Instructional Science
https://doi.org/10.1007/s11251-024-09701-8

ORIGINAL RESEARCH

Comparing effectiveness of exploratory learning activities given before instruction: generating multiple strategies vs. inventing one strategy

Lianda Velić[1] · Marci S. DeCaro[1]

Received: 1 October 2023 / Accepted: 9 December 2024

Abstract

Exploratory learning before instruction typically benefits conceptual understanding compared to traditional instruction-first methods. The current study examined whether different exploration prompts impact students' exploration approaches and learning outcomes, using a quasi-experimental design. Undergraduate students ($N = 164$) in psychological statistics courses were taught the procedure and concepts of standard deviation. Students in the *instruct-first* condition received direct instruction then a practice problem. Students in the explore-first conditions attempted the problem before instruction, with exploration prompts differing between conditions. Students in the *explore-first invent* condition were asked to invent a formula; students in the *explore-first generate* condition were asked to come up with different ways of measuring consistency. Students in the explore-first generate condition scored significantly higher on procedural knowledge (problem solving) than in the explore-first invent condition, conceptual knowledge than in both other conditions, and preparation for future learning (transfer) than in the instruct-first condition. Students in the explore-first invent condition scored no differently on any learning outcomes than in the instruct-first condition. Students given the strategy generation prompt more broadly explored different strategies during the exploration activity, but used fewer correct solution steps than those given the invention prompt. Broader exploration—and not accuracy—was associated with higher conceptual knowledge. Conversely, students in the instruct-first condition used fewer, more accurate, strategies on the activity compared to the explore-first conditions. They also showed greater misconceptions during the activity and posttest, indicating superficial understanding. Both explore-first conditions induced greater awareness of knowledge gaps compared to the instruct-first condition. Generating multiple strategies likely helped students discern important problem features, deepening conceptual structures that supported learning even beyond the initial lesson.

Keywords Exploratory learning · Productive failure · Invention · Strategy generation · Problem solving

✉ Marci S. DeCaro
marci.decaro@louisville.edu

[1] Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY 40292, USA

Published online: 22 January 2025

Introduction

Traditional forms of teaching largely utilize lecture followed by practice (Stains et al., 2018). However, students often fail to develop conceptual understanding compared to more student-centered, active learning methods (Felder & Brent, 2009; Freeman et al., 2014; Prince, 2004; Wegner, 1998). One more student-centered method is *exploratory learning before instruction*. Exploratory learning switches the traditional order by asking students to explore novel content before instruction on the target concepts.

Exploratory learning before instruction is a general term characterizing this two-phase, explore-then-instruct sequence (e.g., Bego et al., 2022; DeCaro & Rittle-Johnson, 2012; Weaver et al., 2018). This sequence is used in studies in several additional research literatures described by different terms (e.g., *problem-solve-instruct* methods, Loibl et al., 2017; Loibl & Rummel, 2014a, 2014b; *productive failure*, Kapur, 2008, 2010, 2011, 2012, 2016; *inventing to prepare for future learning*, Schwartz & Bransford, 1998; Schwartz & Martin, 2004). We use the more general term to reflect that not all exploration activities include problems to solve (cf. Bego et al., 2023; Bush et al., 2023; DeCaro et al., 2022, 2024; Glogger-Frey et al., 2015, Exp. 1; Hieb et al., 2021; Loibl & Leukel, 2023). Studies using this two-phase sequence generally find that an exploration phase with subsequent instruction benefits students' conceptual understanding and transfer to new, related topics more than providing direct instruction first (Darabi et al., 2018; Kapur, 2015, 2016; Loibl et al., 2017; Sinha & Kapur, 2021).

Exploratory learning benefits tend to be selective to higher-level conceptual knowledge (e.g., sense-making, schema formation, transfer; Darabi et al., 2018; Loibl et al., 2017), although some studies have shown benefits for more basic procedural or fluency skills as well (e.g., computing problem-solving steps; DeCaro et al., 2023; Kapur, 2010, 2011; Kapur & Bielaczyc, 2012). However, not all studies find learning benefits (e.g., Chase & Klahr, 2017; Fyfe et al., 2014; Loibl et al., 2020; Nachtigall et al., 2020). Some research suggests that exploratory learning will be less beneficial if mental effort, or cognitive load, is too high during the learning activity (Ashman et al., 2020; Fyfe et al., 2014; Kapur, 2016; Newman & DeCaro, 2019). More research is needed to identify the important design principles and mechanisms that lead to these benefits, beyond domain-specific topics and sub-literatures (Koedinger et al., 2012). Studies that assess factors impacting intermediate learning processes (e.g., during or after the exploration phase) can be especially useful to diagnose when and how exploration impacts learning (Loibl et al., 2023, 2024).

Although research from the literatures listed above all use the same explore-instruct sequence, there are some differences in (a) the ways students are prompted to approach the exploratory activity, (b) whether students' approaches to the learning activity are assessed and connected with learning outcomes, and (c) whether "future learning" (i.e., transfer) is assessed as an outcome measure. The current study examined whether a simple difference in the exploration prompt (i.e., invent one strategy versus generate multiple strategies) impacts how students approach exploring and, therefore, learning outcomes. We also compared both types of exploratory learning to a traditional instruct-first condition using the same materials, in different order. This design provided a baseline comparison to a traditional instructional method and allowed causal evaluation of the benefits of exploring before instruction overall (cf., Bush et al., 2023; DeCaro & Rittle-Johnson, 2012; Loibl et al., 2020; Weaver et al., 2018).

Although each type of exploration prompt has been used in prior studies, they have not been directly compared using the same learning materials, process measures, and assessments. We expected both prompts to support learning mechanisms thought to be important for exploratory learning benefits, except for one key difference. Specifically, we expected that a prompt to generate multiple strategies would promote a wider search through the problem space. This wider search may enable learners to better discern important features of the problem, and enhance the conceptual benefits of exploring. By analyzing the strategies used during problem solving, and surveys of cognitive, metacognitive, and motivational factors, we directly assessed how differences in the exploration prompt and instructional order impact this and other learning mechanisms. We also assessed students' misconceptions during exploration, and subsequent learning outcomes (procedural knowledge, conceptual knowledge, and transfer; Koedinger et al., 2012). This investigation thus provides further insight into whether search through the problem space is critical to exploration. By determining how learning processes are impacted by the exploration phase, we can determine what conditions are most conducive to learning and why (Loibl et al., 2023). More generally, such findings can inform instructors about how the choices they make in designing instructional prompts impacts students' learning processes and outcomes.

Mechanisms of exploratory learning

Exploratory learning is thought to benefit students' conceptual understanding through three primary cognitive and metacognitive mechanisms occurring during and after the first learning phase (Loibl et al., 2017, 2023, 2024). First, by exploring novel problems, students *activate relevant prior knowledge*. Through this process, students better modify or integrate the new knowledge into their existing schemas (Chen & Kalyuga, 2020; Schwartz et al., 2009).

Second, through struggling to find a solution using existing schemas, students become aware of gaps in their understanding of the targeted concept (e.g., Glogger-Frey et al., 2015; Loibl & Rummel, 2014a, 2014b; Schwartz et al., 2007). These *knowledge gaps* may be unobserved in an instruct-first format, which could lead to a false sense of understanding or misconceptions (Kapur, 2016; Schwartz et al., 2007; Tawfik et al., 2015; Wittwer & Renkl, 2008). Students process instruction more deeply when they are aware of their knowledge gaps (VanLehn et al., 2003). Students might also become more motivated to make sense of the new content, increasing their interest and engagement with subsequent instruction (e.g., Belenky & Nokes-Malach, 2012; DeCaro et al., 2015; Kapur, 2016; Lamnina & Chase, 2019; Wise & O'Neill, 2009).

Lastly, by exploring a new problem space, students may begin to explore, identify, and organize *critical problem features* (Kapur & Bielaczyc, 2012; Schwartz et al., 2007, 2011). As they test hypotheses about certain features, students may better understand why these features are, and are not, important (DeCaro & Rittle-Johnson, 2012; Schwartz & Martin, 2004). In contrast, traditional instruction typically highlights these features for students, and students process them superficially (e.g., Kapur, 2016), potentially leading to misconceptions.

These mechanisms likely all contribute, to some extent, to conceptual knowledge change when exploring before instruction (Kapur, 2016; Loibl et al., 2017). Support for these learning mechanisms has been found across studies using the explore-then-instruct sequence in many different learning domains (cf. Loibl et al., 2017). Despite this variety, several studies have been conducted in the domain of statistics, namely teaching students the concepts and procedures of calculating consistency in datasets (i.e., variance, standard deviation, or mean deviation; e.g., Jarosz et al., 2016; Kapur, 2012, 2014a, 2014b, 2015; Loibl & Rummel, 2014b; Loibl et al., 2020; Newman & DeCaro, 2019; Schwartz & Martin, 2004; Wiedmann et al., 2012). Within studies teaching about consistency, the types of exploration activities used vary.

Types of exploration prompts

One difference between studies is the type of prompt used when students are asked to explore a dataset. These prompts might impact how students approach the exploration phase and, thereby, how they learn from instruction (Loibl et al., 2024). In some exploration activities, students are asked to *invent* an index or solution (e.g., Jarosz et al., 2016; Schwartz & Martin, 2004; Schwartz et al., 2011). Invention problems typically include contrasting cases, which create a perceptual space to help students encounter the important features, while keeping less important features constant (e.g., Loibl & Rummel, 2014a, 2014b; Schwartz et al., 2011). For example, Schwartz and Martin (2004) asked students to invent a reliability index for different baseball pitching machines based on graphical representations of where the machine's pitches would land. The invention instruction is intended to help students integrate problem features that might otherwise seem incommensurate—such as using both distance from the mean and sample size to determine the consistency of a sample (Schwartz et al., 2007). By working to synthesize important problem features, students are thought to create conceptual structures, rather than focusing on surface features (Schwartz et al., 2016). This conceptual structure might then improve transfer to other domains (e.g., Schwartz et al., 2007).

Strategy generation prompts build upon invention prompts by asking students to come up with multiple task solutions and approaches, rather than just one. For example, Kapur (2014a) asked students to "design as many measures of consistency as you can" (see also Brand et al., 2023; Hartmann et al., 2021; Kapur, 2012, 2014b, 2015; Loibl et al., 2020; Loibl & Rummel, 2014b; Sinha & Kapur, 2021; Sinha et al., 2021; Trninic et al., 2022). Strategy generation prompts explicitly encourage students to try multiple different representations and sense-making processes (Loibl & Rummel, 2014a; Trninic et al., 2022).

Both types of exploration prompts emphasize the constructivist process of generating solutions for oneself. With both types, students begin by using their prior knowledge and reasoning skills. Because students are not likely to successfully derive the canonical answer, both prompts will likely lead to awareness of knowledge gaps (Kapur, 2010). However, the degree to which students differentiate the problem space might differ based on prompt. Research on problem solving suggests that individuals represent problems, and decide on solution approaches, based on the problem's semantic structure (Thevenot & Oakhill, 2008). Sometimes this initial representation limits solvers' approaches (e.g., Knoblich et al., 1999). Rewording the problem prompt can help learners adjust their approaches to fit the task goals (e.g., DeCaro et al., 2017; Knoblich et al., 1999).

The prompt to invent a strategy suggests that one strategy is the end goal. This framing might lead to more convergent search and refinement processes, whereby students test hypotheses and modify them based on feedback from the environment. The prompt to generate as many strategies as possible suggests that quantity is the goal. This framing might lead to more divergent search and retrieval processes, as the goal is to expand the search space. Thus, invention instructions might lead to a deeper exploration of fewer problem features. Strategy generation instructions might lead to broader exploration of more problem features, including both important and unimportant elements of the problem space (Hartmann et al., 2021; Kapur, 2014a, 2014b, 2015). These learning mechanisms can be evaluated by examining different learning processes and outcomes. The assessment of learning processes and outcomes in studies using strategy generation or invention prompts has some overlap, but also differs to some extent.

Preparation for future learning assessments

Students who understand the deeper underlying concepts should be better able to transfer this knowledge to learn new, related concepts (Kapur, 2010; Marton, 2007; Schwartz & Bransford, 1998; Schwartz et al., 2009; Schwartz et al., 2011). *Future-learning assessments* are one method to assess this transfer ability. These measures include instruction on a new topic (e.g., a passage with a worked example), followed by assessment.

Students who engage in prior exploratory learning typically score higher on these assessments (Bego et al., 2022; Belenky & Nokes-Malach, 2012; Kapur, 2012; Schwartz & Bransford, 1998; Schwartz & Martin, 2004; Schwartz et al., 2011). For example, Schwartz and Martin (2004) found that students who invented a formula for consistency learned the concept of z-scores better as well. Because students more deeply understood the functional relation, they were able to abstract principles that applied to the related domain (Chin et al., 2016; Schwartz et al., 2011). Future-learning assessments are often used in exploratory learning studies that use invention prompts, but these assessments are not typically used in studies using strategy generation prompts.

Productive failure: assessment of activity approaches

Strategy generation prompts tend to be used in research from the *productive failure* literature (e.g., Kapur, 2014a, 2014b, 2015; Trninic et al., 2022; see also Brand et al., 2023; Hartmann et al., 2021; Loibl et al., 2020; Loibl & Rummel, 2014b). Productive failure is the process by which difficult learning conditions can invoke learning processes that deepen understanding. A period of generation (e.g., exploring a novel problem space), enables students to encounter different solution approaches (Kapur, 2008, 2010). During subsequent instruction, students are better able to understand the correct solution, as well as the reasons the incorrect approaches were not correct. Thus, many studies quantify the number of different representation and solution methods students apply during the exploration activity. A greater number of attempted strategies is often associated with higher conceptual learning outcomes (e.g., Kapur, 2014a, 2014b; Kapur et al., 2023; Kapur & Bielaczyc, 2012; but see Hartmann et al., 2022; Loibl & Rummel, 2014a).

In addition to the quantity of solution attempts, some studies investigate the quality of the best solution attempt (e.g., Loibl & Rummel, 2014b; Trninic et al., 2022; Wiedmann et al., 2012). Often, the best solution attempt includes steps toward standard deviation, providing insight in how well students apply or figure out the correct steps to the canonical solution. Quality of solution attempts is associated with learning outcomes in some studies (e.g., Wiedmann et al., 2012) but not others (e.g., Kapur & Bielaczyc, 2012).

Current study

The current study directly compared exploratory learning conditions using invention versus strategy generation prompts. We also compared these explore-first conditions to a more traditional instruct-first condition baseline. We examined process measures (quality and quantity of strategies used, misconceptions during the activity, knowledge gaps, interest, cognitive load) and learning outcomes (procedural, conceptual, and future-learning assessments).

Undergraduate students were introduced to the concept and procedure to calculate consistency (i.e., standard deviation) in statistics, as part of their regular classroom activities. Students in the *explore-first invent* condition were asked to come up with a formula for calculating consistency, before instruction. Students in the *explore-first generate* condition were asked to come up with as many ways to measure consistency as they can, before receiving instruction. Students in the *instruct-first* condition received instruction before practice with the problem. Students in the three conditions completed the exact same materials, with only the order of instruction and the activity prompt differing between conditions. This design allowed us test for the causal effects of (a) exploring before instruction, and (b) the types of exploration prompts used (see Glogger-Frey et al., 2015; Hsu et al., 2015; Loibl et al., 2017; Newman & DeCaro, 2019; Schwartz et al., 2011).

Hypotheses

Learning outcomes

Outcomes measures included procedural knowledge, conceptual knowledge, and future-learning assessments (Fig. 1). Because the benefits of exploring are typically limited to conceptual understanding, we did not predict differences between any conditions on procedural knowledge. However, some research has shown that the benefits of exploring extend to procedural knowledge (e.g., DeCaro et al., 2023; Kapur, 2010, 2011; Kapur & Bielaczyc, 2012).

Based on prior exploratory learning research separately using invention and strategy generation prompts (cf. Loibl et al., 2017), we hypothesized that students in both explore-first conditions would score higher than students in the instruct-first condition on conceptual knowledge and future-learning assessments, regardless of prompt type.

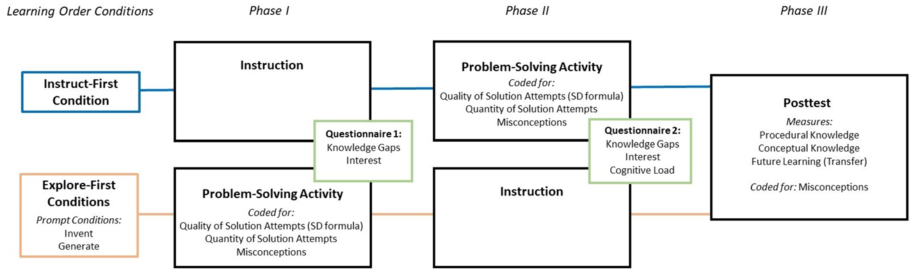


Fig. 1 Experiment procedure and measures

Because prior studies have not directly compared the use of these exploration prompts, we anticipated two possibilities. One possibility is that invention and strategy generation prompts equally improve conceptual knowledge and future learning compared to the instruct-first condition. Both prompt types might lead students to activate relevant prior knowledge, become aware of knowledge gaps, and sufficiently explore the problem features, leading to similar conceptual structures. Another possibility is that strategy generation leads students to explore the problem space more, observing and differentiating critical problem features. In this case, we would expect higher conceptual and future-learning scores in the explore-first generate condition than in the explore-first invent condition.

Overall, the key distinction between the two exploration conditions may lie in the way they prompt students to focus on the problem space and its features. We expected to gain insight into some of these mechanisms by examining students' strategy use on the problem-solving activity and responses to questionnaire items.

Problem-solving activity

Quality and quantity of solutions

We coded students' responses on the problem solving activity (used as exploration or practice), to examine both the quality and quantity of solutions used. *Quality scores* indicate the number of correct steps of the canonical standard deviation formula used during the activity (i.e., how successful students were in their attempts to calculate the formula). Because students in the instruct-first condition were given the formula for standard deviation prior to the activity, we hypothesized that quality scores would be higher than in the explore-first conditions. If students in the explore-first invention condition are more likely to focus narrowly on fewer problem solutions, it is also possible that these students would achieve higher quality scores than students in the explore-first generate condition.

Quantity scores indicate how many different representations and solution methods students attempted during the activity (i.e., how broadly students explored). Because of the prompt to come up with as many strategies as possible, we expected students in the explore-first generate condition to attempt significantly more strategies on the activity than in the other two conditions. Based on prior research, we also expected the number of strategies to correlate with conceptual scores on the posttest (Kapur & Bielaczyc, 2012; Kapur, 2012, 2014a, 2014b; but see Hartmann et al., 2021; Loibl & Rummel, 2014a). Prior research has not connected strategy use with future-learning assessment scores. Whether the relationship between strategy use and conceptual understanding would be strong enough to transfer to future learning was an empirical question.

Misconceptions

As described previously, students who receive instruction then practice may be more likely to process the problem features superficially compared to students who explore first. However, research has not directly tested whether students in an instruct-first condition show greater misconceptions as a result. We additionally coded activities and posttests for evidence of misconceptions in understanding the target concepts. For example, a common error in computing the standard deviation formula is to divide by the mean (M), rather than the sample size (N). This error indicates that students do not understand the rationale (i.e., that dividing by N standardizes the scores across samples varying in size). One concern about exploratory learning is that students will develop and retain misconceptions about the topic (Kirschner et al., 2007). However, exploring before instruction is expected to help students become aware of ways in which their prior knowledge is insufficient (e.g., Kapur, 2010; Loibl et al., 2024). Therefore, we predicted more misconceptions during the activity and posttest in the instruct-first condition than in the explore-first conditions. Because the problem included contrasting cases, guiding students to navigate the most challenging misconceptions, we expected relatively equal misconceptions between the two explore-first conditions.

Questionnaires

Knowledge gaps

We also surveyed students after both the activity and instruction, to assess their perceived knowledge gaps and interest. We hypothesized that students in the explore-first conditions would report higher perceived knowledge gaps than in the instruct-first condition (e.g., Bego et al., 2022; Glogger-Frey et al., 2015; Newman & DeCaro, 2019). We predicted equal perception of knowledge gaps between the explore-first conditions.

Interest

Prior research examining interest has shown mixed results (e.g., Glogger-Frey et al., 2015; Newman & DeCaro, 2019; Sinha et al., 2021; Weaver et al., 2018). Thus, we hypothesized that interest would be equal or higher in the explore-first conditions compared to the instruct-first condition. We did not expect differences between the two explore-first conditions.

Cognitive load

Cognitive load (i.e., mental effort; Paas, 1992) was assessed at the end of the session only, due to an error. Cognitive load is typically examined to gauge whether exploratory learning activities are too cognitively demanding for learners. Prior research suggests that exploratory learning will be less beneficial if the exploration activity is too cognitively demanding (e.g., Ashman et al., 2020; Fyfe et al., 2014; Kapur, 2016; Newman & DeCaro, 2019). This working memory demand depends on a variety of factors, such as capacity and prior knowledge of the learner and complexity of the materials (e.g., Alloway, 2006; Ashman et al., 2020). Thus, we used this measure to help inform our interpretation of the study findings, rather than making a priori hypotheses about this measure. Specifically, this measure was used to ensure that cognitive load was at an appropriate level for our participants. Kalyuga and Singh (2016) argue that exploratory learning activities impact "the intensity of cognitive activity involved in achieving a specific goal of the task" (Kalyuga & Singh, 2016, p. 848). We used a brief "mental effort" measure given in prior research to determine whether cognitive load was affected by instructional order or activity prompt as part of the learning process (Paas, 1992; see also Hsu et al., 2015; Newman & DeCaro, 2019).

Contribution of the study

This research examines whether slight variations in two frequently used exploration prompts may have important impacts on learning processes and outcomes. We expected students to engage in sense-making processes to activate and apply their prior knowledge to the novel problem, and realize gaps in this knowledge, in both explore-first conditions. However, we expected that the specific exploration prompts used would lead students to focus either more narrowly or broadly, impacting the information they explore (e.g., problem features) and the resulting conceptual structures. Thus, this research sheds further light into how conceptual change can be supported by educational activities.

Methods

Participants and design

Participants were undergraduate students ($N = 164$; age $M = 20.54$, $SD = 3.12$; 70.1% women, 25.6% men, 1.3% nonbinary) enrolled in psychology statistics courses at a metropolitan Midwestern U.S. university, with three different instructors of record. Participants completed the study during their regular statistics course lab curriculum. They were not compensated for participation, other than their typical participation credit for attending class. The majority ($n = 132$) of students completed the study online; the rest completed the study in person, based on the modality of their statistics course that semester. Participants were randomly assigned to one of three conditions based on the session in which they participated, in a between-subjects quasi-experimental design: *instruct-first* ($n = 71$; 69 online), *explore-first invent* ($n = 52$; 37 online), or *explore-first generate* ($n = 41$; 27 online). Additional participants were excluded from analyses for not completing the posttest past the first (procedural) essay question ($n = 4$), missing portions of the session (e.g., arriving late or technical issues, $n = 7$), having completed a prior version of the study before ($n = 1$), or not speaking English as their primary language ($n = 9$), due to the large amount of reading material in the study.

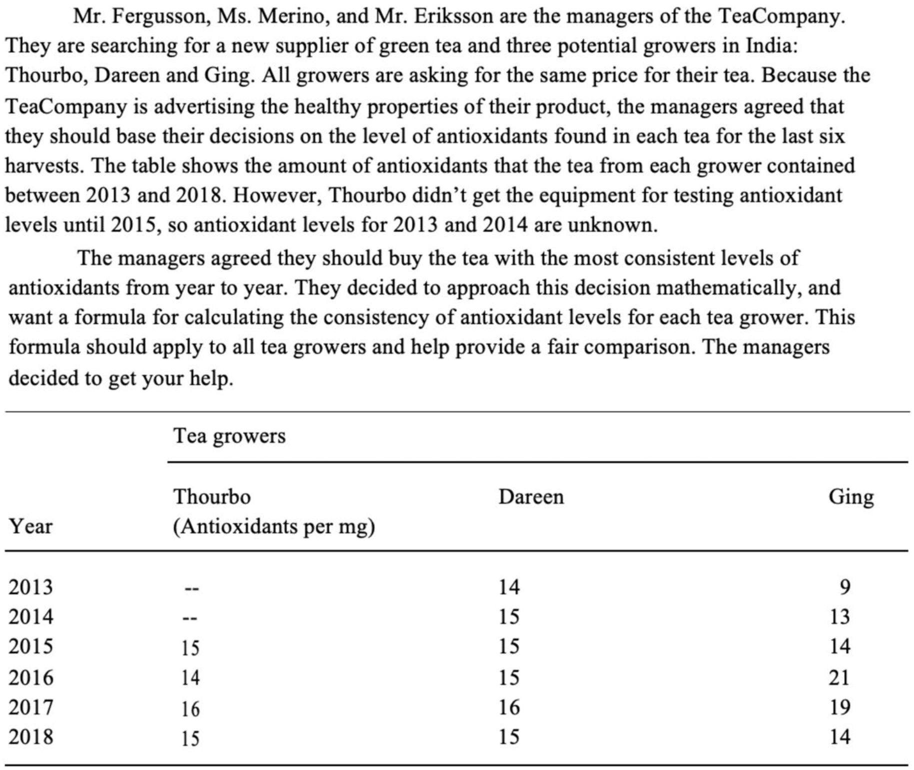
Materials

Instruction

Instruction was provided in a text passage with a worked example explaining procedural and conceptual components to solving standard deviation (Newman & DeCaro, 2019; adapted from Wiedmann et al., 2012). The passage described a problem scenario in which engineers were attempting to determine which of two trampolines has the most consistent levels of bounciness. A table displayed data for inches of rebound for one trampoline, followed by a worked example of computing standard deviation. Explanations and rationales were provided for each step. Then, the second trampoline's data was shown, followed by three practice questions. The first two questions asked students to calculate the standard deviation for the second trampoline, and describe what the value meant. The last question asked students to state which of the two trampolines has the most consistent bounciness.

Problem-solving activity

The problem-solving activity was adapted from Newman and DeCaro (2019) and Wiedmann et al. (2012) (see Fig. 2). Students were given a brief backstory of a tea company that wanted to purchase from a tea grower that had the most consistent levels of antioxidants from year to year. A table was provided of three different tea growers and their antioxidants per mg for each year across five years. In the *instruct-first* condition, students were asked to "use what you have just learned about standard deviation" to determine the most consistent tea-grower (Fig. 2a). In *the explore-first invent* condition, students were asked to "come up with a formula to measure consistency" (Fig. 2b; see Newman & DeCaro, 2019; Schwartz et al., 2011; Wiedmann et al., 2012). In the *explore-first generate* condition, students were asked to "come up with as many different ways to measure consistency as you can" (See Fig. 2c; modified from Kapur, 2014a, 2014b). All students were asked to show their work mathematically.



Mr. Fergusson, Ms. Merino, and Mr. Eriksson are the managers of the TeaCompany. They are searching for a new supplier of green tea and three potential growers in India: Thourbo, Dareen and Ging. All growers are asking for the same price for their tea. Because the TeaCompany is advertising the healthy properties of their product, the managers agreed that they should base their decisions on the level of antioxidants found in each tea for the last six harvests. The table shows the amount of antioxidants that the tea from each grower contained between 2013 and 2018. However, Thourbo didn't get the equipment for testing antioxidant levels until 2015, so antioxidant levels for 2013 and 2014 are unknown.

The managers agreed they should buy the tea with the most consistent levels of antioxidants from year to year. They decided to approach this decision mathematically, and want a formula for calculating the consistency of antioxidant levels for each tea grower. This formula should apply to all tea growers and help provide a fair comparison. The managers decided to get your help.

	Tea growers		
Year	Thourbo (Antioxidants per mg)	Dareen	Ging
2013	--	14	9
2014	--	15	13
2015	15	15	14
2016	14	15	21
2017	16	16	19
2018	15	15	14

Fig. 2 Problem-solving Activity

Activity coding

Problem-solving activities were coded for two components: the quality of the standard deviation formula, and the quantity of attempted solutions in total (e.g., Kapur & Bielaczyc, 2012; Kapur, 2014a, 2014b). For the *quality score*, students received one point for each correct step in the standard deviation formula (for a total of six points). Students could receive up to 18 points if they correctly calculated standard deviation for each of the three tea growers. For the *quantity score*, students' math work was coded based on the number of different strategies they attempted (See Appendix C). Students received one point for each attempted strategy.

Questionnaire

Students were given a questionnaire after each phase of the study (problem solving activity and instruction). After the first phase (instruction for the instruct-first condition and activity for the explore-first conditions), students completed items assessing perceived knowledge gaps (3 items; Cronbach's α = 0.84–0.90; Flynn & Goldsmith, 1999; e.g., "I do not feel very knowledgeable about calculating consistency"), and interest (3 items; α = 0.87–0.89; Ryan, 1982; e.g., "I found this learning activity interesting"; see Newman & DeCaro, 2019). Items for each subscale were intermixed and measured on a 7-point Likert scale (1 = *strongly disagree*; 7 = *strongly agree*). The same questionnaire was again given after the second phase, in addition to a cognitive load item (1 item; Paas, 1992). The cognitive load item was not given after the first phase due to an error. Students were asked to "Please indicate how much mental effort you invested when solving/studying the problem" on a 9-point scale (1 = *very, very low mental effort*; 9 = *very, very high mental effort*). Students additionally completed demographic items and two items asking whether they had prior knowledge of the materials or concept (e.g., "Have you learned about the concept of standard deviation or variance before?").

Posttest

The posttest was designed to primarily assess conceptual knowledge, with one additional procedural knowledge item that included 3.5 possible points (Appendix A). *Procedural knowledge* was evaluated by asking students to execute the correct mathematical sequence to complete a standard deviation problem (Rittle-Johnson & Alibali, 1999). The procedural knowledge item asked students to solve standard deviation for a list of ten numbers. Scores on this item included the number of correct standard deviation steps used out of 6 possible, in the correct order (0.5 points each). An additional half point was given for answers given within 1 point of the correct answer, to allow for rounding or minor computational errors (cf. DeCaro & Rittle-Johnson, 2012).

Because exploratory learning benefits are most commonly found on measures of conceptual understanding, the majority of items assessed this construct (i.e., students' understanding of the underlying principles of standard deviation, or consistency in statistics). *Conceptual knowledge* items included both multiple choice and essay items (α = 0.63; 34 points possible; Appendix A). The five multiple choice questions were adapted from exams given by psychological statistics instructors at our university (2 points each). Essay questions were the same as used by Newman and DeCaro (2019; adapted from Wiedmann et al., 2012; Kapur, 2012). The first question asked students to determine temperature consistency and decide which month an ice hockey tournament should be held. A table was provided listing daily high temperatures for six days for each of two months. Students were instructed to support their decision mathematically. The second item provided the same dataset with a backstory of how one of the values is incorrect, so the decision needed to be revisited. The value was crossed out and the correct value was provided. Students were asked to evaluate whether their previous choice should be changed based on this correction and whether this mistake mattered. Although both items include both problem-solving and conceptual aspects, these components are not independent within each item, and the items primarily require students to provide rationales. Newman and DeCaro (2019) found that these items showed the same pattern of results with each other and with conceptual measures; thus, we included the items on the conceptual subscale. The third item instructed students to explain each component of the standard deviation formula (i.e., x—M, $()^2$, Σ, and $\sqrt{}$) and how it contributes to the concept of standard deviation (see Schwartz & Martin, 2004). Essay responses were scored using the rubric in Appendix B.

Finally, a preparation for *future-learning* (transfer) assessment provided a short passage on standardized scores (z-scores). Students were given brief instruction on calculating and interpreting the concept of standardized scores using a worked example problem about one individual's scores on two athletic activities (adapted from Schwartz & Martin, 2004). Then, students completed two items assessing their understanding of z-scores. First, students were asked to look at a different problem including another individual's scores on two athletic activities, as well as the means and standard deviations within those activities, and determine at which athletic activity the individual performed better. Then, students were given a problem (adapted from Kapur, 2012; Newman & DeCaro, 2019; Wiedmann et al., 2012) asking them to determine which of two students, the top physics or top chemistry student, should receive the best science student award. A table displayed all the scores of top physics and chemistry students for five years, with the mean and standard deviation provided for each subject. This problem is solved by calculating the standardized score of both students and then interpreting and explaining the choice based on the result. Items were scored based on the rubric in Appendix B (15 points possible; Cronbach's $\alpha = 0.77$). A second individual rescored 20% of the posttests; interrater reliability was high, rs = 0.90 to 0.99.

Misconceptions

Participant responses on the problem-solving activity and the posttest were coded for misconceptions, reflecting fundamental misunderstandings about the equation or concept of standard deviation. We added the number of misconceptions given for each participant, using the rubric in Appendix D.

Procedure

Students completed the study as part of their regular statistics course lab section. The session occurred either online or in person, depending on the typical course format. Students were told that the activities would help them learn about concepts relevant to the course material and to try their best, but that their performance on the actual materials would not affect their grade.

Students completed the study over two course sessions, separated by approximately one week. During the first course session, students worked individually to complete a packet including the instruction, problem-solving activity, and questionnaires. Those who completed the study online participated live via the course Learning Management System (*Blackboard Collaborate*). Each section of the packet (i.e., instruction, questionnaires, and activity) was provided through a separate link for that section. The end of each section was marked with a "stop" sign, and students were asked to wait until instructed before continuing. Each section of the packet was timed. During the second session, students completed the posttest and future learning assessment. For both sessions, students working online were asked to work on their own sheet of paper and to upload images of their work. All students were allowed to use a calculator during the study.

Students were randomly assigned to condition based on the lab sessions in which they were enrolled. Students in the *instruct-first* condition worked on the instruction packet (15-min), completed the first questionnaire, worked on the problem-solving activity (18-min), and completed the second questionnaire. Students in the *explore-first* conditions worked on the problem-solving activity (18-min), completed the first questionnaire, worked on the instruction packet (15-min), then completed the second questionnaire. During the second session, students were given 45 min to complete the posttest.

Students participated as part of their regular classroom instruction. Students were informed about the research via an email at the end of the semester and given the option to withdraw their data. All study procedures were approved by the university Institutional Review Board.

Results

Preliminary analyses

Due to the small sample of students taking the class in person, we did not conduct formal analyses of the effect of online or in-person modality. However, in-person and online students showed the same pattern of means across conditions (see Appendix E), with no significant or trending interactions between condition and modality on the posttest scores (procedural, conceptual, future learning), $p\text{s} = 0.507–0.987$. Analyses reported below were collapsed on this factor.

Students' self-reported prior knowledge was low, and no difference in reported prior knowledge was found as a function of condition, $F < 1$, $p = 0.675$, $\eta_p^2 = 0.005$ (instruct-first $M = 1.61$ out of 4, $SE = 0.14$; explore-first invent $M = 1.59$, $SE = 0.17$; explore-first generate $M = 1.79$, $SE = 0.19$). Of the few students (8 total) who reported high prior knowledge, scores remained well below ceiling on the conceptual knowledge scale (range 23.53–94.12%; only two of these students scored above 90%). Thus, our sample did not likely have high prior knowledge of the target concepts.

Learning outcomes

Procedural, conceptual, and future-learning scores are likely correlated (Schneider et al., 2011). However, exploratory learning research treats these outcomes as distinct, because effects tend to be stronger on conceptual and transfer items than procedural. Our hypotheses also differed based on subscale. Thus, separate ANOVAs were used

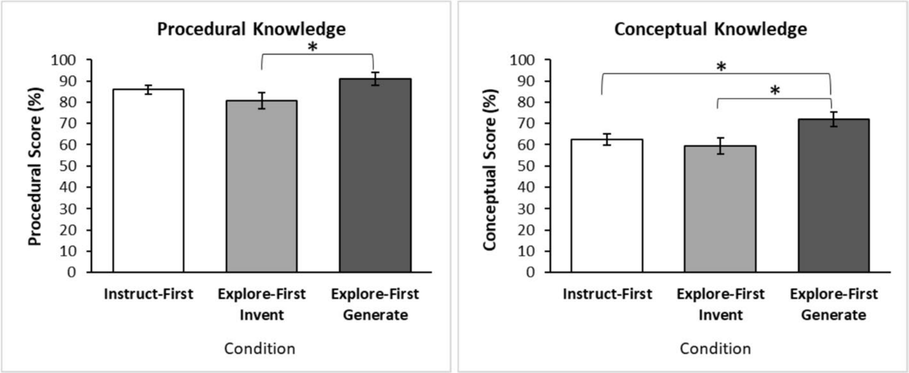


Fig. 3 Mean posttest scores (procedural knowledge, conceptual knowledge) as a function of condition. Error bars represent SEMs

Table 1 Means (standard error in parentheses) of posttest scores and strategy coding as a function of condition

Posttest	Condition		
	Instruct-first	Explore-first invent	Explore-first generate
	Mean (SE)	*Mean (SE)*	*Mean (SE)*
Procedural knowledge (%)	85.92 (2.68)	80.77 (3.13)	90.94 (3.53)
Conceptual knowledge (%)	62.41 (2.83)	59.43 (3.31)	72.08 (3.72)
Transfer (%)	39.09 (3.76)	43.81 (4.39)	53.21 (4.95)
Intervention strategies			
Quality of solution attempts (out of 18)	14.41 (.67)	4.37 (.78)	1.76 (.88)
Quantity of solution attempts	1.18 (.12)	1.71 (.14)	2.95 (.16)
Misconceptions	*Frequency (n)*	*Frequency (n)*	*Frequency (n)*
Activity	32 (71)	7 (52)	5 (41)
Posttest	28 (71)	9 (52)	6 (41)

to examine differences as a function of condition (instruct-first, explore-first generate, explore-first invent) for each posttest subscale (procedural, conceptual, future learning; see Fig. 2). Pairwise differences between conditions were examined using Tukey's LSD.

Procedural knowledge

On the procedural knowledge item, no significant overall effect of condition was found, $F(2,161) = 2.34$, $p = 0.099$, $\eta_p^2 = 0.03$ (Fig. 3, Table 1). Planned comparisons revealed significantly higher procedural scores in the explore-first generate condition compared to the explore-first invent condition, $p = 0.033$, $d = 0.41$. Scores in the instruct-first condition did not differ from scores in either other condition, ps > 0.214, ds = 0.21–0.26.

Conceptual knowledge

On the conceptual knowledge scale, a significant effect of condition was found, $F(2,161)=3.46$, $p=0.034$, $\eta_p^2=0.041$ (Table 1). As shown in Fig. 3, significant differences were found between explore-first generate and instruct-first conditions, $p=0.040$, $d=0.44$, as well as between explore-first generate and explore-first invent conditions, $p=0.012$, $d=0.52$. There were no significant differences between instruct-first and explore-first invent conditions, $p=0.495$, $d=0.11$.

Future learning (transfer)

On the future-learning (transfer) assessment, no overall effect of condition was found, $F(2,161)=2.59$, $p=0.078$, $\eta_p^2=0.03$ (Fig. 4, Table 1). Planned comparisons showed that students in the explore-first generate condition scored significantly higher than students in the instruct-first condition, $p=0.024$, $d=0.43$. There were no significant differences between explore-first invent and instruct-first, $p=0.415$, $d=0.15$, or explore-first generate conditions, $p=0.157$, $d=0.30$.

Problem-solving activity

Quality of solution attempts

For total quality of solution attempt scores on the problem solving activity, students who attempted to calculate standard deviation for the three tea-growers could score up to 18 points (6 points each). There was an overall significant difference between conditions, $F(2,161)=82.08$, $p<0.001$, $\eta_p^2=0.51$ (Fig. 5, Table 1). As expected, students in the instruct-first condition provided higher quality solutions than in the explore-first invent,

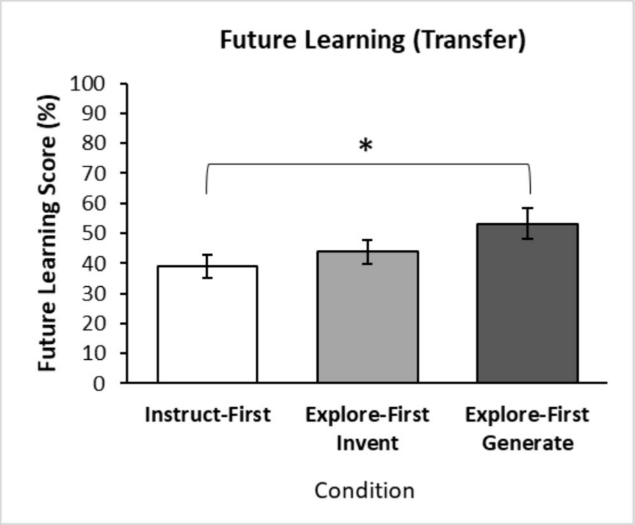
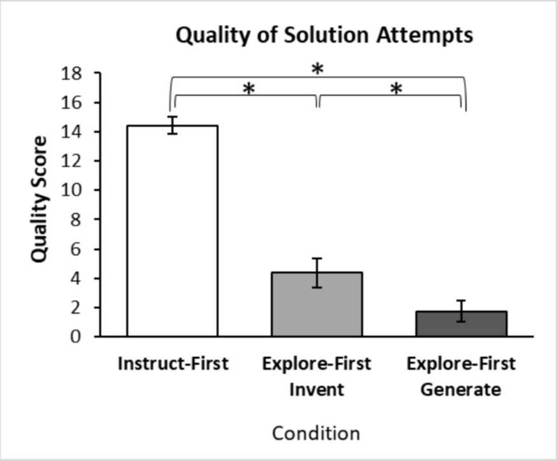


Fig. 4 Mean future-learning scores (transfer) as a function of condition. Error bars represent SEMs

Fig. 5 Quality of solution attempts during the problem-solving activity (Number of correct standard deviation steps out of 6 for each of three datasets). Error bars represent SEMs



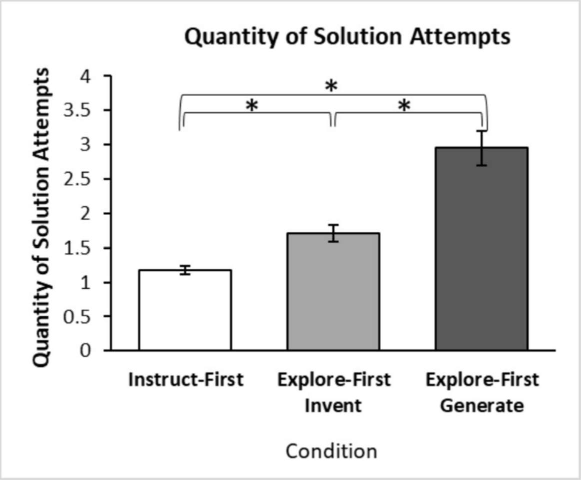
$p < 0.001$, $d = 1.65$., and explore-first generate conditions, $p < 0.001$, $d = 2.63$. Students in the explore-first invent condition used higher quality solutions than in the explore-first generate condition, $p = 0.028$, $d = 0.44$.

The quality of students' standard deviation attempts did not significantly correlate with procedural scores, $r(164) = 0.14$, $p = 0.070$, conceptual scores, $r(164) = 0.08$, $p = 0.343$, or future learning, $r(164) = -0.03$, $p = 0.705$. Regardless of how well students were able to apply or discover the canonical standard deviation formula on the activity, this ability did not translate to greater learning outcomes or preparation for future learning.

Quantity of solution attempts

For quantity of solution attempts, problem-solving activities were coded for the number of different types of solutions attempted by each participant. This coding procedure also served as a manipulation check to ensure students were following the instructions

Fig. 6 Quantity of solution attempts during the problem-solving activity (number of different representations/strategies attempted). Error bars represent SEMs



of using or inventing one solution (i.e., standard deviation method) or generating many solutions. An overall difference in mean solution attempts was found among conditions, $F(2,161) = 39.36$, $p < 0.001$, $\eta_p^2 = 0.33$. As shown in Fig. 6 and Table 1, students in the instruct-first condition attempted fewer solutions than students in the explore-first invent condition, $p = 0.005$, $d = 0.72$, and the explore-first generate condition, $p < 0.001$, $d = 1.46$. Students in the explore-first invent condition attempted fewer solutions than students in the explore-first generate condition, $p < 0.001$, $d = 0.94$.

The quantity of students' solution attempts was significantly correlated with students' conceptual scores, $r(164) = 0.20$, $p = 0.011$. However, the quantity of attempts did not significantly correlate with procedural scores, $r(164) = 0.002$, $p = 0.982$, or future learning, $r(164) = 0.14$, $p = 0.069$. Attempting more solutions was associated with greater conceptual understanding.

Misconceptions

We found an overall difference between conditions on the number of misconceptions shown during the problem-solving activity, $\chi^2(2,164) = 21.24$, $p < 0.001$. In the instruct-first condition, 32 out of 71 students demonstrated misconceptions during the activity. In the explore-first conditions, misconceptions were found for 7 out of 52 students in the invent condition and 5 out of 41 students in the generate condition. This significant pattern of misconceptions extended to the posttest, $\chi^2(2,164) = 11.39$, $p = 0.003$ (see Table 1).

Questionnaire

Phase 1

Descriptive statistics for the questionnaires are shown in Table 2. For the first phase, given after instruction for the instruct-first condition, and after the activity for explore-first conditions, condition impacted students' perceived knowledge gaps, $F(2,160) = 23.34$, $p < 0.001$, $\eta_p^2 = 0.23$. As expected, students in the instruct-first condition reported significantly lower perceived knowledge gaps compared to both the explore-first invent and explore-first

Table 2 Means (standard error in parentheses) of questionnaire items as a function of order of instruction

	Condition		
	Instruct-first	Explore-first invent	Explore-first generate
Phase 1	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
Knowledge gaps	2.66 (.11)	3.76 (.14)	3.63 (.16)
Interest	3.84 (.10)	3.33 (.15)	3.16 (.16)
Phase 2	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
Knowledge gaps	2.36 (.10)	2.64 (.14)	2.38 (.14)
Interest	3.96 (.10)	3.82 (.11)	3.74 (.16)
Cognitive load	5.93 (.22)	5.96 (.26)	5.38 (.33)

Knowledge gaps and interest out of 5 points. Cognitive load out of 9 points generate conditions, $p\text{s} < 0.001$, $d\text{s} = 1.00$–1.14. Students in the two explore-first conditions did not report differences in perceived knowledge gaps, $p = 0.522$, $d = 0.13$.

An effect of condition was also found for reported interest in the first phase, $F(2,160) = 7.51$, $p < 0.001$, $\eta_p^2 = 0.09$. Students in the instruct-first condition reported higher interest compared to both the explore-first invent, $p = 0.005$, $d = 0.52$, and explore-first generate conditions, $p < 0.001$, $d = 0.72$. No significant differences were found between the explore-first conditions, $p = 0.250$, $d = 0.16$.

Phase 2

For the second phase, occurring after the activity for instruct-first condition and after the instruction for the explore-first conditions, these differences went away for both perceived knowledge gaps, $F(2,157) = 1.59$, $p = 0.207$, $\eta_p^2 = 0.02$, and interest, $F < 1$, $p = 0.407$, $\eta_p^2 = 0.01$. Thus, any gaps students perceived in their knowledge, or reduced interest, were equalized between conditions after the instruction and activity were both completed. Additionally, there were no significant differences between conditions in cognitive load (mental effort), $F(2,157) = 1.28$, $p = 0.281$, $\eta_p^2 = 0.02$.

Discussion

We predicted that two commonly-used exploratory learning prompts could lead to different forms of conceptual engagement prior to instruction. The two prompts originated from the *inventing to prepare for future learning* (e.g., Schwartz & Martin, 2004; Schwartz et al., 2011) and *productive failure* (e.g., Kapur, 2014a, 2014b, 2015) literatures. Both literatures posit that exploration activities help students create knowledge structures that deepen their understanding of subsequent instruction (e.g., Schwartz & Bransford, 1998; Schwartz et al., 2007). However, typical studies from each literature focus on different exploration processes. The invention literature highlights that invention prompts help students begin to reconcile disparate problem features (e.g., distance from the mean and sample size in statistics, or mass and velocity in physics; Schwartz et al., 2007, 2011). Research using strategy generation prompts tends to focus on how exploring multiple solution methods and representations helps students differentiate their knowledge (Kapur, 2014a, 2014b, 2015).

Drawing from problem solving research (e.g., DeCaro et al., 2017; Knoblich et al., 1999; Thevenot & Oakhill, 2008), we reasoned that students' conceptual structures might take different forms, depending on the search processes prompted during exploration. We found evidence that invention prompts constrained and deepened students' exploration, whereas strategy generation prompts broadened exploration. Broader—rather than deeper—exploration was associated with greater conceptual understanding and preparation for future learning.

Learning processes

We coded the quality and quantity of strategies students used during the activity, to determine (a) how successful students were at discovering or applying the canonical standard deviation formula, and (b) how broadly they explored the problem space. Unsurprisingly, students in the instruct-first condition, who had just learned the formula, were much more successful in using the formula on the problem solving activity (*quality* scores), with large effect sizes.

Students in the explore-first invent condition derived significantly more of the problem-solving steps than students in the explore-first generate condition (quality scores, small-to-medium effect size). However, invention resulted in fewer total attempted solutions (quantity scores, medium-to-large effect size). These results suggest that invention instructions prompted students to focus and build on a smaller number of solution approaches. Strategy generation instructions prompted a wider search of problem features, but limited how many canonical solution steps students discovered.

Importantly, accuracy (quality) during the problem solving activity was not associated with accuracy on the posttest or future-learning assessment. Instead, the quantity of solutions correlated with conceptual knowledge scores on the posttest. This finding suggests that prompting students to try out multiple approaches helps them to differentiate important problem features, supporting conceptual understanding. Quantity did not significantly correlate with future-learning scores, however, though results trended in that direction. There may be other necessary factors predicting future learning beyond simply generating multiple solutions that were not measured in this study.

Learning outcomes

Even though exploring before instruction increased errors and solution paths compared to an instruct-first approach (cf. Kirschner et al., 2007; Sweller & Chandler, 1994), students' procedural knowledge was unharmed. However, generating solution methods improved procedural knowledge compared to invention.

Exploratory learning is designed to target students' sense-making and development of conceptual schemas (Loibl & Rummel, 2015; Schwartz et al., 2007). We assessed both conceptual knowledge and preparation for future learning. Approximately one week after completing the standard deviation materials, students were given a passage on a new, related topic (standardized z-scores), then asked to complete two problems assessing their learning of the new topic. Consistent with prior research (see Loibl et al., 2017), we found exploratory learning benefits on these measures—but only when students were given strategy generation prompts during the exploration activity. Students in the explore-first generate condition scored significantly higher than in both other conditions on conceptual knowledge, and significantly higher than the instruct-first condition on future learning, with small-to-medium effect sizes. Students in the explore-first invent condition did not score differently than students in the instruct-first condition on either measure.

The null result for invention prompts does not replicate prior studies (e.g., Schwartz & Martin, 2004; Schwartz et al., 2011). Our study differs from prior work using invention prompts in several ways. For example, studies have been typically done with a younger population, do not usually control for the learning materials given between conditions, and are conducted over a longer time span than the current study. Our results suggest that the benefits of strategy generation are stronger than a simple invention prompt, at least with undergraduate students completing a short exploration activity.

These results are consistent with prior studies showing benefits of exploring with strategy generation prompts over an instruct-first condition (e.g., Kapur, 2014a, 2014b, 2015; Trninic et al., 2022). We extend this work by demonstrating that use of strategy generation prompts enhances students' preparation for future learning—more than with invention prompts typically used with future-learning assessments.

Together with the quality and quantity coding, these results suggest that exploring more broadly during exploration better supports students' conceptual understanding and transfer to new learning situations. The strategy generation prompt still requires students to invent, but also pushes students to move beyond the first few solutions that come to mind, likely helping them to encode important problem features (see Gilhooly et al., 2007). Consistent with the productive failure principle, student success at solving the problem appears to be less important than the discovery process itself.

Metacognitive and motivational processes

Both inventing and generating strategies resulted in greater awareness of gaps in students' prior knowledge, compared to the instruct-first condition (large effect size). This pattern replicates prior findings (e.g., Glogger-Frey et al., 2015; Loibl & Rummel, 2014a; Newman & DeCaro, 2019). Any gaps in knowledge perceived after the activity were resolved after instruction.

Metacognitive awareness of knowledge gaps cannot explain differences in learning outcomes between conditions. Students in the explore-first invent condition showed similar knowledge gaps after exploring as in the generate condition, but lower conceptual knowledge scores. Coupled with the strategy coding results, these findings suggest that better differentiating the problem space is required for the conceptual benefits of exploring, likely in addition to increasing metacognitive awareness of knowledge gaps (see Loibl & Rummel, 2014a; Newman & DeCaro, 2019).

Students in the instruct-first condition reported higher interest than students in the explore-first conditions after the first phase (medium-to-large effect size). In the explore-first conditions, reported interest was right at the neutral midpoint of the scale. This result is inconsistent with prior suggestions that interest increases when a task is novel, somewhat complex, and requires personal direction (Rotgans & Schmidt, 2014; Silvia, 2008). This result is also surprising in light of prior exploratory learning studies, which have found that interest is equal or higher in explore-first compared to instruct-first conditions (e.g., Glogger-Frey et al., 2015; Weaver et al., 2018). However, this finding is not particularly surprising considering that students who explored first were more likely to experience failure during the activity.

Taken together with the mixed findings in the literature, this result suggests that interest is not likely driving the learning results. Other affective responses have been measured in the literature that might better explain students' reactions than interest, such as curiosity (Lamnina & Chase, 2019; Loibl & Rummel, 2014a) or surprise (Sinha, 2022). Interest did equalize by the end of the second phase, suggesting that the experience of failure did not dampen interest in the long run (see Hidi & Harackiewicz, 2000).

Cognitive load

Exploratory learning activities are unlikely to benefit learning if they are not matched to learners' abilities (DeCaro et al., 2024; Kapur, 2016). We found no differences in reported cognitive load between conditions, measured once when both phases were complete. Ratings were right around the mid-point of the scale, suggesting that learners did not perceive the exploration experience as too simple or taxing, regardless of prompt type.

Misconceptions

During exploratory learning, students could potentially activate and hold onto misconceptions about the topic (Kirschner et al., 2007). We found the opposite. Students in the instruct-first condition showed significantly greater misconceptions about standard deviation than students in the explore-first conditions, on both the problem solving activity and the posttest. For example, some students divided by the mean (*M*) instead of the sample size (*N*), indicating that they did not understand the purpose of dividing by the number of scores in a sample. Similarly, some students divided by the same sample size for all datasets, even when the datasets did not have equal numbers of scores. Thus, students in the instruct-first condition appeared to be rotely applying formula steps, without understanding the meaning behind the process. For example, they knew they needed to divide, but they did not always know what or why they should divide. This finding directly supports arguments that students given traditional instruction are more likely to passively process the content and develop superficial understanding (Glogger-Frey et al., 2015; Kapur, 2010; Renkl, 1999; Schwartz et al., 2007; Wittwer & Renkl, 2008).

The strategies students use indicate the prior knowledge that was activated (Kapur, 2016). Students in both explore-first conditions showed few misconceptions during the problem solving activity, suggesting that they did not have many misconceptions prior to the learning session. It seems likely that the misconceptions arose due to superficial processing during instruction, but only for students who got instruction first. These findings add evidence that exploratory learning helps to lessen the pitfalls of direct instruction. The number of misconceptions did not differ based on prompt condition, suggesting that both types of exploration helped to reduce misconceptions.

Limitations and future research

By bringing together commonly-used exploratory learning prompts into one study, we are able to determine how these methods impact learning processes and outcomes. This design allows us to begin to connect principles across these prior studies (see Koedinger et al., 2012; Loibl et al., 2024). More work is needed to determine if strategy generation versus invention prompts evoke different processes and learning outcomes with other types of materials, learners, or settings.

Materials

Because our primary focus was on conceptual understanding, we used only one item to assess procedural knowledge. This design choice potentially limited our ability to detect differences on this measure, although we did find some. Our measure of conceptual understanding also was somewhat contaminated by procedural elements, given that students needed to both compute and reason about standard deviation in two of the essay items. Procedural and conceptual understanding develop iteratively, and are correlated with each other (Schneider et al., 2011). Thus, the inclusion of some computational elements was unlikely to limit the conclusions made from this measure. If anything, including both elements should only weaken our ability to find significant results.

However, future research should use a purer measure of conceptual understanding, such as asking students to interpret an already worked-out problem.

Setting

Most exploratory learning studies have been conducted with students during in-person classroom or laboratory settings. Because of the existing course structure at the time of our research, the majority of our students completed the study in an online learning session conducted via live videoconference. A few prior studies have found beneficial effects of exploration in online settings (DeCaro et al., 2023; Hieb et al., 2021; see also Song & Kapur, 2017, for a flipped classroom study in which the instruction was completed online). We did not have a large enough sample to treat learning setting as a between-subjects factor. However, preliminary analyses suggested that the pattern of results was the same between these settings. More research is needed to confirm this lack of difference, and fully compare online and in-person exploratory learning within one study.

Prompts

The primary difference between invention and strategy generation prompts was the goal to come up with one versus multiple solutions. However, another difference between the prompts was the use of the word “formula” in the invention prompt. Students were instructed to “come up with a formula to measure consistency.” In the strategy generation prompt, students were instructed to “come up with as many different ways to measure consistency as you can.” The word “formula” implies that a mathematical calculation is required. Mathematics provides a formal method for students to explicitly connect seemingly incommensurate factors, such as distance from the mean and sample size (Schwartz et al., 2007). However, by focusing on a mathematical solution, the invention prompt may have constricted the search space even more. Students were not only limited in how many solutions but also what type of solution they should search for. The prompts used were taken verbatim from prior research, providing a genuine test of prompts used in prior work. However, more research is needed to know whether it was the restriction to use just one method, or to use a mathematical one, that limited scores in the invention condition.

Additional research is also needed to test whether invention instructions might be more beneficial than strategy generation prompts in certain contexts. For example, there may be circumstances in which the extra search processes of strategy generation might unduly increase learners’ cognitive load (Kirschner et al., 2007). Specifically, strategy generation could be taxing for younger learners (who have lower working memory capacity; Alloway, 2006), learners with too little prior knowledge, or with more complex learning materials (Ashman et al., 2020). In these situations, invention instructions could provide exploration benefits while helping to constrain the problem space. Invention prompts might be more beneficial in some contexts, such as when the cognitive load induced by strategy generation might be too high.

Strategy generation prompts might also be less useful with less guidance, such as without contrasting cases in the problem. Without contrasting cases, students may encounter more unimportant problem features, adding too much extraneous cognitive load.

Cognitive load

Cognitive load was assessed in the current study, because prior work suggests that load may be a boundary condition to the benefits of exploration (e.g., Ashman et al., 2020; Fyfe et al., 2014; Newman & DeCaro, 2019). We found that, at the end of the learning session, students reported equal cognitive load between conditions, around the midpoint of the scale. This finding suggests that cognitive load was not likely an issue in our study. However, due to an error, we did not measure cognitive load after the first learning phase (i.e., after the instruction in the instruct-first condition, or after the activity in the explore-first conditions). It is possible that cognitive load would have been higher in the explore-first conditions after struggle with the activity, and that such load would have also been higher than after students in the instruct-first condition completed the activity. Future research is needed to determine how cognitive load is impacted at the intermediate phase of learning.

Misconceptions

Students' prior misconceptions about consistency were relatively infrequent in this study. Instead, misconceptions arose from instruction. However, students do often carry misconceptions with them into a learning session, in statistics and other domains (e.g., McNeil, 2014; Schwartz et al., 2007). More work is needed to determine if the misconception results generalize to domains in which students hold strong prior misconceptions.

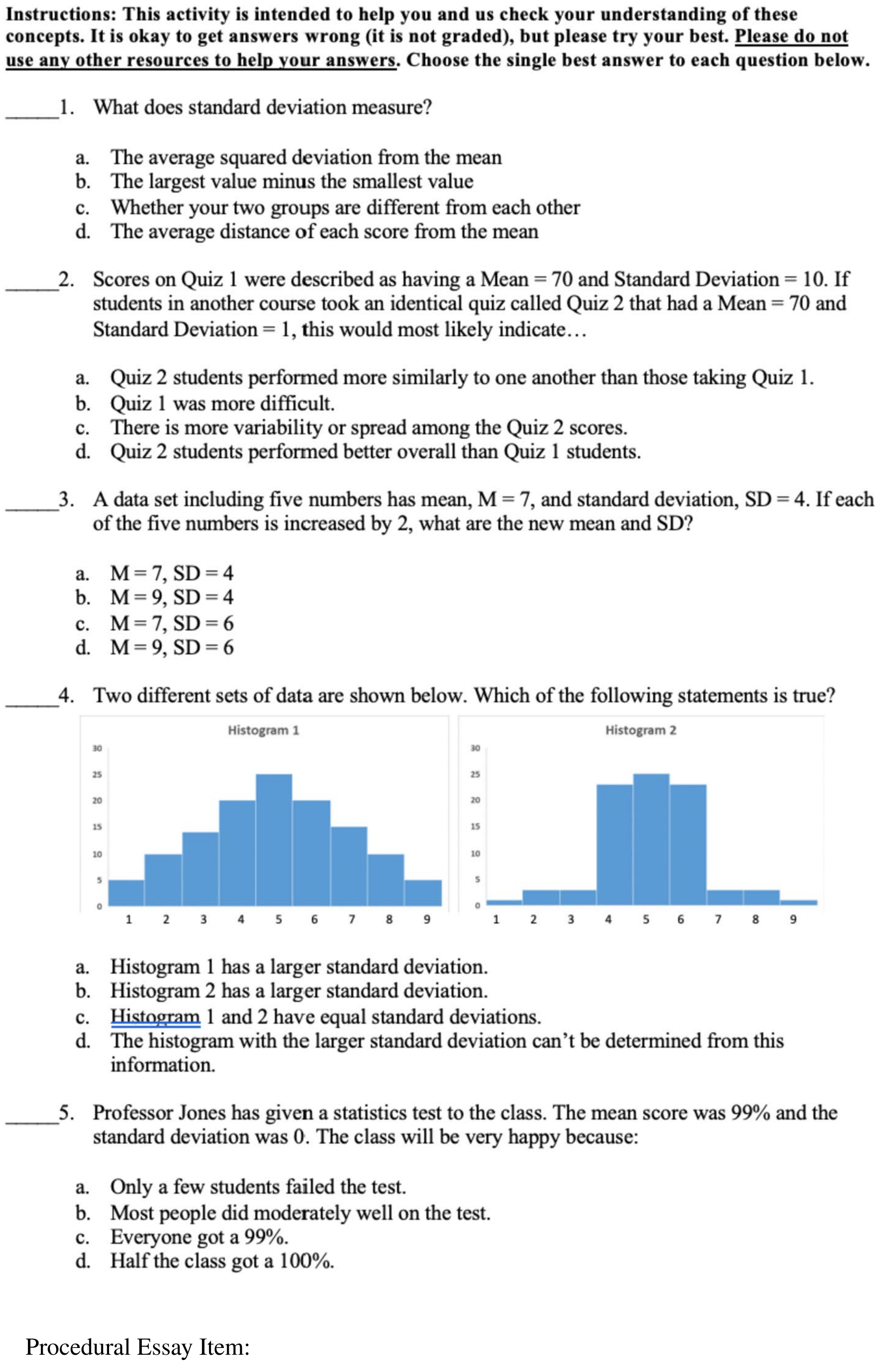
Conclusion

Exploratory learning before instruction is a promising instructional method if the learning goal is to deepen conceptual understanding and transfer. While exploring, students engage in sense-making processes. These processes appear to help them avoid developing misconceptions that arise from the more surface-level processing that may otherwise occur in traditional instruct-then-practice settings.

However, not all exploratory learning conditions enhance learning, including the explore-first invent condition in the current study. One way to help determine whether an instructional design will improve learning is to assess the learning processes during or after the exploration phase, and how these processes connect with different types of learning outcomes (Koedinger et al., 2012; Loibl et al., 2023, 2024). Prompts to generate multiple strategies during exploration appear to facilitate sense-making by helping students consider more problem features. This process appears to be key to exploratory learning benefits. Thus, during exploration, students may benefit more from instructions to widen their discovery processes.

Appendix A

Multiple Choice Items (Conceptual):



Instructions: This activity is intended to help you and us check your understanding of these concepts. It is okay to get answers wrong (it is not graded), but please try your best. Please do not use any other resources to help your answers. Choose the single best answer to each question below.

_____1. What does standard deviation measure?

a. The average squared deviation from the mean
b. The largest value minus the smallest value
c. Whether your two groups are different from each other
d. The average distance of each score from the mean

_____2. Scores on Quiz 1 were described as having a Mean = 70 and Standard Deviation = 10. If students in another course took an identical quiz called Quiz 2 that had a Mean = 70 and Standard Deviation = 1, this would most likely indicate…

a. Quiz 2 students performed more similarly to one another than those taking Quiz 1.
b. Quiz 1 was more difficult.
c. There is more variability or spread among the Quiz 2 scores.
d. Quiz 2 students performed better overall than Quiz 1 students.

_____3. A data set including five numbers has mean, M = 7, and standard deviation, SD = 4. If each of the five numbers is increased by 2, what are the new mean and SD?

a. M = 7, SD = 4
b. M = 9, SD = 4
c. M = 7, SD = 6
d. M = 9, SD = 6

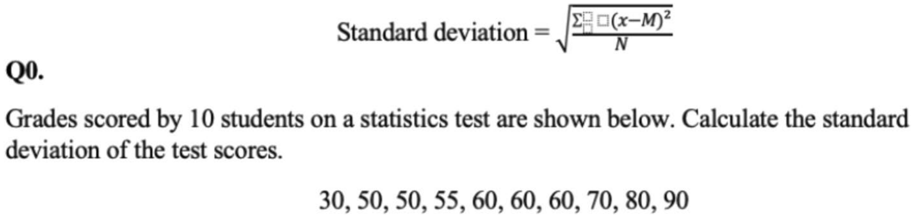
_____4. Two different sets of data are shown below. Which of the following statements is true?

a. Histogram 1 has a larger standard deviation.
b. Histogram 2 has a larger standard deviation.
c. Histogram 1 and 2 have equal standard deviations.
d. The histogram with the larger standard deviation can't be determined from this information.

_____5. Professor Jones has given a statistics test to the class. The mean score was 99% and the standard deviation was 0. The class will be very happy because:

a. Only a few students failed the test.
b. Most people did moderately well on the test.
c. Everyone got a 99%.
d. Half the class got a 100%.

Procedural Essay Item:



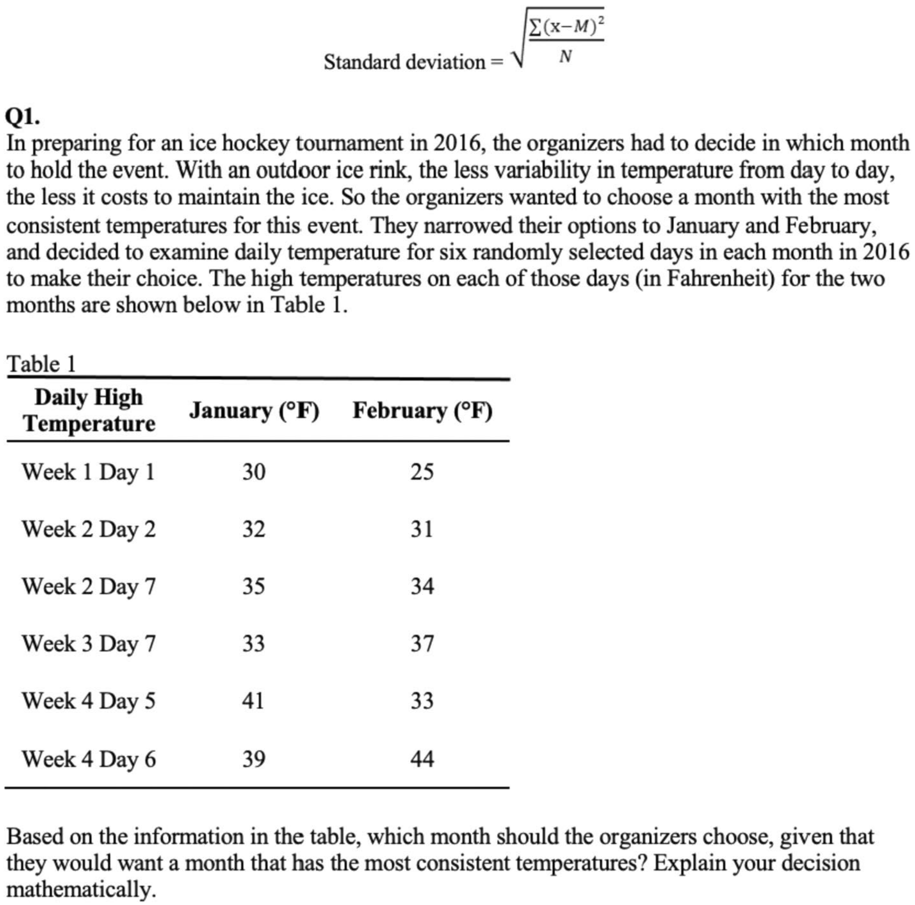
$$\text{Standard deviation} = \sqrt{\frac{\Sigma_{\square}^{\square}\square(x-M)^2}{N}}$$

Q0.

Grades scored by 10 students on a statistics test are shown below. Calculate the standard deviation of the test scores.

30, 50, 50, 55, 60, 60, 60, 70, 80, 90

Conceptual Essay Items:



$$\text{Standard deviation} = \sqrt{\frac{\Sigma(\mathrm{x}-M)^2}{N}}$$

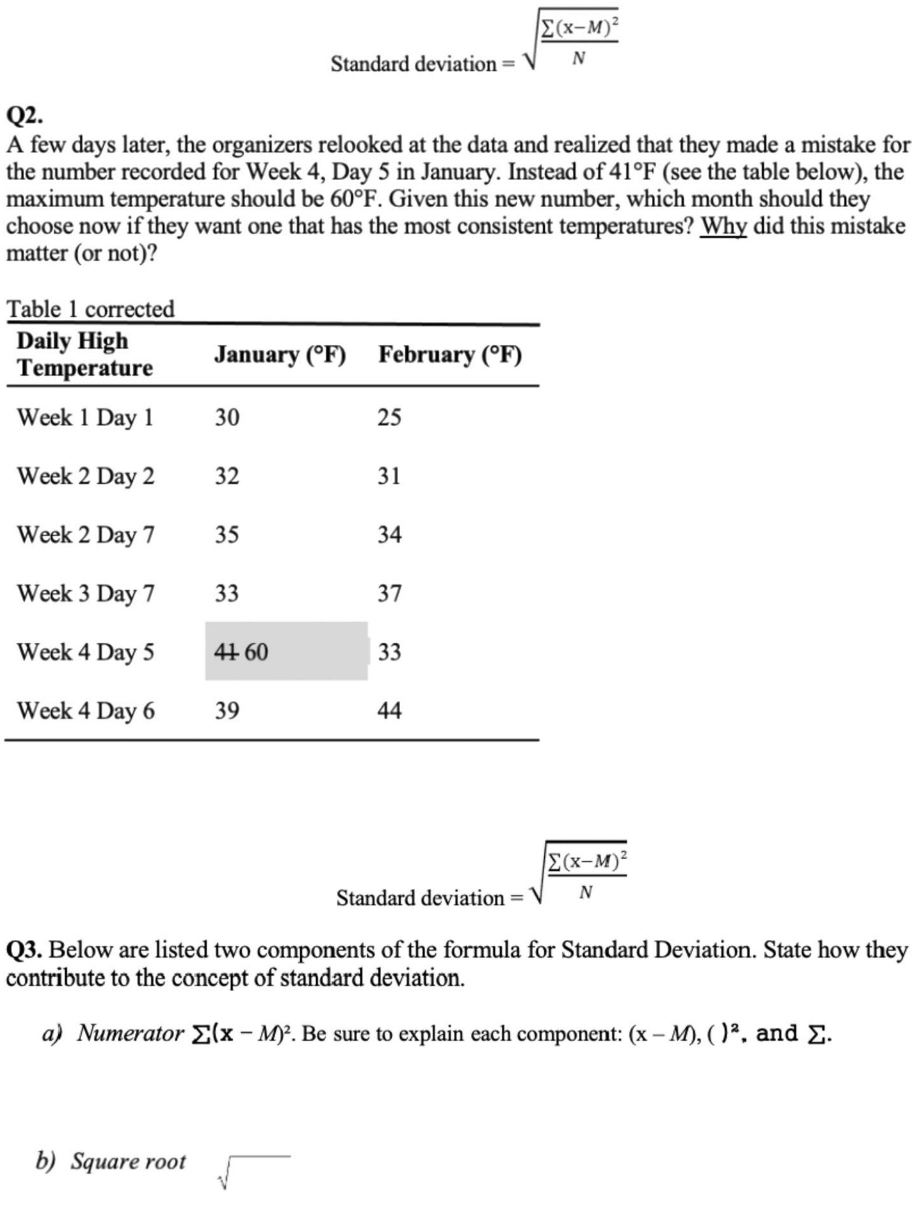
Q1.

In preparing for an ice hockey tournament in 2016, the organizers had to decide in which month to hold the event. With an outdoor ice rink, the less variability in temperature from day to day, the less it costs to maintain the ice. So the organizers wanted to choose a month with the most consistent temperatures for this event. They narrowed their options to January and February, and decided to examine daily temperature for six randomly selected days in each month in 2016 to make their choice. The high temperatures on each of those days (in Fahrenheit) for the two months are shown below in Table 1.

Table 1

Daily High Temperature	**January (°F)**	**February (°F)**
Week 1 Day 1	30	25
Week 2 Day 2	32	31
Week 2 Day 7	35	34
Week 3 Day 7	33	37
Week 4 Day 5	41	33
Week 4 Day 6	39	44

Based on the information in the table, which month should the organizers choose, given that they would want a month that has the most consistent temperatures? Explain your decision mathematically.



$$\text{Standard deviation} = \sqrt{\frac{\sum(x-M)^2}{N}}$$

Q2.
A few days later, the organizers relooked at the data and realized that they made a mistake for the number recorded for Week 4, Day 5 in January. Instead of 41°F (see the table below), the maximum temperature should be 60°F. Given this new number, which month should they choose now if they want one that has the most consistent temperatures? Why did this mistake matter (or not)?

Table 1 corrected

Daily High Temperature	**January (°F)**	**February (°F)**
Week 1 Day 1	30	25
Week 2 Day 2	32	31
Week 2 Day 7	35	34
Week 3 Day 7	33	37
Week 4 Day 5	~~41~~ 60	33
Week 4 Day 6	39	44

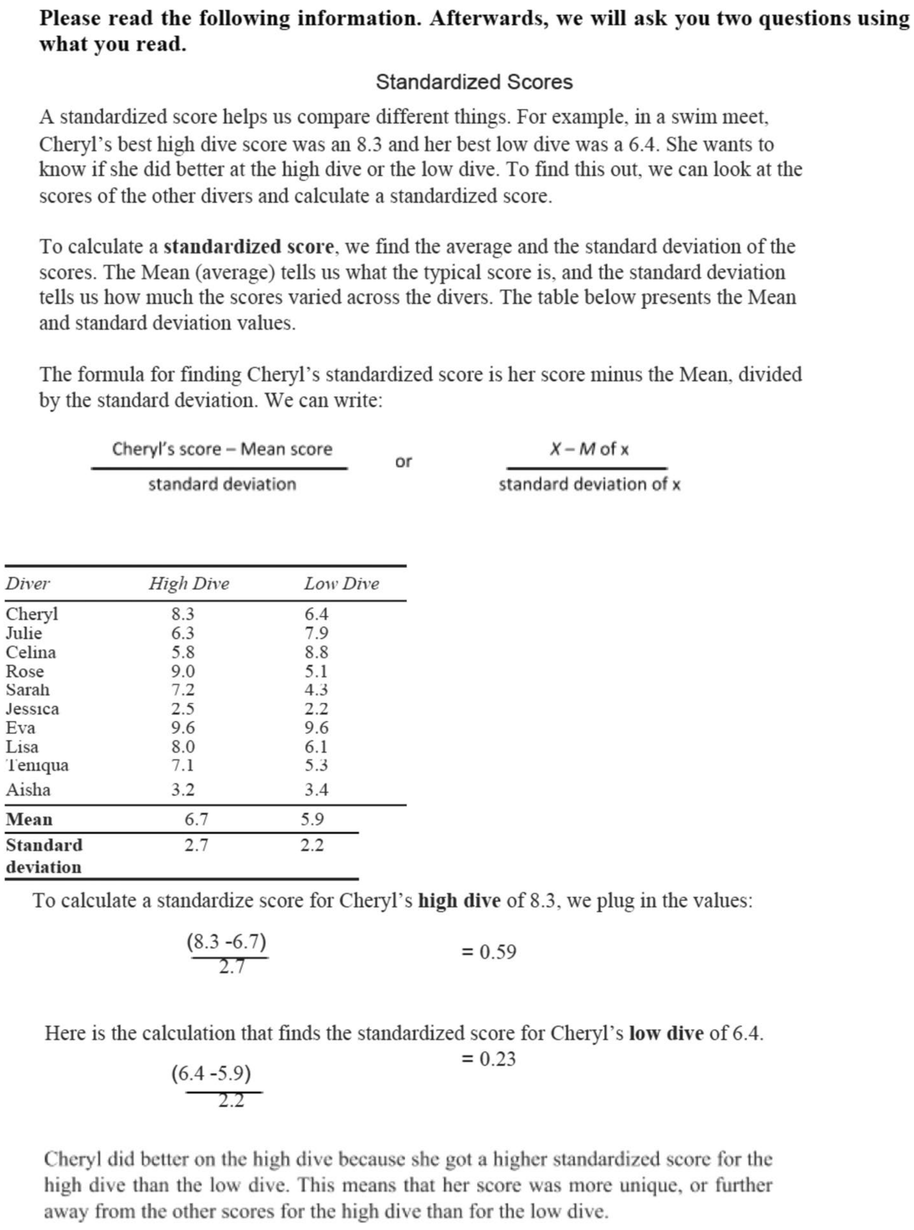
$$\text{Standard deviation} = \sqrt{\frac{\sum(x-M)^2}{N}}$$

Q3. Below are listed two components of the formula for Standard Deviation. State how they contribute to the concept of standard deviation.

a) *Numerator* $\sum(x - M)^2$. Be sure to explain each component: $(x - M)$, $(\ )^2$, and $\sum$.

b) *Square root* $\sqrt{}$

Z-Score Instruction:



Please read the following information. Afterwards, we will ask you two questions using what you read.

Standardized Scores

A standardized score helps us compare different things. For example, in a swim meet, Cheryl's best high dive score was an 8.3 and her best low dive was a 6.4. She wants to know if she did better at the high dive or the low dive. To find this out, we can look at the scores of the other divers and calculate a standardized score.

To calculate a **standardized score**, we find the average and the standard deviation of the scores. The Mean (average) tells us what the typical score is, and the standard deviation tells us how much the scores varied across the divers. The table below presents the Mean and standard deviation values.

The formula for finding Cheryl's standardized score is her score minus the Mean, divided by the standard deviation. We can write:

$$\frac{\text{Cheryl's score} - \text{Mean score}}{\text{standard deviation}} \quad \text{or} \quad \frac{X - M \text{ of x}}{\text{standard deviation of x}}$$

Diver	*High Dive*	*Low Dive*
Cheryl	8.3	6.4
Julie	6.3	7.9
Celina	5.8	8.8
Rose	9.0	5.1
Sarah	7.2	4.3
Jessica	2.5	2.2
Eva	9.6	9.6
Lisa	8.0	6.1
Teniqua	7.1	5.3
Aisha	3.2	3.4
Mean	6.7	5.9
Standard deviation	2.7	2.2

To calculate a standardize score for Cheryl's **high dive** of 8.3, we plug in the values:

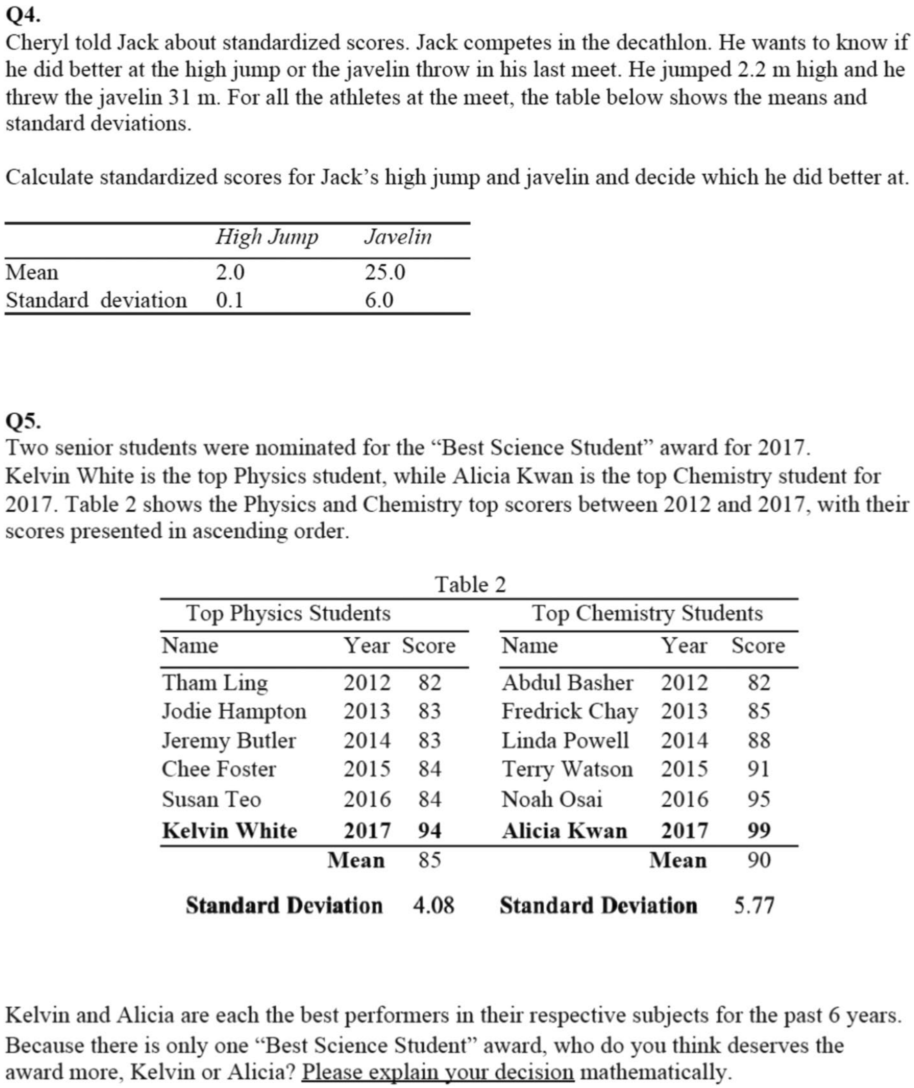
$$\frac{(8.3 \text{ -}6.7)}{2.7} \qquad = 0.59$$

Here is the calculation that finds the standardized score for Cheryl's **low dive** of 6.4.

$$\frac{(6.4 \text{ -}5.9)}{2.2} \qquad = 0.23$$

Cheryl did better on the high dive because she got a higher standardized score for the high dive than the low dive. This means that her score was more unique, or further away from the other scores for the high dive than for the low dive.

Future-Learning Assessment (z-scores):



Q4.

Cheryl told Jack about standardized scores. Jack competes in the decathlon. He wants to know if he did better at the high jump or the javelin throw in his last meet. He jumped 2.2 m high and he threw the javelin 31 m. For all the athletes at the meet, the table below shows the means and standard deviations.

Calculate standardized scores for Jack's high jump and javelin and decide which he did better at.

	High Jump	*Javelin*
Mean	2.0	25.0
Standard deviation	0.1	6.0

Q5.

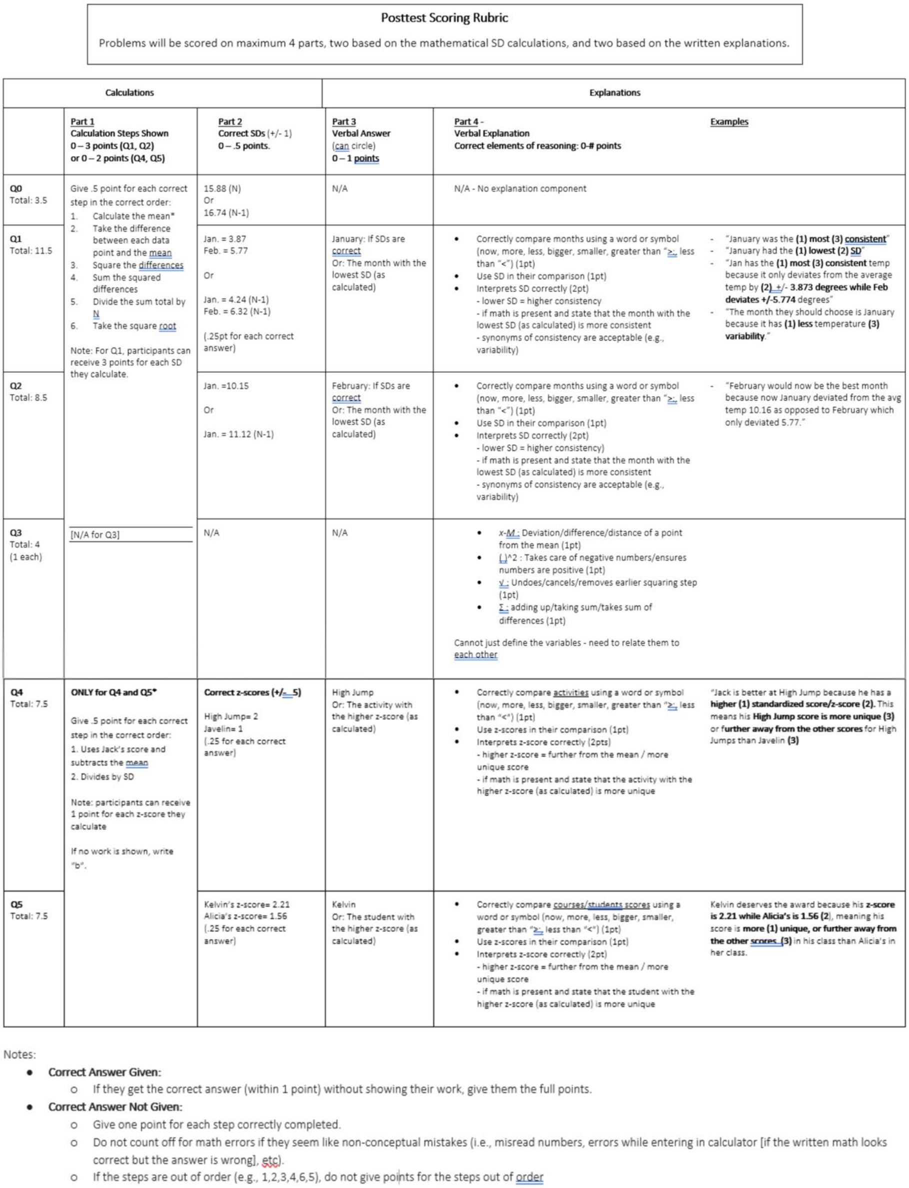
Two senior students were nominated for the "Best Science Student" award for 2017. Kelvin White is the top Physics student, while Alicia Kwan is the top Chemistry student for 2017. Table 2 shows the Physics and Chemistry top scorers between 2012 and 2017, with their scores presented in ascending order.

Table 2

Top Physics Students			Top Chemistry Students		
Name	Year	Score	Name	Year	Score
Tham Ling	2012	82	Abdul Basher	2012	82
Jodie Hampton	2013	83	Fredrick Chay	2013	85
Jeremy Butler	2014	83	Linda Powell	2014	88
Chee Foster	2015	84	Terry Watson	2015	91
Susan Teo	2016	84	Noah Osai	2016	95
Kelvin White	**2017**	**94**	**Alicia Kwan**	**2017**	**99**
	Mean	85		**Mean**	90
Standard Deviation		4.08	**Standard Deviation**		5.77

Kelvin and Alicia are each the best performers in their respective subjects for the past 6 years. Because there is only one "Best Science Student" award, who do you think deserves the award more, Kelvin or Alicia? Please explain your decision mathematically.

Appendix B



Posttest Scoring Rubric

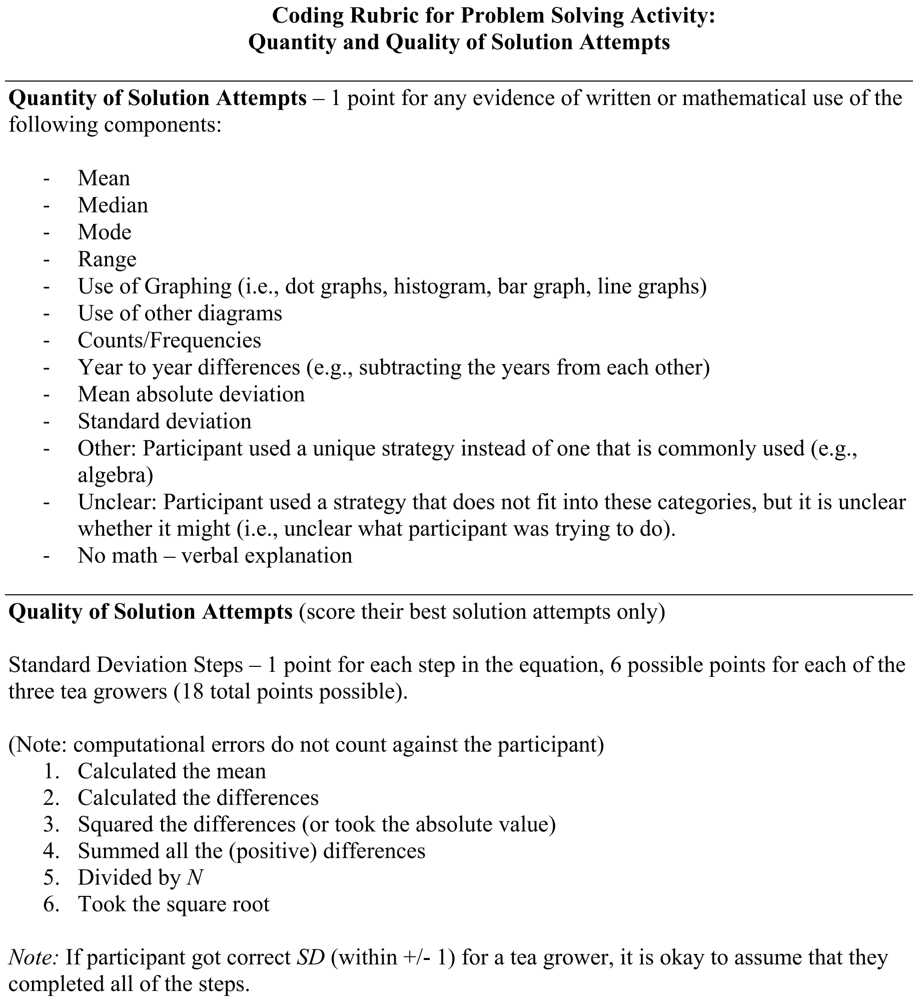
Problems will be scored on maximum 4 parts, two based on the mathematical SD calculations, and two based on the written explanations.

Calculations			Explanations		
	Part 1 **Calculation Steps Shown** **0 – 3 points (Q1, Q2)** **or 0 – 2 points (Q4, Q5)**	**Part 2** **Correct SDs** (+/- 1) **0 – .5 points.**	**Part 3** **Verbal Answer** (can circle) **0 – 1 points**	**Part 4 -** **Verbal Explanation** **Correct elements of reasoning: 0-# points**	**Examples**
Q0 Total: 3.5	Give .5 point for each correct step in the correct order: 1. Calculate the mean* 2. Take the difference between each data point and the mean 3. Square the differences 4. Sum the squared differences 5. Divide the sum total by N 6. Take the square root Note: For Q1, participants can receive 3 points for each SD they calculate.	15.88 (N) Or 16.74 (N-1)	N/A	N/A - No explanation component	
Q1 Total: 11.5		Jan. = 3.87 Feb. = 5.77 Or Jan. = 4.24 (N-1) Feb. = 6.32 (N-1) (.25pt for each correct answer)	January: If SDs are correct Or: The month with the lowest SD (as calculated)	• Correctly compare months using a word or symbol (now, more, less, bigger, smaller, greater than "≥:, less than "<") (1pt) • Use SD in their comparison (1pt) • Interprets SD correctly (2pt) - lower SD = higher consistency - if math is present and state that the month with the lowest SD (as calculated) is more consistent - synonyms of consistency are acceptable (e.g., variability)	- "January was the **(1) most (3) consistent**" - "January had the **(1) lowest (2) SD**" - "Jan has the **(1) most (3) consistent** temp because it only deviates from the average temp by **(2) +/- 3.873 degrees while Feb deviates +/-5.774** degrees" - "The month they should choose is January because it has **(1) less** temperature **(3) variability**."
Q2 Total: 8.5		Jan. =10.15 Or Jan. = 11.12 (N-1)	February: If SDs are correct Or: The month with the lowest SD (as calculated)	• Correctly compare months using a word or symbol (now, more, less, bigger, smaller, greater than "≥:, less than "<") (1pt) • Use SD in their comparison (1pt) • Interprets SD correctly (2pt) - lower SD = higher consistency) - if math is present and state that the month with the lowest SD (as calculated) is more consistent - synonyms of consistency are acceptable (e.g., variability)	- "February would now be the best month because now January deviated from the avg temp 10.16 as opposed to February which only deviated 5.77."
Q3 Total: 4 (1 each)	[N/A for Q3]	N/A	N/A	• *x-M* : Deviation/difference/distance of a point from the mean (1pt) • ( )^2 : Takes care of negative numbers/ensures numbers are positive (1pt) • √ : Undoes/cancels/removes earlier squaring step (1pt) • Σ : adding up/taking sum/takes sum of differences (1pt) Cannot just define the variables - need to relate them to each other	
Q4 Total: 7.5	**ONLY for Q4 and Q5*** Give .5 point for each correct step in the correct order: 1. Uses Jack's score and subtracts the mean 2. Divides by SD Note: participants can receive 1 point for each z-score they calculate If no work is shown, write "b".	**Correct z-scores (+/- .5)** High Jump= 2 Javelin= 1 (.25 for each correct answer)	High Jump Or: The activity with the higher z-score (as calculated)	• Correctly compare activities using a word or symbol (now, more, less, bigger, smaller, greater than "≥:, less than "<") (1pt) • Use z-scores in their comparison (1pt) • Interprets z-score correctly (2pts) - higher z-score = further from the mean / more unique score - if math is present and state that the activity with the higher z-score (as calculated) is more unique	"Jack is better at High Jump because he has a **higher (1) standardized score/z-score (2).** This means his **High Jump score is more unique (3)** or **further away from the other scores** for High Jumps than Javelin **(3)**
Q5 Total: 7.5		Kelvin's z-score= 2.21 Alicia's z-score= 1.56 (.25 for each correct answer)	Kelvin Or: The student with the higher z-score (as calculated)	• Correctly compare courses/students scores using a word or symbol (now, more, less, bigger, smaller, greater than "≥:, less than "<") (1pt) • Use z-scores in their comparison (1pt) • Interprets z-score correctly (2pt) - higher z-score = further from the mean / more unique score - if math is present and state that the student with the higher z-score (as calculated) is more unique	Kelvin deserves the award because his **z-score is 2.21 while Alicia's is 1.56 (2**), meaning his score is **more (1) unique, or further away from the other scores (3)** in his class than Alicia's in her class.

Notes:

- **Correct Answer Given:**
  - If they get the correct answer (within 1 point) without showing their work, give them the full points.
- **Correct Answer Not Given:**
  - Give one point for each step correctly completed.
  - Do not count off for math errors if they seem like non-conceptual mistakes (i.e., misread numbers, errors while entering in calculator [if the written math looks correct but the answer is wrong], etc).
  - If the steps are out of order (e.g., 1,2,3,4,6,5), do not give points for the steps out of order

Appendix C



Coding Rubric for Problem Solving Activity:
Quantity and Quality of Solution Attempts

Quantity of Solution Attempts – 1 point for any evidence of written or mathematical use of the following components:

- Mean
- Median
- Mode
- Range
- Use of Graphing (i.e., dot graphs, histogram, bar graph, line graphs)
- Use of other diagrams
- Counts/Frequencies
- Year to year differences (e.g., subtracting the years from each other)
- Mean absolute deviation
- Standard deviation
- Other: Participant used a unique strategy instead of one that is commonly used (e.g., algebra)
- Unclear: Participant used a strategy that does not fit into these categories, but it is unclear whether it might (i.e., unclear what participant was trying to do).
- No math – verbal explanation

Quality of Solution Attempts (score their best solution attempts only)

Standard Deviation Steps – 1 point for each step in the equation, 6 possible points for each of the three tea growers (18 total points possible).

(Note: computational errors do not count against the participant)

1. Calculated the mean
2. Calculated the differences
3. Squared the differences (or took the absolute value)
4. Summed all the (positive) differences
5. Divided by *N*
6. Took the square root

Note: If participant got correct *SD* (within +/- 1) for a tea grower, it is okay to assume that they completed all of the steps.

This coding scheme is adapted from Loibl et al. (2020) and Wiedmann et al. (2012)

Appendix D: scoring rubric for misconceptions

Misconceptions will show evidence a fundamental error in the concept of standard deviation/consistency. Computational errors **are not** considered misconceptions. Students must show evidence of incorrect conceptual knowledge regarding the formula of standard deviation.

Standard deviation steps:

1. Calculated the mean
2. Calculated the differences
3. Squared the differences (or took the absolute value)
4. Summed all the (positive) differences
5. Divided by *N*
6. Took the square root

Misconception Scoring (for both activity and post-test): 1 point for any written or mathematical evidence of each of the following misconceptions:

- Divide by different *N* (in Step 5; e.g., in tea grower activity, divides by 6 [Thourbo] or divides by 4 [Dareen/Ging]
- Divided by Mean (in Step 5)
- Did Not Divide
- Did Not Square Root
- Did Not Square
- Other: Participant presents a unique misconception instead of one that is commonly used
- Unclear: Participant displays a potential misconception that does not fit into these categories, but it is unclear whether it might (i.e., unclear what participant was trying to do; add a comment to explain).

Appendix E

See Tables3 and4.

Table 3 Means (standard error in parentheses) of posttest scores as a function of order of instruction for participants in-person

	Condition		
	Instruct-first ($n=2$)	Explore-first invent ($n=16$)	Explore-first generate ($n=14$)
Posttest	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
Procedural knowledge (%)	73.81 (15.83)	86.51 (6.05)	87.50 (7.36)
Conceptual knowledge (%)	66.9 (10.99)	66.99 (5.66)	74.22 (5.88)
Transfer (%)	53.33 (10.43)	62.41 (5.65)	67.29 (6.29)

Table 4 Means (standard error in parentheses) of posttest scores and strategy coding as a function of order of instruction for participants online

	Condition		
	Instruct-first ($n=69$)	Explore-first invent ($n=36$)	Explore-first generate ($n=27$)
Posttest	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
Procedural knowledge (%)	85.51 (2.26)	79.54 (4.81)	90.48 (3.58)
Conceptual knowledge (%)	61.57 (2.72)	55.47 (4.56)	69.39 (4.11)
Transfer (%)	38.19 (3.91)	35.23 (4.49)	43.64 (6.67)

Acknowledgements The authors thank Oliva Kaiser, Manooch Saeedi, and Campbell Bego for helpful discussion and assistance with this research, as well as Maria Kondaurova and Christian Stilp for allowing us to work with their classes.

Funding The authors did not receive funding from any organization for the submitted work.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. Materials are included in the Appendices and also available upon request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Alloway, T. P. (2006). How does working memory work in the classroom? *Educational Research and Reviews, 1*(4), 134–139.

Ashman, G., Kalyuga, S., & Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high? *Educational Psychology Review, 32*, 229–247. https://doi.org/10.1007/s10648-019-09500-5

Bego, C. R., Thompson, A., Patrick, R., Chastain, R. J., Hieb, J., Fuselier, L., & DeCaro, M. S. (2023). Exploration with ellipses helps students learn transferrable isometric drawing skills. *Proceedings of the American Society for Engineering Education*.

Bego, C. B., Chastain, R. J., & DeCaro, M. S. (2022). Designing novel activities before instruction: Use of contrasting cases and a rich dataset. *British Journal of Educational Psychology, 93*, 299–317. https://doi.org/10.1111/bjep.12555

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences, 21*(3), 399–432.

Brand, C., Hartmann, C., Loibl, K., & Rummel, N. (2023). Do students learn more from failing alone or in groups? Insights into the effects of collaborative versus individual problem solving in productive failure. *Instructional Science*. https://doi.org/10.1007/s11251-023-09619-7

Bush, J., DeCaro, M. S., & DeCaro, D. A. (2023). Playing a social dilemma game as an exploratory learning activity before instruction improves conceptual understanding. *Journal of Experimental Psychology: Applied.* https://doi.org/10.1037/xap0000470

Chase, C. C., & Klahr, D. (2017). Invention versus direction instruction: For some content, it's a tie. *Journal of Science and Educational Technology, 26*, 582–596.

Chen, O., & Kalyuga, S. (2020). Exploring factors influencing the effectiveness of explicit instruction first and problem-solving first approaches. *European Journal of Psychology of Education, 35*, 607–624. https://doi.org/10.1007/s10212-019-00445-5

Chin, D. B., Chi, M., & Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: Inventing a general solution versus compare and contrast. *Instructional Science, 44*, 177–195. https://doi.org/10.1007/s11251-016-9374-0

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Education Technology Research and Development, 66*, 1101–1118. https://doi.org/10.1007/s11423-018-9579-9

DeCaro, M. S., McClellan, D. K., Powe, A., Franco, D., Chastain, R. J., Hieb, J. L., & Fuselier, L. (2022). Exploring an online simulation before lecture improves undergraduate chemistry learning. *Proceedings of the International Society of the Learning Sciences.*

DeCaro, D. A., DeCaro, M. S., Janssen, M., Lee, A., Graci, A. A., & Flener, D. (2024). Initial regulatory failure helps naïve enforcers learn to create wiser common-pool resource enforcement systems when guided by principles of restorative justice. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0307832

DeCaro, D. A., DeCaro, M. S., & Rittle-Johnson, B. (2015). Achievement motivation and knowledge development during exploratory learning. *Learning and Individual Differences, 37*, 13–26. https://doi.org/10.1016/j.lindif.2014.10.015

DeCaro, M. S., Isaacs, R., Bego, C. R., & Chastain, R. J. (2023). Bringing exploratory learning online: Problem-solving before instruction improves remote undergraduate physics learning. *Frontiers in Education*. https://doi.org/10.3389/feduc.2023.1215975

DeCaro, M.S., Bego, C.R., Velić, L., & Newman, P. (2024). Increasing contrasting cases during exploration or practice problems given before or after instruction. *Instructional Science*. https://doi.org/10.1007/s11251-024-09696-2

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology, 113*, 552–568.

DeCaro, M. S., Van Stockum, C. A., Jr., & Wieth, M. (2017). The relationship between working memory and insight depends on moderators: Reply to Chuderski and Jastrzębski (2017). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 2005–2010. https://doi.org/10.1037/xlm0000460

Felder, R. M., & Brent, R. (2009). Active learning: An introduction. *ASQ Higher Education Brief, 2*, 1–5.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Flynn, L. R., & Goldsmith, R. E. (1999). A short, reliable measure of subjective knowledge. *Journal of Business Research, 46*(1), 57–66. https://doi.org/10.1016/S0148-2963(98)00057-5

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to exploration improves mathematical knowledge. *British Journal of Educational Psychology*. https://doi.org/10.1111/bjep.12035

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology, 98*(4), 611–625. https://doi.org/10.1348/096317907X173421

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction, 39*, 72–87.

Hartmann, C., van Gog, T., & Rummel, N. (2021). Preparatory effects of problem solving versus studying examples prior to instruction. *Instructional Science, 49*(1), 1–21. https://doi.org/10.1007/s11251-020-09528-z

Hartmann, C., van Gog, T., & Rummel, N. (2022). Productive versus vicarious failure: Do students need to fail themselves in order to learn? *Applied Cognitive Psychology, 36*(6), 1219–1233. https://doi.org/10.1002/acp.400

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151–179. https://doi.org/10.3102/00346543070002

Hieb, J., DeCaro, M. S., & Chastain, R. J. (2021). Work in Progress: Exploring before instruction using an online Geogebra™ activity in introductory engineering calculus. *Proceedings of the American Society for Engineering Education.*

Hsu, C.-Y., Kalyuga, S., & Sweller, J. (2015). When should guidance be presented in physics instruction? *Archives of Scientific Psychology, 3*(1), 37–53.

Jarosz, A. F., Goldenberg, O., & Wiley, J. (2016). Learning by invention: Small group discussion activities that support learning in statistics. *Discourse Processes, 54*(4), 285–302. https://doi.org/10.1080/0163853X.2015.1129593

Kalyuga, S., & Singh, A. M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review, 28*, 831–852. https://doi.org/10.1007/s10648-015-9352-0

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2010). Productive Failure in mathematical problem solving. *Instruction Science, 38*(6), 523–550.

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*(4), 651–672.

Kapur, M. (2014a). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences, 23*(4), 651–677.

Kapur, M. (2014b). Productive failure in learning math. *Cognitive Science, 38*, 1008–1022. https://doi.org/10.1111/cogs.12107

Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice, 1*, 51–65. https://doi.org/10.1080/23735082.2015.1002195

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

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

Kapur, M., Saba, J., & Roll, I. (2023). Prior math achievement and inventive production predict learning from productive failure. *npj Science of Learning, 8*, 15. https://doi.org/10.1038/s41539-023-00165-y

Kirschner, P. A., Sweller, J., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist, 42*(2), 115–121.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1534. https://doi.org/10.1037/0278-7393.25.6.1534

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798.

Lamnina, M., & Chase, C. C. (2019). Developing a thirst for knowledge: How uncertainty in the classroom influences curiosity, affect, learning, and transfer. *Contemporary Educational Psychology, 59*, 101785. https://doi.org/10.1016/j.cedpsych.2019.101785

Loibl, K., Leuders, T., Glogger-Frey, I., & Rummel, N. (2023). Cognitive analysis of composite instructional designs: New directions for research on problem-solving prior to instruction. In C. Damșa, M. Borge, E. Koh, & M. Worsley (Eds.), *Proceedings of the 16th international conference on computer-supported collaborative learning—CSCL 2023* (pp. 321–324). International Society of the Learning Sciences.

Loibl, K., Leuders, T., Glogger-Frey, I., & Rummel, N. (2024). CID: a framework for the cognitive analysis of composite instructional designs. *Instructional Science*. https://doi.org/10.1007/s11251-024-09665-9

Loibl, K., & Leukel, C. (2023). Problem-solving prior to instruction in learning motor skills-initial self-determined practice improves javelin throwing performance. *Learning and Instruction, 88*, 101828.

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*, 693–715. https://doi.org/10.1007/s10648-016-9379-x

Loibl, K., & Rummel, N. (2014a). Knowing what you don't know makes failure productive. *Learning and Instruction, 34*, 74–85. https://doi.org/10.1016/j.learninstruc.2014.08.004

Loibl, K., & Rummel, N. (2014b). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*, 305–326. https://doi.org/10.1007/s11251-013-9282-5

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., Tillema, M., Rummel, N., & van Gog, T. (2020). The effect of contrasting cases during problem solving prior to and after instruction. *Instructional Science*. https://doi.org/10.1007/s11251-020-09504-7

Marton, F. (2007). Sameness and difference in transfer. *Journal of the Learning Sciences, 15*(4), 499–535. https://doi.org/10.1207/s15327809jls1504_3

McNeil, N. M. (2014). A change–resistance account of children's difficulties understanding mathematical equivalence. *Child Development Perspectives, 8*(1), 42–47. https://doi.org/10.1111/cdep.12062

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*. https://doi.org/10.1007/s11251-020-09525-2

Newman, P., & DeCaro, M. (2019). Learning by exploring: How much guidance is optimal? *Learning and Instruction, 62*, 49–63. https://doi.org/10.1016/j.learninstruc.2019.05.005

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering. Education, 93*, 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14*(4), 477–488.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175–189. https://doi.org/10.1037/0022-0663.91.1.175

Rotgans, J. I., & Schmidt, H. G. (2014). Situational interest and learning: thirst for knowledge. *Learning and Instruction, 32*, 37–50. https://doi.org/10.1016/j.learninstruc.2014.01.002

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology, 47*(6), 1525.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–522. https://doi.org/10.2307/3233709

Schwartz, D. L., Chase, C. C., Opezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*, 759–775. https://doi.org/10.1037/a0025140

Schwartz, D. L., Lindgren, R., & Lewis, S. (2009). Constructivism in an age of non-constructivist assessments. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure* (pp. 34–61). Routledge/Taylor & Francis Group.

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/s1532690xci2202_1

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 319–344). Lawrence Erlbaum Associates Publishers.

Silvia, P. J. (2008). Interest—the curious emotion. *Current Directions in Psychological Science, 17*(1), 57–60.

Sinha, T. (2022). Enriching problem-solving followed by instruction with explanatory accounts of emotions. *Journal of the Learning Sciences, 31*(2), 151–198. https://doi.org/10.1080/10508406.2021.1964506

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research, 91*, 761–798.

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., & Trninic, D. (2021). Differential benefits of explicit failure-driven and success-driven scaffolding in problem-solving prior to instruction. *Journal of Educational Psychology, 113*(3), 530. https://doi.org/10.1037/edu0000483

Song, Y., & Kapur, M. (2017). How to flip the classroom—"Productive failure or traditional flipped classroom" pedagogical design? *Educational Technology & Society, 20*(1), 292–305.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., Dechenne-Peteres, S. E. , Egan Jr, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgove, A., Palmer, M.S., Plank, K. M., … & Young, A. M. (2018). Anatomy of STEM teaching in north American universities. *Science,* 359(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*, 185–233.

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*, 975–994.

Thevenot, C., & Oakhill, J. (2008). A generalization of the representational change theory from insight to non-insight problems: The case of arithmetic word problems. *Acta Psychologica, 129*(3), 315–324.

Trninic, D., Sinha, T., & Kapur, M. (2022). Comparing the effectiveness of preparatory activities that help undergraduate students learn from instruction. *Learning and Instruction, 82*, 1–12. https://doi.org/10.1016/j.learninstruc.2022.101688

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.

Weaver, J. P., Chastain, R. J., DeCaro, D. A., & DeCaro, M. S. (2018). Reverse the routine: Problem solving before instruction improves conceptual knowledge in undergraduate physics. *Contemporary Educational Psychology, 52*, 36–47. https://doi.org/10.1016/j.cedpsych.2017.12.003

Wegner, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Wiedmann, M., Leach, R. C., Rummel, N., & Wiley, J. (2012). Does group composition affect learning by invention? *Instructional Science, 40*(4), 711–730. https://doi.org/10.1007/s11251-012-9204-y

Wise, A. F., & O'Neill, K. (2009). Beyond more versus less: A reframing of the debate on instructional guidance. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure* (pp. 82–105). Routledge.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*(1), 49–64.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.