The Counterintuitive Interpretations Learned from Putatively Intuitive Simulations

Sebahat Gok (sebgok@iu.edu)
Program in Cognitive Science & Department of Instructional Systems Technology, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Program in Cognitive Science & Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

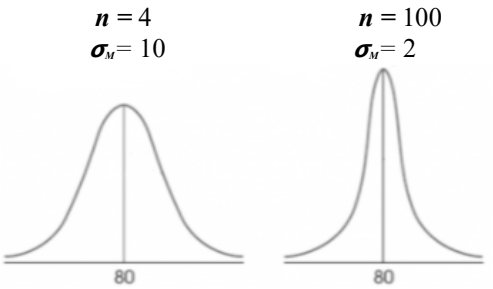
Reasoning about sampling distributions is notably challenging for humans. It has been argued that the complexity involved in sampling processes can be facilitated by interactive computer simulations that allow learners to experiment with variables. In the current study, we compared the effects of learning sampling distributions through a simulation-based learning (SBL) versus direct instruction (DI) method. While both conditions resulted in similar improvement in rule learning and graph identification, neither condition improved more distant transfer of concepts. Furthermore, the simulation-based learning method resulted in unintuitive and surprising kinds of misconceptions about how sample size affects estimation of parameters while the direct instruction group used correct intuitive judgments more often. We argue that similar perceptual properties of different sampling processes in the SBL condition overrode learners' intuitions and led them to make conceptual confusions that they would not typically make. We conclude that conceptually important differences should be grounded in easily interpretable and distinguishable perceptual representations in simulation-based learning methods.

Keywords: education; statistics; learning with simulations; sampling distributions

Introduction

Making sense of statistical inference requires flexible reasoning regarding sampling distributions and the effect of sample size on the properties of a distribution (See Figure 1). However, sampling distributions have been shown to be notoriously difficult for novice learners to grasp (Garfield et al., 2008; Kahneman & Tversky, 1972; Saldanha & Thompson, 2002). Arguing that visualization of sampling distributions facilitates learning of them, some researchers have suggested for learners to be introduced to the topic with interactive computer simulations (Cobb & Moore, 1997; Chance et al., 2004). These simulations simultaneously plot two levels of processes; first, the process of randomly selecting samples from a population; second, collecting the mean of each sample a large number of times, which approximates and visualizes theoretical sampling distributions of means. In simulation-based learning activities, learners typically first generate a prediction of how sampling distributions will be affected by different parameters, and then test their prediction by manipulating the parameters of the simulations.

A **sampling distribution** is a distribution of statistics obtained by selecting all the possible samples of a specific size from a population. If the obtained statistic is the mean of each sample, the distribution is called **sampling distribution of means.** See below two sampling distributions of means with two different sample sizes obtained from the same population with $\mu = 80$, $\sigma = 20$.



Notice that, as n gets larger, the standard deviation of the sampling distribution of means (σ_M) gets smaller, with the sample means tending to approximate population mean (μ) more closely with larger sample size. However, people often believe σ_M gets larger, or stays stable with larger samples (Chance et al., 2004).

Figure 1: Sampling distributions in relation to sample size. The descriptions and figures are adapted from Gravetter et al. (2020).

Prior works support the promises of the simulation-based learning activities described above. First, dynamic visualizations have been found useful for people to see structure in scientific phenomena (Lindgren & Schwartz, 2009). Second, generative activities can help novice learners appreciate the deeper structure of the content, as, otherwise, they do not have the necessary background information to discern the affordances of the domain (Kapur, 2015). Third, interactive simulations can foster a deeper conceptual understanding than non-interactive simulations as they encourage active inquiry for meaning (Evans & Gibbons, 2007; Moreno et al., 2001).

On the other hand, there is also broad evidence to suggest such inquiry-based activities have unintended consequences.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

People often learn best when they are given full and direct instruction by explicit training on rules and their applications when dealing with new information. By contrast, inquiry-based activities can put too challenging demands on a novice learner's working memory capacity (Kirschner et al., 2006; Klahr & Nigam, 2004).

In the current work, we aimed to compare the effects of teaching people with interactive computer simulations (simulation-based learning) to direct instruction (without simulations) on people's reasoning about sampling processes. For simulation-based learning, we adapt the materials from a previous study by Chance et al. (2004) which train people on graph-based questions about sampling distributions and then test them on related graph interpretation and story problems.

Reasoning about Variability of Sampling Distributions

People (correctly) believe that large samples generally allow for more accurate estimates of the population parameters than do small samples. That is, people generally have *size-confidence intuition* (Sedlmeier, 1999). However, even though people are generally able to use size-confidence intuition for questions that ask about individual samples, this understanding does not translate when they are asked the same concept in the context of distributions of sample statistics. Consider the following question originated by Kahneman and Tversky (1972) which was used in several experiments with two different prompts:

Maternity ward problem

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.

Sampling distribution version: For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?[1]

Individual sample version: Which hospital do you think is more likely to find on one day that more than 60% percent of the babies born were boys?[2]

A meta-analysis found that the average correct solution rate for the first version is 33% while the second one is 77% (Sedlemeier & Gigerenzer, 1997) which indicates that humans spontaneously appreciate the impact of sample size on the mean of an individual sample, but not on the variance of sampling distributions. Moreover, the educational literature suggests that misconceptions regarding variability of sampling distributions are also resistant to training (Chance et al., 2004; van Dijke-Drookers, 2021). Following explicit training, learners still often believe the variability in a sample distribution of means does not change with the sample size, or that it gets larger with an increase in sample size. The confusion seems to stem from the lack of understanding that variability in the sampling distribution of means is based on the differences across the means obtained from one individual sample to another.

The Current Study

Given the increasing importance of statistical reasoning in our data-intensive society, and the aforementioned evidence of people's difficulties in reasoning about sampling distributions, it is important to identify effective pedagogical approaches that facilitates people's learning of the concept. In the current work, we train people on the effect of sample size on the mean of samples and the variability of the sampling distributions either with simulation-based learning or direct instruction method. We train the participants on a two-level process:

Level 1 *(mean of individual samples):* Record the mean of a randomly drawn sample with a certain size from a population. Repeat the process with different sample sizes and attend to the direction of the change in the recorded values of the mean.

Level 2 *(standard deviation of sampling distribution of means):* Repeat Level 1 a large number of times and accumulate a collection of sample means. Record the standard deviation of the collection. Attend to the direction of the change in the recorded values of the standard deviation with different sample sizes.

Education research frequently faces the dilemma of varying one variable at a time between conditions to isolate each variable's effect versus comparing instructional methods more holistically, by allowing them to differ along various dimensions, so as to maintain fidelity to their underlying pedagogical models (Schwartz et al., 2011). In this work, we chose the second approach because we are interested in educational interventions that are commonly used to teach students. More specifically, we are interested in investigating whether active exploration of content through interacting with a dynamic visualization tool leads to different learning outcomes than a more traditional mode of instruction. Thus, we compare two pedagogical approaches in our experiment: 'direct instruction (DI)' and 'simulation-based learning (SBL)'.

For the DI condition, we adapted materials and procedures from a popular coursebook used in introduction to statistics classes (Gravetter et al., 2020) and designed the instruction based on the principles of direct instruction. According to the DI method, information that explains the concepts and procedures is provided fully to students from the beginning (Kirschner et al., 2006). The assumption is that working memory is limited, and processing new information is constrained by the working memory capacity. Accordingly,

[1] Note that there are 365 samples for each hospital. The collection of proportion of boys for each sample (day) form an empirical sampling distribution (Sedlmeier, 1999).

[2] The correct answer is the small hospital for either version.

our DI group receives verbal rules and accompanying static visuals first, then attempts to solve related graph problems with feedback. To minimize learners' cognitive load, we first provide Level 1 instruction, then Level 2 instruction.

For SBL condition, we build upon a *predict, observe, explain* (POE) pedagogy. Using POE, students predict the outcome of an event, then they describe their observation, and then finally explain their observation. This approach assumes that when there is a conflict between their prediction and observation, students will reconcile it, which will result in effective conceptual change (White & Gunstone, 1992).

Thus, we compare two instructional conditions. The DI group first receives direct instruction via verbal and pictorial information and then attempts to solve graph problems followed by feedback. The simulation-based learning group first attempts to solve graph interpretation problems with feedback and then compares their given answer to interactive simulations, augmented by guided self-explanation prompts. Informational equivalence in both conditions was approximately equated by a) giving feedback to fill-in the blank self-explanation prompts; when explanation prompts were completed, they produced the same verbal information to the DI condition, and b) providing several static pictures from the simulation for the DI group, which would be similar those observed by SBL-trained learners.

Methods

Undergraduate students participated in a single-session online experiment. The participants were randomly assigned to one of the two training conditions automatically when they started the experiment on their computer. The experiment consisted of pretest, pretraining, training, and posttest phases. The only manipulation between the two conditions occurred at the training phase (See Figure 2). The study was pre-registered on OSF including all materials, analysis plans, and scripts.

Participants

Participants were 141 undergraduate students from the researchers' university. They received partial course credit in exchange for participation. Based on self-reports, their ages were between 18-24, and 68% were female.

Materials

Pretraining Before training manipulations, a pretraining instruction phase targeted concepts that Chance et al. (2004) identified as prerequisites for understanding sampling distributions. This phase covered the following topics: population, sample, mean, standard deviation, and the distribution of sample means. Each topic was presented through verbal information adapted from Gravetter and Wallnau (2013) and were accompanied by histogram graphs. The presentation of each topic was followed by multiple-choice questions on the topic with feedback.

Training The DI condition was introduced the two levels of information separately. For the first level, participants were given the verbal rule that the sample mean will tend to be closer to the population mean as the sample size increases. The rule was exemplified by figures on a 3x3 grid. These figures consisted of screenshots from the simulation which depicted that the sample means got closer to each other and to the population mean with increasing sample size. Below the figures, the verbal rationale was given that the small and large values will tend to average each other out with a larger sample size.

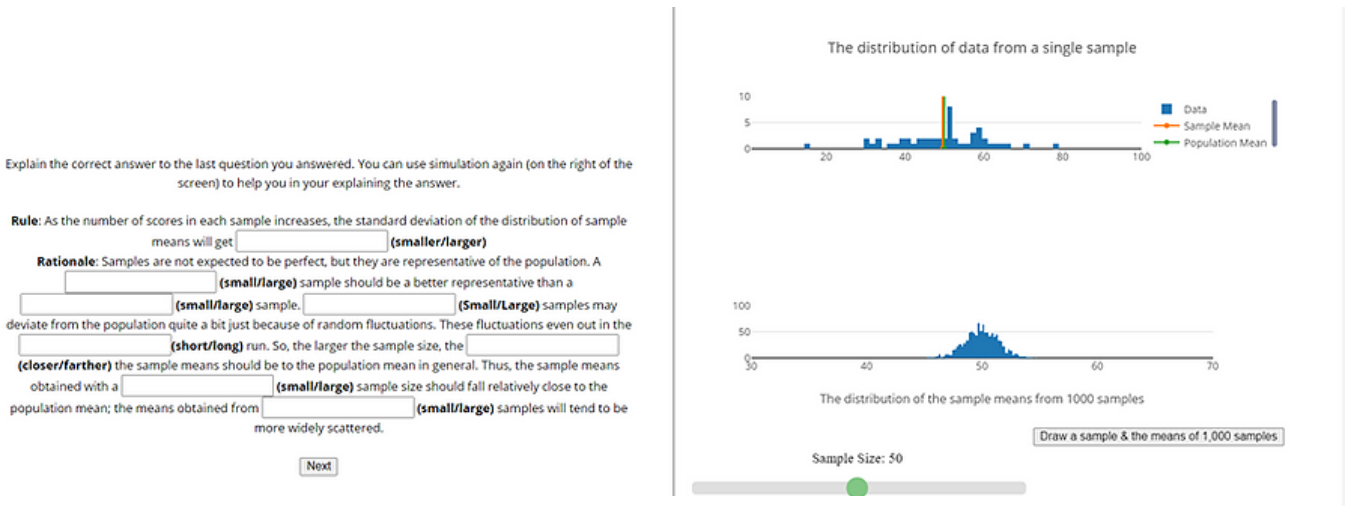


Figure 2: A scene from the training phase of the SBL condition. On the right, the interactive simulation shows the graphs of a single sample (top) and the distribution of means from many separate samples (bottom). On the left, the fill-in-the blank task prompts participants to self-explain the rule and rationale for what they are observing in the simulation. Meanwhile, participants in the DI condition received the same verbal information with accompanying static visuals (i.e., screenshots taken from the simulation).

For the second level, participants were given the rule that the standard deviation of the distribution of sample means will get smaller as the sample size increases. As in the first level, this verbal rule was exemplified by accompanying screenshots taken from the simulation on a 3X3 grid. Below the figures, the verbal rule and rationale for the first level information was repeated and connected to the second level rationale that sample means obtained with a larger sample size should fall relatively close to the population mean.

Then, two multiple-choice questions were posed, each followed by corrective feedback. Each question asked for the identification of the sampling distribution graph with the smaller or larger sample size.

The main difference in the SBL condition was a) the presentation order was reversed (first question, then information); b) the screenshots were replaced with the interactive computer simulation, and the verbal information was changed into self-explanation prompts (See Figure 2). c) Level 1 and Level 2 information was provided on the screen at the same time both through the prompts and simulations. The guided self-explanation prompts were given in fill-in-the blanks form. The participants were not able to proceed to the next stage without completing the form correctly.

Pretest and Posttest Items Pretest and posttest items included 12 identical multiple-choice questions that we classify as isomorphic, transfer, and rule questions. Additionally, post-test included two open-ended questions.

Isomorphic questions consisted of 5 graph questions that had a similar setup to the questions during the training phase. That is, each described a particular population and asked for identification of the sampling distribution graphs with either smaller or larger sample sizes.

Transfer questions had the same structure but with stories that included no graphs; stories analogical to the maternity ward problem; and a graph question related to the empirical law of large numbers.

The two rule questions asked participants to identify the correct rules that were presented in the training session.

Posttest included two additional open-ended items which asked participants to explain the reasons for the rules that were presented at the training sessions.

Scoring of Verbal Data

Prior to the actual study, a pilot study was conducted. The written responses to the two open-ended items at the posttest were coded in a bottom-up manner. A coding-scheme was created based on observed categories of responses based on which the actual study was analyzed. The unit size for coding of the verbal data consisted of each participant's full response to each singular question, which corresponded to one category. 20% percent of the data were independently coded by the first and the second author. The interrater agreement for assigning each response to categories was 85% for the first item with eight categories, 84% for the second item with ten categories. After the two coders discussed the cases of divergence and achieved a mutual agreement on final decisions, the first author completed the coding of all verbal data. The coders were blind to the conditions throughout the coding process.

Results

We compared the time spent on task, measured learning gains for each problem type across conditions and from pre- to posttest, and conducted a verbal analysis of the participants' responses to open-ended items.

Time on Task

We ran a between-subjects t-test on time-on-task. The time spent on the intervention was not significantly different between the two groups (M_{DI} = 21.69 min, M_{SBL} = 23.00 min, $t(136.55) = 0.48$, $p = 0.63$.

Multiple-Choice Items

We ran two separate analyses for each multiple-choice problem type. First, we ran an ANCOVA on the posttest scores with prior knowledge (the sum of the correct answers on pretest and pretraining) as a covariate and the condition (DI vs SBL) as an independent variable. Second, we collapsed the data across the conditions and ran a dependent t-test to measure overall learning from pre- to post-tests (See Figure 3). For each problem type, we present the results from the ANCOVA and t-test, respectively.

For isomorphic problems, 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 ($M = 2.46$, $SD = 1.43$), $t(140) = 7.14$, $p < 0.01$.

For transfer problems, there was not a significant difference between the DI and SBI 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 ($M = 2.45$, $SD = 1.27$), $t(140) = 1.59$, $p = 0.11$.

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

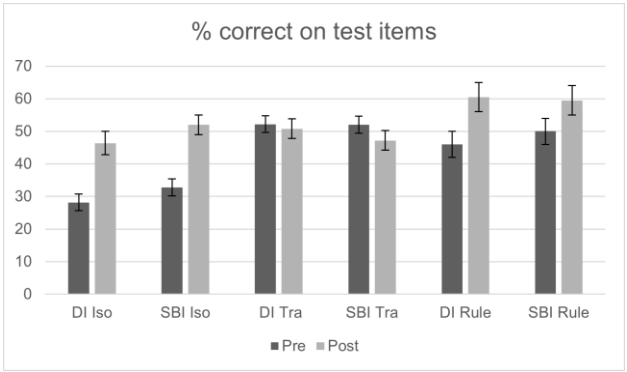


Figure 3: Average percentage of correct answers in pre and post tests for each group. Error bars represent ±1 *SE*.

Open-ended Items

We ran a Pearson's chi-square test on the verbal data from the responses to open-ended questions at posttest. We present the results from the first item and the second item respectively.

The first item prompted participants with the sentence "Sample mean tends to get closer to the population mean as sample size increases. Explain why this is correct." There was a significant association between the condition and the response categories, $\chi^2(7) = 16.08$, $p = 0.02$. We represent the most commonly appearing categories across the conditions (See Table 1).

Table 1: % responses to the first item: Explain why sample mean tends to get closer to the population mean as sample size increases

Response category	DI	SBL
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
Insufficient explanation	15	17

The second item prompted participants with the sentence "As the sample size increases, the distribution of sample means will have a smaller and smaller standard deviation. Explain why this is correct." There was not a significant association between the condition and the response categories, $\chi^2(5) = 6.53$, $p = 0.25$. We represent the most commonly appearing categories across the conditions (See Table 2).

Table 2: % 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	DI	SBL
Insufficient 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

Discussion

In the current study, we compared a simulation-based learning (SBL) method to a direct instruction (DI) method for a sampling distribution task. In the DI group, participants first received direct instruction via verbal and pictorial information and then attempted solving graph problems with feedback, whereas the SBL group first attempted solving graph problems with feedback and then explored their answer through interactive simulations and guided self-explanation prompts. We measured learning through graph problems, story problems, rule statement items, and open-ended explanation items.

Both groups increased performance similarly from pre- to post-test for graph problems while neither group improved on story problems at all, even though both types of questions had a very similar setup (notice that graph questions also contained a story element). Thus, learners were only able to answer story problems when a graph accompanied it which suggests that they mostly gained a superficial understanding of the concepts (i.e., the sampling distribution looks narrower with larger sample size). The current results suggest that it is challenging for learners to transfer their learning when left to their own device even for very similar questions. However, prior work suggests promising techniques to elicit transfer with guidance. Namely, hinting participants learning and transfer tasks are related, or prompting self-explanation of the abstract principle underlying the transfer task, have been shown to increase transfer across analogous examples by helping people generate schema (Loewenstein, 2010). The promise of these techniques should be tested in future sampling distribution studies.

There was some improvement from pre- to post-test for rule questions at similar levels for both groups, however, not an underlying model-based account of the rules. For the first open-ended explanation item, participants in the DI group mostly responded in a way that would be expected without exposure to any training (See Table 1). In other words, they used *size-confidence intuition* (that is, a larger sample is a better representation of the population). Further, a more detailed analysis of the verbal data revealed that participants mostly thought that it was the proportion of the sample to the population size, not the absolute size of the sample that made larger samples more stable, a common misconception previously observed in classroom studies (Garfield et al., 2008). These results suggest that the rationale provided during training (if one has a large sample size, then values smaller and large than the population mean will average each other out, so it will be more likely that the mean will approximate the population mean) was mostly not helpful.

An interesting result is that participants in the SBL group responded with *size-confidence intuition* less often than did the DI group for the first open-ended explanation item (See Table 1). Rather, their responses focused on the variability of the sample. Some believed that outliers, hence, the standard deviation would decrease in a sample with larger sample size (note that students generally mean "observations under the tail" with the word "outliers" (Garfield et al., 2008)). Others

believed that standard deviation of the sample increases with larger samples (note that this answer never appeared in DI condition). Thus, participants in the SBL group responded *as if* they were asked about the variability of sampling distributions instead of the mean of an individual sample at the first item. Note that, if that were the case, the latter explanation (that is, "as sample size increases, the standard deviation increases") would be a typical misconception related to *sampling distribution of means* (Chance et al., 2004). In the current case, however, it is a surprising kind of confusion about *the means of individual samples.* Thus, simulation-based learning method seems to have overridden the *size-confidence intuition* and elicited an unusual and unintuitive kind of misconception. We suspect this resulted from simultaneous engagement with Level 1 (individual sample distribution) and Level 2 (distribution of sample means) graphs which bear highly similar perceptual properties (See Figure 2).

One of the best ways to teach people difficult concepts is to ground them in spatially explicit representations that people's well-honed perceptual routines can process effectively (Goldstone et al., 2010). However, the current work suggests that this strategy comes with a risk − that people will assume that similar perceptual processes entail similar concepts. In the case of our SBL simulations, similar perceptual properties in our Level 1 and Level 2 graphs led participants to make conceptual confusions between these levels that they would not typically make without their salient perceptual similarity.

However, our recommendation is not to avoid perceptual scaffolds for difficult concepts. Rather, we recommend devising perceptual representations so that conceptually important differences have easily and intuitively decipherable perceptual differences as well. As a design implication of this recommendation, we suggest that the collection of *data* at the Level 1 and *sample statistics* at Level 2 graph represented through identical bars and bins be replaced by iconic representations, perceptually differentiated across the two levels. Future work should test learning of sampling processes with perceptual representations that better align with core concepts to further investigate the promises of simulation-based learning.

Open Practices Statement

Pre-registration can be accessed at https://osf.io/rjad4.

References

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

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801-823.

Evans, C., & Gibbons, N. J. (2007). The interactivity effect in multimedia learning. *Computers & Education*, *49*(4), 1147–1160.

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, *2*(2), 265–284.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice*, *1*(1), 51-65.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Lindgren, R., & Schwartz, D. L. (2009). Spatial Learning and Computer Simulations in Science. *International Journal of Science Education*, *31*(3), 419–438.

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. In B. H. Ross (Ed.), *Psychology of learning and motivation*. Academic Press.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177-213.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257-270.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Psychology Press.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*(1), 33-51.

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*, 1-24.

White, R., & Gunstone, R. (1992). *Probing understanding*. Routledge.