UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
In-The-Moment Detrimental Effects of Grounding in a Statistics Simulation

Permalink
https://escholarship.org/uc/item/2w35c09w

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Gok, Sebahat
Goldstone, Robert

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

In-The-Moment Detrimental Effects of Grounding in a Statistics Simulation

Sebahat Gok (sebahgok@gmail.com)[1,2] & Robert L. Goldstone[2]
[1] Center for Connected Learning and Computer-Based Modeling, Northwestern University
[2] Psychological and Brain Sciences & Cognitive Science, Indiana University

Abstract

We developed and experimentally tested three versions of an interactive statistics simulation designed to foster a qualitative understanding of how sample size influences statistical estimation and inference. Guided by a grounded cognition framework, two versions used perceptually rich, manipulable tokens (Token and Token-to-Rectangle) and were compared to a conventional Histogram version. In a laboratory study with undergraduates (n = 252), all groups improved similarly from pretest to posttest. Contrary to our hypotheses, however, the token-based (grounded) conditions showed less productive in-the-moment engagement during simulation use; that is, interactions were more often directed toward perceptually salient token operations rather than sampling-relevant actions. Grounded conditions also showed weaker performance on within-activity conceptual items. These findings suggest that grounding can be seductive rather than supportive in the moment, even when dynamic and perceptually rich representations are aligned with key concepts and avoid noninformative decoration.

Keywords: grounded cognition; statistics education; interactive simulations; sampling distribution

Introduction

Computer simulations are becoming increasingly common in introductory statistics courses (Lock et al., 2025), paralleling earlier computational shifts in science education (Wilensky & Reisman, 2006). In addition to aligning learning activities with the practices of professional scientists and fostering mechanistic thinking, simulations offer an additional pedagogical affordance: Interactive diagrams embedded in computational models recruit learners' perceptual systems to support understanding of formal mathematical concepts (Gok & Goldstone, 2024a).

However, interpreting diagrammatic representations can be difficult for novice learners; therefore, care must be taken when designing them. To illustrate, students sometimes interpret histogram bars as properties of individual data points rather than frequencies or densities (Kaplan et al., 2014). Therefore, several pedagogical simulations are designed with novices' intuitive interpretations of external representations in mind. For example, a popular one, Tinkerplots (Konold & Miller, 2005), uses dot plots as the default plot and allows learners to gradually transition to histograms. Because each dot corresponds to a single data point, this approach accommodates novices' intuitive interpretations.

Yet systematic experimental evidence remains limited on which representational design choices are most effective for novices, and to what extent reducing the demands for representational competence functions as a gateway for learners to master abstract concepts in statistics and, more broadly, in science. The theoretical framework of grounded cognition, which we introduce next, guides our hypotheses about the influence of intuitive external representations on the learning of abstract concepts.

Grounded Cognition Approach to Pedagogical Simulation Design

Grounded cognition rejects the notion that perception and abstract reasoning operate as separate and independent processes (Barsalou, 2008). Instead, it proposes a tight coupling and bidirectional tuning between the two (Goldstone & Barsalou, 1998). Later work uses grounded cognition as a framework for designing external representations that account for the influence of learners' perceptual processes and bodily actions in the acquisition of abstract concepts (Gok & Goldstone, 2024a; Nathan, 2021). One of the design hypotheses, which we particularly focus on in this paper, is that familiar, intuitive, dynamic representations can better support learners' initial sensemaking in simulations, whereas more idealized representations may better support generalization beyond the specific context (Fyfe et al., 2014; Goldstone & Son, 2005).

Building on this prior work, here, we develop and experimentally test a statistics simulation designed to connect learners' intuition and perception with formal statistical ideas. We combine grounded cognition (Gok & Goldstone, 2024a) with a design-based research approach (Barab & Squire, 2004), iterating on multiple simulation versions over several years through think-aloud interviews and controlled experiments (Gok & Goldstone, 2022; 2024b). In this paper, we report our latest design with its experimental evaluation.

We first introduce the target concept, then describe key design elements and the alternative design conditions, and finally present learning outcomes alongside action log analyses as a window into the learning processes associated with each condition.

The Target Concept: Sample Size's Influence on Statistical Distributions

Our intervention focused on fostering a qualitative understanding of the influence of sample size on statistical distributions across three levels. The core ideas are as follows:

At the **single-sample level (level 1)**, small samples tend to be noisier. As the sample size increases, the sample typically

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

becomes more representative of the population, and summary statistics computed from it become more stable.

Then, one can consider the **distribution of sample means (level 2)** generated by repeatedly sampling from the population distribution and tallying each sample's mean**.** The key points here are that increasing the sample size **(a)** makes the distribution narrower – sample means cluster more tightly around the population mean, and **(b)** reduces the proportion of extreme values in the tails.

These ideas connect directly to **statistical power and inference (level 3).** When the sampling distribution is wide (as with small samples), true effects are harder to detect because sampling variability can mask the signal. As the sample size increases and the sampling distribution tightens, the same effect size becomes easier to distinguish from noise. Consequently, larger samples increase statistical power and improve the likelihood of detecting a real effect

The Methods

Participants

A total of 291 undergraduate students from the Psychology course participant pool completed the study in exchange for course credits (Token-to-Rectangles: n = 86; Token: n = 108; Histogram: n = 97). After applying a pre-specified attention check, the final sample included 252 participants (Token-to-Rectangle: n = 74; Token: n = 92; Histogram: n = 86).

Participants were, on average, 19 years old (SD = 2). Gender was 75% female, 24% male, with one person genderqueer or questioning and one preferring not to report. Self-reported ethnicity was predominantly White (73%), with Asian (12%), Hispanic (7%), Black (5%), and the remainder were multiracial/other categories.

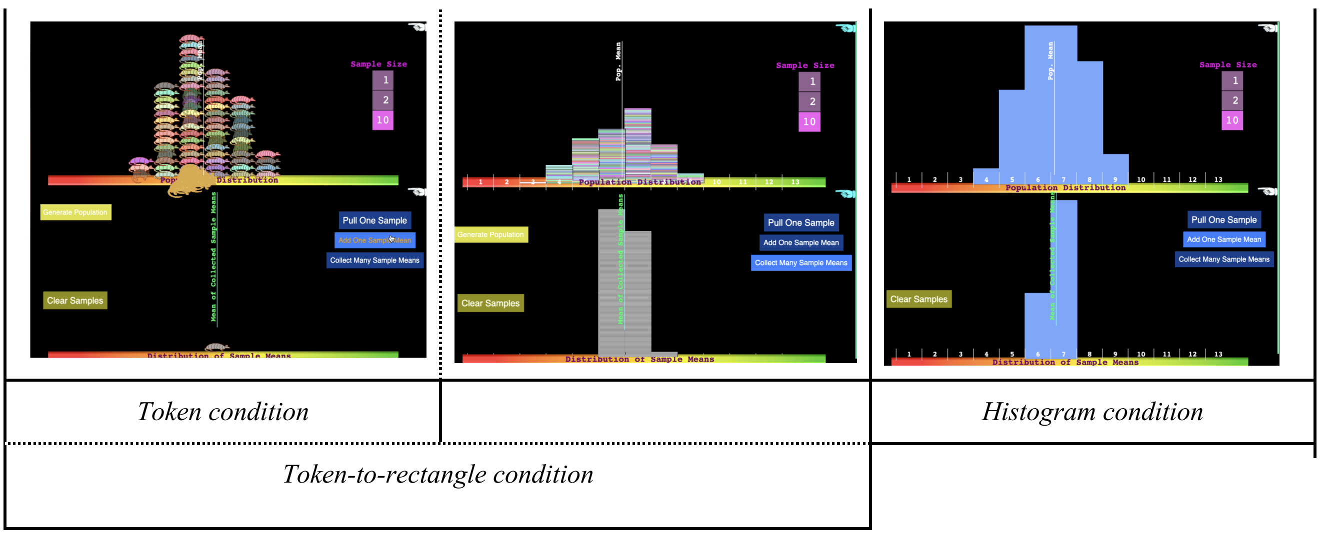

Figure 1. Screenshots from the conditions

Simulation Design and Conditions

We developed one interactive statistics simulation per condition. Across all three simulations, learners interacted with the same underlying population and completed identical sampling tasks: generating a population, drawing random samples of different sizes (Level 1), collecting sample means to build a sampling distribution (Level 2a), using an interactive threshold marker to read tail percentages from the sampling distribution (Level 2b), and then responding to a statistical inference question (Level 3). Conditions differed only in the representational form used to display populations, samples, and sampling distributions (See Figure 1). (web link to the experimental script will appear here at the unmasked version after peer review).

Token condition Graphs were shown using token-based stacks (armadillo icons). Each token represented a data point, with numerical values indicated by stripes on armadillo bodies that increased systematically from left to right. Tokens were assigned random colors, but sample means were visually distinguished using gray to emphasize the distinction between the two.

In addition to these visual features, the condition included three interactive animations: (a) when learners collected a sample mean, sampled tokens moved toward the mean location and combined into a mean token; (b) hovering over a token enlarged it so that its stripes could be counted more easily; and (c) learners could drag individual tokens to manipulate the population distribution directly. We refer to the collection of all these design choices in this condition as the *grounding* approach.

Token-to-Rectangle condition This condition was identical to the Token condition, except that midway through the activities, the armadillo tokens transitioned into rectangle tokens, thereby more closely resembling a histogram representation, consistent with the concreteness fading principle (Fyfe et al., 2014) (Table 3 shows the stage at which the transition happened).

Histogram condition Graphs were presented as single-color conventional histograms. It did not include any animations.

Procedures

The study was conducted as a self-directed computer-based session in a research laboratory.

A. Cover story The experiment began with a cover story about armadillos in Paraguayan forests, including a known average stripe count per armadillo. Participants were then instructed to determine whether the armadillos observed on campus plausibly came from the same population, based on their average stripe count **(Level 3)**.

B. Pretest Next, participants completed a 10-item multiple-choice pretest assessing the target concepts. Items included questions on *interpreting graphs, verbal conceptual understanding,* and *story-based reasoning.* Each question targeted either **Level 1** or **Level 2.**

C. Simulation activities Participants were randomly assigned to one of the three conditions and completed three types of computer-based simulation activities:

C1. Simulation functions (guided walkthrough). Activities began with a guided walkthrough to familiarize participants with the simulation's core functions. For this purpose, prompts were provided to guide participants to specific actions. e.g., "Choose a sample size (1, 2, 10). Click Pull One Sample to extract a random sample from the population."

The walkthrough covered all core functions (i.e., generating a population, drawing random samples of different sizes, collecting sample means to construct a sampling distribution, and using the interactive bar to read tail percentages). Immediately before each prompt, relevant key concepts (e.g., sample, mean) and their roles in statistical investigation were briefly introduced. Log data from this phase was also used as an attention check; that is, participants who did not complete the required actions were excluded from data analyses.

Then, the participants were asked 5 questions that gauged their understanding of simulation functionalities, such as "What does the 'Pop. Mean' line show?" with options: "The average height of the armadillos", "The geographical boundary of the armadillos' habitat", "The average number of stripes on the population of armadillos."

C2. Free Exploration. After the guided walkthrough, participants completed a brief free-exploration phase (2–5 min) during which they could interact with the simulation as they chose. They were instructed to focus on understanding how sample size affects sample means and to record their observations in an open-response text box.

C3. Predict-Test-Explain (PTE). Participants then completed a set of multiple-choice Predict–Test–Explain (PTE) prompts in which they (P) generated predictions, (T) tested those predictions by interacting with the simulation, and (E) selected a mechanistic explanation for the observed outcomes. The PTE sequence spanned all three conceptual levels described above.

Immediately prior to the Level 3 prompts, the Token-to-Rectangles condition transitioned from armadillo tokens to rectangle representations (see Figure 1, middle panel).

D. Direct Instruction Next, participants received a brief, computer-based explanatory text with static diagrams that explicitly explained the target concepts. This section summarized the concepts outlined in the simulation activities and then asked participants 5 explicit rule questions (e.g., "What happens to the distribution of sample means' variability as the sample size increases?").

E. Posttest The experiment ended with a posttest that was isomorphic to the pretest and counterbalanced across participants.

Logged Data

All participant interactions with the simulation were logged as follows:

PopulationAnimations Hover/drag interactions with individual tokens (Token and Token-to-Rectangle conditions only)

BarMoves Movements of the threshold marker ("bar")

SamplingActions Includes picking sample size, drawing individual samples, adding individual sample means, and collecting many sample means. These actions were originally logged separately but are summed for the purposes of this paper.

TotalActions Total count of logged actions (sum of the above). The simulation also included a "clear" action that resets the screen. 'Clear' action was logged and included in the TotalActions, but we exclude it from separate analyses because it is not theoretically informative.

Results

We first report improvements from pre- to posttest with their log-data correlates. We then report participants' self-ratings. Finally, we present the remaining results in the order they occurred during the study: performance on the simulation-functions phase, interaction patterns during free exploration, Predict–Test–Explain (PTE) responses, and explicit rule questions.

Posttest

A mixed ANOVA with Time (pre vs. post) as a within-subjects factor and Condition (Token, Token-to-Rectangles, Histogram) as a between-subjects factor showed a main effect of Time, $F(1, 249) = 30.57$, $p < .001$. There was no main effect of Condition, $F(2, 249) = 1.01$, $p = .36$, and no Time × Condition interaction, $F(2, 249) = 0.22$, $p = .81$.

These results indicate that the intervention improved overall performance, but there was no significant effect of condition (See Table 1 for descriptives).

Posttest and interaction logs correlations

We divided logged actions (those during the guided walkthrough phase) into task-relevant actions (those explicitly instructed) and exploratory actions (task-irrelevant at that moment). Posttest score was not reliably related to exploratory actions, $r = -.05$, $t(250) = -0.72$, $p = .47$, but was positively associated with task-relevant actions, $r = .21$, $t(250) = 3.49$, $p < .001$, and with total logged actions throughout the experiment, $r = .13$, $t(249) = 2.06$, $p = .04$. [1]Descriptively, these results suggest that the association was strongest for task-relevant actions.

Self-ratings

Self-ratings indicated moderate perceived difficulty and moderate helpfulness for both the simulation and the direct instruction, with no significant differences across conditions (all ps > .05; see Table 1 for descriptive statistics).

Table 1: Main descriptives by condition (M, SD)

Measures	Token	Token-Rect	Histog
Pre Score	5.99(1.66)	5.78(1.74)	6.16(1.53)
Post Score	6.53(1.86)	6.51(1.75)	6.79(1.91)
Pretest difficulty	3.38(0.80)	3.53(0.83)	3.40(0.79)
Simulation difficulty	3.12(0.90)	3.03(0.95)	2.93(0.94)
Simulation helpfulness	3.54(1.00)	3.65(0.88)	3.65(0.97)
Direct instruction difficulty	2.89(0.83)	2.84(0.83)	2.79(0.70)
Direct instruction helpfulness	3.80(0.79)	3.84(0.86)	3.59(1.09)
Posttest difficulty	2.76(0.82)	2.80(0.72)	2.67(0.85)
Overall helpfulness	3.71(0.76)	3.76(0.81)	3.69(0.86)
Simulation functions	3.80(0.99)	3.58(1.12)	3.82(1.02)

Simulation functions

Performance on the simulation-functions questions did not differ by condition, $F(2, 249) = 1.32$, $p = .27$ (see Table 1 for descriptives).

Free exploration

We analyzed participants' logged behaviors during the free exploration phase. Although the Histogram condition included fewer available action types (i.e., no *populationAnimations*), a Kruskal–Wallis rank-sum test of group medians revealed no significant differences in the total number of actions across conditions, $\chi^2(2) = 0.51$, $p = .77$.

However, conditions differed in the types of actions participants performed. As expected by design, significantly more *populationAnimations* were logged in the token-based conditions ($\chi^2(2) = 130.71$, $p < .0001$). Interestingly, this additional action affordance provided to token-based conditions appears to have come with trade-offs: a Kruskal–Wallis test revealed a significant effect of condition on participants' *samplingActions* ($\chi^2(2) = 18.38$, $p < .0001$). Follow-up pairwise Wilcoxon tests showed that participants in the Histogram condition engaged in significantly more *SamplingActions* than those in either the Token ($p < .001$) or Token-to-Rectangle ($p < .001$) conditions. No difference was observed between the two token-based conditions ($p = .90$).

Overall, these results suggest that the token-based conditions devoted substantial attention to population animations, which led them to engage relatively less in sampling relevant actions (i.e., drawing samples, drawing means, and collecting means) compared to Histogram participants (See Table 2).

Table 2: Median number of logs during free exploration

Log Data	Token	Token-Rect	Histog
PopulationAnimations	**15**	**12**	**0**
BarMoves	14	12	16
SamplingActions	**7**	**8.5**	**21**
TotalActions	49.5	47	61

PTE (Predict-Test-Explain)

To analyze the Predict–Test–Explain (PTE) sequence, we used two models. For the Predict vs. Test phases, we analyzed item scores using a mixed ANOVA with phase (Predict vs. Test) as a within-subject factor and condition as a between-subjects factor, focusing on the condition × phase interaction to test whether the Predict–Test difference varied across conditions. For the Explanation phase, we fit logistic regression models predicting whether participants selected the correct explanation (0/1) for each condition, with Token-to-Rectangles as the reference group (See Table for descriptive statistics).

[1] Spearman and Kendall correlations showed the same pattern.

Table 3: PTE (Predict-Test-Explain) (M, SD)

Level	Stage	Token	Token-Rect	Histog
1	Pred	0.86(0.35)	0.76(0.43)	0.81(0.39)
	Test	0.80(0.40)	0.73(0.45)	0.81(0.39)
	Exp	0.28(0.45)	0.27(0.45)	0.29(0.46)
2a	Pred	0.72(0.44)	0.78(0.41)	0.86(0.34)
	Test	0.79(0.40)	0.70(0.46)	0.74(0.43)
	Exp	0.69(0.46)	0.59(0.49)	0.78(0.42)
	Transition for Token-to-Rectangle condition			
2b	Pred	0.57(0.49)	0.39(0.49)	0.58(0.49)
	Test	0.45(0.50)	0.61(0.49)	0.58(0.50)
	Exp	0.72(0.45)	0.59(0.49)	0.77(0.43)
3		0.33(0.47)	0.43(0.50)	0.48(0.50)

Level 1 (single-sample)

Prediction vs Test There was no reliable effect of condition, $F(2, 249) = 1.42$, $p = .24$, no main effect of phase, $F(1, 249) = 1.18$, $p = .28$, and no condition × phase interaction, $F(2, 249) = 0.42$, $p = .66$ ($\eta g^2 s \leq .009$)

Explanation Neither the Token condition ($b = 0.06$, $SE = 0.35$, $p = .86$) nor the Histogram condition ($b = 0.10$, $SE = 0.35$, $p = .77$) differed from the reference condition.

Level 2a (sample-means)

Prediction vs Test There was a condition × phase interaction, $F(2, 249) = 4.10$, $p = .018$, $\eta g^2 = .009$. Follow-up tests of the Predict–Test difference within each condition indicated a significant difference in Histogram $t(249) = 2.41$, $p = .01$, but no significant difference in Token-to-Rectangles, $t(249) = 1.56$, $p = .12$, or Token.

Explanation The histogram condition was more likely to select the correct explanation than Token-to-Rectangles ($b = 0.88$, $SE = 0.35$, $p = .01$). The Token condition did not differ from Token-to-Rectangles ($b = 0.44$, $SE = 0.33$, $p = .18$).

Level 2b (sample-means)

Prediction vs Test There was a significant condition × phase interaction, $F(2, 249) = 5.81$, $p = .003$, $\eta g^2 = .019$, indicating that the Predict–Test difference varied by condition. Follow-up tests showed a significant difference in Token-to-Rectangles, $t(249) = -2.93$, $p = .004$, but not in Token, $t(249) = 1.81$, $p = .072$, or Histogram, $t(249) = 0.00$, $p = 1.0$

Explanation The histogram condition was more likely to select the correct explanation than Token-to-Rectangles ($b = 0.81$, $SE = 0.35$, $p = .02$). The Token condition did not differ reliably from Token-to-Rectangles ($b = 0.55$, $SE = 0.33$, $p = .09$).

Level 3 (inference)

Accuracy on the inference question did not differ significantly by condition, $F(2, 249) = 2.22$, $p = .11$.

Discussion

In this study, we developed and experimentally tested three representational versions of an interactive statistics simulation designed to support a qualitative understanding of the influence of sample size on statistical distributions and inference. The grounded cognition framework (Gok & Goldstone, 2024a) informed token-based designs that employed perceptually rich, manipulable representations to support initial sensemaking. We compared a concrete Token condition, a conventional Histogram condition, and a concreteness-fading variant (Fyfe & Nathan, 2019) that transitioned from tokens to an idealized rectangle representation (Token-to-Rectangles).

Across conditions, participants improved from pretest to posttest, indicating that the overall intervention supported learning. However, learning gains did not differ reliably by representational format. Process measures, in contrast, indicated several significant differences in how the representational format affected participants' engagement with the simulation and their in-the-moment understanding of it. Critically, these patterns were often in the opposite direction of what we predicted, as detailed below.

First, we provided Token-based groups (Token and Token-to-Rectangle) with animations that would increase interactivity and the perceptual and dynamic salience of population properties and mean collection. However, rather than productively grounding learners' understanding of statistical processes, these design choices appear to have diverted participants' attention away from the core sampling tasks. The token-based conditions yielded fewer sampling actions than the histogram condition. This suggests that learners may be willing to invest only a limited amount of time in exploring simulations, regardless of the number of available features, which necessarily entails trade-offs for the *grounding* approach. In this case, the added interactivity may have redirected engagement toward less instructionally central aspects of the simulation, rather than increasing overall time-on-task or deepening engagement with the critical statistical processes.

Second, we expected the Token-based graphs to show an advantage in the Predict-Test-Explain (PTE) sequences, based on the hypothesized benefits of perceptual grounding. These conditions not only featured more concrete representations but also aligned more closely with the wording of the PTE questions. For example, several questions referred to features such as "the stripes of armadillos," which were visually depicted in the Token-based graphs but not in the Histogram condition. Despite this alignment disadvantage, the Histogram condition consistently outperformed the Token-based groups on both the Predict and Explain components of the PTE sequences, particularly at Level 2, where we had expected grounding to be most beneficial. This level required coordination of mean and variability information, which the grounded representations were designed to support.

Consistent with our results, we did not anticipate condition differences at Level 1 (sample size's influence on the single sample mean) because the focus was on intuitively accessible judgments. We also did not expect grounding to yield an advantage at Level 3 (sample size's influence on statistical inference), given its highly abstract nature and the number of mental inferences it requires. The observed advantage of the Histogram condition at Level 2 (sample size's influence on the distribution of sample means), however, suggests that the grounded token designs were less effective even for the mid-level conceptual coordination they were intended to support. Specifically for the Token-to-Rectangle condition, the transition moment was particularly disruptive, as indicated by the drop in performance immediately following the transition (See Table 3).

Importantly, we designed the token-based approaches with the known pitfalls of concreteness in mind (e.g., Son & Goldstone, 2009). The grounding approach here aimed to avoid distracting features and to constrain animations to concept-relevant operations (e.g., visualizing averaging), and it was iterated across multiple design cycles, with interviews and experiments reported elsewhere (Gok & Goldstone, 2022; Gok & Goldstone, 2024b). Despite these efforts, concreteness still did not yield immediate benefits and, in some cases, appeared to redirect attention toward visually engaging interactions rather than sampling-relevant actions, which undermined initial learning. Together, these results echo prior findings that concreteness can be seductive rather than supportive (Sundarajan & Adesope, 2020), and it can be difficult to design grounded representations that are reliably helpful.

These results contrast with our earlier small-sample think-aloud study (Gok & Goldstone, 2022), in which students appeared to grasp the target concepts more readily through the grounded simulation. One possibility to explain the discrepancy is that human facilitation may matter to direct attention to concept-relevant features in perceptually and dynamically rich representations. Future work should test this possibility.

Although token-based designs appeared less productive in-the-moment, there was no evidence that they were worse at posttest. None of the representational formats appeared to provide a reliable gateway to the more complex inferential reasoning required at Level 3, nor did they yield any advantage at posttest. This suggests that representational format may matter less than other supports (e.g., explicit/direct instruction; repeated practice) once learners reach the most abstract layer of the concept.

At the same time, the dissociation between within-activity performance and posttest outcomes may reflect an offsetting downstream benefit of grounded representations. Token-based conditions may have served as referential anchors that made the subsequent instructional text easier to learn from, consistent with a preparation-for-future-learning account (Schwartz & Bransford, 1998), compensating for reduced legibility during the activity itself. This interpretation can also explain another dissociation in the process data: task-relevant actions were most strongly associated with posttest scores, and the Histogram condition produced the most task-relevant actions, yet this did not translate into a posttest advantage. Future work should directly test these competing explanations, for example, by experimentally manipulating the presence and timing of the instructional text.

References

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 Psychology.* *59*(1), 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

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,* 9-25. https://doi.org/10.1007/s10648-014-9249-3

Fyfe, E. R., & Nathan, M. J. (2019). Making "concreteness fading" more concrete as a theory of instruction for promoting transfer. *Educational Review, 71*(4), 403–422. https://doi.org/10.1080/00131911.2018.1424116

Gok, S., & Goldstone, R. L. (2022). Going beyond formalisms: A grounded and embodied learning approach to the design of pedagogical statistics simulations. *The Journal of Applied Instructional Design, 11*(4). https://doi.org/10.59668/423.8196

Gok, S., & Goldstone, R. L. (2024a). How do students reason about statistical sampling with computer simulations? An integrative review from a grounded cognition perspective. *Cognitive Research: Principles & Implications. 9*(33). 1-29. https://doi.org/10.1186/s41235-024-00561-x

Gok, S., & Goldstone, R. L. (2024b, April). Using simulations to ground statistical inference in perception and action. Poster presented at the 2024 American Educational Research Association Meeting. Philadelphia, PA.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65*(2-3), 231-262. https://doi.org/10.1016/S0010-0277(97)00047-4

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

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
Konold, C., & Miller, C.D. (2005). TinkerPlots: *Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.
Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2025). *Statistics: Unlocking the power of data.* John Wiley & Sons.
Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475-5223. https://doi.org/10.1207/s1532690xci1604_4
Son, J. Y., & Goldstone, R. L. (2009). Contextualization in perspective. *Cognition and Instruction*, *27*(1), 51-89. https://doi.org/10.1080/07370000802584539
Sundararajan, N., Adesope, O. Keep it coherent: A meta-analysis of the seductive details effect. *Educ Psychol Rev* 32, 707–734 (2020). https://doi.org/10.1007/s10648-020-09522-4
Wilensky, U., & Reisman, K. (2006). Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—An embodied modeling approach. *Cognition and Instruction*, *24*(2), 171–209. https://doi.org/10.1207/s1532690xci2402_1