The Role of Prior Knowledge in Effects of Embodied Pedagogies on Learning

A PREPRINT

Icy (Yunyi) Zhang
Department of Educational Psychology, University of Wisconsin-Madison

Alice Xu
Department of Psychology, University of California, Los Angeles

Craig Enders
Department of Psychology, University of California, Los Angeles

Ji Y. Son
Department of Psychology, University of California, Los Angeles

James W. Stigler
Department of Psychology, University of California, Los Angeles

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Icy Zhang, UW Madison Educational Psychology, 53706. Email: icy.zhang@wisc.edu

The study design, hypotheses, and analytic plans were not pre-registered. The study data and analytic code are open access at:

https://osf.io/ntsr2/?view_only=ce650267f407451a9ea26abcb428d8f7

Abstract

Background: Embodied learning pedagogies are increasingly popular, but our understanding of how embodied pedagogies enhance learning of complex concepts in higher education over longer periods of time remains nascent, particularly how learners' prior knowledge might interact with different types of embodied pedagogies. We propose the Performing First Hypothesis as an organizing framework, which posits that learners with lower prior knowledge, because they lack the foundational understanding needed to make sense of abstract representations, will benefit from actively performing actions. In contrast, learners who have higher prior knowledge will learn better from observing the actions as it enables more opportunity for connection-making and transfer.

Methods: We implemented and compared two types of embodied pedagogy: (1) perform, in which students actively participated in hands-on activities, and (2) observe, in which students observed others doing hands-on activities. The activities were implemented as the lab component of a 10-week college-level introductory statistics course. 227 college students paired up and participated in the study. One half of the pair was randomly assigned to the Perform condition, where they themselves did the hands-on activities, and the other half was assigned to the Observe condition, where they observed their partner performing those activities.

Findings: An analysis of students' midterm and final exam scores showed a significant interaction between prior knowledge and type of pedagogy in the hypothesized direction. Interestingly, only self-rated prior knowledge—not other factors such as math anxiety or how well they had performed in math in previous classes—significantly moderated the effectiveness of the embodied pedagogies. Multilevel analysis of weekly tests supported the exam findings,

demonstrating that low-prior-knowledge learners benefited more from physically performing the activities, whereas high-prior-knowledge learners benefited more from observing the activities. **Contribution:** The findings support the Performing First Hypothesis, suggest new directions for research on the mechanisms and applications of embodied pedagogies in complex learning environments, and inform real-world educational practices.

***Keywords*:** embodied cognition, learning, data science education, prior knowledge, hands-on curriculum, longitudinal research

Educational Impact and Implications Statement

This study found that learners with low prior knowledge benefited more from actively performing embodied actions (e.g., gesture, object manipulation, and hand drawing), while those with high prior knowledge gained more from observing a classmate perform them. These findings bridge theory, research, and practice in three ways. First, they demonstrate the effectiveness of embodied pedagogies beyond short-term learning of a single concept, highlighting their role in developing complex knowledge of a multitude of concepts over time. Second, the study does not merely ask whether embodiment impacts learning but instead asks how different embodied activities can be effectively implemented in classrooms, providing actionable insights for instructors and content developers. Finally, it elucidates the role of individual differences, particularly prior knowledge, in shaping embodied learning experiences, offering insights for more personalized instruction.

The Role of Prior Knowledge in Effects of Embodied Pedagogies on Learning

Despite the long-standing argument that conceptual ideas originate from our sensorimotor experience in the world (Hume, 1938; Locke, 1948; Quine & Ullian, 1978), many educational practices emphasize the mastery of formalized knowledge, such as symbols and equations, before practical experiences (see Nathan, 2012 for a review). While this approach can be effective for some learners or for learning some concepts, it relies on the assumption that learners will be able to connect formal notation and abstract concepts to their meanings in some other way that does not require experiential or direct experience (Goldstone et al., 2008; Martin, 2009). Research, however, shows that many novices struggle to construct meaningful and interconnected understandings of such abstractions. Learners may grasp formal concepts at a surface level yet fail to link them with other representations or real-world situations (Lau & Yuen, 2008; Nathan, 2012; West & Ross, 2002), which in turn limits their ability to transfer knowledge to novel contexts (Bransford & Schwartz, 1999; Stigler et al., 2010). In contrast, pedagogies that ground abstract ideas in concrete, physical experiences before introducing formalized instruction—an approach often referred to as *concreteness fading*—tend to produce more flexible and generalizable abstract knowledge (e.g., Fyfe et al., 2015; Goldstone & Son, 2005; McNeil & Fyfe, 2012).

Concreteness fading is an example of embodied pedagogies. Embodied pedagogies–which involve using the body to acquire even abstract concepts and skills through learners' enactment, observation, and mental simulation of bodily actions (Clark, 2008; Tytler & Prain, 2013; Wilson, 2002)–have emerged as a promising approach to fostering transferable knowledge. These pedagogies emerge from theories of embodied cognition that assume concepts are grounded, that is, given meaning and intertwined with action and perception (Barsalou, 1999,

2008; Borghi & Pecher, 2011; Clark, 2008; Golonka & Wilson, 2012; Lakoff & Johnson, 2008; Trudeau & Dixon, 2007). Instead of solely emphasizing the storage and processing of knowledge based on abstract, symbolic concepts, theories of embodied cognition emphasize the role of sensorimotor experiences even in learning highly abstract concepts such as those prevalent in Science, Technology, Engineering, and Mathematics (STEM). When we do not carefully incorporate embodied actions into instruction, we may be overlooking a key lever of improving cognitive processes (Abrahamson et al., 2020; Sullivan, 2018).

Research on embodied pedagogies show an increase in learning when learners directly act on external manipulatives or representations (Sommerville & Woodward, 2010; Wilson, 2002), represent ideas in gesture or actions (Congdon & Goldin-Meadow, 2021; Zhang et al., 2021), or simply watch someone else produce gestures or object manipulations (Cook et al., 2024; Zhang et al., 2022). Integrating embodied pedagogies into the instruction can offer students the grounding they need in order to develop a meaningful and robust mental representation of abstract symbols (Laurillard , 2001; Trudeau & Dixon, 2007).

Although these embodied pedagogies and frameworks have been well-researched, mastering complex real-world domains requires more extensive interaction with a broad array of interconnected concepts over longer periods of time than is typically offered in laboratory studies or a single classroom observation (Fries et al., 2021). Below, we identify three key gaps between research and practice in the realm of embodied learning.

Gaps Between Current Research and Practice

This paper aims to advance our understanding of embodied pedagogies by addressing three gaps between current research and teaching practice. First, questions remain about the

generalizability of embodied pedagogies beyond the learning of a single concept. Most existing evidence comes from short-term laboratory studies, where embodied pedagogies have been shown to facilitate learning at the individual concept level. Although emerging longitudinal research has provided important insights (e.g. Kosmas & Zaphiris, 2023), these studies are largely observational or mixed-methods and focused primarily on children in lower grade levels, who are cognitively and developmentally different from college students. The applicability and effectiveness of embodied pedagogies in a longitudinal, college-level classroom setting—where students must learn a multitude of interrelated concepts over time—remain unclear.

The second gap between research and practice is a practical one: how should teachers implement embodied activities in classroom settings? Although many studies demonstrate that embodied pedagogies outperform less embodied approaches (for reviews, see Nathan, 2021; Novack & Goldin-Meadow, 2015), these findings provide little concrete guidance for instructional decision-making (Fugate et al., 2018). For example, some studies have students engage in bodily movements (Johnson-Glenberg et al., 2014), produce gestures (Broaders et al., 2007), manipulate objects (Rosenbaum et al., 2012), or draw (Zhang et al., 2025), while others focus on learners observing actions performed by instructors or peers (Cook et al., 2024; Goldin-Meadow et al., 2012; Zhang et al., 2022, 2025). Yet these studies typically compare embodied strategies to non-embodied ones, rather than to each other, leaving unanswered which forms of embodiment are most effective for classroom practice. As a result, teachers are left to navigate a fragmented body of evidence, with little guidance even on basic questions such as whether students should physically perform actions or simply observe them through, for example, a live demonstration. Scholars have emphasized the importance of asking research questions that are more directly relevant to practitioners (Farley-Ripple et al., 2018) and of translating findings into

actionable insights for educators (Alibali et al., 2024). For theories of embodied cognition to meaningfully shape education, research must move beyond asking *whether* embodiment supports learning to examining *how* different forms of embodiment can be implemented effectively in classrooms.

The third gap we face in our effort to bridge research and practice concerns individual differences among learners, particularly the diversity of their prior knowledge. In lab studies, prior knowledge is usually treated as a covariate and only a few studies have considered how prior knowledge could interact with embodied pedagogies (e.g., Congdon et al., 2018; Cook et al., 2024; Guarino & Wakefield, 2020; Zacharia et al., 2012). Even laboratory experiments that do consider prior knowledge may not be a good model for investigating how learners develop a deep understanding of abstract concepts in STEM domains because prior knowledge may have different effects when examined over longer periods of time (Brod, 2021; Simonsmeier et al., 2022; Thompson & Zamboanga, 2003). The critical point is that when learners come into the class for an embodied learning experience, their initial understanding of the concepts differs. We need research to understand how different types of embodied pedagogies, such as performing and observing actions, might affect diverse learners. No study to our knowledge has directly investigated how prior knowledge might moderate the effect of different types of embodied pedagogies in complex domains over a significant period of time.

Research Question and Present Study

Addressing these three gaps requires moving beyond short-term laboratory investigations to classroom contexts, where students engage with multiple interconnected concepts over extended periods of time. To advance the field and bridge these gaps, our study focuses on the following research question: *How does a longitudinal intervention, implemented with two forms*

of embodied pedagogies (performing versus observing), affect learners with varying levels of prior knowledge?

We situate the current investigation in the domain of statistics and data science for two reasons. First, there is a long documented difficulty of teaching statistics and data science, whether that being helping students understand individual concepts such as probability distribution (Zhang et al., 2025), sampling distribution (Chance et al., 2004) and confidence interval (Henriques, 2016) or develop more intuitive understanding of simulation-based techniques such as randomization and bootstrapping (Zhang et al., 2022). Exploring the effectiveness of embodied pedagogies, especially among diverse learners, in the domain of statistics can yield practical insights into the teaching and learning of statistics. Second, given the inherently abstract nature of statistical and data science concepts, exploring the effectiveness of embodied pedagogies in a longitudinal context offers theoretical significance. Prior research has shown that grounding abstract ideas, such as randomness, in hands-on manipulations of physical datasets can enhance students' comprehension of these concepts and their associated statistical inferences in controlled laboratory settings (Zhang et al., 2022). Extending this research to a longitudinal classroom study will provide deeper insight into the role of embodiment in the development of abstract knowledge in a complex, multi-concept domain.

We begin to bridge the first gap by conducting a randomized control trial in a 10-week introductory statistics course. This allows us to examine how embodied pedagogies operate over a realistic time course of learning complex concepts that require many weeks of instruction.

To address the second gap and provide actionable guidance for instructors, this study compares two different implementations of embodied pedagogy: students performing the actions

themselves versus observing others perform actions. This comparison helps clarify whether direct engagement is necessary or if observation is enough in complex learning domains.

Finally, in consideration of individual differences among learners, this study aims to understand the experiences of students with a wide range of prior educational experiences. Because students taking a core college course such as introductory statistics come in with varying levels of prior knowledge, we can examine how performing versus observing actions interacts with prior knowledge. Our study raises the possibility that certain kinds of embodied pedagogy can be particularly effective for students with low prior knowledge of statistics, algebraic notation, or programming.

While the design of the present study helps bridge methodological gaps, it is also an attempt to provide a new theoretical explanation, called the Performing First framework, of how and why different types of embodied pedagogy might interact with learners' prior knowledge. Before we present our Performing First framework, we will show how it is rooted in simulation-based theories of embodied cognition and review empirical findings related to performing versus observing embodied activities for learners with varying levels of prior knowledge.

Theoretical Framework and Empirical Evidence

Various theoretical frameworks have offered valuable insights into how embodied pedagogies support learning, ranging from accounts of specific pedagogical mechanisms to broader theories of cognition (e.g., Barsalou, 2008; Beilock & Goldin-Meadow, 2010; Hostetter & Alibali, 2008; Nathan, 2017; Nathan & Alibali, 2021). One influential mechanism is mapping, which helps learners transfer knowledge across analogous situations by leveraging sensorimotor experiences as a base analog for understanding new concepts or problems (Nathan & Alibali, 2021; Novack & Goldin-Meadow, 2017; Zhang et al., 2024b). Novack and Goldin-Meadow

(2017) argued that gestures support generalization and knowledge transfer not only through their physical properties but also by serving as representational tools—acting as meaningful analogies for mapping abstract ideas. Nathan and Alibali (2021) proposed an embodied theory of transfer in mathematics learning, emphasizing the role of mapping familiar ways of perceiving and acting from past experiences to new contexts.

However, mapping alone has limitations when applied to complex domains. In domains such as statistics, the base analog does not often exist or it may only exist for a subset of students. All students need to have a meaningful understanding of the embodied analogy before any sort of mapping could occur (Barsalou, 1999), but the degree of support they need differs. While mapping may facilitate the learning of individual concepts if learners have acquired the initial base analog, when it comes to learning in complex domains, we need a theory of learning that explains how those sensorimotor foundations should be introduced, mapped over time, and eventually connected to each other. Such a learning theory will enable more informed design of novel interventions and testable predictions about their effectiveness across diverse learners. Revisiting broader accounts of how embodiment supports learning may provide useful insights toward this goal.

Insights from Simulation-based Theories of Embodied Cognition

One of the most classic and popular classes of theories, the simulation-based theories of embodied cognition, provides a useful lens for understanding how the sensorimotor foundation of abstract concepts can be efficiently introduced, emphasizing on the difference between performing actions and observing them for different learners. According to perceptual symbol systems theory (Barsalou, 1999), a prominent simulation-based theory, when learners engage in a perception-action experience, it activates a network of neurons, creating multimodal frames

that serve as mental representations of that experience (Zhang et al., 2024a). These frames integrate information from several modalities (e.g., haptics, vision, touch, language). Later, when learners observe or simply imagine bodily actions, similar to those they engaged in, these multimodal frames are reactivated, allowing them to mentally simulate elements of their original learning experience (Pezzulo & Calvi, 2011). Empirical evidence for this comes from research showing that premotor processes involved in planning and preparing gestures related to mental abacus enhanced learners' visuospatial representations and supported subsequent problem-solving (Brooks et al., 2018).

The assumption that performing creates multimodal frames–which learners can later access as perceptual and action-based simulations–suggests that performing might be particularly effective for learners with low prior knowledge. However, this idea has not been explicitly articulated in existing embodied learning theories. The gesture-as-simulated-action (GSA) framework proposes that action production depends on perceptual and motor simulations (Hostetter & Alibali, 2010). Yet, much less is known about the reverse direction: how engaging in action production might enhance learners' ability to later simulate those actions mentally.

Moreover, for learners with greater prior knowledge or emerging expertise, embodied cognition theories have not clearly specified predictions regarding the relative effectiveness of performing versus observing instructional approaches. To begin developing a working theoretical account of how prior knowledge may interact with different embodied pedagogies, we turn to empirical findings from the embodied learning literature for guidance.

Empirical Evidence of How Prior Knowledge Moderates the Effect of Embodied Pedagogies

Although the specific interaction between prior knowledge and these two embodied pedagogies (performing vs. observing) has not been experimentally tested, some empirical work

comparing other types of embodied pedagogies provides insights into what this interaction might look like. For example, Congdon et al. (2018) compared gesture and object manipulation to see whether learners with low prior knowledge might understand one but not the other. Because object manipulation enables learners to directly interact with external objects, it is more perceptible and potentially more comprehensible by novice learners. Congdon and colleagues found that when students learning linear units of measure were randomly assigned to one of the two embodied pedagogies, those with low prior knowledge benefited from object manipulation but not from gesture. In contrast, children with higher prior knowledge benefited equally from both types of embodied pedagogies.

This difference suggests that whereas *object manipulation* can be more directly experienced and is thus more meaningful to novice learners, *gesture* might only activate embodied knowledge for learners who have the prior knowledge required to make meaningful connections between the gesture and the targeted abstract concepts. Higher prior knowledge makes performing less perceptible embodied pedagogies meaningful to learners, whereas learners with low prior knowledge need more perceptible performing experiences, such as manipulating the concrete objects, to understand the meaning behind the actions.

This study by Congdon et al. (2018) highlighted the importance of prior knowledge, but both of the embodied pedagogies included in the Congdon et al. (2018) study involved physical performance. We do not know the effect that merely *observing* embodied pedagogies would produce for learners with different levels of prior knowledge. We also do not know whether learners with high prior knowledge will benefit more from observing embodied pedagogies or whether more concrete and perceptible embodied pedagogies are always better for learning. For

example, more concrete pedagogies might diminish learners' ability to transfer knowledge to solve new problems in novel contexts (Novack et al., 2014).

In addition, just as the first gap we have identified between research and practice, the insights generated by measuring participants' knowledge right before and right after one highly controlled lab experiment may not be sufficient to generate insights about learning in authentic, complex STEM domains over time. Because developing expertise in STEM domains is a lengthy process, a single intervention teaching one concept may not accurately reveal the complexity of learning over time. In order to conduct longitudinal research that examines how prior knowledge interacts with the learning of a multitude of concepts, we need a working theory to explicitly model how the effectiveness of embodied learning will vary based on learners' expertise.

The Performing First Hypothesis

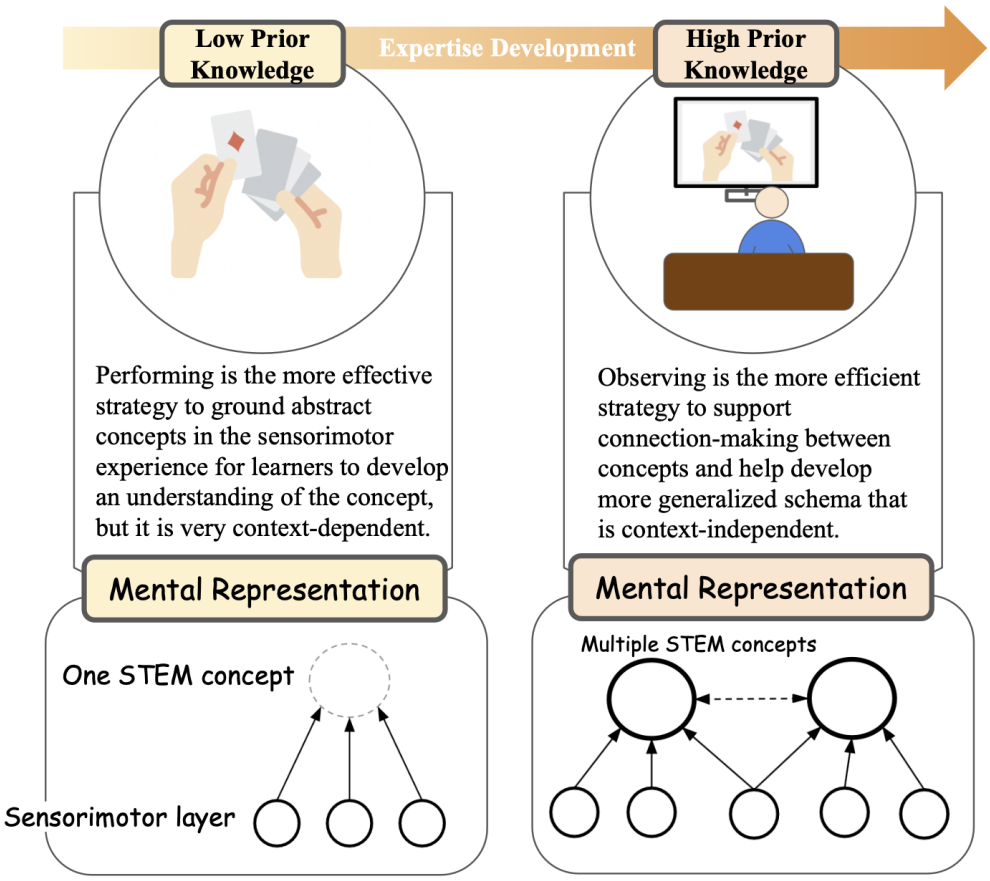
We propose a working framework, the Performing First Hypothesis, about the effectiveness of different types of embodied pedagogies when learners have different levels of expertise (Authors et al., under review). This hypothesis has two complementary parts: (1) for learners with low prior knowledge, activities that involve direct sensorimotor engagement are more effective than the observation of such activities whereas (2) for learners with high prior knowledge, simply observing such sensorimotor activities can be more helpful than direct engagement.

Whereas most research on educational interventions finds a positive correlation between students' prior knowledge of the concepts and their performance after the intervention, this hypothesis posits that performing embodied activities might attenuate or even eliminate the correlation between prior knowledge and performance after the intervention for learners.

In the following two sections, we present evidence supporting each part of the Performing First framework. Figure 1 illustrates the mental representations that explain why learners with different levels of prior learning may benefit either from performing or observing. These representations provide the cognitive mechanisms for patterns we might observe in our study. For the more extended discussion of the Performing First Framework, see Authors et al. (under review).

Figure 1

A summary of the Performing First Framework



Performing Can be More Helpful for Learners with Low Prior Knowledge

The advantage of performing is especially pronounced for learners with less prior knowledge or experience. The perceptual symbol systems theory posits that the usefulness of

these simulations depends on the initial sensorimotor engagement (Barsalou, 2008), which suggests that performing actions might be especially impactful for learners who lack existing sensorimotor grounding. For these learners, performing actions is necessary to ground the abstract concept in familiar bodily actions (Fyfe et al., 2015; Lin et al., 2016; McNeil & Fyfe, 2012), allowing them to construct their own knowledge (Martin, 2009) and building the meaningful multimodal frames to be utilized in later simulations (Barsalou, 1999). In contrast, for learners who already have relevant prior knowledge, observing actions can trigger existing simulations, enabling them to make use of their prior understanding without the need for physical activity.

One line of evidence comes from infant research. In a classic study, Sommerville and colleagues (2005) provided evidence from infants who were randomly assigned to get varied experience in grasping objects. Infants who could not yet grasp objects were given "sticky" velcro mittens, allowing them to experience what it would be like to reach out and grasp an object. These infants, having now experienced the grasping action themselves, found observing an experimenter perform a similar action more interesting than infants without the first-hand experience (Gerson & Woodward, 2014). The initial sensorimotor experiences of grasping may have made the observations of others grasping more meaningful.

Neurological research on mirror neurons (Rizzolatti & Craighero, 2004) further supports the importance of initial engagement through action. Mirror neurons are a class of motor neurons that discharge when a person acts, but they also discharge when a person observes an action–provided the observer understands the goal of the action (Calvo-Merino, 2013; Gazzola et al., 2007). Performing the actions may provide crucial insight into the goals and meaning behind the action necessary for activating these neural pathways later.

The role of prior knowledge becomes even more evident in research on embodied learning of abstract and difficult concepts in educational settings. For instance, Goldin-Meadow and colleagues (2012) directly compared the effects of performing versus observing gestures on a mental rotation task. In their study, six-year-old, novice learners were asked to judge whether two shapes at different angles of rotation were the same or different. Children who performed gestures that mimicked the rotation of the figures outperformed those who simply observed someone else performing the same gestures on a video clip.

Whereas Goldin-Meadow et al. (2012) did not explicitly examine the heterogeneity of learners' prior knowledge, Zacharia and colleagues (2012) demonstrated that learners with low prior knowledge benefited more from physically manipulating objects than from virtually manipulating objects on a computer screen, which is lower in sensorimotor engagement. Kindergarteners who had a correct understanding of a balance scale (i.e., high prior knowledge) and those who did not (i.e., low prior knowledge) were randomly assigned to either physically interact with a balance scale and objects of different weights or virtually interact with those same objects on a screen. For high-prior-knowledge learners, performance on using the balance scale did not differ based on whether they engaged in physical or virtual manipulation. However, low-prior-knowledge learners showed a different pattern of results: those who performed the physical manipulation significantly outperformed those who performed the virtual manipulation. Notably, the low-prior-knowledge participants who engaged in virtual manipulation did not improve from the pre- to posttest, whereas those who physically manipulated the objects demonstrated the largest gains.

Similarly, Novack and colleagues (2014) found a similar pattern of results when examining the effects of relatively more concrete versus abstract gesture on children's

understanding of mathematical equivalence. While instruction with abstract gestures generally led to better performance on novel problems, a secondary analysis of the data suggested an interaction between prior knowledge and the type of instruction (Congdon & Goldin-Meadow, 2021). Children with low pretest scores learned the least from abstract gesture instruction compared to children with high pretest scores who received the same instruction and those with low pretest scores who received concrete gesture instruction. The findings suggest that concrete gesture instruction, which might be considered relatively more meaningful and grounded, can be especially useful and interpretable for low-prior-knowledge learners. It is worth clarifying that although symbols such as the equal sign are perceptual objects, they are considered "abstract" because they are symbolic entities that have an arbitrary relation to their referents and can thus stand for a variety of superficially dissimilar ideas. Embodied pedagogy might be particularly helpful in creating meaningful referents (e.g., simulations) that can be drawn upon when viewing those symbols later.

Although these studies examine children, prior knowledge also plays a crucial role in learning among adults. For example, Grotelueschen (1979) found that adults with low prior knowledge benefited more from an instructional sequence that transitioned from concrete to abstract concepts, whereas adults with higher prior knowledge benefited more from materials that were abstract throughout. This suggests that the level of prior knowledge may determine the extent to which learners need concrete, sensorimotor experiences to build understanding before moving to more abstract forms of learning.

In summary, our own experience as well as research by others indicates that the same embodied pedagogy may be experienced as meaningful by some learners but not by others. In

particular, direct physical engagement might be an especially effective strategy for learners with low prior knowledge in the domain.

Observing Can be More Helpful for Learners with High Prior Knowledge

We established that performing actions relevant to the concept being learned helps low-prior-knowledge learners anchor abstract ideas in meaningful bodily experiences. However, these grounded representations tend to be highly context-dependent (Son & Goldstone, 2009). Nathan and Alibali (2021) posit that transfer happens when people connect their past experiences from one situation to a novel context, by applying familiar ways of perceiving and acting to create continuity between physical experiences and novel situations. Thus, we reason that once learners develop a foundational understanding, action observation might become a more efficient strategy for facilitating connection-making and knowledge transfer, enabling them to apply what they have learned in new contexts.

Cognitive load theory provides theoretical insights for this conjecture. Cognitive load theory suggests that, due to humans' limited cognitive capacity, it is crucial to allocate as much cognitive resource as possible to learning itself (Chandler & Sweller, 1996; Pyers et al., 2021; Sweller, 2010). Other cognitive effort that is irrelevant to the actual learning, but also part of the learning process (which often a result of the instructional design), is considered extraneous load (Kalyuga, 2011; Sweller et al., 2011). But, what is considered extraneous for learning is not the same for different learners. For novice learners, the physical effort involved in performing activities is not extraneous. However, for more experienced learners, because observing is sufficient to activate previously stored neural connections (Barsalou, 1999), this physical effort is extraneous to the learning process. These high-prior-knowledge learners may benefit more from

observing the activity, as it frees up cognitive resources that can be directed towards deeper understanding of the concept itself.

The work on concreteness-fading, complement, the expertise-reversal effect, and transfer together poses a similar conundrum: a more physically grounded and concrete pedagogy is more useful at the beginning, but it might increase the effort required for more experienced learners to generalize and transfer (Belenky & Schalk, 2014; Fyfe & Nathan, 2019; Kaminski et al., 2008; Lin et al., 2016; Son & Goldstone, 2009). When learners are fully engaged in physically performing the activity (e.g. interacting with manipulatives or gesturing), they might fall short of opportunities to transfer the knowledge outside of this specific context. Restricting learners to direct physical engagement even when they have established some fundamental knowledge might cause their knowledge representation to be completely context-dependent (Son & Goldstone, 2009) and restrict transfer (Kaminski et al., 2008). The complement literature has suggested that an overload of details and contextual elements in concrete representations (performing actions involves an excess amount of contextual details) can cause misunderstandings or trigger the use of incorrect schemas based on superficial traits (Harp & Mayer, 1998), which can hinder generalization and transfer (Ainsworth, 2006; Belenky & Schalk, 2014; Goldstone & Son, 2005). Lastly, although not directly related to embodied learning, the expertise reversal effect—the finding that instructional methods effective for novices may hinder experts (Kalyuga et al., 2003)—supports this conjecture. For advanced learners, removing redundant representations or step-by-step solutions can be more beneficial than including them (Renkl & Atkinson, 2007; Wittwer & Renkl, 2008).

This also circles back to the point of cognitive load capacity. When learners are observing the actions, they can simulate the actions and activate their mental representations with

less effort. The freed-up cognitive resources can be used for making connections between this concept and other concepts, which enhances transfer. Research on conceptual transfer has found that observing led to lower error ratio on the transfer tasks than performing (Chalmers & Rosenbaum, 1974). College students who only observed the first set of tasks were more flexible when the transfer tasks required them to make a reverse association between the items they were trained to associate in the first task. Chalmers and Rosenbaum (1974) argued that observing affords this flexibility because it helps to detach the learning from the specific learning condition, which in turn facilitates transfer. Although not directly contrasted with a performing group, more recent research has also shown that observing actions such as gesture or hand drawing can improve transfer (e.g., Cook et al., 2013; Fiorella & Mayer, 2016).

One critical re-analysis of a recent study done by Zhang et al. (2024) supports the argument that compared with low-prior-knowledge learners, high-prior-knowledge learners can more effectively benefit from observing. Zhang et al. (2025) conducted a personalized mediation analysis on college students with varying levels of prior knowledge in data science. The study found that high-prior-knowledge learners benefited more from observing an embodied instructional video than their low-prior-knowledge counterparts. Specifically, observing these instructional videos enhanced their learning by fostering stronger visuospatial representations, further supporting the role of observation in advanced learning stages.

We argue that for learners who have already established some grounding of the knowledge, the observing pedagogy may serve as a productive middle ground between bodily engagement and abstract instruction. Observation enables learners to more readily activate previously formed embodied representations (Calvo-Merino et al., 2005) while simultaneously

distancing them from the immediate physical context, thereby preparing them to construct more abstract, generalized schemas.

Yet, the role of prior knowledge in embodied learning is complex, as not all studies align perfectly with these findings. For example, Wakefield and James (2015) found that some learners may need a certain level of prior knowledge to begin to benefit from embodied pedagogies. They assigned children to learn the concept of a palindrome using either speech-only instruction, speech+gesture match instruction, or speech+gesture mismatch instruction. Only children with high phonological ability—those with greater prior knowledge in the domain—benefited from performing gestures compared to speech-only instruction. Guarino and Wakefield (2020) suggested that this result might reflect a developmental stage in which children need a foundational understanding for gestures to be meaningful. This study introduces the possibility of a "sweet spot" in which learners have just enough prior knowledge to benefit from embodied learning experiences. All together, more theoretical development and empirical work is needed to investigate how embodiment impacts learners with different levels of prior knowledge.

The Current Study

In the current study, we developed a supplementary lab curriculum employing embodied pedagogies, and then implemented the curriculum over a 10-week term as part of a college-level introductory course in statistics and data science. The course was taught using a CourseKata interactive online textbook, available for preview at https://coursekata.org (Son & Stigler, 2024). The book takes a modeling approach to statistics and teaches students to analyze data using R as well as techniques such as randomization, bootstrapping, and simulation. (For more information

about the CourseKata book and the principles guiding its design see Stigler et al., 2020; Son et al., 2021; Fries et al., 2021.)

In the mandatory once-per-week lab sessions, students were randomly paired with lab partners for the duration of the course. Within each pair, roles were assigned randomly: one student as the performer, engaging directly in embodied activities, and the other as the recorder/observer, observing and recording the performer's hand movements using a smartphone. When the performers were engaging in the actions, observers were asked to attend closely to their partner's actions, document the activity, and offer verbal—but not physical—feedback. The recording task was intended both to focus observers' attention on the performer's movements and to occupy their hands, reducing the likelihood of involuntarily assisting with the actions. This arrangement created a distinctive form of "observation": rather than observing solely for their own learning, observers were also tasked with documenting their partner's performance.

We examined students' weekly pre- and post-test scores as well as their performance on summative assessments (e.g. midterm and final exams). We hypothesized that there would be an interaction between the type of embodied intervention and the participants' prior knowledge. We expected that performers with lower prior knowledge would outperform observers who also had lower prior knowledge, whereas observers with higher prior knowledge would outperform performers with higher prior knowledge.

Methods

Participants

Participants were students enrolled in an introductory psychological statistics course at a large public research university. Participants were pre-psychology majors and this course is a

prerequisite before students could formally declare their major. The class was structured as two weekly lectures and a lab session. The experimental interventions took place once a week for nine weeks during the 50-minute in-person lab sessions. Students received class participation points for attending lab sessions but were allowed to miss one lab session without losing any participation points. All participants were given informed consent at the beginning of the first lab and may choose to remove their data from the analyses. As research conducted in established educational settings evaluating two types of instructional strategies, the study was exempt by the university's institutional review board (IRB). The two types of instructional pedagogies have both been shown to improve learning and would not have an adverse impact on students' learning.

A priori power analysis was not conducted, because this classroom intervention intended to include all students enrolled in the course except for those who did not agree to the data-sharing agreement of the course or dropped the class. There were 236 students enrolled in the course. After removing the students who opted to not share their data or who dropped the course, the final sample size was 227. Out of these students, 171 students self-identified as female (75%), 49 as male (22%), and 7 as non-binary (3%). The self-reported race and ethnicity of these students were as follows: 100 Asian or Asian American (44%), 8 Black or African American (4%), 47 Hispanic, Latino or Spanish origin (21%), 13 Middle Eastern or North African (6%), 46 White (20%), and 13 mixed/multi races (6%).

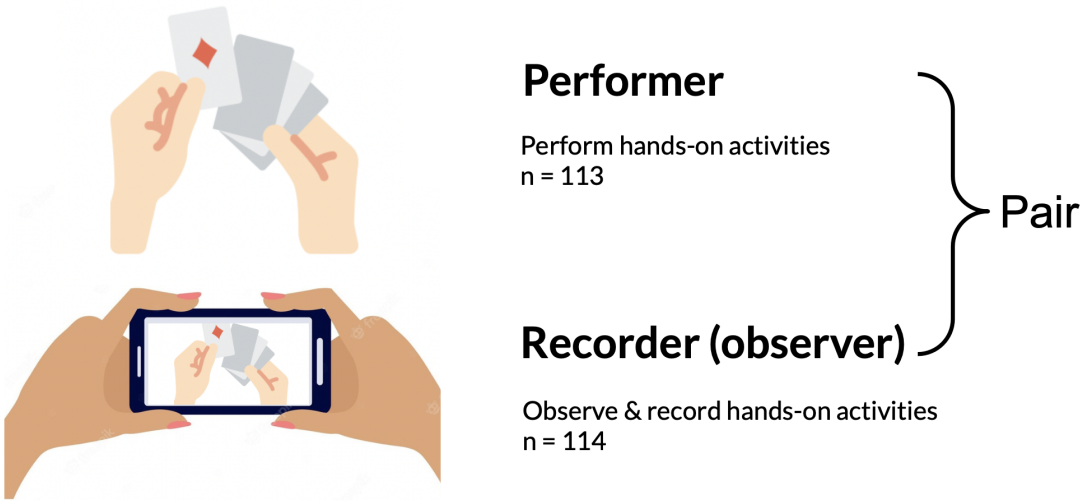
Design and Procedure

Participants were randomly assigned a lab partner at the beginning of the course. Within each student dyad, students were randomly assigned to either the perform condition (n = 113) or

the Observe/Record condition (n = 114). Each student's role and partner stayed the same for the entire course (Figure 2).

Figure 2

Condition and Dyad



Participants in the *Perform* condition (performers) carried out embodied instructional actions designed to support learning. Those in the *Observe* condition (observers) closely watched their partner's actions while recording them on a smartphone. For short activities, observers recorded the entire sequence, whereas for longer, multi-step activities they were asked to capture only the beginning and end, which provided a clear structure for their role while also preventing them from physically carrying out the tasks. During each activity, observers provided verbal feedback to their partners, followed by a structured discussion between the pair. To contextualize these role assignments, we provided a cover story: students were told that their video recordings would supply valuable hand-movement data to support the design of a pedagogical agent. If a student's partner could not make it to a lab, the student was temporarily paired with another student while maintaining the originally assigned role.

Starting from the second week, students began each lab by taking a set of practice questions on their laptops. They received a link to a Qualtrics survey and were given 6-10 minutes to complete the questions, with the time varying based on the number of questions. Half of these questions were designed as a delayed posttest assessment to measure what students had learned from the previous week. The other half were designed as a pretest to assess what students might already know about the concepts they were going to be taught in the current lab. (Because week 2 was the first lab activity, all questions from that week were pretest questions.) Each week's beginning Qualtrics survey also included a confirmation of the student's assigned role (performer or observer), serving as a reminder for the students.A teaching assistant verified that the reported roles matched the experimental assignments and followed up via email with students if discrepancies occurred. Additionally, to ensure observers carried out their tasks correctly, they were required to upload their video recordings to the course website after each lab. A teaching assistant monitored these submissions and emailed students who failed to upload. In rare cases where recordings were lost due to technical issues, the assistant contacted the performer to confirm that the activities had been completed as intended. No concerns emerged from these communications regarding adherence to the assigned roles.

After taking the assessments, students engaged in a lesson that incorporated embodied activities in the forms of object manipulation, gesture, and drawing. The instructor provided detailed steps of instruction through powerpoint slides and Jupyter Notebooks for both the performer and the observer. In some cases, an earlier activity served as a foundation for an activity in a later week. For example, students cut out a piece of paper with a small data table and used the pieces of paper to "shuffle" a variable and "resample" observations (i.e., sampling with replacement). In a later lab session, students used what they had learned about shuffle and

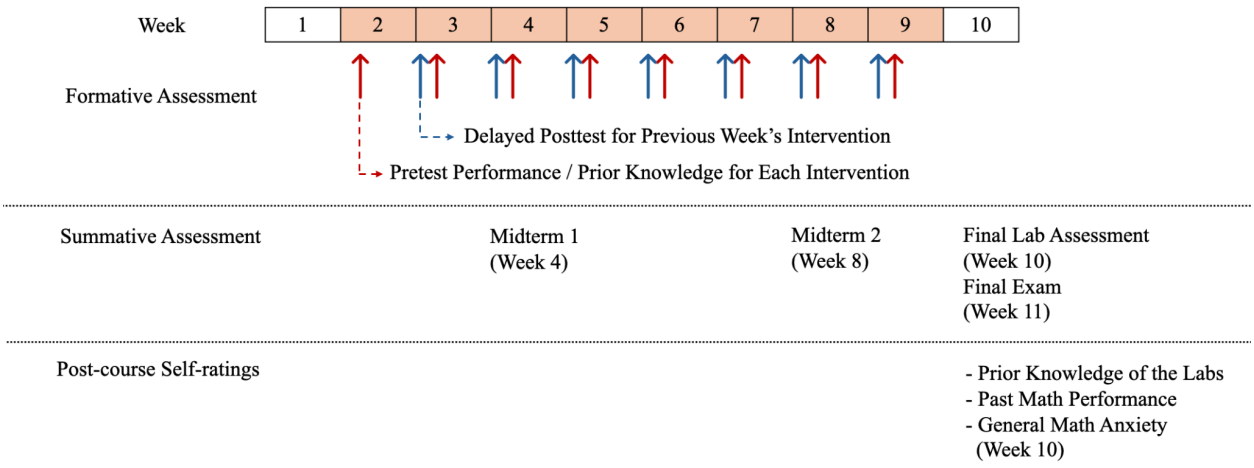
resample in a hands-on activity in which they constructed sampling distributions. Based on their randomly assigned condition, students either performed these activities or observed and recorded their partner performing them.

During each lab session, the lead experimenter and at least two trained learning assistants (with three assistants present in most cases) circulated the classroom to provide support. Students were encouraged to raise their hands if they had questions about the embodied activities, such as what to do or how to perform them. In addition to answering questions, the learning assistants closely monitored the activities and intervened when necessary—for example, if neither student or more than one student in a pair was performing the task, if an observer was disengaged (e.g., recording with a smartphone without attending to their partner's actions), or if a performer carried out the task incorrectly. By the second week, most students understood their roles, and interventions from learning assistants to stop observers from participating were rarely necessary. Although the supervision was not entirely foolproof, the learning assistants' support played a crucial role in maintaining the fidelity of the intervention. After the lab, students in the Observe condition uploaded the videos they took of the hands-on activities to their course assignment page as a confirmation of their lab attendance.

During the course, there were two midterm exams. The first midterm was administered in week 4 and the second midterm was administered in week 8. At the end of the course, participants completed their final lab assessment and a lab exit survey in the 10th week and took a final exam in the 11th week. Figure 3 illustrates when the assessments and ratings were administered.

Figure 3

Timing of Each Assessment



Note. The colored weeks are the weeks that they received an embodied intervention.

Materials

Each lab was taught using a combination of Jupyter Notebooks, PowerPoint slides, and worksheets. A description of the lab schedule and activities is included in Appendix A.

Embodied Pedagogies

The embodied activities in our intervention were not uniform but represented distinct forms of embodied pedagogies. These included (1) object manipulation, such as cutting and moving paper pieces to illustrate randomization and bootstrapping or shifting a physical sampling distribution to explore confidence intervals; (2) hand-drawing, where learners created durable, external visuospatial traces of statistical concepts, such as sketching the sum of squares for the empty model and the two-group model; and (3) gesture, in which learners used their hands to represent abstract operations in real time, such as gesturing to compare the empty model and two-group model.

Decisions about which embodied pedagogy to implement were informed by joint discussions among the experimenter, course instructor, and textbook author, as well as by

classroom feasibility and existing taxonomies of embodiment (e.g., Johnson-Glenberg et al., 2016; Skulmowski & Rey, 2018). These frameworks emphasize features such as task integration, congruency, and sensorimotor engagement, which we followed by designing actions that were tightly linked to the instructional goals in a conceptually congruent way.

We also drew on empirical findings highlighting the benefits of making actions concrete and contextually meaningful (Congdon et al., 2018; Nathan, 2012). When feasible, we prioritized object manipulation (e.g., in Weeks 2, 3, 4, and 9). For example, when teaching confidence intervals, we printed a sampling distribution on transparency paper to create a manipulable visual, enabling performers to physically move the distribution on a worksheet to determine the lower and upper bounds of the interval. When direct manipulation was not feasible (e.g., when teaching modeling-related concepts), we combined gesture- and drawing-based instruction. For example, performers drew the sums of squares for the total, model, and error in a two-group model and used gestures to reason about whether the data reflected a large or small sum of squares.

Measures

Formative Assessments (Practice Questions)

Prior knowledge Assessments. Half of the practice questions administered to students at the beginning of each lab served as a pretest of prior knowledge related to the day's activities. A complete list of the prior knowledge assessment questions can be accessed on the study's OSF page (https://osf.io/ntsr2/?view_only=ce650267f407451a9ea26abcb428d8f7).

Delayed Post-Test Assessment. The other half of the practice questions answered at the start of each lab were designed as delayed post-test assessments of concepts worked on during the previous week's lab. It is worth noting that each week's concepts were organized according to

the course schedule. Due to this nature, some weeks' concepts built off of an earlier week whereas others did not. (For a complete list of the questions see https://osf.io/ntsr2/?view_only=ce650267f407451a9ea26abcb428d8f7). The prior knowledge and delayed post-test assessments contained distinct sets of questions, except for the delayed post-test assessment in week 3 and the prior knowledge assessment in week 4. Since both weeks focused on simulating the data-generating process using the shuffle() and resample() functions, we used the full set of questions from the beginning of week 4 as both the delayed post-test for week 3 and the prior knowledge assessment for week 4. Because the weekly pretests targeted different concepts, they were not designed to establish measurement invariance over time. Thus, scores across weeks should not be interpreted as directly comparable indicators of a single construct. Instead, they provide snapshots of students' topic-specific knowledge at different points in the course.

The prior-knowledge assessment and the delayed post-test assessment were co-developed by the lead experimenter, the course instructor, and the author of *Coursekata* (the textbook). The assessments evaluated not only students' comprehension of the concepts taught in each lab but also their ability to apply these concepts in statistical reasoning and inference. While some questions incorporated visuospatial components, the assessments themselves were intentionally not embodied. This choice reflects the goal of ensuring that, although students learned the concepts through embodied experiences, they could flexibly apply their knowledge in non-embodied contexts.

Scoring of Formative Assessment Questions. Four trained coders graded participants' responses to the practice questions, with two coding each response as a means of assessing inter-rater reliability. All coding was conducted blind to condition.

The coding rubric for each week was collaboratively developed during a weekly meeting between the four coders and the lead researcher, informed by a preliminary analysis of a subsample of responses. Each question was scored out of one point, with half points awarded for responses that were correct but incomplete or contained minor misunderstandings. To ensure consistency, the coding team first reached a consensus on the essential components of a complete and accurate response for each question (i.e., the criteria for a full score of 1). They then randomly selected 20 student responses to determine whether a partial score of 0.5 was necessary for distinguishing different levels of understanding. If a partial score was deemed appropriate, the team established specific rubric categories for responses that would receive a 0.5.

If the discrepancy rate between two coders on a given question exceeded 20%, they met with the lead researcher to review and revise the coding rubric. Discrepancies arose primarily in three situations. First, during the initial two weeks, coders struggled to remain consistent when grading ambiguous responses. Some tended to be stricter, while others were more lenient. To calibrate expectations, weekly rubric meetings were used not only to refine the rubric but also to review sample responses together and establish shared benchmarks for leniency, supplemented with example gradings. Second, some discrepancies reflected limitations of the rubric itself. When criteria lacked sufficient detail (e.g., simply stating “demonstrates understanding of a two-group model”), coders often diverged in their judgments. This was especially common for questions initially scored on a binary 0/1 scale: some responses were correct but incomplete, prompting one coder to treat the omission as minor while the other considered it critical. To resolve this, the rubric was revised to specify required elements more explicitly or, when appropriate, introduce a 0.5 score for partially correct responses. Finally, disagreements sometimes stemmed from students’ use of course-specific terminology unfamiliar to the coders.

In such cases, the lead experimenter clarified the terms and expanded the rubric with alternative phrasings to capture the same concept. Following revisions to the rubric, coders recoded the responses, after which interrater reliability was assessed across all weeks using quadratic-weighted Cohen's kappa. This statistic accounts for partial agreement by giving more weight to larger disagreements (e.g., a rating of 1 vs. 0) and less to smaller ones (e.g., 1 vs. 0.5). The resulting kappa value indicated substantial agreement between coders, $\kappa = 0.79$. Week-by-week kappa values are provided in Appendix B.

Once interrater reliability was established, items with discrepancy rates below 20% were jointly reviewed by the two coders to resolve differences and arrive at a final score. For any cases where consensus could not be reached, which was extremely rare, the lead experimenter was brought in to facilitate resolution through a joint discussion. During these meetings, the lead first reviewed the rubric and student response without seeing the coders' scores, highlighting text corresponding to each rubric element. Each coder then presented their reasoning, after which the lead clarified where their thresholds diverged and mapped the discrepancies to the rubric to make a final determination. The coders were given the opportunity to discuss and confirm agreement with the decision, which did not raise further questions in these rare instances. These agreed-upon scores were used in the analyses.

Summative Assessments

Final Lab Assessment. The final lab assessment was given to students in-person in week 10 at the last lab they were required to attend. It was composed of 28 questions (17 open-response, five multiple-choice, and six drawing questions). It was structured as a paper and pencil test. The same four coders scored students' responses using the same grading approach.

Midterm One and Midterm Two Performance. Midterm One consisted of 20 questions (15 multiple-choice and five open-response/coding questions). The exam assessed students' mastery of content covered in Chapters 1 to 5 in the book.

Midterm Two had 25 questions (18 multiple-choice and seven open-response/coding), which assessed students' mastery of content covered in Chapters 1 to 9 in the book.

The two midterms were graded by the two teaching assistants for the course, blind to the experimental condition students were assigned to. Each question was worth one point, with half points given to open-response questions that were incomplete or partially correct.

Final Exam. The final exam consisted of 44 questions designed to assess concepts covered in the entire book (Chapters 1 to 12). Out of the 44 questions, 29 were multiple choice and 15 were open response or coding questions. The same two teaching assistants graded the final exam following the same approach as their grading of the midterms.

Self-ratings at the End of the Course

Self-rated Overall Prior Knowledge of Lab Content. We asked students, at the end of the course, to respond retrospectively to the question: "What percentage of concepts covered in the labs are concepts you already knew?" Students could choose 0%, 20%, 40%, 60%, 80% or 100%.

Past Math Performance. Previous math performance was self-rated by students at the end of the course on the same survey. Students rated how much they agreed with the statement, "My math/statistics marks were lower compared to other subjects", by choosing among five possible choices: strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree.

General Math Anxiety. Students' general math anxiety was measured at the same time as their past math performance, using the Single-Item Math Anxiety Scale (SIMA; Núñez-Peña

et al., 2014). Students responded to the question, "On a scale of 1 to 6 (1 being not anxious at all and 6 being extremely anxious), how math anxious are you?"

Analyses

We tested our main hypothesis with two types of analyses. Analysis of CoVariance (ANCOVA) was used to examine whether students' self-ratings (i.e., their general prior knowledge, past math performance, and math anxiety) moderated the effect of condition on the summative assessments (the two midterms, final lab assessment, and final exam). Specifically, we hypothesized that the perform pedagogy would benefit low-prior-knowledge learners more than high-prior knowledge learners and the observe pedagogy would be more beneficial for high-prior-knowledge learners than low-prior-knowledge learners. We also expected this interaction to be weak for the earlier summative assessments (e.g., midterm one) because the intervention had just begun, but stronger for assessments used later in the course. Lastly, we hypothesized that other self-ratings such as math anxiety would not show a significant interaction with condition.

Next, to investigate whether students' pretest performance moderated the effect of condition on the weekly formative delayed posttest assessments, we fitted a three-level multilevel model (MLM) with the weekly pretest and delayed posttests (i.e. the repeated measures data). The first level was the repeated measures nested within students. The second level was the students, who are nested in dyads (i.e. the third level). Below, we show the equation for the overall model:

$$Posttest_{ijk} = \gamma_{000} + \gamma_{010}Condition_{jk} + \gamma_{020}Pretest_{jk}^{b.cgm} + \gamma_{030}Condition_{jk} * Pretest_{jk}^{b.cgm} + \gamma_{100}Pretest_{ijk}^{w} + \gamma_{110}Pretest_{ijk}^{w} * Condition_{jk} + \gamma_{200}Time_{ijk} + \gamma_{210}Condition_{jk} * Time_{ijk} + u_{00k} + r_{0jk} + e_{ijk}$$

Table 1

Multilevel Model Fixed and Random Effects Summary

Model Term	Interpretation
Fixed Effects	
γ_{000}	Predicted posttest score for a student with 0 on all predictors. For example, the predicted posttest score in the first week for an observer with a level-1 pretest score that equals to the mean of his or her all pretest scores and a level-2 cluster mean (i.e., the average of one person's pretest scores across all weeks) that equals to the grand mean of all students' pretest scores
γ_{010}	Effect of condition (i.e., the mean difference of posttest scores between the performers and the observers) in the first week when the level-1 pretest score of a student is equal to the mean of this student's all pretest scores, and the level-2 cluster mean is equal to the grand mean of all students' pretest scores
γ_{020}	Effect of pretest score at the between-person (group-mean) level for the Observe group
γ_{030}	Level-2-interaction: does the effect of condition vary by each student's average pretest score? A slope difference for every one-point increase in the mean of one students' pretest scores.
γ_{100}	Effect of pretest score at the within-person (centered) level for the Observe group
γ_{110}	Cross-level-interaction: does the within-person pretest effect differ by condition? A slope difference for every one-point increase in the level-1 pretest scores.
γ_{200}	Effect of time (i.e. week) for the Observe group when the level-1 pretest score of a student is equal to the mean of this student's all pretest scores, and the level-2 cluster mean is equal to the grand mean of all students' pretest scores
γ_{210}	Does the effect of the intervention differ each week based on content? A slope difference for every one-point increase in week.
Random Effects	
u_{00k}	Random intercept for dyad/pair; captures residual between-dyad effect
r_{0jk}	Random intercept for individual; captures residual between-individual effect
e_{ijk}	Residual error at each measurement occasion

Table 1 presents an interpretation of each coefficient in the equation above. Notably, we partitioned the variation of students' pretest performance into $Pretest_{jk}^{b.cgm}$, which stands for the mean of each student's pretest performance (i.e. the between-person variation; level-2) and $Pretest_{ijk}^{w}$, which stands for how each student's pretest performance varied from week to week (i.e. the within-person variation; level-1). The equations for each level are fully presented in Appendix G. It is important to note that although our analytic framework included "time" (week) as a variable, the models were not specified as growth curve models. Because each weekly assessment targeted a different statistical concept (e.g., randomization, two-group model), they do not constitute repeated measures of a single construct. The time variable was included solely to account for the possibility that the intervention effect might differ depending on the week and corresponding content.

Based on our hypothesis, we were particularly interested in γ_{030} and γ_{110}. Because the variation in pretest performance has both level-1 and level-2 variation, the interaction between pretest performance and condition (perform versus observe) was partitioned into the interaction between condition and each individual's average pretest performance (i.e., γ_{030}), which we will refer to as the level-2 interaction, and the interaction between condition and the variation within each individual's pretest performance (γ_{110}), which we will refer to as the cross-level interaction. We planned to test for both the level-2 interaction and the cross-level interaction. Because students' understanding of each week's concept would naturally fluctuate (e.g. a student with a high prior knowledge of week 3's concept might have a low prior knowledge of week 6's concept), we specifically expected a significant cross-level interaction between condition and each individual's pretest performance.

Missing Data

Each student was assigned a partner ID to identify their lab partner. While the majority of students retained the same partner throughout the study, in instances where a student had more than one lab partner they were assigned the partner ID of the student with whom they were paired most frequently. When students missed a lab, or if their partner was switched during a lab, their data for that week (i.e., their corresponding pre and delayed posttest scores for that week) was treated as missing. Because the quality of interaction might have changed after a partner switch, we treated one student's pre and delayed posttest scores as missing when there was a partner switch. Students who withdrew from the class midway, were not included in any data analyses.

The two most frequent reasons for missing a lab were sickness and family emergency. But there might be other reasons too. When the reasons for missingness were not entirely certain (which is most of the time in longitudinal classroom research), it is theoretically possible that unobserved test scores were related to the likelihood of missingness (e.g., low-achieving students choosing not to participate in the lab due to a lack of sufficient knowledge or motivation to participate). If the probability of missing labs depended on students' prior scores (which were observed), but not directly on their actual lab performance (i.e. the pre and delayed test scores), then the missingness was conditionally missing at random (MAR) given prior scores. Thus, we employed missing data-handling procedures based on the assumption that the missingness mechanism is conditionally MAR, meaning that once a student's observed data are accounted for, the likelihood of missing values does not depend on the unobserved data (Graham, 2009). Missing data for the multilevel analysis were imputed using the Markov Chain Monte Carlo (MCMC) algorithm as implemented in the Blimp application (Enders, 2017). For additional

context on the Bayesian imputation using Blimp, readers are encouraged to consult Enders (2022).

Transparency and Openness

We provide details on our sample size determination, any data exclusions, all experimental manipulations, and all measures used in the study, following JARS guidelines (Kazak, 2018). All data, analysis code and syntax, and research materials are publicly available through the Open Science Framework at https://osf.io/ntsr2/?view_only=ce650267f407451a9ea26abcb428d8f7. Data analysis was conducted using R (version 4.0.0; R Core Team, 2019) with the supernova package (version 3.0.0; Blake et al., 2024). This study was not pre-registered, and its design and analysis were not formally documented in advance.

Results

Analyses of Summative Assessments

The sample size for this part of the analysis is 217 (i.e., the number of students who completed the post survey), except for the analysis using the final lab assessment as the outcome variable. Because the final lab assessment was given in-person in their last structured lab session and some students missed the last lab session, there were 22 more missing cases for the final lab assessment (N = 195). Table 2 shows the descriptive statistics of students' performance on the summative assessments by condition.

Table 2

Descriptive Statistics of Summative Assessments by Condition

Assessment	Condition	N	Mean	SD

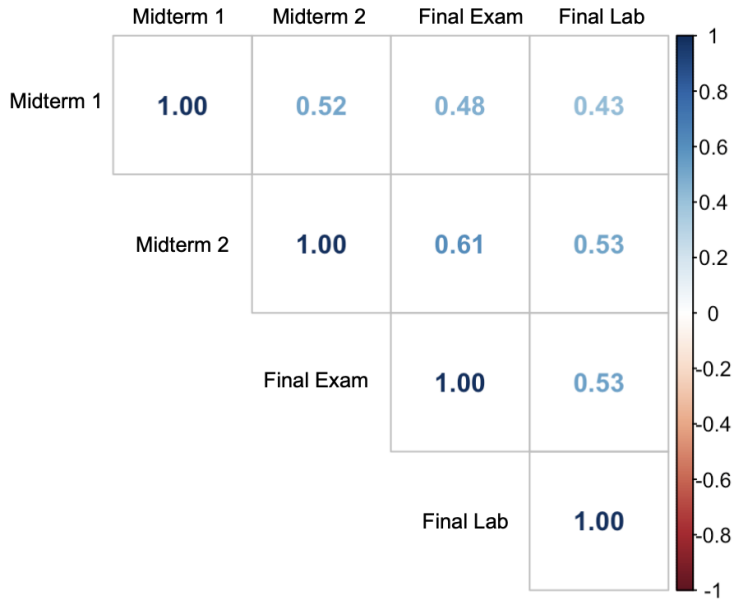
Midterm 1	Perform	110	95.35	7.05
	Observe	107	94.41	8.61
Midterm 2	Perform	110	89.44	14.00
	Observe	107	89.43	9.63
Final exam	Perform	110	89.74	8.62
	Observe	107	89.94	10.46
Final lab	Perform	97	23.94	6.87
	Observe	98	22.89	8.06

Psychometric Evidence of the Summative Assessments

Figure 4 shows the correlation between the summative assessments. The correlation between the summative assessments are moderately strong, indicating some consistency of students' performance on the assessments. We also evaluated the internal consistency of the summative assessments in measuring students' overall performance in the course using Cronbach's alpha, which showed a high level of reliability ($\alpha = 0.81$).

Figure 4

Correlation Between the Summative Assessments



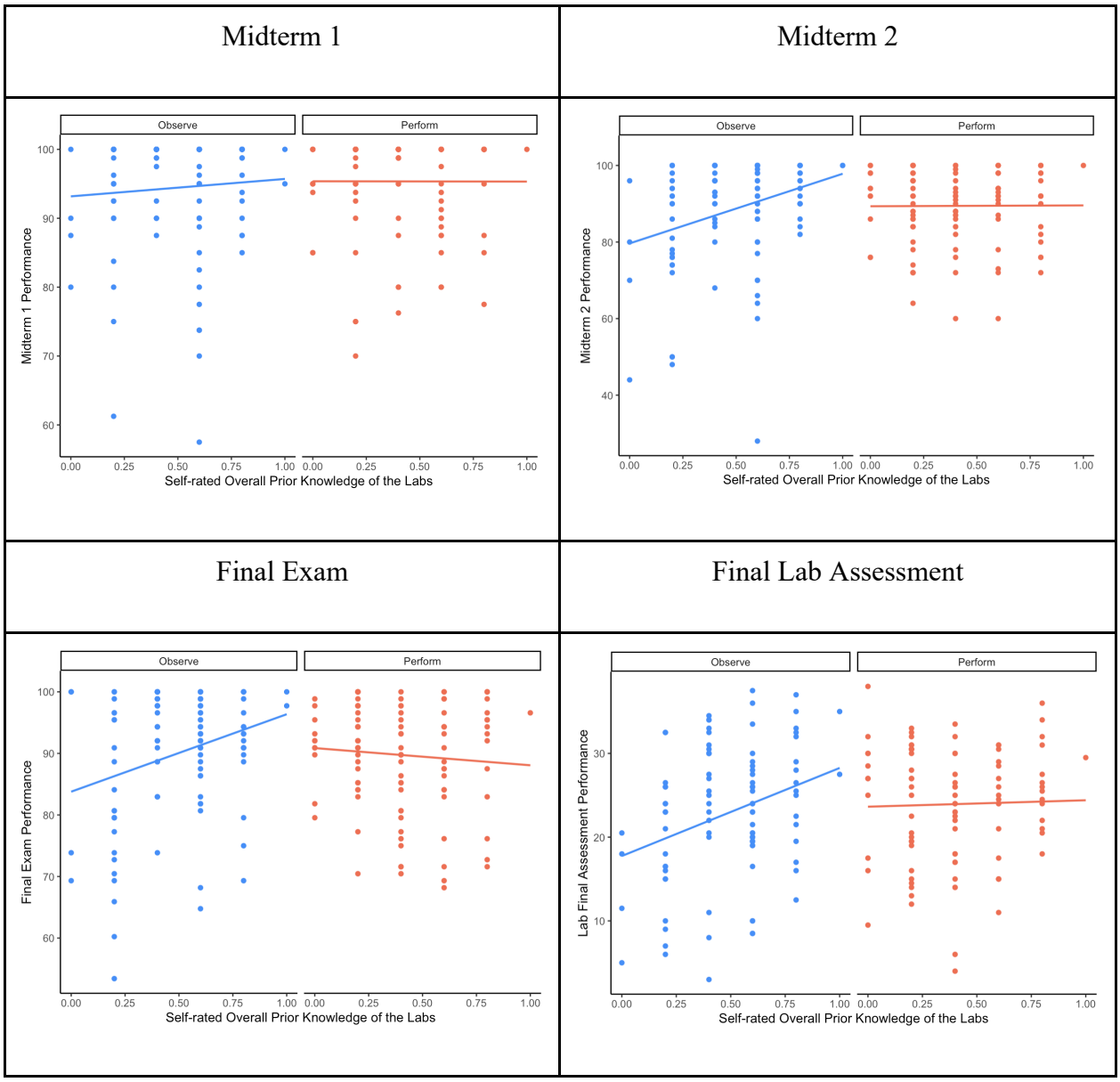
Interactions of Prior Knowledge and Condition

There were no significant main effects of condition on any of the summative assessments (midterm 1, midterm 2, the final exam, or the final lab assessment). This result was not changed by controlling for students' self-rated overall prior knowledge of the interventions (see Appendix C for the complete results).

Figure 5 shows the relationship between condition and self-rated prior knowledge predicting midterm 1, midterm 2, final exam, and final lab assessment. With the exception of midterm 1, all three other graphs showed a similar pattern of results: in the Observe condition, students' self-rated prior knowledge showed a positive correlation with their performance on the assessment, but such a correlation was not evident in the Perform condition.

Figure 5

The Relationship Between Condition and Self-Rated Prior Knowledge for Each of the Four Summative Assessments



We tested whether there was a significant interaction between prior knowledge and condition for each of the four outcome variables using ANCOVA. Statistical analysis was carried out using R (R Core Team, 2019). Due to students dropping the class or not attending the session, four participants were removed from the analysis for midterm 1, midterm 2, and the final exam, and 26 participants were removed from the analysis for the lab final assessment.

Table 3

ANCOVA Outputs Using Each Summative Assessment As the Outcome Variable

Outcome variable	*predictor*	*Estimates*	*df*	*F*	*CI*	*p*
Midterm 1	(Intercept)	93.18			89.67 – 96.68	**<0.001**
(N = 217)	condition	2.19	1	0.926	-2.30 – 6.69	0.337
	pre	2.53	1	0.594	-3.94 – 8.99	0.442
	condition*pre	-2.57	1	0.334	-11.36 – 6.21	0.564
Midterm 2	(Intercept)	79.64			74.46 – 84.83	**<0.001**
(N = 217)	condition	9.70	1	8.281	3.06 – 16.35	**0.004**
	pre	18.19	1	14.047	8.62 – 27.75	**<0.001**
	condition*pre	-17.97	1	7.430	-30.96 – -4.97	**0.007**
Final Exam	(Intercept)	83.78			79.61 – 87.95	**<0.001**
(N = 217)	condition	12.58	1	6.821	4.89 – 20.27	**0.001**
	pre	7.08	1	10.391	1.74 – 12.43	**0.010**
	condition*pre	-15.37	1	8.410	-25.82 – -4.92	**0.004**
Final Lab	(Intercept)	17.75			14.35 – 21.15	**<0.001**
(N = 195)	condition	10.51	1	7.120	4.23 – 16.80	**0.001**
	pre	5.89	1	10.882	1.54 – 10.24	**0.008**
	condition*pre	-9.74	1	5.048	-18.30 – -1.19	**0.026**

Note. pre refers to students' self-rated prior knowledge of the labs.

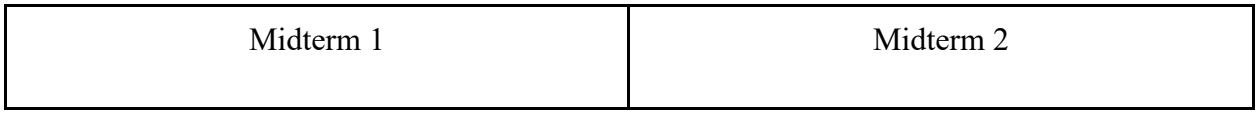
The ANOVA results using each summative assessment as the outcome variable is summarized in Table 3. Importantly, there was no significant interaction between self-rated prior knowledge of the labs and condition for midterm 1 ($F(1, 213) = 0.33$, $p = .564$). However, there was a significant interaction between students' self-rated prior knowledge and condition for

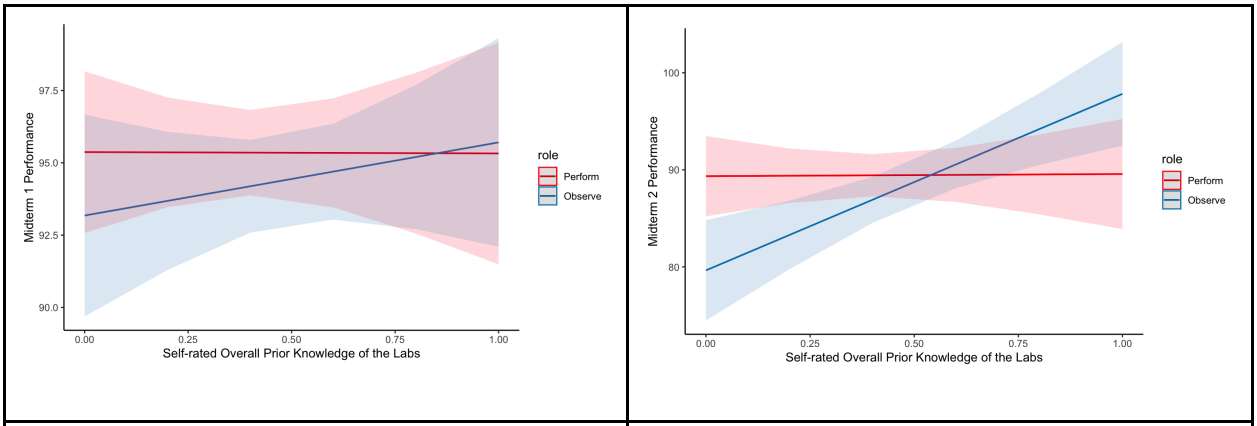
midterm 2 ($F(1, 213) = 7.43$, $p = .007$), the final exam ($F(1, 213) = 8.41$, $p = .004$), and the final lab assessment ($F(1, 191) = 5.05$, $p = .026$). A complete table of results for the four models is included in Appendix D.

Model predictions and 95% confidence intervals (CIs) are plotted in Figure 6, a suggested way to interpret significant interactions by Garofalo et al. (2022) and other researchers in the field (e.g. Rosnow, 1989). From the graphs, we can see that for the three models where we detected significant interactions, the red and blue lines cross each other at approximately the 50% rating of prior knowledge. For students with lower self-rated prior knowledge (lower than or equal to 20%), those in the Perform condition showed reasonable evidence of superior learning in midterm 2, the final exam, and the final lab assessment than those in the Observe condition. This is evidenced by the fact that less than 25% of the shaded 95% CI for Perform model predictions overlaps with the 95% CI for the model predictions for the Observe group (Cumming & Fidler, 2009). Similarly, for students with higher self-rated prior knowledge (higher than or equal to 80%), in contrast, those in the Observe condition achieved significantly better learning outcomes than students in the Perform condition (this difference is less evident in the final lab assessment).

Figure 6

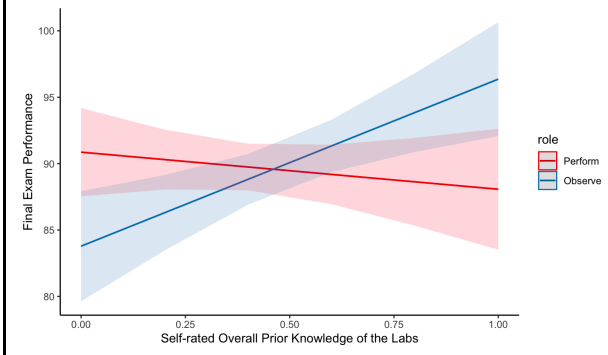
Model by Self-Rated Prior Knowledge and Condition for each Summative Assessment





Final Exam

Final Lab Assessment



Given that prior knowledge often predicts posttest performance in learning contexts, we explored whether this relationship would differ between the Perform and Observe conditions—specifically, whether the correlation would be weaker or absent in the Perform group, and whether the slopes of the two groups would significantly differ. We conducted exploratory analysis using simple slope analyses to examine the correlation between pretest and each summative assessment at the two levels of condition and organized the results in Table 4. The results revealed that, except for midterm 1, when students were in the Perform group, the relationship between self-rated prior knowledge and the assessment was not significant. However, when students were in the Observe group, self-rated prior knowledge significantly

predicted performance on the assessment, as shown by the lower bond of the 95% confidence CI being larger than 0.

Table 4

A Summary of the Simple Slope Analysis

Assessment	Condition	Slope	SE	df	Lower CI	Upper CI
Midterm 1	Perform	-0.05	3.02	213	-5.99	5.90
	Observe	2.53	3.28	213	-3.94	8.99
Midterm 2	Perform	0.22	4.46	213	-8.57	9.01
	Observe	18.19	4.85	213	8.62	27.75
Final exam	Perform	-2.79	3.59	213	-9.86	4.28
	Observe	12.58	3.90	213	4.89	20.27
Final lab	Perform	0.77	2.94	191	-5.03	6.57
	Observe	10.51	3.19	191	4.23	16.80

Note. CI is at 95%.

Interactions of Past Math Performance and Condition

In none of the summative assessments was the interaction between condition and self-rated past math performance significant ($p > .05$). The visualizations for relationship between self-rated past math performance and the four summative assessments broken down by condition are shown in Appendix E.

Interactions between Math Anxiety and Condition

There was in general a negative relationship between math anxiety and performance regardless of the condition. We did not find any significant interaction between math anxiety and

condition for any course assessment ($p > .05$). The visualizations for the relationship between students' self-rated math anxiety and each summative assessment separated by condition are shown in Appendix F.

Analyses thus far provide support for the Performing First hypothesis. Students who rated themselves as lower in prior knowledge for the lab activities benefited more from performing relative to observing than did students with higher prior knowledge. But these analyses measure prior knowledge using students' self-ratings. We next use multilevel modeling to see if the same effects can be spotted using formative assessment data, collected prior to each lab, as a more targeted and objective measure of prior knowledge.

Analyses of Formative Assessments

Raw Scores and Baseline Equivalence

The raw weekly data's mean and standard deviation is shown in Table 5. We examined students' scores in the pre-intervention assessment in week 2, before they were randomly assigned to conditions or received any intervention to make sure there were no significant pre-existing differences of understanding. Because the distribution of students' scores in week 2's pretest was skewed to the left (Figure 7), we performed Mann-Whitney U test using the stats package in R (version 3.6.2; R Core Team, 2019) to determine whether there is a difference in students' pre-intervention performance scores between students who were in the Perform condition and those in the Observe condition. The results indicated no significant difference in the pre-intervention assessment in week 2 between students in the Perform group and the Observe group, W = 5637, p = .659.

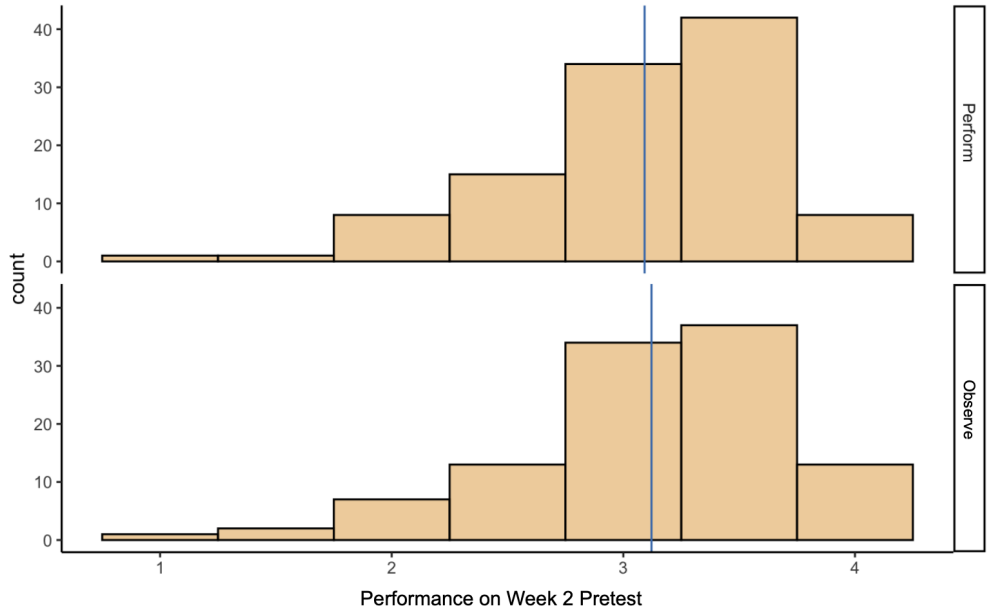
Table 5

The Mean and SD of Each Week's Pretest and Posttest Scores

Week	Type	N	Overall Mean	Overall SD	Condition	Group mean	Group SD
2	Pretest	216	3.11	0.59	Perform	3.09	0.57
					Observe	3.12	0.61
2	Posttest	200	2.61	1.25	Perform	2.59	1.13
					Observe	2.62	1.37
3	Pretest	200	2.88	1.14	Perform	2.93	1.09
					Observe	2.83	1.2
3	Posttest	209	2.13	1.31	Perform	2.01	1.26
					Observe	2.24	1.35
4	Pretest	209	2.13	1.31	Perform	2.01	1.26
					Observe	2.24	1.35
4	Posttest	215	2.93	2.02	Perform	2.98	2.07
					Observe	2.89	1.98
5	Pretest	215	2.14	1.18	Perform	2.15	1.26
					Observe	2.13	1.1
5	Posttest	207	4.92	1.93	Perform	4.92	2.04
					Observe	4.91	1.82
6	Pretest	207	4.7	2.23	Perform	4.43	2.12
					Observe	4.98	2.32
6	Posttest	208	3.61	1.59	Perform	3.72	1.61
					Observe	3.5	1.57
7	Pretest	208	1.52	1.19	Perform	1.53	1.21
					Observe	1.5	1.17
7	Posttest	205	1.15	0.81	Perform	1.22	0.72
					Observe	1.08	0.87
8	Pretest	205	3.01	1.91	Perform	2.83	1.85
					Observe	3.17	1.96
8	Posttest	204	2.45	1.48	Perform	2.52	1.49
					Observe	2.37	1.48

Figure 7

The Distribution of Students' Performance on Week 2 Pretest by Condition

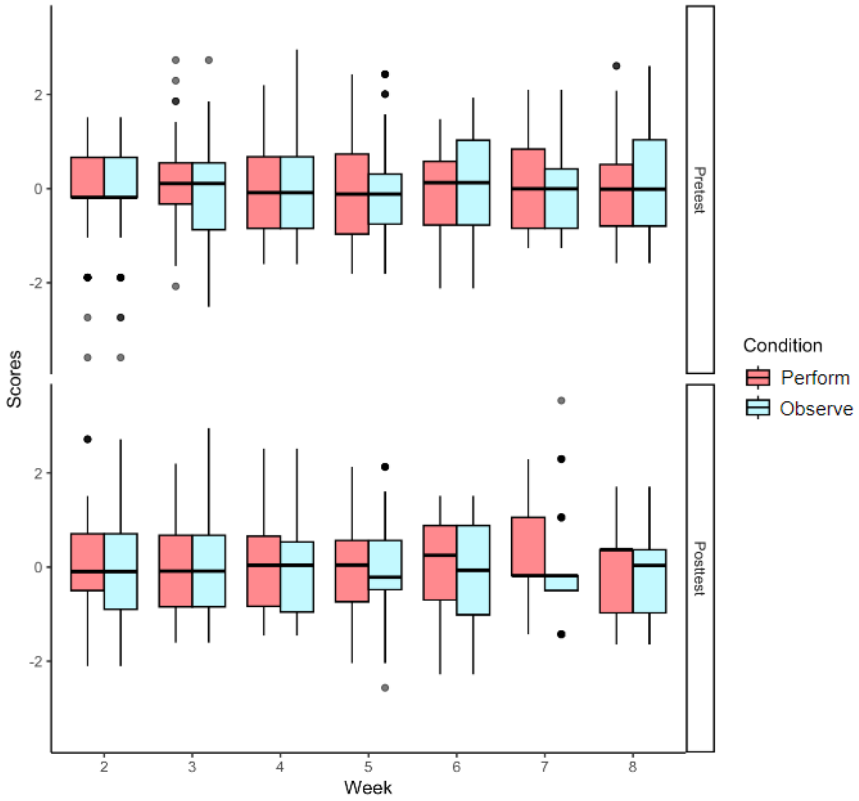


Preprocessing and Plotting the Data

Because each week's pre- and post-test questions might differ in difficulty, each of the eight pretests and eight posttests were standardized so that the average score on each was 0. We plotted the standardized scores for each week to explore the differences between the two conditions in each week (Figure 8). From the graph, we can see that the Perform and Observe groups did not differ from each other on pretest questions. They also did not differ a lot from each other on the post-tests in the early weeks, but starting in week 5, the Perform group seemed to score a bit higher than the Observe group.

Figure 8

Standardized Pretest and Posttest Scores for each Week by Condition

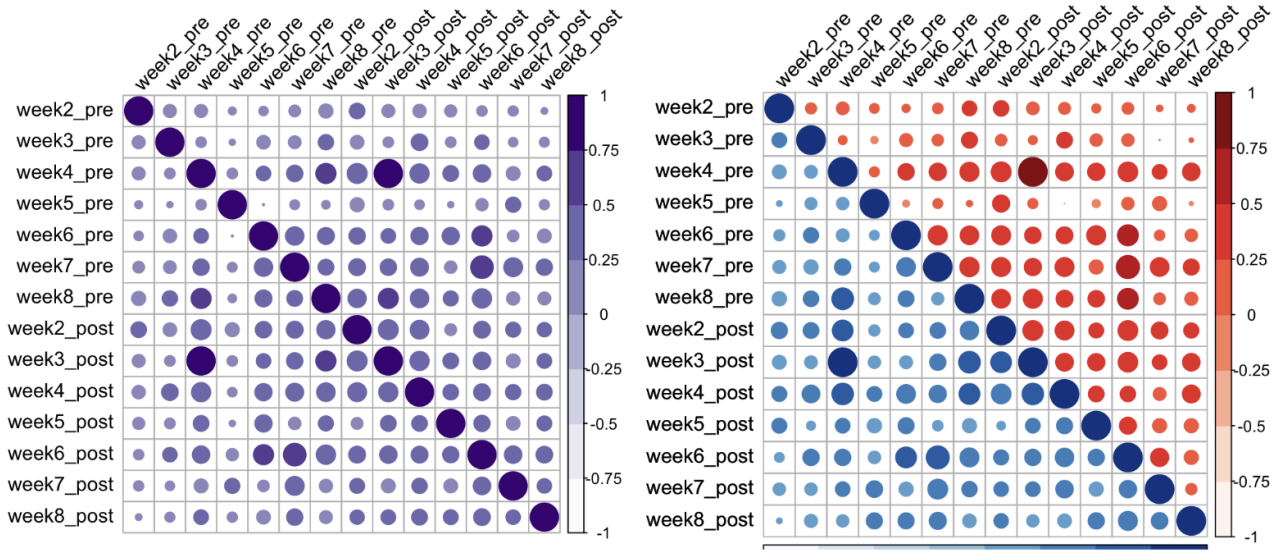


Psychometric Evidence of Weekly Tests

We plotted the overall correlation between each test (left of Figure 9) and the correlation separated by condition (right of Figure 9). The red plot represents the correlation in the Perform group and the blue plot represents the correlation in the Observe group. In general, the correlation between the prior knowledge assessment and delayed posttest assessment was stronger in the Observe group than the Perform group, as shown by the right graph having larger circles in the bottom left corner than the top right corner. In addition, the internal consistency of the assessment was examined using Cronbach's alpha, which showed a high level of reliability ($\alpha = 0.84$).

Figure 9

The Correlation Between Each Test (Left) and the Correlation Separated by Condition (Right).



Note. For the right graph, the red plot is for the Perform group and the blue plot is for the Observe group.

Multilevel Model

Given that the pre-test performance and post-test performance were not just collected once but repeatedly, there were three levels of analysis (i.e., repeated measures, student, and dyad). The sixteen repeated measures were nested within students and students were nested in dyads. Thus, we calculated the intraclass correlation coefficient (ICC) first for the post-test performance at each level followed by a similar analysis of pre-test performance. 68.8% of the total variability in post-test performance is due to repeated measures (level-1). Between-person mean differences (level-2) accounted for 23.8% of the variability in post-test performance. In other words, the expected correlation between two post-test scores of the same student is 23.8%, which is considered large in educational datasets (Hedges & Hedberg, 2007). Lastly, between-dyad mean differences (level-3) accounted for 7% of the variability in post-test performance.

While this percentage may seem relatively small, it was enough to warrant the inclusion of a third level for dyads in the model (Musca et al., 2011).

Similarly, for pre-test performance, repeated measures accounted for 77.6% of the total variability in pre-test scores. Between-person mean differences (level-2) accounted for 16.3% of the variability in pre-test performance. Lastly, between-dyad mean differences (level-3) accounted for 5.7% of the variability in pre-test performance.

Missing Data. Missing data analysis and imputation was conducted using the Blimp application. Overall, the weekly pretest scores had 8.4% missing data and delayed posttest scores had 9.2% missing. Specifically, 85% of the level-1 scores were complete. Out of the missing level-1 scores, 5.9% were missing delayed posttest scores only, 5.1% were missing pretest only, and 3.3% were missing both.

Interaction between Pretest and Condition

We used the MCMC algorithm to fit the model using the Blimp application (Enders, 2017). Bayesian estimation treats parameters as random variables and employs MCMC to sample parameters from a series of posterior distributions. Recent computer simulation results have demonstrated this approach using MCMC to be especially effective for multilevel models with random coefficients and interaction effects (Enders et al., 2020). MCMC estimation (2 chains, 20,000 iterations per chain, 10,000 burn-in) showed good convergence ($\hat{R} < 1.05$). Table 6 shows the coefficients of the output. As a reminder, the pretest performance was partitioned into two variables, separately capturing the between-people variability and within-person variability in pretest performance. There was no significant level-2 interaction between condition and pretest performance (median = 0.09, 95% credible interval = [-0.15, 0.34], but there was a

significant cross-level interaction between condition and pretest (median = -0.12, 95% credible interval = [-0.22, -0.01]). We visualized the cross-level interaction in Figure 10.

Figure 10

The Cross-level Interaction between Condition and Each Students' Variation in Pretest Performance

The significant expected difference between within-person slopes for an observer and a performer of -0.12 means that the correlation between pretest and posttest scores of the same person was weaker for performers. The Bayesian credible interval indicates that there was a 95% probability that the true difference between within-person slopes for someone in the Observe condition versus the Perform condition would lie within [-0.22, -0.01].

Table 6

Results from MLM

Variances	Estimates	StdDev	2.5% CI	97.5% CI	N effective

Level 2 Variance (intercept)	0.006	0.008	0.000	0.029	109.956
Level 3 Variance (intercept)	0.009	0.010	0.001	0.037	143.922
Residual Variance	0.679	0.027	0.629	0.734	10926.904
Coefficients	Estimates	StdDev	2.5% CI	97.5% CI	N effective
Intercept	0.039	0.096	-0.150	0.228	1286.796
Condition.Exp	-0.100	0.129	-0.350	0.154	2411.983
Pretest	0.107	0.037	0.034	0.181	9133.976
Time	-0.022	0.015	-0.052	0.008	14841.879
Pretest.mean[ID]	1.062	0.139	0.794	1.342	719.325
Pretest.mean[Dyad]	-0.003	0.312	-0.593	0.631	552.061
Condition.Exp*Pretest	-0.114	0.053	-0.218	-0.012	9510.082
Condition.Exp*Time	0.044	0.022	0.001	0.086	14499.221
Condition.Exp*Pretest .mean[ID]	0.178	0.187	-0.180	0.556	1170.244
Condition.Exp*Pretest .mean[Dyad]	-0.270	0.417	-1.103	0.549	679.409

* CI stands for the credible interval.

We next probed the significant cross-level interaction at the two levels of condition. For students in the Observe condition, the slope coefficient between pretest performance and delayed posttest performance was positive and statistically significant (median = 0.11, 95% credible interval = [0.03, 0.18]). This finding suggests that for students in the Observe condition, if they scored higher in that week's pretest, they were predicted to also score higher in the delayed posttest of that week's intervention. However, for students in the Perform condition, the slope coefficient between pretest performance and delayed posttest performance was not statistically significant (median = -0.01, 95% credible interval = [-0.08, 0.07]).

Discussion

In the current project, we designed and implemented an embodied lab curriculum in a college-level introductory statistics course taught using R programming. Central to our endeavor was to develop and test a working theory that began to address three crucial gaps between embodied learning research and practice. In two sets of analyses, we demonstrated the pivotal role of prior knowledge in moderating the effect of performing versus observing actions and gestures. First, we found that learners' self-ratings (at the end of the course) of their overall prior knowledge of the concepts covered in the labs significantly moderated the effect of condition on their midterm 2, final exam, and final lab assessment performance. Specifically, when students had low prior knowledge, performing was more effective than observing; when students had high prior knowledge, observing was more effective than performing. Additionally, the slopes for the correlation between self-rated prior knowledge and these three summative assessments were significantly positive for students who observed the embodied activities and were non-significant for students who engaged in direct physical performance.

We did not find self-rated prior knowledge to moderate the effect of condition on the first midterm, for which there are three potential explanations. First, there may be a ceiling effect in students' performance in midterm 1. Second, the overall prior knowledge rating may not accurately capture students' prior knowledge in the first weeks of the course. Third, and more related to the investigation of why embodied pedagogies are important in learning, is that the absence of this moderating effect can possibly be attributed to the students having experienced only two lab sessions by that point.

Students might need to practice performing actions and gestures more than two times in order to construct mental representations of concepts that are robust enough to be connected to other concepts, an argument supported by the practicing connections hypothesis (Fries et al., 2021). In weeks 2 and 3 (i.e., the two weeks before the midterm) students learned to enact or observe hands-on shuffling versus resampling of a dataset. In the later weeks of the course, students used what they had learned about shuffle and resample to engage in hands-on activities related to creating and manipulating sampling distributions. In this way, the mental representations they acquired in earlier weeks were further developed and connected. This is one of the strengths of conducting an intervention longitudinally because such an approach allows us to detect effects that would not have been detected in an intervention that lasted only two sessions.

We also found that other factors, such as students' previous mathematics performance and mathematics anxiety, did not significantly moderate the effect of condition. This suggests that what matters is not whether learners had more positive experiences with mathematics in the past, but whether they came into the course with already-developed mental representations related to the concepts that were the focus of the intervention. Past literature has also highlighted

how actions can lead to conceptual insights and the development of more robust mental representations (Brooks et al., 2018; Broaders et al., 2007; Cartmill et al., 2012; Goldin-Meadow et al., 2009). Our findings contribute to this literature by positing a unique role of learners' prior knowledge and putting forward the Performing First Hypothesis that serves as a framework for such investigations.

Still, it is worth noting that students answered this question retrospectively. This retrospective pretest methodology has been extensively used and tested in instructional practices research (e.g. Lam & Bengo, 2003). But, when it comes to students' actual content knowledge, retrospective self-ratings of knowledge can be very different from their actual understanding of concepts. For example, Moore and Tananis (2009) found that students often overestimate their pre-test understanding of core concepts to be taught. In addition, although the potential confounding variables we measured did not moderate the effect of condition, it is possible that students' rating of prior knowledge is impacted by other experience in the intervention instead of the other way around. Thus, we urge our audience to view this finding in terms of students' self-rating of the percentage of concepts covered in the labs that they already knew.

Because there must be differences between this self-rating and actual understanding, we conducted another set of analyses with the MLM, which used students' performance on each week's pretest to measure their prior knowledge. The three-level MLM revealed a significant cross-level interaction between the type of embodied intervention and learners' prior knowledge. Specifically, when learners observed hands-on activities, there was a significant correlation between their prior knowledge and their performance on a delayed posttest. Conversely, when learners engaged in physical activities themselves, this correlation between prior knowledge and posttest performance was not evident. This suggests that while prior knowledge typically

predicts post-intervention performance, active participation in physical tasks disrupts this predictive relationship, possibly due to the different cognitive processes involved in physical versus observational learning.

Taken as a whole, the reduced influence of prior knowledge when learners actively perform embodied tasks delivers an important message for educators. In diverse classrooms, relying solely on teacher demonstrations risks widening existing gaps: students with higher prior knowledge are more likely to benefit, while those with less prior knowledge may struggle to keep up. By contrast, when all students physically engage in the activities themselves, prior knowledge plays a smaller role in shaping outcomes. Hands-on participation appears to "level the playing field," giving students with lower prior knowledge a better chance to catch up and narrowing disparities between novices and more advanced learners. From an equity perspective, this suggests that when learners possess relatively low prior knowledge of the content, classroom embodied pedagogies might be more inclusive if they extend beyond teacher modeling to include active, hands-on performance by students.

The findings together support the Performing First Hypothesis, which hypothesized that prior knowledge of the concepts to be learned—and not other factors—would moderate the impact of varying levels of embodiment (i.e., performing versus observing) on learning outcomes. It is important to acknowledge that, although the mechanisms proposed in the Performing First Hypothesis offer a plausible explanation for our findings, they remain speculative without direct process data. To fully clarify how performing versus observing benefits different learners, future research must track how factors such as cognitive load, the quality of sensorimotor grounding, and connection-making evolve over time. Rather than presenting definitive conclusions, we view the current framework as an agenda-setting

opportunity: it identifies critical variables that longitudinal and experimental studies can investigate to advance a mechanistic understanding of how embodied pedagogies support expertise development. For example, future work could develop methods to measure learners' cognitive load during instruction and test its potential mediating role. Future work could also deliberately seek evidence to further elucidate this mechanism, for instance, by exploring the type and quality of mental representations formed under various embodied and non-embodied pedagogical approaches among diverse learners. Past research has revealed mixed evidence pertaining to the mental representations underlying embodied pedagogies (Alibali et al., 2000; Wagner et al., 2004). Preliminary evidence suggests that more embodied pedagogies foster enhanced visuospatial representations of the concepts taught compared to both less embodied and non-embodied approaches (Zhang et al., 2024b). These mental representations appear to mediate the influence of differing pedagogical strategies on students' learning outcomes.

The findings from the current study also suggest new directions for embodied learning research. Many studies have demonstrated that embodied interventions (i.e., interventions that leverage bodily experience in some form) benefit learning of a single concept when compared with abstract instructions that do not involve the body (ohnson-Glenberg & Megowan-Romanowicz, 2017; Trudeau & Dixon, 2007; Zhang et al., 2022). The current study suggests more nuanced questions of what and when - what type of embodied pedagogies are the most effective, and when are they the most effective? Answering these new questions calls for more longitudinal designs and curriculums that could afford these designs.

In the current longitudinal study, we focused on the distinction between performing and observing hands-on activities, because this would offer practical insights for teachers and other practitioners working in the classrooms. We hypothesized that there is something inherently

unique about “performing” activities, a concept that is supported by classical developmental psychology. Piaget, for example, highlighted the sensorimotor stage as foundational for later higher-order thinking and abstract concept processing (Piaget, 1983; Piaget & Inhelder, 1969). While beginning with a clear distinction between performing and observing provides a useful framework, we acknowledge that rigidly adhering to this dichotomy may oversimplify the complex nature of embodied pedagogies. We also recognize the potential value in exploring other types of embodied pedagogies, including whole-body movements, mixed or augmented reality manipulations, and directed actions, as these too may offer significant educational benefits (e.g., Johnson-Glenberg et al., 2017; Zhang et al., 2021). We encourage future studies to leverage this initial distinction to further investigate the relationship between prior knowledge and other forms of embodied pedagogies to fully elucidate the question of “what” types of pedagogies and “when” in learners’ knowledge development would embodied pedagogies be beneficial. Although a binary comparison of non-embodied versus embodied interventions has been studied and proven effective in lab settings, future studies could also incorporate a non-embodied control group to enable conclusions about the broader added value of embodiment compared to traditional, non-embodied instruction and allow for a more comprehensive analysis of the effectiveness of different types of embodied pedagogies. For example, comparing the physical and observational engagement inherent in embodied conditions with more abstract instructional methods, such as diagrams or written explanations, in a longitudinal setting could help elucidate the unique contributions of embodied pedagogies to learning over time.

We also note that the Observe condition in this study was based on peer modeling, a specific form of observational learning in which students watch a peer perform a task. Prior research has shown that peer modeling can positively impact learning (Eghterafi et al., 2022;

Schunk, 1987; Ledford & Wolery, 2015). However, peer modeling differs from expert modeling, in which students observe an instructor or expert demonstrate a task. This distinction is important because it may influence how observing supports learners with different levels of prior knowledge. In our study, higher-prior-knowledge learners benefited more from observing their peers than lower-prior-knowledge learners. By contrast, a substantial body of cognitive load research suggests that observing experts particularly helps novice learners, since expert demonstrations—whether through written worked examples (van Gog et al., 2004) or dynamic formats such as video modeling and live demonstrations (Braaksma et al., 2004; Wouters et al., 2009)—reduce extraneous cognitive load and thereby support novice learning (Paas & van Merriënboer, 1994; Sweller et al., 2011). It is therefore possible that if our comparison had involved expert modeling rather than peer modeling, the results might have differed. This tension between CLT's traditional predictions and our peer-modeling design is not sufficiently explored in the current study and points to a promising avenue for future research.

At the same time, the peer modelling in this study was distinctive in that students in the Observe condition were not only watching their partner but were also tasked with recording their partner's actions, creating a unique kind of observation and peer interaction that might be misaligned with other forms of observational learning. Unlike traditional observational learning contexts—where observation is directed primarily toward one's own learning—our design required students to document their partner's actions for the stated purpose of supporting the development of a pedagogical agent. Even though observers were not required to record continuously, this demand may have altered the way they attended to the task, introducing mechanisms different from those assumed in prior observational learning studies.

Still, because prior knowledge is typically a strong predictor of posttest performance, it is unlikely that the positive slope observed between pretest and posttest in the Observe condition can be attributed solely to this documenting requirement. It is possible, however, the documenting task may have magnified the role of prior knowledge—allowing high-prior-knowledge learners to benefit while making it harder for low-prior-knowledge learners to learn—thus contributing to the steeper slope in the Observe condition. Again, incorporating a non-embodied control group in future research could help clarify whether the dual task of observing plus documenting disproportionately burdens learners with lower prior knowledge. Future research might also be interested in examining more systematically the benefits and drawbacks of different paradigms of observational learning.

We are not the only researchers who have started to pay attention to the role prior knowledge plays in learners' response to embodied pedagogies. Previous findings in the literature has suggested that more embodied learning pedagogies might benefit learners with low prior knowledge more than learners with high prior knowledge (e.g., Cook et al., 2024, Congdon et al., 2018; Congdon & Goldin-Meadow, 2021). For example, in addition to the study by Congdon et al., 2018 reviewed in the introduction, a recent study by Cook and colleagues (2024) randomly assigned second- and third-grade students to watch an instructional video about mathematical equivalence that either included gestures or not. The researchers found a complex interaction between learners' prior knowledge (as demonstrated by their pretest strategy), the condition (gesture vs. no gesture), the question type (whether the question was a conceptual question or procedural question) and the way they asked the question (whether there was interference from prior knowledge). Although it is difficult to interpret the exact meaning of the four-way

interaction, the findings suggest that the effect of gesture depends on the learners' prior knowledge and the nature of the questions.

A central question under our proposed framework concerns how prior knowledge is measured and analytically treated. More work is needed to determine which aspects of prior knowledge most strongly moderate the effects of embodied pedagogies. While our study focused on content-specific, conceptual prior knowledge, other dimensions—such as past sensorimotor experience and physical intuitions (e.g., Congdon et al., 2025), working memory (e.g., Congdon, 2024), and spatial reasoning abilities (e.g., Lee-Cultura & Giannakos, 2020)—may also play critical roles.

In addition, we treated prior knowledge as a continuous variable, but future research should examine whether its moderating role reflects a dose–response relationship or a threshold effect. Simulation-based accounts of embodied learning suggest that what matters is whether learners have constructed multimodal frames that can be reactivated later. If so, once learners cross a certain threshold of prior knowledge, extensive physical engagement may no longer be necessary for them to benefit, and they may gain as much from pedagogies that involve less sensorimotor engagement. Below and above this threshold, variability in prior knowledge may matter less. If this conjecture holds, identifying the threshold point and treating prior knowledge as a binary variable could be a more productive approach. Addressing this issue will require drawing on both philosophical and cognitive discussions about the nature of grounded mental representations and how they can be evaluated.

Future research should also seek greater triangulation in measuring learners' prior knowledge. In the present work, we used retrospective self-report and repeated weekly quizzes as proxies. While these approaches provide useful information, self-reports are subject to biases,

and the weekly assessments were designed to capture concept-specific knowledge rather than a single construct across time. This limits their utility as a unified proxy for prior knowledge on a trajectory of knowledge growth.

Building on this work, future research could design repeated measures with evidence of measurement invariance—for example, developing a core set of items that captures both conceptual understanding and sensorimotor grounding, which students could complete at multiple points throughout a course. Doing so will enable understanding of how students' knowledge develops over time through different types of pedagogies using analyses such as the growth curve model.

We also recognize that several of the future directions we propose—such as implementing standardized repeated measures to establish measurement invariance, including a non-embodied control group in classroom settings, or having students passively observe without giving them an additional task—may come at the cost of ecological validity. For instance, designing repeated measures that assess the same constructs over time could conflict with the dynamic, concept-specific progression of learning in an actual course. Likewise, asking students to function solely as passive observers, or removing them entirely from embodied experiences, may limit the kinds of rich, interactive learning that characterize effective instruction. These methodological refinements would undoubtedly enhance comparability and internal validity, but they risk undermining the authenticity and efficacy of the classroom context. Future research will therefore need to carefully balance this trade-off between methodological rigor and ecological validity to ensure that findings are both reliable and educationally meaningful.

These are just the beginning of such an investigation to explore the intricate relationship between prior knowledge and different types of embodied pedagogies. Our Performing First

theory provides a framework for future research, enabling researchers to develop testable hypotheses and experiments that will advance the field. Once we establish a validated method for measuring and analyzing prior knowledge, a natural next step is to explore how we model knowledge development over time. The paradigm used in the current investigation might also inspire future work to more systematically explore the longitudinal impact of other factors such as working memory (e.g., Congdon, 2024) and spatial ability (e.g., Lee-Cultura & Giannakos, 2020).

In the present investigation, we examined the longitudinal impact of embodied pedagogies within a college-level statistics and data science course—an especially relevant context due to the inherent abstraction of statistical concepts and the persistent challenges associated with teaching them. While prior lab-based studies have demonstrated the effectiveness of embodied approaches in domains such as physics and chemistry (e.g., Johnson-Glenberg et al., 2017; 2023), it is essential to test these pedagogies in authentic classroom environments, where learning conditions are more variable and complex. Indeed, pedagogical techniques that are successful in controlled settings may not always generalize. For instance, although concreteness fading has shown consistent benefits in the lab, Kokkonen et al. (2022) found that its application in a real-world science class—teaching electromagnetic induction—was no more effective than the reverse instructional sequence.

To enhance the generalizability and practical relevance of the current findings, future studies should include longitudinal field experiments conducted in real classrooms across a range of STEM and non-STEM domains. Investigating embodied pedagogies in diverse, authentic learning environments is critical, given that instructional strategies effective in one subject may not translate easily to another. Furthermore, although our classroom-based design provides

strong ecological validity, it also introduces certain limitations—for example, we could not ensure that all students in the performance condition executed the tasks accurately or that observers remained fully engaged throughout the activities.

Lastly, in a similar vein of delivering actionable insights to teaching and instructional design, we want to highlight the importance of tailoring educational strategies in STEM teaching and learning. We also want to be explicit about how to tailor teaching based on learners' prior knowledge. Although teacher demonstrations are much easier to implement in real classrooms, for novices it seems more beneficial to incorporate active participation in physical performance into the curriculum to facilitate a deeper understanding of the concepts. After students have gained meaningful hands-on experience with the concepts, teachers can then switch to teacher demonstrations or perhaps even more abstract instruction. Future research should continue exploring the nuanced ways in which different types of embodied pedagogies impact learners at different time points of their knowledge development. This could lead to more refined strategies and curriculums that cater to the diverse needs of learners in higher education, particularly in STEM domains.

References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020, August). The future of embodied design for mathematics teaching and learning. In *Frontiers in Education* (Vol. 5, p. 147). Frontiers Media SA.

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

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593-613.

Alibali, M.W., Matthews, P.G., Rodrigues, J., Meng, R., Vest, N.A., Jay, V., Menendez, D., Murray, J.O., Donovan, A.M., Anthony, L.E. & McNeil, N.M (2024). A bird's-eye view of research practices in mathematical cognition, learning, and instruction: Reimagining the status quo. *Journal of Experimental Child Psychology*, *248*, 106056.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577-660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617-645.

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological science*, *21*(11), 1605-1610.

Belenky, D. M., & Schalk, L. (2014). The effects of idealized and grounded materials on learning, transfer, and interest: An organizing framework for categorizing external knowledge representations. *Educational Psychology Review*, *26*, 27-50.

Borghi, A. M., & Pecher, D. (2011). Introduction to the special topic embodied and grounded cognition. *Frontiers in Psychology*, *2*, 187.

Braaksma, M. A., Rijlaarsdam, G., Van den Bergh, H., & van Hout-Wolters, B. H. M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and instruction*, *22*(1), 1-36.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, *24*(1), 61-100.

Bransford, J. D., & Stein, B. S. (1993). The IDEAL problem solver.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539.

Brod, G. (2021). Toward an understanding of when prior knowledge helps or hinders learning. *npj Science of Learning*, *6*(1), 24.

Brooks, N. B., Barner, D., Frank, M., & Goldin-Meadow, S. (2018). The role of gesture in supporting mental representations: The case of mental abacus arithmetic. *Cognitive Science*, *42*(2), 554-575.

Calvo-Merino, B. (2013). Neural mechanisms for action observation. *People watching: Social, perceptual, and neurophysiological studies of body perception*, 283-303.

Cartmill, E. A., Beilock, S., & Goldin-Meadow, S. (2012). A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 129-143.

Chakraborty, J., & Esposito, A. G. (2024). Adult Learners Self-Derive New Knowledge through Integration of Novel Information and Prior Knowledge and Are more Successful with Reactivation. *Mind, Brain, and Education*.

Chalmers, D. K., & Rosenbaum, M. E. (1974). Learning by observing versus learning by doing. *Journal of Educational Psychology, 66*(2), 216–224. https://doi.org/10.1037/h0036271

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Dordrecht: Springer Netherlands.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied cognitive psychology*, *10*(2), 151-170.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. MIT Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA.

Congdon, E. L., & Goldin-Meadow, S. (2021). Mechanisms of embodied learning through gestures and actions: Lessons from development. *Handbook of embodied psychology: Thinking, feeling, and acting*, 527-546.

Congdon, E. L., Kwon, M. K., & Levine, S. C. (2018). Learning to measure through action and gesture: Children's prior knowledge matters. *Cognition*, *180*, 182-190.

Congdon, E. L., Novack, M. A., & Wakefield, E. M. (2025). Exploring individual differences: A case for measuring children's spontaneous gesture production as a predictor of learning from gesture instruction. *Topics in Cognitive Science*, *17*(3), 569-585.

Congdon, E. L. (2024). Individual differences in working memory predict the efficacy of experimenter-manipulated gestures in first-grade children. *Child Development*, *95*(5), 1494-1507.

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child development*, *84*(6), 1863-1871.

Cook, S. W., Wernette, E. M., Valentine, M., Aldugom, M., Pruner, T., & Fenn, K. M. (2024). How Prior Knowledge, Gesture Instruction, and Interference After Instruction Interact to Influence Learning of Mathematical Equivalence. *Cognitive Science*, *48*(2), e13412.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 15-26.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal.* Oxford University Press.

Eghterafi, W., Tucker, M. C., Zhang, I., & Son, J. Y. (2022). Effect of Feedback with Video-Based Peer Modeling on Learning and Self-Efficacy. *Online Learning*, *26*(2), 1-21.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological methods*, *25*(1), 88.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, *98*, 4-18.

Enders, C. K. (2022). *Applied missing data analysis*(2nd ed.). New York: Guilford press.

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235-245.

Fiorella, L., & Mayer, R. E. (2016). Effects of observing the instructor draw diagrams on learning from multimedia messages. *Journal of Educational Psychology*, *108*(4), 528.

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2021). Practicing connections: A framework to guide instructional design for developing understanding in complex domains. *Educational Psychology Review*, *33*(2), 739-762.

Fu, Y., & Franz, E. A. (2014). Viewer perspective in the mirroring of actions. *Experimental Brain Research*, *232*, 3665-3674.

Fugate, J. M., Macrine, S. L., & Cipriano, C. (2018). The role of embodied cognition for transforming learning. *International Journal of School & Educational Psychology*, *7*(4), 274-288.

Fyfe, E. R., & Nathan, M. J. (2019). Making "concreteness fading" more concrete as a theory of instruction for promoting transfer. *Educational Review*, *71*(4), 403-422.

Fyfe, E. R., McNeil, N. M., & Borjas, S. (2015). Benefits of "concreteness fading" for children's mathematics understanding. *Learning and Instruction*, *35*, 104-120.

Garofalo, S., Giovagnoli, S., Orsoni, M., Starita, F., & Benassi, M. (2022). Interaction effect: Are you doing the right thing?. *PLoS One*, *17*(7), e0271668.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, *35*(4), 1674-1684.

Gerson, S. A., & Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child development*, *85*(1), 264-277.

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*(5), 234-241.

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*(3), 267-272.

Goldin-Meadow, S., Levine, S. C., Zinchenko, E., Yip, T. K., Hemani, N., & Factor, L. (2012). Doing gesture promotes learning a mental transformation task better than seeing gesture. *Developmental Science*, *15*(6), 876-884.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the learning sciences*, *14*(1), 69-110.

Goldstone, R. L., Landy, D., & Son, J. Y. (2008). A well grounded education: The role of perception in science and mathematics. *Symbols, embodiment, and meaning*, 327-355.

Golonka, S., & Wilson, A. D. (2012). Gibson's ecological approach. *Avant: Trends in Interdisciplinary Studies*, *3*(2), 40-53.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of pragmatics*, *32*(10), 1489-1522.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*(1), 549-576.

Grotelueschen, A. D. (1979). Introductory material structure and prior knowledge effects on adult learning. *Adult Education*, *29*(2), 75-82.

Guarino, K. F., & Wakefield, E. M. (2020). Teaching analogical reasoning with co-speech gesture shows children where to look, but only boosts learning for some. *Frontiers in Psychology*, *11*, 575628.

Hall, R., & Nemirovsky, R. (2012). Introduction to the special issue: Modalities of body engagement in mathematical activity and learning. *Journal of the learning sciences*, *21*(2), 207-215.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of educational psychology*, *90*(3), 414.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Henriques, A. (2016). Students' difficulties in understanding of confidence intervals. *The Teaching and Learning of Statistics: International Perspectives*, 129-138.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review*, *15*, 495-514.

Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language*, *63*(2), 245-257.

Hume, D. (1938). *An Abstract of a Treatise of Human Nature, 1740*. CUP Archive.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, **14**, S103-S109.

Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education. *Cognitive Research: Principles and Implications*, *2*(1), 1-28.

Johnson-Glenberg, M. C., Birchfield, D. A., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion-capture environments: Two science studies. *Journal of Educational Psychology*, *106*(1), 86.

Johnson-Glenberg, M. C., Yu, C. S., Liu, F., Amador, C., Bao, Y., Yu, S., & LiKamWa, R. (2023). Embodied mixed reality with passive haptics in STEM education: randomized control study with chemistry titration. *Frontiers in Virtual Reality*, *4*, 1047833.

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*, 209-215.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational Psychologist, 38*, 23–31.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational psychology review*, *23*, 1-19.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224-234.

Kokkonen, T., Lichtenberger, A., & Schalk, L. (2022). Concreteness fading in learning secondary school physics concepts. *Learning and Instruction*, *77*, 101524.

Kosmas, P., & Zaphiris, P. (2023). Improving students' learning performance through Technology-Enhanced Embodied Learning: A four-year investigation in classrooms. *Education and Information Technologies*, *28*(9), 11051-11074.

Kosslyn, S. M. (2005). Mental images and the brain. *Cognitive Neuro- psychology*, **22**, 333-347.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lam, T. C., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *The American Journal of Evaluation*, *24*(1), 65-80.

Lau, W. W. F. & Yuen, A. H. K. (2008). Exploring the effects of gender and learning styles on computer programming performance: Implications for programming pedagogy. British Journal of Educational Technology, doi: 10.1111/j.1467-8535.2008.00847.x

Laurillard, D. (2001). E-university-what have we learned?. *International Journal of Management Education*, *1*, 3-7.

Laurillard, D. (2013). *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge.

Lee-Cultura, S., & Giannakos, M. (2020). Embodied interaction and spatial skills: A systematic review of empirical studies. *Interacting with Computers*, *32*(4), 331-366.

Lin, Y. I., Son, J. Y., & Rudd, J. A. (2016). Asymmetric translation between multiple representations in chemistry. *International Journal of Science Education*, *38*(4), 644-662.

Lindgren, R., Morphew, J. W., Kang, J., Planey, J., & Mestre, J. P. (2022). Learning and transfer effects of embodied simulations targeting crosscutting concepts in science. *Journal of Educational Psychology*, *114*(3), 462.

Locke, J. (1948). An essay concerning human understanding, 1690.

Martin, T. (2009). A theory of physically distributed learning: How external environments and internal states interact in mathematics learning. *Child Development Perspectives*, *3*(3), 140-144.

McNeil, N. M., & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction*, *22*(6), 440-448.

Moore, D., & Tananis, C. A. (2009). Measuring change in a short-term educational program using a retrospective pretest design. *American Journal of Evaluation*, *30*(2), 189-202.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in psychology*, *2*, 74.

Nathan, M. J., & Alibali, M. W. (2021). An embodied theory of transfer of mathematical learning. *Transfer of learning: Progressive perspectives for mathematics education and related fields*, 27-58.

Nathan, M. J. (2012). Rethinking formalisms in formal education. *Educational Psychologist*, *47*(2), 125-148.

Nathan, M. J. (2017). One function of gesture is to make new ideas. *Why gesture*, 175-196.

Nathan, M. J. (2021). *Foundations of embodied learning: A paradigm for education*. Routledge.

Novack, M., & Goldin-Meadow, S. (2015). Learning from gesture: How our hands change our minds. *Educational Psychology Review*, *27*, 405-412.

Novack, M. A., & Goldin-Meadow, S. (2017). Gesture as representational action: A paper about function. *Psychonomic Bulletin & Review*, *24*(3), 652-665.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction: Using the hands to learn math. *Psychological Science*, *25*(4), 903-910.

Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2014). The Single-Item Math Anxiety scale (SIMA): An alternative way of measuring mathematical anxiety. *Personality and Individual Differences*, *60*, S75-S76.

Open Science Framework (2025). https://osf.io/ntsr2/?view_only=ce650267f407451a9ea26abcb428d8f7

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, *6*(4), 351-371.

Pezzulo, G., & Calvi, G. (2011). Computational explorations of perceptual symbol systems theory. *New Ideas in Psychology*, *29*(3), 275-297.

Piaget, J. (1983). Piaget's theory. *Handbook of child psychology*, *1*(4).

Pyers, J. E., Magid, R., Gollan, T. H., & Emmorey, K. (2021). Gesture helps, only if you need it: Inhibiting gesture reduces tip-of-the-tongue resolution for those with weak short-term memory. *Cognitive Science*, *45*(1), e12914.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (Vol. 2). New York: Random House.

Raab, M., & Araújo, D. (2019). Embodied cognition with and without mental representations: The case of embodied choices in sports. *Frontiers in Psychology*, *10*, 1825.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Renkl, A., & Atkinson, R. K. (2007). An example order for cognitive skill acquisition. In F. E. Ritter, J. Nerb, E. Lehtinen, & T. M. O'Shea (Eds.), *In order to learn: How the sequence of topics influences learning* (pp. 95–105). New York: Oxford University Press.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169-192.

Rosenbaum, D. A., Chapman, K. M., Weigelt, M., Weiss, D. J., & van der Wel, R. (2012). Cognition, action, and object manipulation. *Psychological Bulletin*, *138*(5), 924.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, *105*(1), 143.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, *57*(1), 31-54.

Skulmowski, A., & Rey, G. D. (2018). Embodied learning: introducing a taxonomy based on bodily engagement and task integration. *Cognitive research: principles and implications*, *3*(1), 6.

Sommerville, J. A., & Woodward, A. L. (2010). The link between action production and action processing in infancy. *Naturalizing intention in action*, 67-89.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1-B11.

Son, J. Y., & Goldstone, R. L. (2009). Contextualization in perspective. *Cognition and Instruction*, *27*(1), 51-89.

Son, J. Y., & Stigler, J. W. (2024). Statistics and data science: A modeling approach . https://coursekata.org/

Son, J. Y., Blake, A. B., Fries, L., & Stigler, J. W. (2021). Modeling first: Applying learning science to the teaching of introductory statistics. *Journal of Statistics and Data Science Education*, *29*(1), 4-21.

Stigler, J. W., Givvin, K. B., & Thompson, B. J. (2010). What community college developmental mathematics students understand about mathematics. *MathAMATYC Educator*, *1*(3), 4-16.

Stigler, J. W., Son, J. Y., Givvin, K. B., Blake, A. B., Fries, L., Shaw, S. T., & Tucker, M. C. (2020). The Better Book approach for education research and development. *Teachers College Record*, *122*(9), 1-32.

Sullivan, J. V. (2018). Learning and embodied cognition: A review and proposal. *Psychology Learning & Teaching*, *17*(2), 128-143.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). The expertise reversal effect. In *Cognitive load theory* (pp. 155-170). New York, NY: Springer New York.

Sweller, J., Ayres, P., Kalyuga, S., Sweller, J., Ayres, P., & Kalyuga, S. (2011). Intrinsic and extraneous cognitive load. *Cognitive load theory*, 57-69.

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Harvard University Press.

Thompson, R. A., & Zamboanga, B. L. (2003). Prior knowledge and its relevance to student achievement in introduction to psychology. *Teaching of Psychology*, *30*(2), 96-101.

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research*, 64, 37–54. https://doi.org/10.3102/00346543064001037

Trudeau, J. J., & Dixon, J. A. (2007). Embodiment and abstraction: Actions create relational representations. *Psychonomic bulletin & review*, *14*, 994-1000.

Tytler, R., & Prain, V. (2013). Representation construction to support conceptual change. In *International handbook of research on conceptual change* (pp. 560-579). Routledge.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2004). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional science*, *32*(1), 83-98.

Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial?. *Journal of Memory and Language*, *50*(4), 395-407.

Wakefield, E. M., & James, K. H. (2015). Effects of learning with gesture on children's understanding of a new language concept. *Developmental Psychology*, *51*(8), 1105.

West, M., & Ross, S. (2002). Retaining females in computer science: A new look at a persistent problem. JCSC, 17(5), 1-7.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*, 625-636.

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.

Wouters, P., Paas, F., & van Merriënboer, J. J. (2009). Observational learning from animated models: Effects of modality and reflection on transfer. *Contemporary Educational Psychology*, *34*(1), 1-8.

Zacharia, Z. C., Loizou, E., & Papaevripidou, M. (2012). Is physicality an important aspect of learning through science experimentation among kindergarten students?. *Early Childhood Research Quarterly*, *27*(3), 447-457.

Zhang, I., Givvin, K. B., Sipple, J. M., Son, J. Y., & Stigler, J. W. (2021). Instructed hand movements affect students' learning of an abstract concept from video. *Cognitive Science*, *45*(2), e12940.

Zhang, I., Guo, X. H., Son, J. Y., Blank, I. A., & Stigler, J. W. (2025). Watching videos of a drawing hand improves students' understanding of the normal probability distribution. *Memory & Cognition*, *53*(1), 262-281.

Zhang, I. Y., Gray, M. E., Cheng, A. X., Son, J. Y., & Stigler, J. W. (2024b). Representational-mapping strategies improve learning from an online statistics textbook. *Journal of Experimental Psychology: Applied*, *30*(2), 293.

Zhang, I. Y., Tucker, M. C., & Stigler, J. W. (2022). Watching a hands-on activity improves students' understanding of randomness. *Computers & Education*, *186*, 104545.

Zhang, I. Y., Xu, A., Son, J. Y., & Stigler, J. W. (2024a). Watching hands move enhances learning from concrete and dynamic visualizations. *Journal of Experimental Psychology: General*.

Appendix A: lab schedule

Week	Content
Week 1	General introduction
Week 2	Cutting out and shuffling a dataset to understand shuffle ()
Week 3	Cutting out and sampling with replacement to understand resample()
Week 4	Comparing shuffle() and resample()
Week 5	Drawing and gesturing to understand the empty model and the two-group model
Week 6	Drawing and gesturing to understand the three-group model
Week 7	Drawing and gesturing to understand the regression model, and compare it with group models
Week 8	Using shuffle () and resample () to simulate sampling distributions
Week 9	Hands-on construction of the confidence interval
Week 10	Lab general assessment and debrief

Appendix B: Interrater Reliability (Quadratic-weighted Cohen's Kappa) by Weeks

Week	2	3	4	5	6	7	8	9
κ	0.70	0.79	0.85	0.81	0.82	0.86	0.76	0.86

Appendix C

Model: midterm1 ~ role + concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	93.86			91.25 – 96.47	**<0.001**
role [Perform]	1.04	1	0.909	-1.11 – 3.18	0.342
concept known	1.13	1	0.261	-3.24 – 5.50	0.610
Observations	217				
R^2 / R^2 adjusted	0.005 / -0.005				

Model: midterm2 ~ role + concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	84.41			80.48 – 88.34	**<0.001**
role [Perform]	1.63	1	0.992	-1.59 – 4.85	0.320
concept known	8.45	1	6.426	1.88 – 15.02	**0.012**
Observations	217				
R^2 / R^2 adjusted	0.030 / 0.021				

Model: final exam ~ role + concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	87.86			84.70 – 91.03	**<0.001**
role [Perform]	0.17	1	0.017	-2.42 – 2.77	0.896
concept known	4.25	1	2.500	-1.05 – 9.54	0.115
Observations	217				
R^2 / R^2 adjusted	0.012 / 0.002				

Model: final lab assessment ~ role + concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	20.32			17.75 – 22.89	**<0.001**
role [Perform]	1.55	1	2.061	-0.58 – 3.68	0.153
concept known	5.25	1	5.780	0.94 – 9.56	**0.017**
Observations	195				
R^2 / R^2 adjusted	0.034 / 0.024				

Appendix D

Model: midterm1 ~ role * concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	93.18			89.67 – 96.68	**<0.001**
role [Perform]	2.19	1	0.926	-2.30 – 6.69	0.337
concept known	2.53	1	0.594	-3.94 – 8.99	0.442
role [Perform] * concept known	-2.57	1	0.334	-11.36 – 6.21	0.564
Observations	217				
R^2 / R^2 adjusted	0.006 / -0.008				

Model: midterm2 ~ role * concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	79.64			74.46 – 84.83	**<0.001**
role [Perform]	9.70	1	8.281	3.06 – 16.35	**0.004**
concept known	18.19	1	14.047	8.62 – 27.75	**<0.001**
role [Perform] * concept known	-17.97	1	7.430	-30.96 – -4.97	**0.007**
Observations	217				
R^2 / R^2 adjusted	0.063 / 0.050				

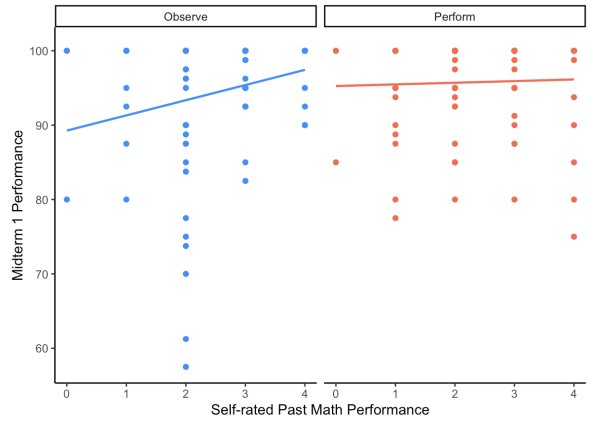
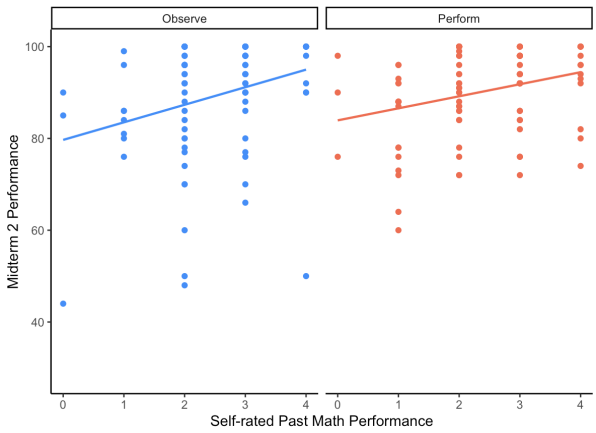
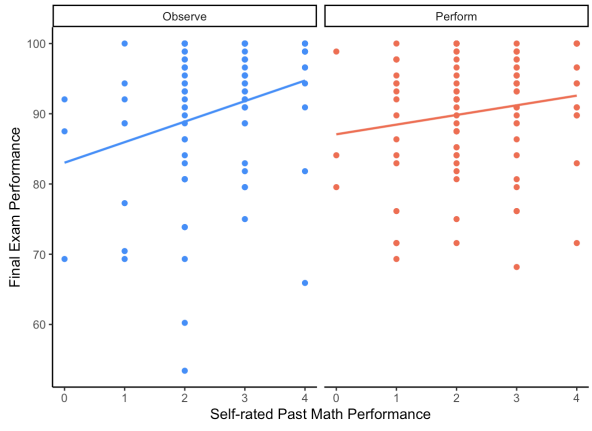
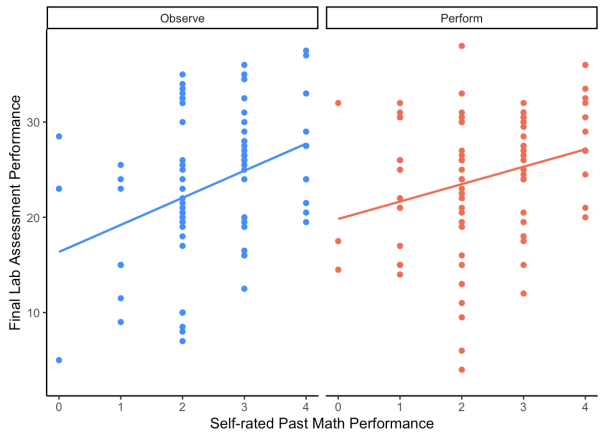
Model: final exam ~ role * concept_known

Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	83.78			79.61 – 87.95	**<0.001**
concept known	12.58	1	6.821	4.89 – 20.27	**0.001**
role [Perform]	7.08	1	10.391	1.74 – 12.43	**0.010**
concept known * role [Perform]	-15.37	1	8.410	-25.82 – -4.92	**0.004**
Observations	217				
R^2 / R^2 adjusted	0.049 / 0.036				

Model: final lab assessment ~ role * concept_known

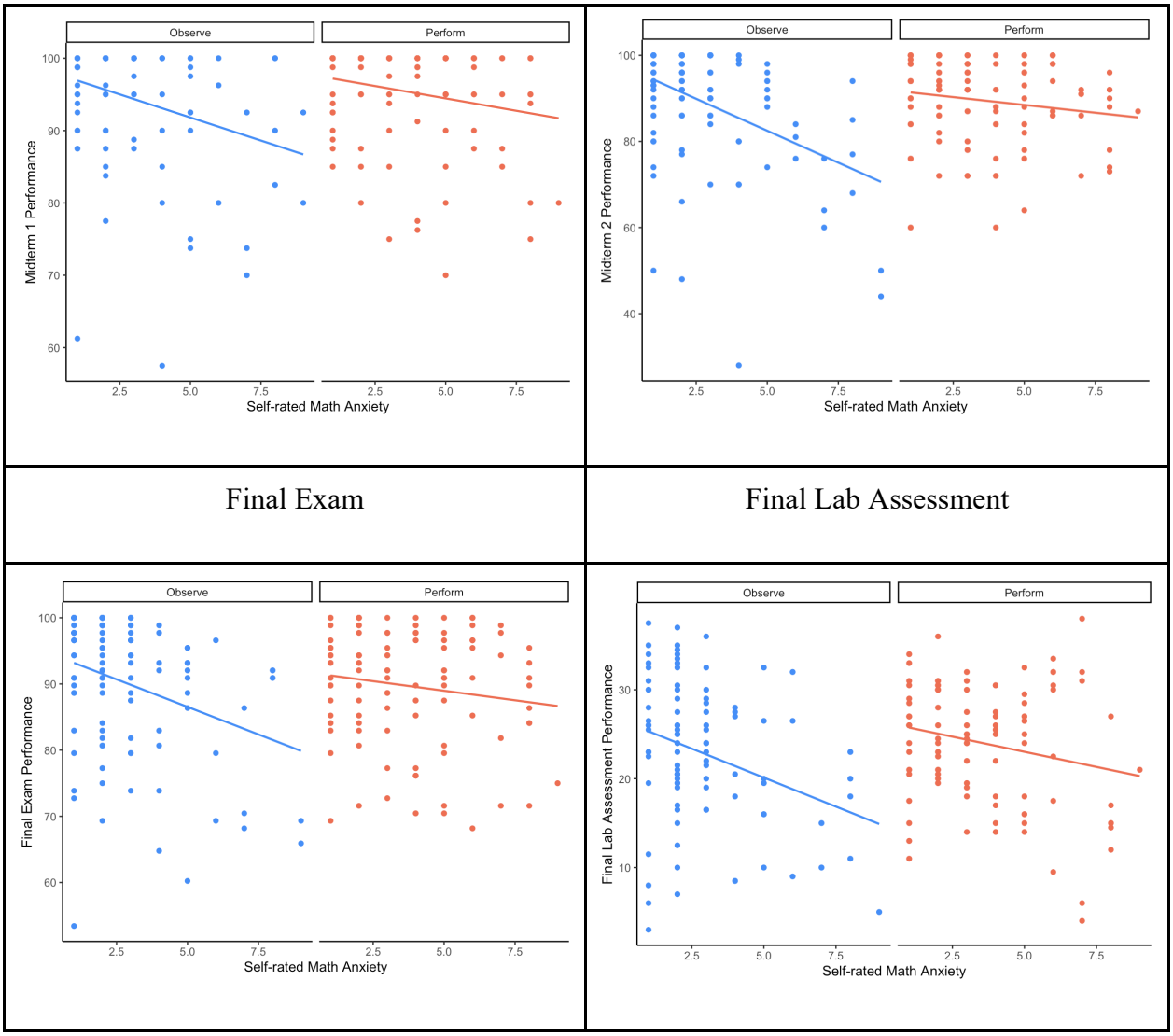
Predictors	*Estimates*	*df*	*F*	*CI*	*p*
(Intercept)	17.75			14.35 – 21.15	**<0.001**
concept known	10.51	1	7.120	4.23 – 16.80	**0.001**
role [Perform]	5.89	1	10.882	1.54 – 10.24	**0.008**
concept known * role [Perform]	-9.74	1	5.048	-18.30 – -1.19	**0.026**
Observations	195				
R^2 / R^2 adjusted	0.059 / 0.044				

Appendix E: The Relationship Between Condition and Self-rated Past Math Performance for Each of the Four Summative Assessments

Midterm 1	Midterm 2
	
Final Exam	**Final Lab Assessment**
	

Appendix F: The Relationship Between Condition and Self-rated Math Anxiety for Each of the Four Summative Assessments

Midterm 1	Midterm 2



Appendix G: equation for each level

Level 1:

$$Posttest_{ijk} = \pi_{0jk} + \pi_{1jk}(Pretest^{w}_{ijk}) + \pi_{2jk}Time_{ijk} + e_{ijk}$$

Level 2:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}Condition_{jk} + \beta_{02k}Pretest^{b.cgm}_{jk} + \beta_{03k}Condition_{jk} * Pretest^{b.cgm}_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}Condition_{jk}$$

$$\pi_{2jk} = \beta_{20k} + \beta_{21k}Condition_{jk}$$

Level 3:

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{03k} = \gamma_{030}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{21k} = \gamma_{210}$$