Going Beyond Formalisms

A Grounded and Embodied Learning Approach to the Design of Pedagogical Statistics Simulations

Sebahat Gok & Robert L. Goldstone

Data Literacy   Grounded and Embodied Learning   Statistical Simulations

Computer-based interactive simulations that model the processes of sampling from a population are increasingly being used in data literacy education. However, these simulations are often summarized by graphs designed from the point of view of experts which makes them difficult for novices to grasp. In our ongoing design-based research project, we build and test alternative sampling simulations to the standard ones. Based on a grounded and embodied learning perspective, the core to our design position is that difficult and abstract sampling concepts and processes should: be grounded in familiar objects that are intuitive to interpret, incorporate concrete animations that spontaneously activate learners' gestures, and be accompanied by verbal instruction for a deeply integrated learning. Here, we report the results from the initial two phases of our project. In the first iteration, through an online experiment (N=126), we show that superficial perceptual elements in a standard simulation can lead to misinterpretation of concepts. In the second iteration, we pilot test a new grounded simulation with think-aloud interviews (N=9). We reflect on the complementary affordances of visual models, verbal instruction, and learners' gestures in fostering integrated and deep understanding of concepts.

Introduction

Data literacy, defined as "competence in making sense of data, including management, modeling, interpretation, and presentation of data in critical manner" (Gehrke et al., 2021, p. 201), has gained attention in K-12 (Common Core State Standards, 2022) and undergraduate education (de Veaux et al., 2017) with the increasing reliance on data for critical decision-making processes in today's society. At both levels, curricular programs aim to equip students with the skills of making informed decisions and solving personal and societal problems based on data. To this end, students are

expected to make inferences from observed samples which require flexible reasoning of statistical concepts such as population, sampling, variation, and randomness (Adhikari et al., 2021; Gehrke et al. 2021). However, statistical concepts can be notoriously difficult to grasp (Hokor, 2022; Schwartz et al., 1998; Tversky & Kahneman, 1972). To make statistical inference accessible to students, educators have developed interactive computer simulations (e.g., Donoghue et al., 2021; Sutherland & Ridgway, 2017). The simulations combine the affordances of multiple forms of media by integrating interactive dynamic graphs, verbal tags, and explanations. Immersed in their rich representations, students engage in practices of modeling real-world phenomena through inquiry-based investigations (Pfannkuch et al., 2018).

While the simulations are enthusiastically welcomed by the education community, the empirical evidence from the existing simulations are highly variable in terms of whether simulations bring considerable learning gains for statistical concepts. Some studies incorporating simulations in classrooms found dismal results. With simulations, students still lacked a causal understanding of probabilistic processes, and simulations sometimes even caused additional misconceptions because they led to incorrect interpretations of graphs (Ben-zvi et al., 2012; Lehrer & Schauble, 2017; Maxara & Biehler, 2010). A few studies, however, reported positive learning gains associated with simulations (Jacob & Doerr, 2014; van Dikje-Drogers et al., 2021). The variations in the results suggest there is more to understand about how simulations are most effective in supporting novices' learning of data concepts and processes. In this ongoing design-based research project, the researchers aim to help meet this need by building, comparing, and testing alternative types of sampling simulations. This article reports the results from initial iterations with a focus on design decisions regarding the visual, dynamic, and verbal features of the simulations from a grounded and embodied learning perspective.

Before introducing this work, however, it is important to understand the traditional solution to which we are seeking an alternative. Many current simulations use, what we call, *generic representations of data*. In the next section, we describe these types of representations, discuss their limitations, and then, introduce how we re-envision the data representations that form the foundations for the current work.

Generic Visual Representations of Data

Simulations that model sampling processes, the focus in this study, typically adopt generic visual representations such as histograms and pie charts (See figure 1). We call such graphs *generic* as they are conventional representations depicting concepts and relations in ways that strip away the real visual attributes of the data they represent. Relying on generic graphs for teaching new concepts often has limitations. While such graphs can be effective data representation tools for experts, they are not always suitable to the developmental needs of novice students (Nathan, 2021). Students often have poor understanding of generic graphs. In histograms, students often confuse what the horizontal and vertical dimensions mean. For example, they think flatness of histograms indicate low variability, or that the X-axis indicates chronological order even when there is no time-related variable in the data (Kaplan et al., 2014). Such misconceptions have been shown to be resistant to training.

Figure 1

Examples of Standard Simulations
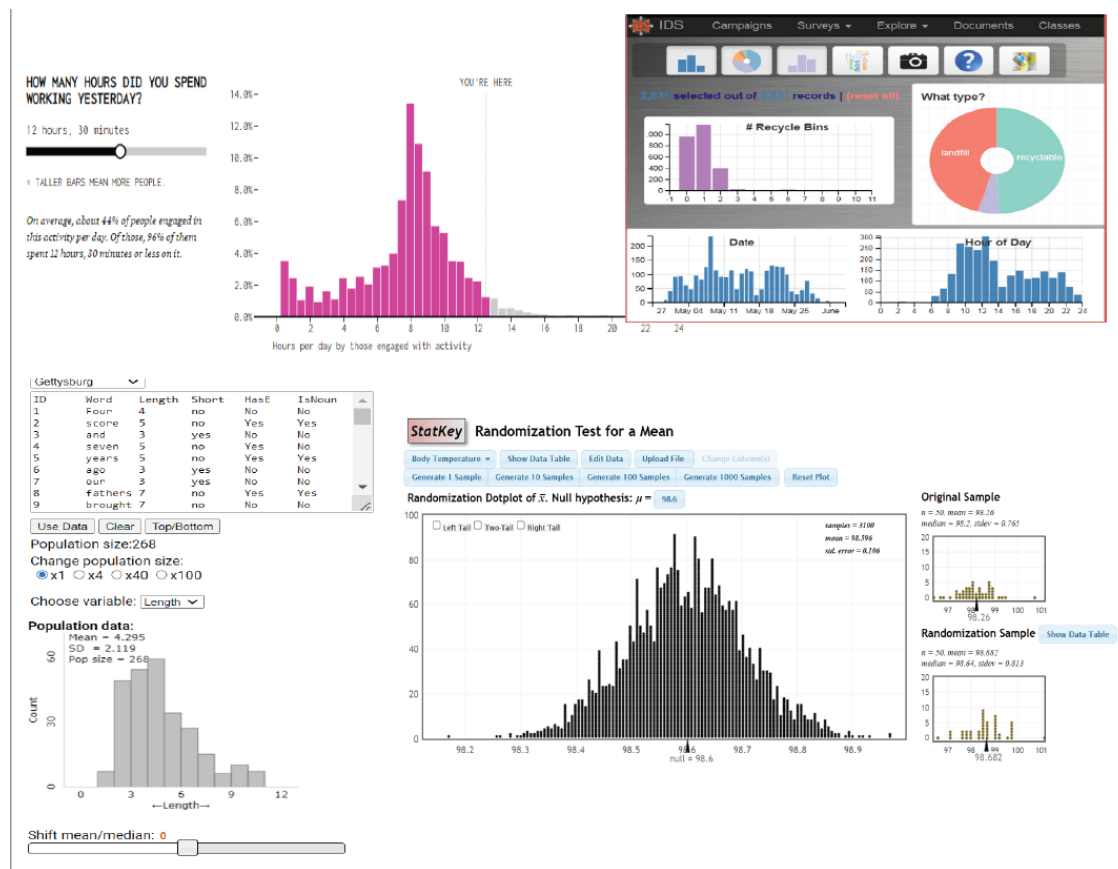
Note. A representative sample of popular and modern pedagogical statistical simulations used in K-12 and undergraduate education (On top left; from the Bootstrap program by Brown University team, on top right; from the Introduction to Data Science curriculum team by UCLA College of Statistics, on bottom left; from Rossman and Chance, on bottom right, from the team of Locks). Note that each depicts data in a highly similar fashion. The same type of histograms are used whether one aims to depict the distribution of people, cities, recycle bins, or abstract process of sampling. We will call the simulations which employ such generic histograms standard simulations in the rest of the paper.

The *standard simulations* are *ungrounded* in the sense that they are not connected to things students already know from their primary, personal, and real-world experiences. For novices, these types of representations often require numerous mental inferences (Rau, 2017) and learners' failure to learn with them might indicate a lack of understanding of the representations rather than the ideas behind them. (Nathan, 2021). For graphs to be effective, students should be able to understand how they depict information (Rau, 2017), and the representations should be grounded in learners' familiar experiences (Nathan, 2021). To this end, the current study goes beyond generic graphs and re-envisions how data representations from simulations can be made more accessible to novices through the lens of the *grounded and embodied learning framework* with the overarching goal of reducing the entry barriers to data literacy.

Grounded and Embodied Learning Framework

Earlier theories of cognition postulated meaning occurs from the manipulation and combination of mental symbols (Anderson, 1996; Newell & Simon, 1956). According to these *symbolic cognition* accounts, reasoning is guided by abstract rules dissociated from modal systems and bodies. This approach assumes people can effectively be trained with abstract versions of rules, and they will later be able to transfer what they have learned by applying the acquired symbolic rules to distant domains (Smith et al., 1992). Symbolic cognition theories have had a large influence on education. In statistics education, the focus has traditionally been on mastery of abstract rules and mathematical formalisms with little attention paid to learners' perceptions, bodily actions, and interactions with their environment

(e.g., Lovett et al., 2008; Nisbett et al., 1987). The abstract rules, however, are often hard to grasp, easy to misapply and forget, because they are often disconnected from students' life experiences (Nathan, 2021).

In later years, symbolic accounts of cognition have been challenged by different strands in cognitive science such as *grounded cognition*, *embodied cognition, dynamic cognition, and situated cognition*, to name a few (for a review, see Barsalou, 2008). Even though the strands differ from each other about their view on the nature of the mind, they all have commonly argued for closer relationships between abstract thoughts, sensorimotor systems, and situated activity. Recently, Nathan (2021) has combined these modern strands of cognition under the umbrella term of "grounded and embodied learning (GEL) framework" to offer a new lens for the design of educational environments. The GEL framework premises meaning arises from the relationship of a person's actions and the affordances of the particular situation the individual is in. The affordances of a situation depend on the individual's goals, personal learning history, and the cultural norms they acquired. One's primary experiences, perceptions, gestures, and body are central to how one makes sense of the world (Nathan, 2021). Based on this thesis, the GEL framework suggests educational experiences should be designed in a way that connects abstract ideas and representations to students' lived experiences, including perceptions and body-based interactions, for meaningful learning to occur (Nathan, 2021).

We are not the first to investigate what opportunities a grounded and embodied learning approach can offer for the design of multimedia learning experiences. Abrahamson (2012) designed physical random device generators that tap into students' pre-analytic perceptual judgments to teach compound events. Loy (2021) designed static lined-up graphs for students to engage in hypothesis testing based on the perceptual differences between graphs before conducting mathematical analysis. Zhang et al. (2022) designed instructional statistics videos in which students observed hand movements of an instructor drawing normal distribution graphs, and were instructed to mimic the positions of central tendencies in the graphs with their hand movements (Zhang et al., 2021). These studies either reported statistically significant learning gains from their design approach (Zhang et al., 2021, 2022) or qualitatively meaningful learning experiences (Abrahamson, 2012; Loy, 2021). Building upon this body of literature, we design sampling simulations enriched with icons that draw on people's ability to easily track frequency information from such representations (Brase, 2008), and animations that represent mathematical computations dynamically and spatially. While interpreting the graphs, the students mimic these spatial movements with their gestures (hand and finger movements) which can indicate deeper engagement than just "seeing the graph" (Gerofsky, 2011, p. 245). In the next section, we detail our approach to multimedia learning for the design of computer simulations at the current project.

'Re-thinking' Multimedia Learning from the Lens of Grounded and Embodied Learning Framework

Multimedia theory traditionally distinguishes between learning from pictures and words, noting that these two modalities offer complementary advantages and disadvantages (Mayer, 2009). Pictures excel at representing analog, continuous, and rich information, particularly about spatial relations (Hegarty, 2011), whereas words excel at conveying discrete, symbolic interpretations (Dingemanse et al., 2015). In many domains, both modalities are needed for effective learning. nstructional strategies that promote their integration. for example through concurrent presentation of both, are often particularly effective (Mayer & Anderson, 1992).

Moving beyond this classic distinction between words and pictures, researchers have further investigated pictorial representations to understand the unique affordances of graphical displays that are static, animated, and interactive. Relative to static images, animations offer advantages when learning involves understanding how variables change over time and space (Ploetzner et al., 2020), but they can come at the cost of discouraging students from constructing their own mental model of a situation (Mayer et al., 2005). Relative to "canned" animations, simulations offer learners "live," dynamically computed sequences of images that often incorporate interactivity (National Research Council, 2011). This interactivity can promote learning directly by having learners come to understand how parameters they control influence how the simulation unfolds, and indirectly by increasing learners' motivation to learn and engage (Magana et al., 2022).

Our interest in simulations for teaching concepts related to statistical inference stems from their promise in helping learners to create their own internal mental models (Boyle et al., 2014). In our observations of students learning mathematical principles using computer simulations, students often actively interpret the interactions of the elements while interacting with the simulations (Goldstone & Wilensky, 2008). These interpretations are firmly "grounded in the particular simulation with which they are interacting" (p. 480). However, as the interpretations are highly selective, perspectival, and idealized (Goldstone & Sakomoto, 2003; Goldstone & Son, 2005), they can apply to two situations which look apparently dissimilar. Students' understandings are thus grounded in a specific concrete context, yet also transferable to new contexts. This route to transfer of learning through interpretation of concrete simulations is often more effective than approaches that stress formalisms. Formalisms such as logical expressions or algebraic notation do unify disparate situations under a common formalism, but seeing the applicability of a formalism to a situation is notoriously difficult for students (Nathan, 2012). Instead, when simulations are paired with lesson plans that guide learners to notice useful patterns, learners come to perceive these patterns, and the perceptual routines that the learners acquire along the way naturally carry over to new situations (Goldstone et al., 2017). Far transfer based on noticing a shared formalism is rare (Day & Goldstone, 2012). By contrast, people who train their perceptual systems to find a pattern often automatically use their trained perceptual system in new contexts (Kellman & Massey, 2013).

The Authors' Positionality

Researchers' beliefs about the nature of social reality, knowledge, and how we interact with the world impact their research process (Holmes, 2020). Whether researchers are aware of it or not, these ontological and epistemological beliefs have influences on how they conduct their research and how they interpret their results. Therefore, an explicit acknowledgment of the author position is important for readers to make a better-informed judgment about the research process.

The research and design team of this project consists of the two authors of this article. The first author is a graduate student of instructional design and cognitive science with 10 years of teaching and design experience in different fields spanning science, educational technology, and foreign language education. The second author is a professor of cognitive science whose expertise lies in the bidirectional relationships of human perception and cognition. While his earlier research focused on basic processes involved in human cognition, in later years, his worked broadened to apply cognitive theories to the design of educational technologies in mathematics and science classrooms. Based on this body of work, he has developed computer-based mathematics tutors which are used in K-12 educational institutions nationwide.

The researchers believe the study of human cognition can provide insights for the design of educational technologies, and, reciprocally, the results from educational work can further enrichen our understanding of human cognition. Our situatedness in the field of cognitive science and related expertise has directly influenced the theoretical framework we adopted, and our institutional affiliation has influenced the choice of participants through the opportunities available to us. Furthermore, the first author's training in instructional design has influenced the iterative design choices of the project based on a design-based research framework.

The first author designed the instructional activities that accompanied the simulations while the second author designed the simulations based on his expertise in perceptual and conceptual learning. In the light of reading of the related work in the field, the first author identified the research questions, conducted experiments, interviews, and the data analysis, and both authors met weekly to discuss the project, their interpretations of the results, and what steps to take next. Concurrently, we also received feedback from our informal meetings with colleagues who were statistics educators, cognitive scientists, and instructional designers, which might have influenced our design choices in addition to analysis of the empirical results.

The Current Study

The overarching research question is whether difficult concepts in statistical reasoning related to sampling can be successfully learned by incorporating a grounded and embodied learning perspective to computer-based simulations. Through a design-based research methodology (Barab & Squire, 2004), we employed iterative design cycles of the

simulation as a multimedia artifact by combining it with instructional texts. We present findings from two design iterations and several design decisions employed to elicit a well-grounded understanding of sampling for inference making in statistics.
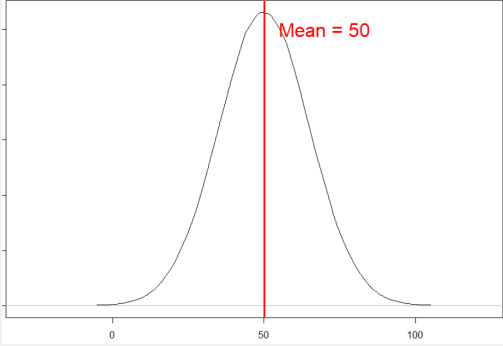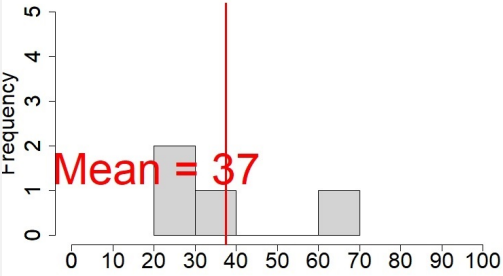
The Design Iterations

In the first iteration, we aimed to test what *standard simulations* commonly used for data literacy and statistics education offer (See figure 1). To this end, we emulated the common visual features of these simulations. We tested our standard simulation against a more traditional teaching method that does not employ simulations through a controlled experiment. In the second iteration, we went beyond the standard features of statistical simulations and investigated the promises of an innovative grounded simulation through think-aloud interviews with students. Before presenting our iterations, we first overview the subject domain we focused on in the current study.

The Subject Domain: Distribution of Sample Means

Learning objectives focused on the topic of distribution of sample means. For statistical inference, students need to be able to flexibly reason about distributions of sample means and how sample size affects their properties. We overview the key sampling processes, the rules of sampling processes with their rationale, and common student misconceptions below in Table 1.
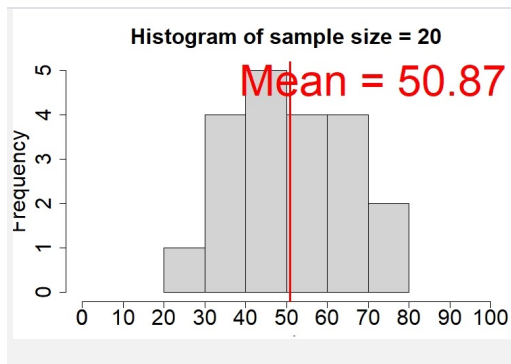
Table 1

The Summary of the Subject Domain

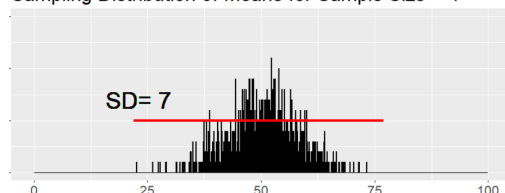| Sampling Processes	Rules and rationales	Misconceptions
An example of a normally distributed population graph with mean = 50. In real life, we often cannot collect information from the whole population, therefore, we draw samples from it. Two different samples with size 4 and 20 are drawn below to point out the importance of sample size in estimating population mean.	Rule: As sample size increases, it is more likely that the sample mean gets closer to the population mean. Rationale: For a larger sample, it is less likely that all values are low or high. As a result, low and high values cancel each other out.	Students often know that larger samples generally allow for a better estimate of the population mean. However, they (wrongly) believe that it is the sample size's proportion to population, not its absolute size, that affects this estimation

Sampling Processes	Rules and rationales	Misconceptions

Histogram of sample size = 20

Mean = 50.87

(Histogram with Frequency on y-axis (0–5) and x-axis 0–100, bars centered around 50 with a red vertical line at the mean)

A random sample of a specific size is taken from the population and its mean value is calculated and then recorded. Another random sample with the same size is then taken and its value is recorded. This process is repeated many times. In other words, means of many random samples of a specific size are collected from the population. This collection of sample means is called the distribution of means. See below two different distributions of means with sample size 4 and 20.
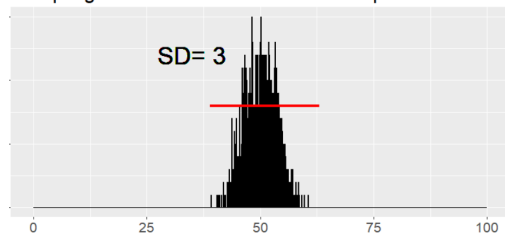
As sample size gets larger, the standard deviation of the distribution of means gets smaller. Rationale: Remember that the distribution of means is a collection of means of random samples of a particular size. With a larger sample size, more of the sample means will be closer to the actual mean, which will decrease the standard deviation of the distributions of means.

This rule is notoriously difficult for students to grasp. Students often confuse sampling distributions of means with an individual sample. Students often believe standard deviation increases or stays stable with larger samples.

Sampling Distribution of Means for Sample Size = 4

SD= 7

(Distribution plot, x-axis 0–100, with a red horizontal line indicating SD around the mean ~50)

Sampling Distribution of Means for Sample Size = 20

SD= 3

(Distribution plot, x-axis 0–100, narrower spread with a red horizontal line indicating SD around the mean ~50)

Note. A review of sampling processes that are depicted in sampling simulations (in the left column), the rules with their rationales that explain these processes (in the middle column), and what students often wrongly believe about these processes (in the right column).

The First Iteration: Comparing a Standard Simulation to a Non-simulation Method

In the first iteration, we aimed to gauge whether a standard simulation that emulates popular simulations would provide different learning experiences than a more traditional teaching method that is based on direct instruction through static text and images.

Participant

141 undergraduate students from the researchers' university participated in a one-hour online study to receive participation credits in an introductory Psychology course. Based on self-reports, their ages were between 18 and24,

68% were female, and 65% were white. We expected the experiment to be relevant for students' learning goals as they were required to take at least one statistics course to complete their undergraduate degree.

The Design of Conditions

For the non-simulation condition, we designed a traditional mode of instruction through a computer program. The students first received direct instruction via verbal and pictorial information, and then, attempted to solve graph problems followed by feedback. For the standard simulation condition, we mimicked the features of popular pedagogical simulations overviewed in the first section (see Figure 1). That is, we placed an individual sample distribution on the top of the screen, and the distribution of sample means on the bottom, both expressed as generic histograms.

Students' simultaneous engagement with visual representations and verbal explanations can lead to deeper learning through the integration of their intuitive understanding from visuals and more formal and explicit ways of understanding from language (Aleven & Koedigner, 2002). To this end, we combined interactive dynamic visualizations and verbal explanation prompts together to provide the opportunity for students to integrate visual and verbal information together (See Figure 2).

Figure 2

The Standard Simulation from the First Iteration
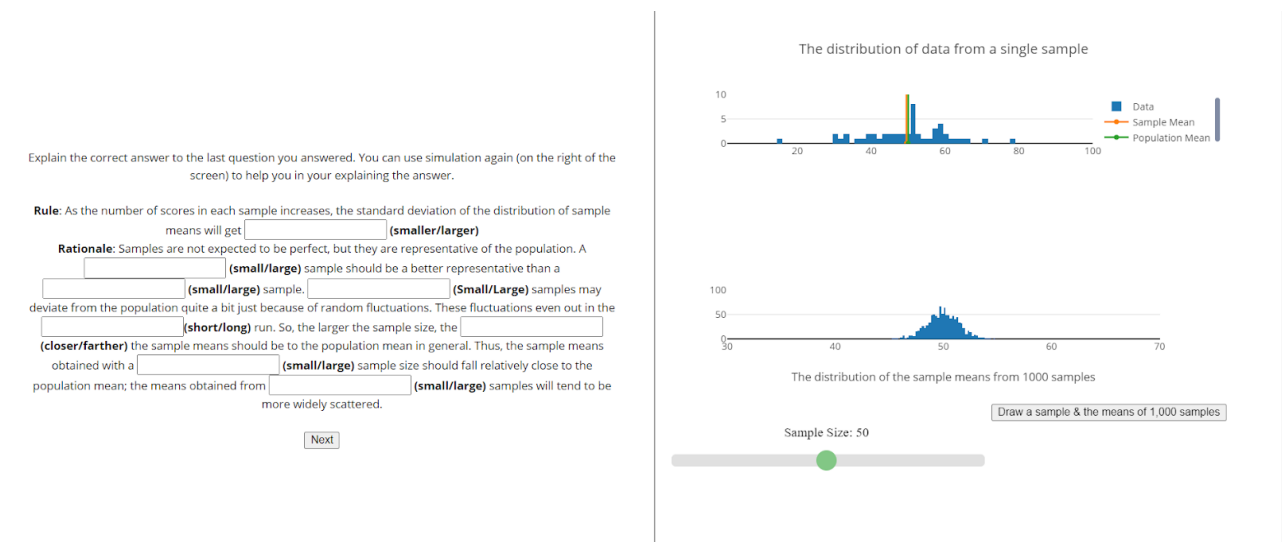
Note. A screenshot from our standard simulation with self-explanation prompts. The simulation group first attempted to solve graph interpretation problems and then explored the correct solution with interactive simulations, augmented by guided self-explanation prompts.

Hypothesis

The previous work advocating use of simulations for teaching sampling distributions argued simulations foster a deeper conceptual understanding of sampling concept (Chance et al., 2004; Cobb & Moore, 1997). Based on this body of work, we hypothesized that the *standard simulation* condition would have significantly better performance in the posttest than the non-simulation condition.

Research Design and Procedures

In an online computer-based experiment, the participants were randomly assigned to one of two groups: Simulation vs. non-simulation group. The intervention consisted of a pretest, learning, and post-test phases. Pretest and post-test

items included 12 identical multiple-choice questions classified as graph, story problems, and rule questions. Additionally, the post-test included two open-ended rule-explanation questions (See Table 2).

Table 2

Example Items

Example Graph Item (5 questions in total):
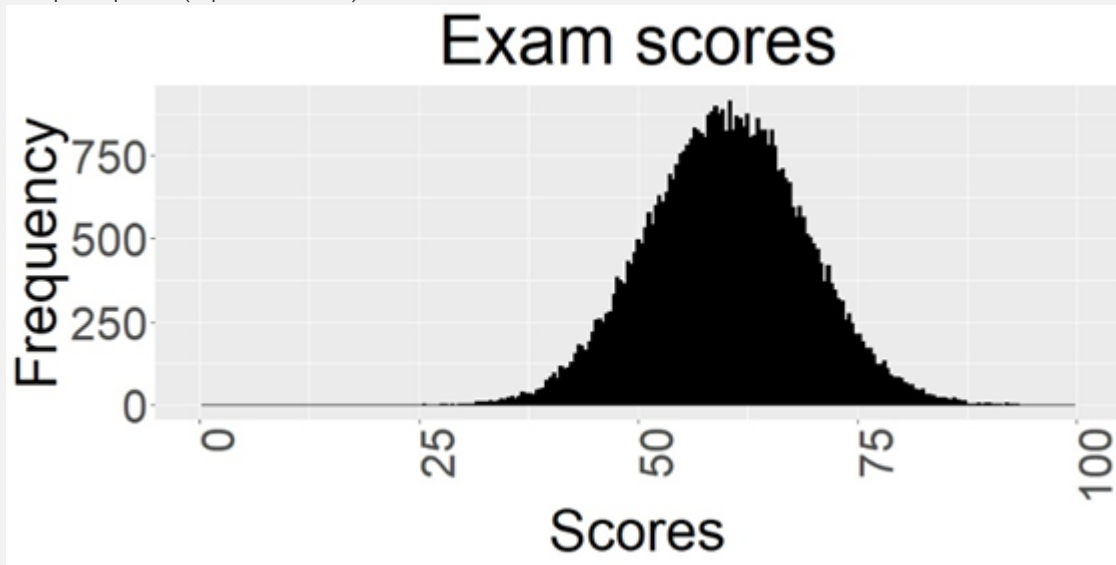
The population distribution for an exam score is displayed above. Below, you see two distributions of the sample means for random samples drawn from the population. One comes from a distribution with sample size of 2. The other comes from a distribution with sample size of 15. Which distribution comes from a situation where the sample size is 15? A. B.*
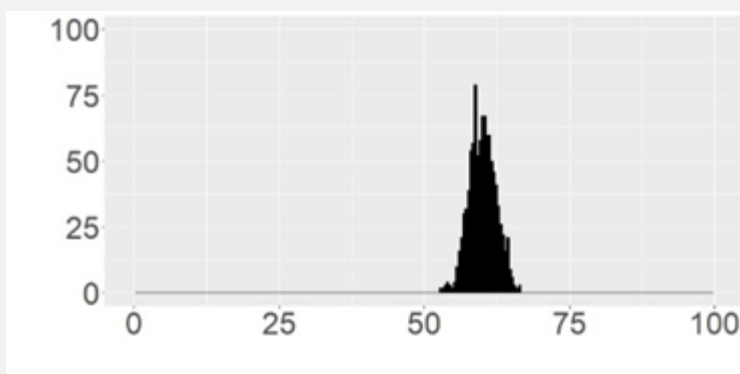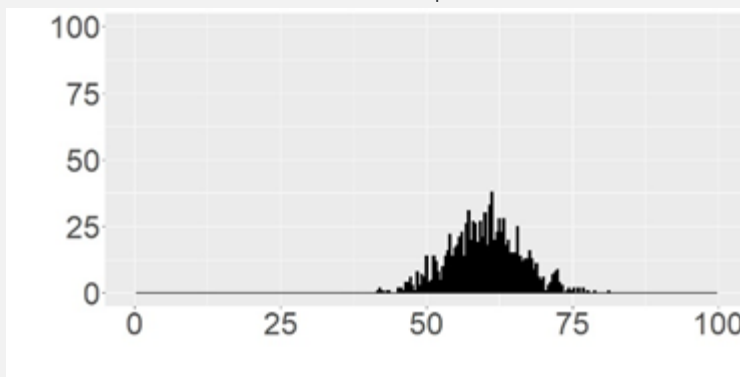
Example Story Problem Item (5 questions in total): American males must register at a local post office when they turn 18. In addition to other information, the height of each male is obtained. The national mean (average) height for 18-year old males is 69 inches (5 ft. 9 in.). Every day for one year, about 5 men registered at a small post office and about 50 men registered at a large post office. At the end of each day, a clerk at each post

office computed and recorded the mean height of the men who registered there that day.One day, you will visit one of the offices. You want to find the office where the mean height of the men is closer to that of the population's. Which office should you go to increase your chances?A. You should go to the small office.B. You should go to the large office.*C. Both have equal chancesD. There is no basis for predicting which post office would have more chances.

Example Rule Question (2 questions in total):Consider any possible population of values and all of the samples of a specific size (n) that can be taken from that population. Below are four statements about the distribution of the sample means. Which one is CORRECT?A. As the sample size increases, the distribution of sample means will have a smaller and smaller standard deviation.*B. As the sample size increases, the distribution of sample means will have a larger and larger standard deviation.C. No matter what the sample size is, the distribution of sample means will have the same standard deviation.D. As the sample size increases, the distribution of sample means will have a similar standard deviation to that of the population.

Explanation questions (2 questions) The sample mean tends to get closer to the population mean as sample size increases. Explain why this is correct._____As the sample size increases, the distribution of sample means will have a smaller and smaller standard deviation. Explain why this is correct._____

Note. '*' identifies the correct answer.

Scoring of verbal data

We conducted a pilot study prior to the actual study. The two authors applied inductive coding to the responses of the rule explanation questions to create the coding scheme. The authors discussed the codes and ensured coding agreement. The data from the actual study was analyzed based on this coding scheme. The response to each question constituted the unit size, which corresponded to one category. Both authors independently coded 20% of data. The interrater agreement for assigning each response to categories was 85% for the first item, and 84% for the second item. After the two authors discussed the differing categorization and achieved a mutual agreement, the first author completed the coding of all data. The authors were blind to which conditions the data were obtained from during the complete coding process.

Results

We measured learning gains for each question type separately. For each problem type (except verbal explanation questions), we ran two statistical analyses. First, ANCOVA on the post-test scores with prior knowledge as a covariate and the condition (simulation vs. non-simulation) as independent variable. Second, we collapsed the conditions and ran a paired t-test to measure overall learning gain from pre- to post-test. (See Table 3).

At graph questions, there was not a significant effect of condition, $F(1, 138) = 1.01$, $p = 0.31$. However, there was an overall learning gain from pre (M = 1.53, SD = 1.12) to posttest (M = 2.46, SD = 1.43), $t(140) = 7.14$, $p < 0.01$.

At story problems, there was not a significant difference between the conditions, $F(1, 138) = 1.89$, $p = 0.17$. Further, there was not any significant difference between pre (M = 2.60, SD = 1.15) and posttest performance (M = 2.45, SD = 1.27), $t(140) = 1.59$, $p = 0.11$.

At rule problems, there was not a significant difference between the conditions, $F(1, 138) = 0.19$, $p = 0.65$. However, there was a significant learning gain from pre- (M = .96, SD = 0.69) to post-test (M = 1.20, SD = 0.77), $t(140) = 3.61$, $p < 0.01$.

Table 3

Iteration 1: Pre and Posttest Scores

	Graph Questions		Story Problems		Rule Questions	
	Sim	Non-sim	Sim	Non-sim	Sim	Non-sim
Pretest	32.8%	28.2%	52%	52.2%	50%	46%
Posttest	52%	46.4%	47.2%	50.8%	59.5%	60.5%

Note. Average percentage of correct answers in pre and post test for simulation (Sim) and non-simulation (Non-sim) group.

For verbal responses, we ran a Pearson's chi-square test on the coded responses. For the first item (See Table 2), "Explain why the sample mean tends to get closer to the population mean as sample size increases", there was a significant association between the response categories and the condition ($\chi 2(7) = 16.08$, p = 0.02).

Table 4

Percentage Responses to the First Item: "Explain why the sample mean tends to get closer to the population mean as sample size increases."

Response category	Non-Sim group	Sim group
Larger sample is a better representation of the population*	55%	37%
Bigger sample size results in less likelihood and/or impact of outliers	10%	18%
As sample size increases, the standard deviation increases	0%	12%
Nonsense explanation	15%	17%

Note. '*' identifies the correct explanation.

For the second item, "Explain why the standard deviation of the distribution of sample means will get smaller as sample size increases", there was not a significant association between response categories and the condition ($\chi 2(5) = 6.53$, p = 0.25; see Table 5).

Table 5

Percentage Responses to the Second Item: "Explain why the standard deviation of the distribution of sample means will get smaller as sample sizes increases."

Response category	Non-Sim group	Sim group
Nonsense explanation	60%	52%
More sample means are closer to the population mean as sample size increases.*	15%	21%
More data are closer to the average as sample size increases	8%	11%
A larger sample size leads to less likelihood and/or impact of outliers in data	11%	4%

Note. '*' identifies the correct explanation.

Discussion

In the first iteration, we ran an experiment to compare a simulation-based vs. non-simulation learning method for a sampling distribution task. In the non-simulation group, participants first received direct instruction with verbal information and pictures and then, solved graph problems with feedback. The simulation group first attempted solving graph problems with feedback and then, explored the solution through interactive simulations accompanied by self-explanation prompts. We measured learning with four types of test items: graph problems, story problems, rule statement items, and open-ended explanation items.

Both groups increased performance at similar levels from pre- to post-test for graph problems. However, neither improved their learning at story problems. These results suggest both groups gained mostly a superficial understanding of the concepts by attending to the physical features of the graphs (e.g., "the distribution of sample means will look narrower when the sample size is larger").

The groups improved their learning of rules to a similar level from pre- to post-test. However, neither were able to explain the rationale of the rule. For the first open-ended explanation item ("Explain why sample mean tends to get

closer to the population mean as sample size increases"), an answer that would indicate understanding of the sampling process could be "it is less likely that all numbers will be low or high for a large sample. As a result, it is more likely that low and high numbers will average each other out in larger samples". Rather, students mostly gave a superficial response that would be expected without any exposure to instruction, such as "larger samples are a better representation of the population". 55% of the students in the non-simulation group and 37% of the students in the simulation group gave this kind of explanation for the item.

Unfortunately, the simulation group (37%) used this superficial, but nevertheless correct, explanation less than the non-simulation group (55%) and instead, more often displayed misconceptions in their explanation. Some of them (12 %) believed that the standard deviation would increase with larger samples (note that this answer never appeared in the non-simulation condition). This is a surprising kind of explanation about why larger samples tend to give a better estimation of the population mean. Thus, the simulation-based learning method has overridden some students' intuition and caused an unusual type of misconception.

Prior work with simulations sheds light on this interesting result. Adams et al. (2008) found in their physics simulations that when students see items that look superficially similar (or different) to each other (such as shape and color), they believe this superficial similarity (or difference) also meant a deeper conceptual similarity (or difference). In the domain of sampling simulations, van Dijke-Droogers et al. (2021) observed that simply differentiating the color and shape of the sampling distribution graph from individual sample graphs decreased students' conceptual confusions. In the light of this evidence, we believe students' confusion in the current study resulted from interacting simultaneously with the individual sample distribution (on top) and the distribution of sample means (on bottom) graphs which look similar to each other (See figure 2). Students learned the rule that the standard deviation of the distribution of sample means changes while engaging with the graph at the bottom, but wrongly associated this rule with the single sample graph at the top. Given that the graphs looked like each other, students wrongly believed this visual similarity also meant conceptual similarity. The confusing perceptual aspects of the graphs might also explain why combining them with self-explanations did not lead to better learning unlike in the previous work (Aleven & Koedinger, 2002).

We caution that the results from this iteration might not apply to all *standard simulations* used in statistics education. This iteration had important limitations that constrain generalizability. First, the experiment lasted for an hour which might be shorter than the time allocated to this topic in real classrooms. Furthermore, in real classroom settings, students might have the additional opportunity to revisit the topic of sampling distributions several times through practice. Therefore, the mediocre overall results might be attributable to the limited exposure of students to the material. Second, the study took place as an online experiment. Students might have made less effort in the study than they would typically make if they engaged with the simulation in the classroom in the company of their peers and teacher. Nevertheless, the fact that there were still differences between the two design conditions in terms of learning outcomes even at a short experimental manipulation suggest that it is worthwhile paying attention to the specific design choices in simulations.

The Second Iteration: Developing a Grounded Sampling Simulation

The findings from the first iteration suggested that the standard simulation was not particularly helpful in fostering conceptual understanding, and overall, resulted in similarly mediocre results compared to a more traditional form of teaching. Moreover, the simulation even created an additional misconception. Our interpretation is that representing different types of distributions by similar, generic histograms causes conceptual confusion. The advantage of the standard kinds of histograms (shown in figures 1 and 2) is that the same graphical format can be used to represent a huge variety of different types and structures of data. The disadvantage is that important differences between these structures are obscured.

For the second iteration, we aimed to engage students' natural perceptual learning capabilities more effectively through a grounded cognition perspective. Inspired by the notion that complex cognition is grounded in well-learned perceptual

and bodily processes, we employed concrete and familiar visual design elements to foster better sense making of abstract sampling processes.

However, we did not solely rely on visuals for effective teaching of abstract and difficult concepts. As in the first iteration, we aimed to combine the affordances of verbal and visual information for more integrated and deeper learning of concepts. To this end, we designed a paper task sheet to guide students' interactions with the simulation.

The Design of the Grounded Simulation and Task Sheet

Based on the grounded and embodied learning framework (Goldstone et al., 2010; Nathan, 2021), in our second iteration, we aimed for abstract concepts and processes to be represented with familiar objects and concrete animations in the simulation (See the simulation at https://pcl.sitehost.iu.edu/robsexperiments/tests&examples/tokenSampling/iteration2.html). This core design principle came into play with three main design choices. First, we replaced the standard bars and bins of the histograms with icons sitting on top of each other to ease the representational competence required for grasping histograms. Each icon represents a single instance of the population or a single mean taken from a sample. Second, in order to avoid the confusion that happened in the first iteration, we graphically differentiated the representation of the distribution of sample means from actual observations by using different icons for each. Third, we dynamically animated the statistical processes such as calculating the means from the sample so that students could construct a spatial representation corresponding to the process of "cancellation of low and high scores" while trying to understand the distribution of means (See table 6).

Table 6
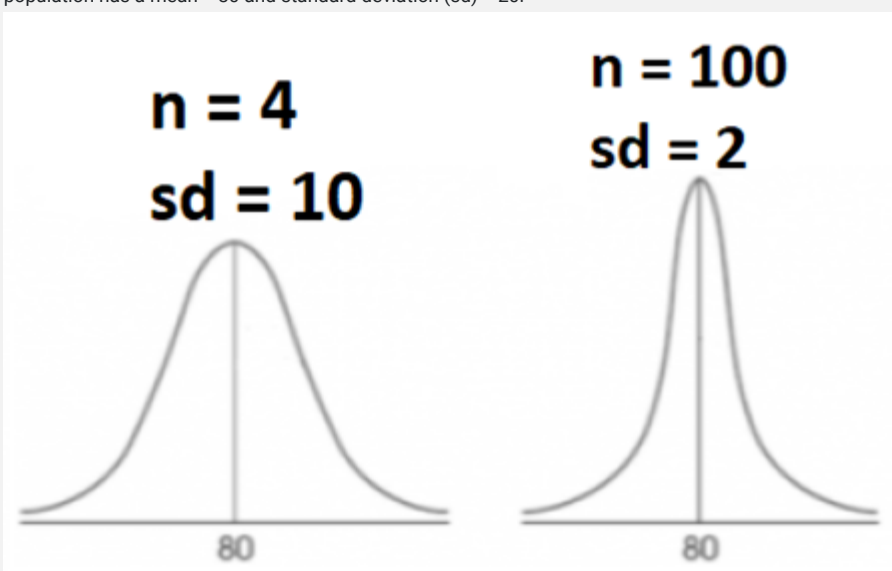
Visual Design Elements in the Grounded Simulation

Design principles	Instantiation of the principle	What feature of the standard histogram is replaced?
1. Use icon-based histograms to ease the representational competence required for understanding histograms	representing each individual data point by an icon of a gear	bars that summarize a collection of data
	representing the value of each data point by the number of spikes on a gear	position on the X-axis
	grounding frequency naturally in terms of counts of icons sitting on top of each other in a column	frequency or density on the Y-axis
	individual colors to reinforce that each icon is representing a unique individual in the population	uni-colored graph
2. Visually distinguish between graphs that depict different processes	representing the collection of means with a white color, which is different from the colors of the actual gears, and with the letter "M" inside the gear to emphasize what is collected here is "means of different sets of sampled gears", not the gears themselves.	Identical visualizations of the histograms of the distribution of a sample and the collection of means of samples
3. Dynamically and spatially animate statistical processes	randomly sampled observations fall down from the population to the sample graph	NA
	when a sample falls down, their original place in the population flashes, which graphically depicts the concept of sampling with replacement.	NA
	Animating the process of computing the means from the sample observations by having the sampled gears converge to their mean, so that students can construct a spatial representation corresponding to the process that "low and high numbers average each other out".	NA

The accompanying task sheet included two phases of activities. In the first phase, the activities guided students during their interaction with the simulation with prediction and test questions, drawing tasks, and oral reasoning tasks. In the second phase, the students stopped interacting with the simulation and read information from the paper which

explained the rules with text and histograms. After they finished reading, they answered post-test questions which required answering graph questions and story problems and giving rationales for the rules, as in the first iteration (See Table 7). The students discussed their responses aloud with the interviewer. The interviewer did not give feedback on students' answers; however, she explained any unclarities with the instruction and probed students for further explanation.

Table 7

Paper Task Sheet

During the interaction with the simulation

Activity	Example item
Predicting and testing (with drawing and written explanations)	Think about if we were to take 2 gears randomly from the population, find their average (mean) number of teeth, and record the average. If we repeatedly did this and collected a list of 2-gear averages, how would this collection look like? What would be the range of averages we would see? Draw the diagram below in the blank space above 3a.Predict: Do the same task we did above, but this time think about taking 10 gears at a time instead of 2. Draw the diagram below in the blank space above 3b. Means of sample gears (n=2) Means of sample gears (n=10) Compare your two diagrams above. What happens as the sample size increases? Write your thoughts briefly _____Test: Now, go to the simulation and collect sample means with sizes 2 and 10. Do you observe any changes as the sample size increases? _____Conclusion: After seeing the simulation, do you change your thoughts? _____
Further reasoning questions	Wisdom of the crowdIn 1906, British scientist Sir Francis Galton asked 787 villagers to guess the weight of an ox. None of them got the right answer, but when Galton averaged their guesses, he arrived at a near perfect estimate. Often the average of many people's guesses is closer to the actual number than most individuals' guesses. Why? Unusual population distributionsWhat happens to the collection of sample means when we sample from a population with two distinct clumps? Population sizeWhat would happen if we reduce/increase population size? Does sampling 10 from a population of 50 still come as close to the mean as sampling from a population of 100?

After the interaction with the simulation

Information sheet	Below are two sampling distributions of means with two different sample sizes obtained from the same population. The population has a mean = 80 and standard deviation (sd) = 20.
	Notice that as n gets larger, the standard deviation of the distribution of sample means gets smaller, with the sample means tending to approximate the population mean more closely with larger sample size.
Rule explanation	Identical open-ended items with the first study to gauge students' reasoning about the rules
Graph questions	Similar items with the first study that involve identifying the sampling distribution of means histograms with smaller vs larger sample size.

Story problems	Maternity task A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%; sometimes lower. Which hospital do you think is more likely to find on one day that more than 60% percent of the babies born were boys? a) Large hospital b) Small hospital* c) They are the sameMedical survey A medical survey is being held to study some factors pertaining to coronary diseases. Two teams are collecting data. One checks three men a day, and the other checks one man a day. These men are chosen randomly from the population. Each man's height is measured during the checkup. The average height of adult males is 5 ft 10 in., and there are as many men whose height is above average as there are men whose height is below average. The team checking three men a day ranks them with respect to their height, and counts the days on which the average height of men is more than 5 ft 11 m. The other team merely counts the days on which the man they checked was taller than 5 ft 11 in. Which team do you think counted more such days?Team checking one man*

Note. '*' identifies the correct answer.

Participants

Nine undergraduate students from the researchers' university participated in a one-hour face-to-face study at the exchange of course credits for an introductory psychology course. The participants were different individuals from the ones in the first iteration. Based on self-reports, their ages were between 18 and 24; six were female and three were male; and their majors were psychology (N=3), business school (N=2), human development (N=1), interior design (N=1), biology (N=1), and finance (N=1). Six students had been introduced to basic statistics topics in high school or college while three of them reported no prior statistics background.

Research Design

The study took place as a think-aloud interview study with individual students at a research laboratory. The first author and each individual student sat next to each other across a desktop computer. The student interacted with the simulation displayed on the computer while reading the directions from the paper task sheet on the table. The interviewer repeatedly told the student to think aloud during this process, asked further follow-up questions, gauged students' understanding of the tasks, and responded to any clarification questions from the student.

Data Analysis

We transcribed the interview recordings and subjected them to content analysis by applying inductive coding to raw data. The authors met frequently and revised the codebook until they ensured coding agreement before reporting.

Findings

The first impressions of the graph. Students were mostly able to interpret the new simulation without much difficulty. Seven students correctly identified the change in the horizontal dimension by pointing out the increasing number of spikes on the gears from left to right. On the other hand, two students suggested gears were jumping higher or there was a little wave from left to right. For those two students with no previous statistics background, the random and irrelevant features of the graph overshadowed conceptually important patterns.

Learning with predict and test strategy

The predicting and testing pedagogy was largely successful. Except for one or two students for each predict-and-test question, the students were able to reach the correct conclusion about how sample size affects the distribution of samples and sample means after interacting with the graphs. For the remaining few cases, students still misinterpreted the pattern even after testing it on the graph, and sometimes it took a few tests with different graphs for them to identify the important patterns. For example, in one case, a student had predicted that mean would increase with larger sample size. When he tested his prediction on the screen, even though he saw that the mean tended to be closer to the population mean with a larger sample size, he missed this pattern. Instead, he suggested that the mean stays the same whatever the sample size is. This student, however, was able to reach the correct conclusion after experimenting with a second graph. Another was distracted by the unrelated and random patterns such as the colors of the dots, the speed

of how gears fall, and the shape they make when falling. This student finally reached the correct conclusion after the interviewer directed her attention to the relevant feature which was the sample mean line.

Figure 3

An Example Student Drawing
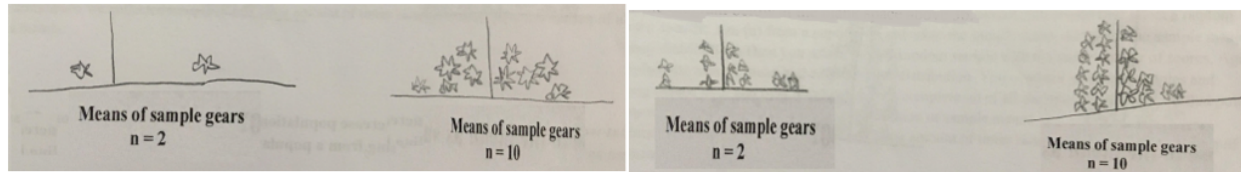
Note. A student's drawings when *predicting* how the graph is going to look (on the left) and their updated graph after *testing* it with the simulation (on the right). The student seems to have *diminished* their confusion between the *number of the gears* and *the number of the means of gears* from predicting to testing phase.

Grasping the process that "low and high numbers average out for larger samples"

One of the important learning objectives was for students to grasp the process that larger samples tend to estimate the population mean better because it is more likely that low and high numbers average out as sample size increases. To this aim, we animated this process as shown in Table 6. The results suggest that students mostly showed a quick understanding of this process. For example, for the wisdom of crowd question (See table 7), six students successfully suggested that lower and higher numbers average out to the correct weight of the ox. More interestingly, when asked if the question was related to what they saw in the simulation, the responses suggested that students were not aware of the connection. This is consistent with previous results suggesting that learners can benefit from an earlier situation when presented with a subsequent analog even when they show no explicit awareness of the connection between the two (Day & Goldstone, 2011).

Separate affordances of simulation vs. task sheet

After the students had finished several inquiry-based activities with the simulation, they read a paper task sheet which mimicked a textbook. This task sheet explicitly stated the rules the students had discovered when interacting with the simulation. Next, the students solved graph problems that asked them to identify histograms in relation to sample size. When solving these problems, they referred to the rule they read on the paper instead of their own discoveries with the simulation such as "That graph belongs to the smaller sample size because of the rule on the paper". When they were trying to reason about a rule, though, the students referred to the visual aspects of the simulation, for example, "Oh, the standard deviation is smaller for the graph with the larger sample because the means should closely sit on top of each other".

Post-test performance

The aggregate percentage of correct answers were 72% for the graph questions and 78% for the story problems (we did not ask rule selection questions in this iteration). Furthermore, as in the first iteration, we analyzed the verbal data from the rule explanation questions (See table 8 and 9). Students mostly showed a good understanding of the relationship between the sample size and the standard deviation of the distribution of sample means.

Table 8

Percentage Responses to the First Item: "Explain why sample mean tends to get closer to the population mean as sample size increases."

Response category	number (%)

Small and large numbers average each other out.**	2 (22.22%)
More points allow a more accurate average.*	2 (22.22%)
There is less variability	2 (22.22%)
There is a larger pool of numbers.	1 (11.11%)
Sample becomes more proportional to the entire population.	1 (11.11%)
More sample means are closer to the population mean as sample size increases	1 (11.11 %)

"**" The ideal explanation *"*"* Correct, but superficial explanation

Table 9

Percentage Responses to the Second Item: "Explain why the standard deviation of the distribution of sample means will get smaller as sample sizes increases."

Response category	number (%)
More sample means are closer to the population mean as sample size increases.*	7 (77.77%)
There is a larger range.	1 (11.11%)
Insufficient explanation	1 (11.11%)

*Note. **'* identifies the correct explanation.

Discussion

In the second iteration, we investigated students' learning of sampling concepts through our new grounded simulation. Augmented by a task sheet, students first predicted how the sample size would affect the features of the distributions and then, experimented with the simulation to test their prediction. After this inquiry-based activity, the students read information on the sheet which explicitly stated the rules that they had discovered while interacting with the simulation. As in the first iteration, the students' learning was measured in a post-test through graph questions, story problems, and rule explanation items. The results suggest students overall made few conceptual confusions, and they were able to explain causal mechanisms of sampling processes even without any statistics background. A large percentage of students provided correct answers to graph problems, showed good reasoning at story problems, and wrote quality explanations of rules.

An important caution is that this iteration was designed as an initial study to test our new simulation with a small sample size (n=9). While the results show initial promise of designing a simulation based on grounded and embodied learning considerations, strong generalized conclusions should not be drawn considering the nature of the study and lack of control conditions. Our next research goal is to compare a standard and grounded simulation through a rigorous controlled experiment with a larger sample size to allow direct comparison between different approaches to simulations. Still, the initial results reveal important insights about students' interactions with the simulation, the role of verbal information that accompanies the simulation, and the choice of instructional activities in which the simulation is situated.

The simulation and the task sheet had separate affordances that contributed to overall learning. What was the contribution of the visual elements of the simulation? Our results suggest perceptual features do not contribute to learning, only at superficial levels. Perceptual learning can be powerful and deep when the perceptual features are carefully designed in a way that reveals the important concepts. More specifically, students' observations of the falling gears collapsing into a mean facilitated their understanding of the sampling process. Students transferred what they had observed from these animations to their discussions of the story problems as their verbal explanations suggested. More interestingly, students were apparently unaware of the structural connections between the simulation and

subsequent stories, which supports prior findings by Day and Goldstone (2011, 2012) which suggested that perceptual learning is a powerful, automatic, and implicit mechanism for transfer.

Furthermore, the first author observed during interviews that students often used gestures that mimic the animations while they were discussing their reasoning. For example, they brought two outstretched hands or fingers together while talking about the process of obtaining a mean from a sample. Prior research suggests that gestures ease the understanding of abstract and difficult mathematical concepts and processes (Nathan et al., 2021), and students' embodiment of the graphs suggest deeper engagement than merely watching them (Gerofsky, 2011). Similarly, we suspect animations, and the students' spontaneous gestures mimicking these animations, might have had facilitative effects on their understanding of the difficult concept of sampling distribution.

What about the role of verbal information in the task sheet? The results suggest that providing students with verbally explicit rules, which mimicked a typical textbook, had important influences on their learning. When discussing the rules, students often referred to the rules stated on the task sheet even though they had themselves discovered these rules while experimenting with the simulation. It is possible that perceptual learning through simulations was rather implicit, and the task sheet served as an explicit verbal memory aid. The verbal form of information serves as a tool for discourse. However, verbal information can be rather rote and inert when students do not understand their use (Aleven & Koedigner, 2002). In this case, the abstract verbal rules were first grounded in concrete visual animations. In other words, perceptual learning might have served as a grounding for meaningful learning of explicit verbal information.

We are further improving the design of our simulation for a third iteration by implementing design principles from embodied learning (Alibali & Nathan, 2018) and concreteness fading (Fyfe et al., 2014). We mentioned students spontaneously used gestures that mimicked the animations they saw on the screen. A follow-up question is whether asking students to gesture helps them to understand the sampling concepts better? To answer this question, we are currently designing a task in which students are instructed to mimic the animations with their hands when they are watching a video of the simulation.

While our icon-based approach to sampling simulations suggests promising results, an important caveat is such concrete representations can place limits on students' transfer of their learning in some cases (Goldstone & Sakamoto, 2003). For example, students learning sampling processes through our iconic graphs might not apply their knowledge when they encounter a generic histogram in a textbook. Therefore, a more promising approach might be to start with concrete representations to make statistics more accessible to novices, and then gradually fade them into more idealized ones so that students can effectively use the generic histograms that statisticians typically use. To this end, in a third iteration, we have combined the grounded and standard graphs into a single simulation. Over the course of training, the richer, more contextualized depiction is replaced with simpler rectangles (see the simulation at https://pcl.sitehost.iu.edu/robsexperiments/tests&examples/tokenSampling/iteration3.html).

Overall, we conclude that a deliberate design approach informed by theory and empirical testing is important for providing incremental improvements to pedagogical simulations. We also believe that instructional activities that situate the simulation are at least as important as its design elements. In the final section, we reflect on design aspects of the statistics simulations and the accompanying instructional activities based on the results from our two iterations.

General Discussion

With the accessibility of modern technologies, computer-based interactive simulations are becoming increasingly common in data literacy education with the purpose of making statistics concepts accessible to novice students. We argued these simulations often do not deliver on their promises because they are designed from an expert perspective; that is, they do not ground the data representations in students' primary experiences, their perceptions, or bodily actions. As an alternative, we proposed a novel statistics simulation designed based on aGEL framework. In the current work, we specifically focused on icon-based graph representations. Our data suggests that our icon-based, dynamically animated graph shows initial promise in making difficult sampling concepts more accessible to students. Similarly, prior

research has shown iconic representations to dramatically improve people's learning of base-rate concepts (Brase, 2014). From an ecological perspective, iconic representations tap into people's ability to effortlessly track frequency information as it approximates the presentations of frequencies of objects people see in their everyday environment (Brase & Hill, 2015).

The initial results from the current work suggest several dimensions to test with simulations for future work. We view simulations to be complements to textbooks and lectures rather than their replacements. Prior work suggests active inquiry helps students learn more effectively from the subsequent instruction (Schwartz & Martin, 2004). Linguistic materials and verbal instructions turn intuitive and implicit kinds of learning gained from the simulations into explicit and verbalizable tools for powerful and effective discourse. Accordingly, we designed simulation activities to precede the verbal information sheet, which served as a more standard form of instruction. Future research and design studies should further explore how to combine different forms of media as complements to simulation.

An important result from the think-aloud interviews was students often did not see what was happening in the simulations objectively, rather they saw it as a combination of their prior beliefs and what was happening on the screen. Providing several experimentation opportunities with simulations helped students gradually update their prior beliefs in the direction of actual results. An important implication for future research and design is to aim to better understand the relationship between students' prior beliefs and the role of repeated practice in inquiry activities with simulation.

In the current work, we focused on the combination of icon-based graphs, spatial representations of mathematical processes, and gestures as one possible way of instantiating grounded and embodied learning for sampling simulations. However, the GEL framework is not only constrained to individual student's perception and body-based activities. The GEL framework views social and cultural experiences to be critical components of *grounding* scientific conceptualizations. Accordingly, future work should consider different instantiations of sampling simulations based on students' personal and social experiences for a complementary perspective on grounding.

Conclusion

Overall, the three iterations of the statistical sampling simulation have underscored the pedagogical benefits of providing grounded models for learners. It may be tempting to prioritize equations and summary rules when teaching statistical concepts because these formalisms are designed to be broadly applicable to an unlimited number of scenarios. However, the problem with these generic formalisms is that future scenarios do not clearly present themselves to the learner as being governed by the formalisms. Instead, what is needed is for learners to develop new ways of seeing future scenarios as instances of what they have previously learned. Perceptually grounded simulations provide the kind of experiences learners need to develop these new ways of seeing and interpreting. Accordingly, we encourage instructional designers and teachers to resist the tendency to put perception and deep understanding in opposition. Superficial appearances can indeed be misleading, but not all perceptions are superficial. Perceptual and interactive models offer promise in promoting grounded understanding and transfer that go beyond those achievable with formalisms because they change how learners naturally see their world.

References

Abrahamson, D. (2014). Rethinking probability education: perceptual judgment as epistemic resource. In E. J. Chernoff & B. Sriraman (Eds.). *Probabilistic Thinking: Presenting Plural Perspectives* (pp. 239–260). Springer. https://doi.org/10.1007/978-94-007-7155-0_13

Adams, W. K., Reid, S., LeMaster, R., McKagan, S. B., Perkins, K. K., Dubson, M., & Wieman, C. E. (2008). A study of educational simulations part i - engagement and learning. *Journal of Interactive Learning Research*, 19(3), 397–419.

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer based cognitive tutor. *Cognitive science*, 26(2), 147-179. https://doi.org/10.1207/s15516709cog2602_1

Alibali, M. W., & Nathan, M. J. (2018). Embodied cognition in learning and teaching: action, observation, and imagination. In F. Fischer, C. E. Hmelo-Silver, S. Goldman, & P. Reimann (Eds.). *International Handbook of the Learning Sciences* (pp. 75-85). Routledge. https://doi.org/10.4324/9781315617572-810.4324/9781315617572-8

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355–365. https://doi.org/10.1037/0003-066X.51.4.355

Barab, S., & Squire, K. (2004). Design-based research: putting a stake in the ground. *Journal of the Learning Sciences*, 13(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics*, 88(3), 291–303. https://doi.org/10.1007/s10649-015-9593-3

Boyle, E. A., MacArthur, E. W., Connolly, T. M., Hainey, T., Manea, M., Karki, A. & van Rosmalen, P. (2014). A narrative literature review of games, animations and simulations to teach research methods and statistics, *Computers & Education*, 74, 1-14. https://doi.org/10.1016/j.compedu.2014.01.004

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, 15(2), 284-289. https://doi.org/10.3758/PBR.15.2.284

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 6, 340. https://doi.org/10.3389/fpsyg.2015.00340

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking (pp. 295-323)*. Springer. https://doi.org/10.1007/1-4020-2278-6_13://doi.org/10.1007/1-4020-2278-6_13

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723

Common Core State Standards Initiative. (2022). Retrieved July 19, 2022, from http://www.corestandards.org/

Day, S. B., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 37, 551-567. https://psycnet.apa.org/doi/10.1037/a0022333

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist,* 47, 153-176. https://doi.org/10.1080/00461520.2012.696438

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., . . . Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences*, 19(10), 603–615. https://doi.org/10.1016/j.tics.2015.07.013

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education*, 29(1), S27-S39. https://doi-

org.proxyiub.uits.iu.edu/10.1080/10691898.2020.1860725

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, 26(1), 9–25. https://doi.org/10.1007/s10648-014-9249-3

Gehrke, M., Kistler, T., Lübke, K., Markgraf, N., Krol, B., & Sauer, S. (2021). Statistics education from a data-centric perspective. *Teaching Statistics*, 43(1), 201–215. https://doi.org/10.1111/test.12264

Gerofsky, S. (2011). Seeing the graph vs. being the graph: gesture, engagement and awareness in school mathematics. In G. Stam & M. Ishino (Eds.). *Integrating Gestures* (pp. 245–256). John Benjamins.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science,* 2, 265-284. https://doi.org/10.1111/j.1756-8765.2009.01055.x

Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science*, 1-8. https://doi.org/10.1177/0963721417704888

Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology,* 46, 414-466. https://doi.org/10.1016/S0010-0285(02)00519-4

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences,* 14, 69-110. https://doi.org/10.1207/s15327809jls1401_4

Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *The Journal of the Learning Sciences,* 17, 465-516. https://doi.org/10.1080/10508400802394898

Gravetter, F. J., Wallnau, L. B., Forzano, L. A. B., & Witnauer, J. E. (2020). *Essentials of statistics for the behavioral sciences*. Cengage Learning.

Hegarty, M. (2011), The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science,* 3, 446-474. https://doi.org/10.1111/j.1756-8765.2011.01150.x

Hokor, E. K., Apawu, J., Owusu-Ansah, N. A., & Agormor, S. (2021). Preservice teachers' misconceptions in solving probabilistic problems. *Pedagogical Research*, 7(1), em0112. https://doi.org/10.29333/pr/11441

Holmes, A. G. D. (2020). Researcher positionality--a consideration of its influence and place in qualitative research--a new researcher guide. *Shanlax International Journal of Education*, 8(4), 1-10. https://doi.org/10.34293/education.v8i4.3232

Jacob, B., & Doerr, H. M. (2014). Statistical reasoning with the sampling distribution. *Quadrante*, 23(2), 123–146. https://doi.org/10.48489/quadrante.22905

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, 22(2). https://doi.org/10.1080/10691898.2014.11889701

Kellman, P. J., & Massey, C. M. (2013) Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.). *The Psychology of Learning and Motivation* (Vol. 58, pp. 117-165). Elsevier Inc.

Lehrer, R., & Schauble, L. (2017). Children's conceptions of sampling in local ecosystems investigations. *Science Education*, 101(6), 968-984. https://doi.org/10.1002/sce.21297

Lovett, M., Meyer, O., & Thille, C. (2008). The open learning initiative: measuring the effectiveness of the oli statistics course in accelerating student learning. *Journal of Interactive Media in Education*, 1-16. http://jime.open.ac.uk/2008/14

Loy, A. (2021). Bringing visual inference to the classroom. *Journal of Statistics and Data Science Education*, 29(2), 171– 182. https://doi.org/10.1080/26939169.2021.1920866

Magana, A. J., Hwang, J., Feng, S., Rebello, S., Zu, T., & Kao, D. (2022). Emotional and cognitive effects of learning with computer simulations and computer videogames. *Journal of Computer Assisted Learning*, 38(3), 875– 891. https://doi.org/10.1111/jcal.12654

Maxara, C., & Biehler, R. (2010). Students' understanding and reasoning about sample size and the law of large numbers after a computer-intensive introductory course on stochastics. In C. Reading (Ed.). *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics* (pp. 1-6). International Statistical Institute. https://iase-web.org/documents/papers/icots8/ICOTS8_3C2_MAXARA.pdf

Mayer R.E. (2009). *Multimedia Learning*. Cambridge Univ. Press.

Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology,* 84(4), 444–452.

Mayer, R. E., Hegarty, M., Mayer, S., & Campbell, J. (2005). When Static Media Promote Active Learning: Annotated Illustrations Versus Narrated Animations in Multimedia Instruction. *Journal of Experimental Psychology: Applied,* 11(4), 256–265. https://psycnet.apa.org/doi/10.1037/1076-898X.11.4.256

Nathan, M. J. (2012). Rethinking formalisms in formal education. *Educational Psychologist,* 47(2), 125-148. https://doi.org/10.1080/00461520.2012.667063

Nathan, M. J. (2021). *Foundations of Embodied Learning: A Paradigm for Education*. Routledge. https://doi.org/10.4324/9780429329098

Nathan, M. J., Schenck, K. E., Vinsonhaler, R., Michaelis, J. E., Swart, M. I., & Walkington, C. (2021). Embodied geometric reasoning: Dynamic gestures during intuition, insight, and proof. *Journal of Educational Psychology*, 113(5), 929– 948. https://doi.org/10.1037/edu0000638

National Research Council (2011). *Learning Science Through Computer Games and Simulations*. The National Academies Press. https://doi.org/10.17226/13078

Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. *IRE Transactions on information theory*, 2(3), 61-79.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science*, 238(4827), 625-631, https://www.science.org/doi/abs/10.1126/science.3672116

Ploetzner, R., Berney, S., & Bétrancourt, M. (2020). A review of learning demands in instructional animations: The educational effectiveness of animations unfolds if the features of change need to be learned. *J. Comput. Assist. Learn.* 36(6), 838-860. https://doi.org/10.1111/jcal.12476

Rau, M. A. (2017). Conditions for the effectiveness of multiple visual representations in enhancing stem learning. *Educational Psychology Review*, 29(4), 717–761. https://doi.org/10.1007/s10648-016-9365-3

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. Lajoie (Ed.). *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12*, 233-273. Routledge.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129-184. https://doi-org.proxyiub.uits.iu.edu/10.1207/s1532690xci2202_1

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, 16(1), 1–40. https://doi.org/10.1207/s15516709cog1601_1

Sutherland, S., & Ridgway, J. (2017). Interactive visualizations and statistical literacy. *Statistics Education Research Journal*, 16(1), 26-30. https://doi.org/10.52041/serj.v16i1.21010.52041/serj.v16i1.210

van Dijke-Droogers, M., Drijvers, P., & Bakker, A. (2021). Introducing statistical inference: design of a theoretically and empirically based learning trajectory. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-021-10208-8

Zhang, I. Y., 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. https://doi.org/10.1111/cogs.12940

Zhang, I. Y., Tucker, M. C., & Stigler, J. W. (2022). Watching a hands-on activity improves students' understanding of randomness. *Computers & Education*, 185. https://doi.org/10.1016/j.compedu.2022.104545

Sebahat Gok

Indiana University

Sebahat Gok is a Ph.D. Candidate in the program of Cognitive Science and the department of Instructional Systems Technology at Indiana University. Her current research interests focus on application of the theories of grounded and embodied cognition to educational technologies.
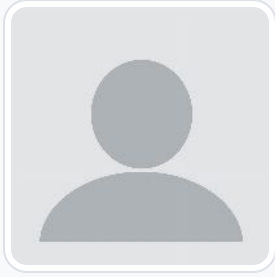
Robert L. Goldstone

Indiana University

Robert Goldstone is a Distinguished Professor in the program of Cognitive Science and the department of Psychological and Brain Sciences at Indiana University. His research interests include concept learning and representation, perceptual learning, educational applications of cognitive science, decision making, collective behavior, and computational modeling of human cognition.