgments of future drawing performance, with a greater tendency toward overconfidence (see Fig. 7).

4.3.3. *Accuracy of prospective comprehension judgments*

Table 4 also presents descriptive statistics for the accuracy of students' prospective judgments for the explanation test. There were no significant group by time interactions for absolute accuracy, $F(1,114) = 0.03$, $p = .873$, or confidence bias, $F(1,114) = 0.82$, $p = .368$. This suggests that comparing one's drawing to provided visuals did not affect the accuracy of students' judgments of comprehension.

4.4. *Types of comparisons made among the compare group*

Next, we explored the type of comparisons made by participants in the compare group while processing the instructional visuals. Participants made a mean of 6.1 comparisons ($SD = 1.6$). A paired samples *t*-tests indicated that participants made significantly more surface-level comparisons ($M = 3.5$; $SD = 1.9$) than conceptual comparisons. ($M = 2.6$; $SD = 1.7$), $t(59) = 2.27$, $p = .013$, $d = .29$. However, the total number of surface-level comparisons or conceptual comparisons (or the proportion of conceptual to surface comparisons) did not significantly correlate with any of the judgment accuracy measures, performance on the final drawing test, or performance on the explanation test (all p's $>$ .05). This suggests making more conceptual comparisons was generally not associated with better metacognitive or learning outcomes (see Discussion).

To provide further insight into the comparison process, we also explored the frequency of six specific types of surface-level differences that students might have mentioned (see Table 5): (a) the use of color, (b) the left/right orientation of the heart, (c) the dimensionality of the visuals (2D vs. 3D), (d) the relative size/location of the structures (e.g., the location/size of the lungs), (e) the level of detail presented in the visuals, and (f) the number of visuals generated vs. provided. For drawing 1, the most common surface-level differences included relative size or location, left/right orientation, and level of detail. For drawing 2, the most common surface-level differences were level of detail, relative size or location, and number of drawings. Interestingly, students almost never mentioned factors like the use of color or 3D in the instructional visuals even though students' drawings were all in black and white and students overwhelmingly created their drawings in 2D. Overall, these exploratory data help specify which surface-level differences between their drawings and the instructional visuals were most salient to students during the comparison process.

Table 4
Accuracy of Students' Prospective Judgments of Drawing and Explanation Test Performance.

Group	Drawing Test: Absolute Accuracy				Drawing Test: Confidence Bias			
	Learning Phase		Restudy Phase		Learning Phase		Restudy Phase	
	M	SD	M	SD	M	SD	M	SD
Compare	0.88	0.7	1.20	0.8	0.72	0.9	1.14	0.9
Control	0.93	0.7	0.90	0.7	0.33	1.1	0.51	1.0
	Explanation Test: Absolute Accuracy				Explanation Test: Confidence Bias			
	Learning Phase		Restudy Phase		Learning Phase		Restudy Phase	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Compare	0.88	0.6	0.91	0.7	0.69	0.8	0.79	0.8
Control	1.05	0.7	1.10	0.7	0.65	1.1	0.86	1.0

Note. Absolute accuracy: values reflect absolute difference between perceived score and actual score. Confidence bias: values reflect the signed difference between perceived score and actual score (positive values reflect overconfidence; negative values reflect underconfidence).

4.5. *Correlations among drawing and comprehension performance*

Finally, we examined the correlations among the drawing performance measures and the explanation test. As expected, drawing performance during the learning phase was positively associated with drawing performance during the test phase ($r = .65$, $p < .001$). Furthermore, in line with prior work, drawing quality during the learning phase and the test phase were both positively associated with performance on the explanation test ($r = .39$, $p < .001$ and $r = .50$, $p <$ .001, respectively). This is notable because, as reported above, comparing to instructional visuals affected participants' judgments of drawing but did not affect their judgments of comprehension.

5. Discussion

Prior research suggests students need instructional support to experience the benefits of learning by drawing. One common form of instructional support is to prompt students to compare their drawings to instructional visuals, such as those created by textbook illustrators.

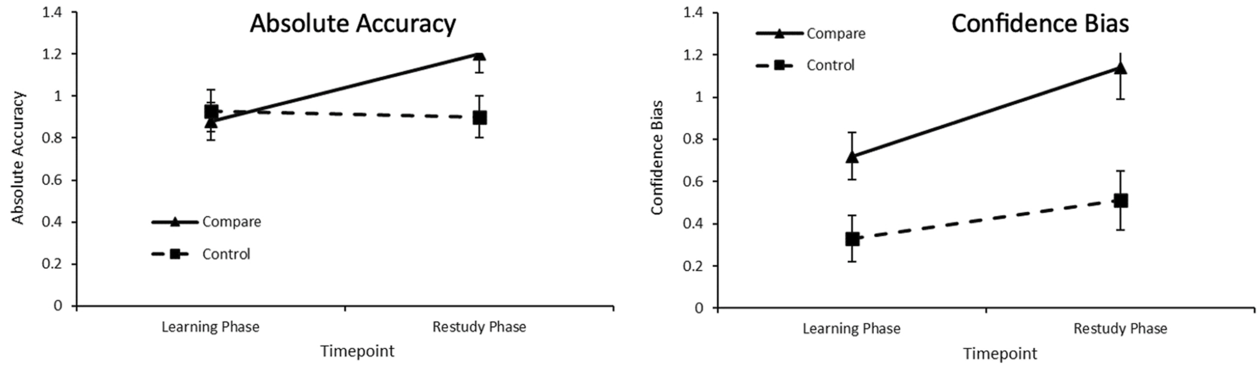

Fig. 7. Absolute Accuracy (left) and Confidence Bias (right) for Prospective Judgments of Drawing Performance. Note. Error bars represent standard error of the mean. For absolute accuracy (left), greater values reflect worse accuracy. For confidence bias (right), greater values reflect higher overconfidence.

Table 5
Frequency of Different Types of Superficial Comparisons Between Generated Drawings and Instructional Visuals Among the Compare Group.

Type of Superficial Difference	% for Drawing 1	% for Drawing 2
Use of color	2 %	2 %
Left/right heart orientation	30 %	4 %
Dimensionality (2D/3D)	0 %	2 %
Relative size/location	57 %	21 %
Level of detail	32 %	34 %
Number of drawings	5 %	20 %

Note: Percentage (%) is out of 56 total participants in the compare group.

While some prior studies have found benefits of comparing generative drawings to instructional visuals, other studies have found null or even negative effects. We hypothesized that these mixed findings may be explained by how instructional visuals affect students' metacognitive judgments during learning. According to the comparison facilitation hypothesis, instructional visuals should help students identify and correct knowledge gaps, thereby supporting higher judgment accuracy and learning outcomes. In contrast, according to the comparison interference hypothesis, instructional visuals may cause students to focus on superficial metacognitive cues that lead to lower judgment accuracy and learning outcomes. We tested these competing hypotheses by examining how comparing generative and instructional visuals affects students' perceived and actual drawing and comprehension performance. Specifically, we examined changes in the accuracy of students' retrospective and prospective judgments of drawing performance and students' prospective judgments of comprehension performance.

The results provide support for the comparison interference hypothesis: comparing one's drawings to instructional visuals generally led to worse drawing judgment accuracy and performance. First, comparing to instructional visuals caused students to make less accurate *retrospective* judgments about the quality of the drawings they created during the learning phase. Specifically, the compare group exhibited relatively more underconfidence: they perceived their drawings as lower quality than they actually were. Second, comparing to instructional visuals caused students to make less accurate *prospective* judgments about how well they would create new drawings on the final drawing test. Specifically, students in the compare group exhibited worse absolute accuracy, with a trend towards overconfidence: they thought they would perform better on the final drawing test than they actually did.

These findings suggest students in the compare group noticed differences between their drawings and the instructional visuals but focused on differences that were superficial or otherwise irrelevant to their understanding, which may have interfered with their judgment accuracy. Indeed, analysis of students' responses to the comparison prompts revealed that students generated significantly more irrelevant or *surface-level* comparisons than relevant or *conceptual* comparisons. This is consistent with Zhang and Fiorella's (2023) model of learning from errors, which distinguishes between *surface* errors, requiring only detection of salient differences between one's responses and provided reference information, and *deep* errors, which require actively self-explaining why one's responses are incorrect at a conceptual level. Students noticed many surface-level differences, such as those related to the specific style or conventions used in the instructional visuals (e.g., left/right orientation of the heart, relative size/location, level of detail), which might have contributed to underconfidence in their retrospective judgments of drawing quality. Having noticed these surface-level differences, students might have subsequently been relatively overconfident about their ability to create high-quality drawings on the subsequent post-test.

It is important to note that the explanation provided above is inferred from the pattern of students' judgments, actual performance, and the types of comparisons students made with the provided visuals. We did not find a significant relationship between the type of comparisons students made (surface level or conceptual) and their judgment accuracy. However, this may be due to several factors. First, judgment accuracy was already relatively low among the students in the compare group. Second, most participants made relatively few comparisons overall and reported a mix of irrelevant surface-level errors as well as (to a lesser extent) relevant conceptual errors. Thus, there may have been a restriction of range in participants' judgment accuracy and the quantity and quality of comparisons they made, making it difficult to detect how the use of different metacognitive cues relates to judgment accuracy. Third, students might not reliably report all the differences they notice between their drawings and the provided visuals, and further, the comparisons they include might not represent the information they used to make their judgments. For example, students might have used features of the provided visuals, rather than comparisons with their own drawings per se, to make their judgments. We relied on their responses to the comparison prompts and did not explicitly ask students how they made their retrospective or prospective judgments, though prior research suggests that if we did, such self-reports might be unreliable (e.g., Griffin et al., 2019). Finally, even the conceptual comparisons students made may have been relatively superficial. That is, students might have *detected* a conceptual error, but that does not necessarily mean they understood and *corrected* the error by revising their existing knowledge (Zhang & Fiorella, 2023). Overall, the provided visuals clearly interfered with students' retrospective and prospective drawing judgments, suggesting students focused on features of the instructional visuals that were not diagnostic of the conceptual fidelity of their drawings or their understanding of the circulatory system.

Interestingly, while the instructional visuals negatively affected the accuracy of students' drawing judgments, they did not affect the accuracy of students' prospective judgments of comprehension. In other words, students apparently used information from the instructional visuals to inform their drawing judgments but not their comprehension judgments (also see Jaeger & Fiorella, 2023). On one hand, this is not surprising because students were explicitly asked to compare the provided visuals to their own drawings. On the other hand, this finding is notable because it suggests that students did not view the perceived quality of their drawings as a reflection of their comprehension, even though, consistent with prior research (e.g., Fiorella & Kuhlmann, 2020; Schwamborn et al., 2010), drawing quality and comprehension performance were strongly correlated. Similar to the explanation above, one possibility is that comparing their drawings to instructional visuals caused students to focus on their perceived ability to *reproduce* the visuals as presented (including irrelevant features of the visuals) rather than their understanding of the underlying concepts conveyed in the visuals. This was despite explicit instructions to participants that the purpose of the study and the drawing task was to focus on one's understanding of how the circulatory system works.

5.1. Practical implications

On a practical level, the findings provide implications for the effective implementation of learning by drawing. Most notably, the findings suggest that instructors cannot assume students will productively compare their drawings to conventional instructional visuals, even when students are explicitly prompted to compare how their drawings differ from them. In fact, this study suggested instructional visuals can produce *worse* metacognitive and learning outcomes than not receiving instructional visuals at all. This does not imply that instructors should not provide visuals that students can use as feedback, but it suggests students need even more explicit support to encourage them to detect and correct the conceptual differences between their drawings and the provided visuals. For example, instructors might need to direct students to engage in comparison of specific conceptual errors in their drawings or provide focused prompts that require students to self-explain their errors (e.g., Jaeger et al., 2020). Students might also benefit from an opportunity to revise their drawings based on the provided visuals (e.g., Van Meter, 2001) to ensure that they have corrected any errors, though students still likely need support to ensure they focus on correcting conceptual rather than surface-level errors.

Another practical consideration is the appearance and style of the provided visuals. We used instructional visuals designed by a professional illustrator to have high conceptual fidelity and limit extraneous details (e.g., see Butcher, 2006). These types of visuals are commonly used in instructional materials such as textbooks and are typically effective at supporting students' understanding from text (e.g., see Mayer & Fiorella, 2022). Nonetheless, results from the present study suggest they also contain specific conventions of style that might interfere with learning when integrated with learner-generated drawing activities. Overall, this study highlights the potential downsides of asking students to compare their own drawings to instructional visuals, particularly because it may cause students to focus on superficial differences between the two representations.

5.2. Limitations and future directions

One limitation of the study is that students were not given specific instructions on how to compare their drawings to the instructional visuals. Consequently, students might not have been aware that they should focus on conceptual differences rather than superficial ones. However, all students did receive explicit instructions that the goal of the study and the drawing tasks was to improve their *understanding* of how the circulatory system works. Furthermore, the prospective and retrospective drawing judgment prompts explicitly asked students to judge the quality of their drawings (retrospective) and their future drawing performance (prospective) in terms of their drawings' 'completeness and accuracy.' Finally, the instructional visuals were representative of the types of visuals students would see in biology textbooks. Thus, the drawing instructions and learning materials were authentic to how students typically learn about the circulatory system. Nonetheless, as mentioned above, the results do indicate that students focused on irrelevant features of the instructional visuals, suggesting students need alternative or more explicit support to encourage productive comparisons between their drawings and the provided visuals. One direction for future research is to consider alternative formats of the provided visuals that more closely match the style of student drawings. Simplified or 'layman' style provided drawings might make it easier for students to focus on the conceptual differences in their drawings. However, this approach also has its challenges because there is considerable within-student variability in the appearance of students' drawings. Thus, future research should also investigate whether additional interventions, such as more focused prompts to compare or self-explain, or a more targeted focus on comparing conceptual differences, would facilitate rather than interfere with metacognitive and learning outcomes.

Similarly, future research should examine how explicit support during the drawing process interacts with the effects of comparing generated and provided instructional visuals. One possibility is that students in the present study created drawings that contained many differences from the instructional visuals, including aesthetic and conceptual differences. Consequently, students may have been overwhelmed during the comparison process and primarily noticed the salient surface-level differences. Providing students with scaffolding during the drawing process, such as partially completed drawings (e.g., Jaeger et al., 2020; Schmeck et al., 2014), might ensure that students' drawings look more perceptually and conceptually similar to the provided visuals, thereby facilitating the comparison process.

There were also some limitations associated with our implementation of the judgments of learning. First, students made retrospective judgments about each of their drawings separately, but they made prospective judgments about their final drawing performance in general. Thus, we could not distinguish prospective judgment accuracy for each of the two drawings separately. However, the pattern of retrospective judgments and actual drawing performance suggests that comparing to instructional visuals had similar effects for both drawings. Second, we did not counterbalance the different types of judgments of learning, so students always made prospective judgments of drawing before their judgments of comprehension. Thus, we cannot rule out that making one judgment type affected subsequent judgments. However, one notable result from the study is that comparison affected prospective drawing judgment accuracy but not prospective comprehension judgment accuracy, suggesting that students did distinguish between the different types of judgments rather than make similar comprehension judgments following their drawing judgments.

Future research should also consider other ways of measuring the types of comparisons students make between their drawings and provided visuals. In the present study, participants typed their comparison into a textbox. This method provided insight into the types of information students focused on during the comparison process, but some types of comparisons (particularly involving visual information) might be challenging to communicate via text alone. A future study could video record participants' speech and gestures (e.g., pointing to specific elements of the generated or provided visual) to provide a more comprehensive understanding of what differences students focus on and where they might need instructional support.

More broadly, this study should be extended with additional learning materials, student populations, and within more authentic classroom settings. In a laboratory context, students might not have been sufficiently motivated to invest the effort required to make deeper conceptual comparisons between their drawings and the provided visuals. Furthermore, other research suggests drawing activities themselves present unique cognitive, metacognitive, and motivational barriers (see Fiorella, 2023), such as drawing requiring considerable time and effort, students holding unproductive beliefs about the utility of drawing, and drawing being relatively less effective with younger learners. Future research should systematically examine how individual student characteristics such as age, prior knowledge, and motivation might moderate how students respond to provided visual comparison tasks during learning by drawing.

Author Note

This research was supported by grants from the National Science Foundation awarded to Logan Fiorella (1955348) and Allison Jaeger (1956466).

Ethics statement

The authors declare no conflicts of interest. This study was approved by and conducted in accordance with the authors' Institutional Review Board (IRB).

CRediT authorship contribution statement

Logan Fiorella: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Allison J. Jaeger:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition. **Alexis Capobianco:** Writing – review & editing, Formal analysis, Data curation. **Anna Burnett:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We include a link to the data in the manuscript.

Appendix A

Text on Human Circulatory System

The human circulatory system

Part 1

The heart consists of four chambers. On the top are the left atrium and the right atrium. These chambers receive blood returning from the body and the lungs. On the bottom are the left ventricle and right ventricle. These chambers pump blood to the body and the lungs.

Although it is a single organ, the heart functions as a double pump. The right side receives relatively oxygen-poor blood from the body through the large superior and inferior venae cavae and pumps it out through the pulmonary trunk. The pulmonary trunk splits into the right and left pulmonary arteries, which carry blood to the lungs. Here, oxygen is picked up and carbon dioxide is unloaded. Oxygen-rich blood drains from the lungs and returns to the left side of the heart through the pulmonary veins. This system of circulation, from the right side of the heart to the lungs and back to the left side of the heart, is called the pulmonary circulation. Its only function is to carry blood to the lungs for gas exchange and then return it to the heart.

Blood that returns to the left side of the heart is pumped out into the aorta. Systemic arteries branch out from the aorta to supply blood to all body tissues. Oxygen-poor blood circulates from the tissues back to the right atrium via the superior or inferior vena cava of the higher or lower body. This second circuit, from the left side of the heart through the body tissues and back to the right side of the heart, is called the systemic circulation. It supplies oxygen- and nutrient-rich blood to all body organs. Because the left ventricle pumps blood into the systemic circulation throughout the whole body, its walls are substantially thicker than those of the right ventricle, and it is a much more powerful pump.

Inside the lungs are many tiny air sacs called alveoli which are wrapped in a network of capillaries. As blood flows through the capillaries, differences in the concentrations of oxygen and carbon dioxide cause these gases to move between the blood and alveoli. These gases pass easily back and forth through the thin walls of the capillaries in the lungs and the body. Oxygen levels are higher in the alveoli, which causes oxygen molecules to diffuse across the membranes into the blood. In contrast, carbon dioxide levels are higher in the blood, which causes carbon dioxide molecules to diffuse from the capillaries into the alveoli.

Blood makes a complete circuit from the right side of the heart through the lungs, back through the left side of the heart and out the aorta every 2.5 s when a person is at rest. During exercise, the blood can travel this short loop in about 1 s. The blood protein called hemoglobin is largely responsible for how fast the exchange between oxygen and carbon dioxide occurs. Each molecule of hemoglobin carries four molecules of oxygen to the tissues of the body and gives blood its bright red color. As blood is drained of oxygen, the molecules of hemoglobin lose their red stain and become purple or blue.

Part 2

The heart is equipped with four valves, which allow blood to flow in only one direction through the heart chambers—from the atria through the ventricles and out the arteries leaving the heart. The left and right atrioventricular (AV) valves are located between the atrial and ventricular chambers on each side. These valves prevent backflow into the atria when the ventricles contract. When the heart is relaxed and blood is passively filling its chambers, the AV-valve flaps hang limply into the ventricles. As the ventricles contract, they press on the blood in their chambers, and the pressure inside the ventricles begins to rise. This forces the AV-valve flaps upward, closing the valves and preventing backflow.

The semilunar valves guard the bases of the two large arteries leaving the ventricles. When the ventricles contract and force blood out of the heart, the valves are forced open and flatten against the walls of the arteries. When the ventricles relax, the blood begins to flow backward toward the heart, and the valves close to prevent blood from reentering the heart. As the atrioventricular and semilunar valves open and close in response to pressure changes in the heart, they force blood to continually move through the heart in one direction.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2024.102277.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198.

Ainsworth, S. E., & Scheiter, K. (2021). Learning by drawing visual representations: Potential, purposes, and practical implications. *Current Directions in Psychological Science, 30*(1), 61–67.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Reviews of Psychology, 64*, 417–444.

Bobek, E., & Tversky, B. (2016). Creating visual explanations improves learning. *Cognitive Research: Principles and Implications, 1*, 27.

Brod, G. (2021). Generative learning: Which strategies for which age? *Educational Psychology Review., 33*(4), 1295–1318.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182–197.

Cardwell, B. A., Lindsay, D. S., Forster, K., & Garry, M. (2017). Uninformative photos can increase people's perceived knowledge of complicated processes. *Journal of Applied Research in Memory and Cognition, 6*, 244–252.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology, 35*(1), 59–74.

Fiorella, L. (2023). Making sense of generative learning. *Educational Psychology Review, 35*(50), 1–42.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741.

Fiorella, L., & Jaeger, J. (2023). *Are there metacognitive benefits of learner- and instructor-generated visualizations?* Applied Cognitive Psychology: Advance online publication.

Fiorella, L., & Kuhlmann. (2020). Creating drawings enhances learning by teaching. *Journal of Educational Psychology, 112*(4), 811–822.

Fiorella, L., & Zhang, Q. (2018). Drawing boundary conditions for learning by drawing. *Educational Psychology Review, 30*, 1115–1137.

Gilbert, J. K. (2008). Visualization: An emergent field of practice and enquiry in science education. In J. Gilbert, M. Reiner, & M. Nakhleh (Eds.), *Visualization: Theory and practice in science education* (pp. 3–24). Springer.

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619–646). Cambridge University Press.

Hegarty, M., Carpenter, P. A., & Just, M. A. (1991). Diagrams in the comprehension of scientific texts. In R. Barr, M. L., Kamil, P., & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 641-668). New York: Longman.

Ikeda, K., Kitagami, S., Takahashi, T., Hattori, Y., & Ito, Y. (2013). Neuroscientific information bias in metacomprehension: The effect of brain images on metacomprehension judgment of neuroscience research. *Psychonomic Bulletin & Review, 20*(6), 1357–1363.

Jaeger, A. J., & Fiorella, L. (2023). *Metacognitive effects of instructional visuals: The role of cue use and judgment type*. Metacognition and Learning: Advance online publication.

Jaeger, A. J., Marzano, J., & Shipley, T. F. (2020). When seeing what's wrong makes you right: The effect of incorrect examples on 3D diagram learning. *Applied Cognitive Psychology, 34*, 844–861.

Jaeger, A. J., Velazquez, M., Dawdanow, A., & Shipley, T. (2018). Using sketching to reduce memory for seductive details. *Journal of Educational Psychology, 110*, 899–916.

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34*, 58–73.

Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction, 13*(2), 205–226.

Leopold, C., & Leutner, D. (2012). Science text comprehension: Drawing, main idea selection, and summarizing as learning strategies. *Learning and Instruction, 22*(1), 16–26.

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior, 25*(2), 284–289.

Leutner, D., & Schmeck, A. (2022). The drawing principle in multimedia learning. In R. E. Mayer, & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd edition, pp. 360–369). Cambridge University Press.

Mayer, R. E., & Fiorella, L. (2022). *Eds*. The Cambridge handbook of multimedia learning: Third edition. Cambridge University Press.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465–489.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778.

Novick, L. R. (2006). The importance of both diagrammatic conventions and domain-specific knowledge for diagram literacy in science: The hierarchy as an illustrative case. In D. Barker-Plummer, R. Cox, & N. Swoboda (Eds.), *Diagrams, LNAI 4045* (pp. 1–11). Springer.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgments reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698–711.

Schmeck, A., Mayer, R. E., Opfermann, M., Pfeiffer, V., & Leutner, D. (2014). Drawing pictures during learning from scientific text: Testing the generative drawing effect and the prognostic drawing effect. *Contemporary Educational Psychology, 39*(4), 275–286.

Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., & Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. *Journal of Educational Psychology, 102*(4), 872–879.

Schwamborn, A., Thillmann, H., Opfermann, M., & Leutner, D. (2011). Cognitive load and instructionally supported learning with provided and learner-generated visualizations. *Computers in Human Behavior, 27*(1), 89–93.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Thiede, K. W., Wright, K. L., Hagenah, S., Wenner, J., Abbott, J., & Arechiga, A. (2022). Drawing to improve metacomprehension accuracy. *Learning and Instruction, 77*, Article 101541.

van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Mapping and drawing to improve students' and teachers' monitoring and regulation of students' learning from text: Current findings and future directions. *Educational Psychology Review, 32*, 951–977.

Van Meter, P. (2001). Drawing construction as a strategy for learning from text. *Journal of Educational Psychology, 93*(1), 129–140.

Van Meter, P., & Firetto, C. M. (2013). Cognitive model of drawing construction. In G. Schraw, M. T. McCrudden, & D. Robinson (Eds.), *Learning through visual displays* (pp. 247–280). Information Age Publishing.

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology, 33*(1), 9–19.

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist, 24*(4), 345–376.

Zhang, Q., & Fiorella, L. (2019). Role of generated and provided visuals in supporting learning from scientific text. *Contemporary Educational Psychology, 59*, Article 101808.

Zhang, Q., & Fiorella, L. (2021). Learning by drawing: When is it worth the time and effort? *Contemporary Educational Psychology, 66*, Article 101990.

Zhang, Q., & Fiorella, L. (2023). An integrated model of learning from errors. *Educational Psychologist, 58*(1), 18–34.