UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Experientially Grounded Learning About the Roles of Variability, Sample Size, and Difference Between Means in Statistical Reasoning

Permalink

https://escholarship.org/uc/item/2gq7d6xm

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors

Yu, Jingqi
Goldstone, Robert L
Landy, David

Publication Date

2018

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

Experientially Grounded Learning About the Roles of Variability, Sample Size, and Difference Between Means in Statistical Reasoning

Jingqi Yu **Robert L. Goldstone** **David Landy**
jy45@iu.edu rgoldsto@indiana.edu dhlandy@gmail.com
Psychological and Brain Sciences, Indiana university, Bloomington, IN 47405

Abstract

Despite its omnipresence in this information-laden society, statistics is hard. The present study explored the applicability of a grounded cognition approach to learning basic statistical concepts. Participants in 2 experiments interacted with perceptually rich computer simulations designed to foster understanding of the relations between fundamental statistical concepts and to promote the ability to reason with statistics. During training, participants were asked to estimate the probability of two samples coming from the same population, with sample size, variability, and difference between means independently manipulated. The amount of learning during training was measured by the difference between participants' confidence judgments and those of an Ideal Observer. The amount of transfer was assessed by the increase in accuracy from a pretest to a posttest. Learning and transfer were observed when tailored guidance was given along with the perceptually salient properties. Implications of our quantitative measures of human sensitivity to statistical concepts were discussed.

Keywords: grounded cognition; statistical inferences; statistics education; variability; sample size; mean

Introduction

In this information-laden society, the ability to reason with statistical ideas and make sense of statistical information, has become increasingly crucial and desirable. Thanks to the Internet, statistical information is now everywhere and easily accessible. In many work-related and everyday contexts, statistical literacy is no longer optional since it facilitates basic communication. For example, picking a product on Amazon usually involves a comparison between multiple alternatives varying on different statistical dimensions, such as average ratings, total number of reviews, and underlying review distribution. Thus, statistics is no longer a language that only statisticians and scientists speak and understand, nor does it exist merely in some academic domains that care about statistical significance. The omnipresence of data makes statistical literacy a necessity that helps individuals not only confidently navigate in a sea of numbers, but understand social and natural phenomena more accurately.

Despite its generally acknowledge importance and the great effort made to promote statistics education, statistics is hard to learn. Compounding the difficulty is the prevalence of statistics anxiety (e.g., Zeidner, 1991). On the bright side, technological enhancements enable new ways of presenting materials otherwise not feasible. One of the most common implementations is to create perceptually rich stimuli instantiating various types of interactions that students can draw upon when learning new concepts. This focus exemplifies a grounded approach to learning (Black, 2010).

Taken together, we are interested in whether and how perceptually grounded interaction can foster statistical reasoning. This question, to our best knowledge, has not been explicitly documented. In the following paper, we discuss common difficulties and misconceptions regarding statistical reasoning. We then present existing efforts on applying a grounded approach to learning. Finally, we introduce the computer simulation we developed to explore how grounding can be applied to bolster statistical reasoning.

Difficulties in Statistical Reasoning

The definition of statistical reasoning takes many forms, but generally, it refers to the way people reason with statistical ideas and make sense of statistical information (Garfield & Gal, 1999). One of the most robust phenomena in statistics education is that while students can successfully implement procedures for computing statistics, they have trouble applying essentially the same statistics in applications assessing their conceptual understanding (Gardner & Hudson, 1999). This gap between conceptual and procedural understanding is interpreted by some researchers as due to an overemphasis on calculating aggregated values and plotting graphs for sample data (e.g., Sorto, 2006). In other words, statistics to many students are still only about describing the properties of a given data set, but not generalizing beyond the specific data set to infer what future data sets would be likely or unlikely. Moreover, successful statistical reasoning requires an integrated understanding of fundamental statistical concepts, which unfortunately many learners lack. An inaccurate or incomplete understanding of basic statistical concepts interferes with proper reasoning, such as sampling (e.g., Watson, 2004), variation (e.g., Cobb, McClain, & Gravemeijer, 2003). Lastly, statistical reasoning always involves reasoning with uncertainty whose difficulty has been widely cataloged (Tversky & Kahneman, 1974). Hence, it is unsurprising to see lasting difficulties in making sense of diverse statistical phenomena.

A Grounded Cognition Perspective on Education

Dewey states (1986), "There is an intimate and necessary relation between the process of actual experience and education." This assertion echoes the gist of a grounded cognition perspective, an idea that environment and bodily experiences are of great importance to the development of cognitive processes (e.g., Barsalou, 1999). Therefore, many

attempted have been made to develop perceptually-rich manipulatives as an aid to scaffold students' conceptual learning. Research in this line has suggested three steps involved in a grounded cognition approach to learning: a) have a perceptually grounded experience, b) learn to imagine the perceptually grounded experience, and c) imagine the experience when learning from symbolic materials (Black, 2010). Successful attempts to apply a grounded approach to education have been documented in a wide variety of fields, such as mathematics (Suh, Moyer, & Heo, 2005) and physics (Zacharia, 2007). The common goal of these applications is to help learners develop a "feel" for what they are learning (Black, 2010)

Present Study

In the current work, we advocate a token representation in which each individual datum's measurement is shown by an intuitive visualization in the same visual dimension. For example, the height of a manufactured object is indicated by its height on the screen. So far the closest design to our proposal of token representation might be the Reese's Pieces Samples applet in the Rossman/Chance Applet Collection (Rossman & Chance, 2004) in which circles are colored in different shades of yellow that resemble actual Reese's pieces, but this coloration is simply used to separate targeted pieces from non-targeted pieces with no intention to suggest variation among samples. Because observers can quickly and accurately compute ensemble statistics about a display (e.g., Alvarez & Oliva, 2008). Thus, there is good reason to believe that learners are capable of visually aggregating tokens to compute aggregated values of interest. Moreover, immediate feedback is included in our manipulative because repeatedly producing credible data that is inconsistent with a learner's current understanding can support reflective change of the underlying misconception (e.g., Chin & Brewer, 1993). The use of token representation also allows us to investigate how perceptual scaffolds with special and generic features influence the effectiveness of perceptual grounding, and how they affect transfer of learning (if applicable). Not all physical properties are created equal. Neuropsychological studies have supported location's uniqueness. The location property is processed independently from other properties on other dimensions (Humphreys,1981).

Experiment 1

Experiment 1 had two goals: on the one hand, on the one hand, we were interested in identifying and confirming some common misconceptions that college students have regarding statistical inference; on the other hand, we would like to get insights for developing a grounded simulation of population sampling to tackle these misconceptions.

Participants We recruited 141 undergraduates at Indiana University, Bloomington in exchange for course credit.

Stimuli The experiment included three parts: pretest, training, and posttest. Both pretest and posttest probed the relations between standard deviation, mean, and sample size and their effects on confidence judgements. The test pool contained 12 three or four-option multiple choice questions. Each test had six randomly chosen questions. The pretest was used to assess students' statistical reasoning prior to interacting with simulations and the posttest was used to detect any changes in their statistical reasoning. It was a mixed design with conditions (color and location, discussed below) being a between-subject variable and factor levels being within-subject variables.

Cover stories. Our cover stories took place in a factory where two machines produce products (widgets or balls) under one of the three distinctive settings (i.e., different levels of means) on any given day, but their settings change from trial to trial. On some days the two machines had the same settings whereas on other days they did not. Products' consistency depended on which operator was in charge, sometimes with little variation of the products, sometimes moderate variation, and other times large variation (i.e., different levels of variability). After each day, different sized samples of products were presented to the examiner (i.e., learner) to estimate the probability that the two machines were set to the same setting on that day. The cover story was explicit that there were three levels for each of three variables: means, standard deviation, and sample size. To avoid any misleading interpretations, we limited the use of numerical information and standard statistical language. Two cover stories for two underlying visual appearances, color and location, were created with the same gist.

Visuals. We picked two easily recognizable visual properties: color and location. For each visual dimension, the mean, variance, and sample sizes could be visually determined without any numerical information being required. A color condition (Figure 1 upper panel) featured green circle widgets at three distinctive average lightness levels in an RGB color space: (0, 100, 0), (0, 120, 0), and (0, 140, 0) (greater G values produce lighter greens). Because the greenness level was the average value of a population, G values of individual widgets were deviated from the mean by an amount specified by the standard deviation (19, 38, 64). Sample size was represented by the number of widgets shown. A location condition (Figure 1 bottom panel) was identical to the color location except that we used bouncing ball heights as our tokens. We used bouncing heights (the highest point a bouncing ball reaches after it hits the floor) to convey location information.

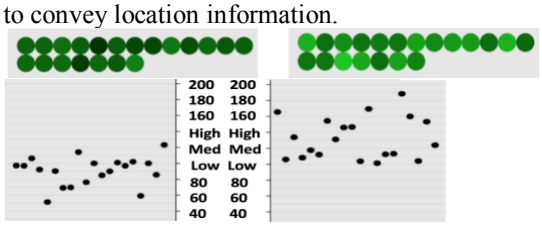

Figure 1. Two examples of a sample trial with *left mean* = 100, *right mean* = 140, *sample size* = 20, and *standard deviation* = 19. Upper panel (a sample color trial): on average, the left side widget's shade of green (100) is darker than the right-side widget's shade of green (140). Bottom panel (a sample location trial): on

average, the left side balls' bouncing height (100) was lower than the right-side balls' bouncing height (140).

Feedback. Trial by trial feedback included three parts: The first part was how close a learner's guess (P_{guess}) was to the Ideal Observer's (P_{actual}), along with how many points were earned $X_{earned} = 100 - |P_{actual} \times 100 - P_{guess} \times 100|$. The Ideal Observer was a Bayesian model created under the assumption that it always behaved rationally and gave perfect probability estimates. Its confidence judgements were made with the same information made available to participants. P_{actual} was calculated by Markov Chain Monte Carlo sampling. Hence, the closer a learner's guess was to that of the Ideal Observer, the more points were earned. The second part was a facial expression. This face initially seemed to be anticipating a response, and would then present various levels of happiness depending on how close a guess was (the closer the happier, see Figure 2). The final part was the information of underlying settings, including level of means, standard deviation, sample size (in plain language matching the cover story), and a larger collection of objects produced under each setting. The feedback was designed to encourage participants to attempt to adjust their guesses to maximize their performance.

Figure 2. Facial expression gradient as a function of how close a participant's guess was to the Ideal Observer's estimate.

Training session. The training session had 144 trials, with each trial asking participants to judge the probability that the two machines were given the same setting based on self-drawn samples (statistical inference). The probability was translated into confidence (in terms of the two settings being the same or different, not in terms of accuracy in their judgment). Confidence estimates ranged from 0% (definitely different) to 100% (definitely same), with 50% indicating complete uncertainty (increments of 1%). The samples were manipulated in a 3 (difference between means) × 3 (standard deviations) × 3 (sample sizes) repeated measure design. Each factor ranged across three levels featuring low, medium, and high values: difference between means (0, 20, 40), standard deviation (19, 38, 64), and sample size (5,10,20). Each token was normally distributed with a mean of one of the mean levels, and a standard deviation of one of the standard deviation levels. To make judgments simpler, on any given trial, standard deviation and sample size were the same between two samples while means may or may not differ. Although the ground truth was not knowable by either the learners or the Ideal Observer, we had half *same* and half *different* trials. The numbers of trials with each level of sample size and standard deviation were equalized accordingly.

Procedure Participants first completed six multiple choice questions in the pretest with feedback on overall accuracy upon completion, followed by a cover story as well as tutorial matching the condition they were randomly assigned into. They were then instructed to press the "Step 1 Draw Samples" button to draw two separate samples from behind a curtain (contained within gray rectangles) and to move a slider to indicate their probability judgment (Step 2). Step 3 (submit guess) and Step 4 (reveal setting information) were designed to provide feedback. Participants then pressed "Step 5 Next Game" to start a new trial, repeating the same five steps for 144 times. On each trial, their guess, the Ideal Observer's estimate, and points earned were recorded. Figure 3 shows a complete feedback page in a color condition (the location condition was similar). Upon completing the 144th trial, participants were given their total score out of the maximum possible score 14400. They then completed a posttest containing the six remaining questions from the question pool. Questions were randomized and overall accuracy was given after completion.

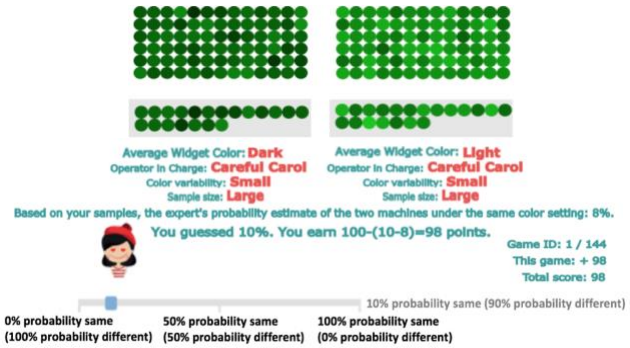

Figure 3. Experiment 1's complete feedback page (a color condition).

Results & Discussions

We calculated a correlation for each participant's guesses and the Ideal Observer' estimates during the training session. Those whose correlations were two standard deviations below the mean were excluded (N = 1). Hence, 140 participants were included in the following analysis ($N_{color} = 69, N_{location} = 71$).

Difficulties & Misconceptions. Analyses of accuracy with respect to the Ideal Observer and sensitivity to standard deviation, sample size, and difference between means were conducted to reveal the influence of sample factors on participants' statistical reasoning.

Accuracy analysis. Answers were transformed to reflect confidence along the direction of the ground truth. Because the slider incremented from 0% - 100% in the direction of sameness and decremented from 100% - 0% in the direction of difference, values on a same trial were kept as they were and value on a different trial were flipped on the 100-point scale. For example, an answer of a 60% was coded as 60 on a same trial and coded as 40 (100-60) on a different trial. As demonstrated in Figure 4, the Ideal Observer's estimates differed significantly from one level to another of a factor, so did the averages of participants' guesses, *p*s < .001

except that their judgments barely changed from a medium to a large sample size, p = .18. Sample factors revealed a similar pattern: participants were consistently more conservative than the Ideal Observer.

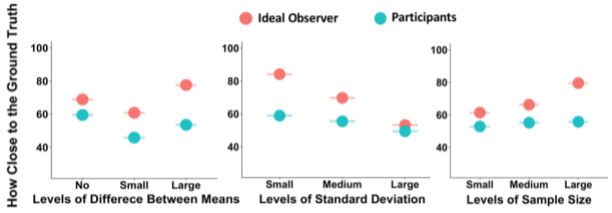

Figure 4. Aggregated participants' guesses and ideal estimates. Error bars represent one standard error of the mean.

Sensitivity analysis. Sensitivity was measured by how steeply estimates changed as the level of sample factors changed. For difference between means, we used original estimates. For sample size and standard deviation, we measured estimates' deviations from 50%. Because 50% indicates complete uncertainty, greater deviations regardless of the direction from 50% suggest greater confidence. This is important, because factors like increasing sample size should increase confidence in an answer, but the specific direction depends on whether the means are truly different. A zero (large) difference between means is typically associated with a higher probability of sameness (difference). Thus, we used original responses to measure while and a large difference between means are typically associated with greater confidence.

As suggested by Figure 5, for each sample factor, interactions between levels of factors and estimators were observed, ps < .001. Indeed, the Ideal Observer responded to changes in factors more steeply than participants. We used the steepness of change as an approximation of sensitivity. Hence, we interpreted these significant interactions as suggesting that participants were less sensitive than the Ideal Observer with respect to changes in standard deviation, sample size, and difference between means. A ratio (R) of participants' sensitivity to the Ideal Observer's sensitivity was calculated for each sample factor to compare relative influence of each factor assuming that the Ideal Observer reacted perfectly rational to changes in statistical information embedded in the token representation. R = 1 implied that participants were influenced by a sample factor as was the ideal Observer. R > 1 implied over sensitivity (giving too much weight to a factor) and R < 1 implied under sensitivity (giving too little weight to a factor). A one-way analysis of variance (ANOVA) revealed a significant difference between three ratios, $F(2,417)$ = 25.41, p < .001. Specifically, participants gave significantly more consideration to difference between means (M = .28, SD = .23) than standard deviation (M = .14, SD = .20), $t(278)$ = 5.32, p < .001 and sample size (M = .09, SD = .25), $t(278)$ = 6.54, p < .001. Standard deviation had a marginally greater impact on participants than sample size, $t(278)$ = 1.87, p = .063. Thus, as demonstrated in Figure 5, it is clear that participants were influenced by difference between means the most, followed by standard deviation, and then sample size. Despite being most sensitive to difference between means, the participants (M = -.33, SD = .27) were not nearly as influenced by this factor as was the Ideal Observer (M = -1.16), $t(419)$ = 63.73, p < .001.

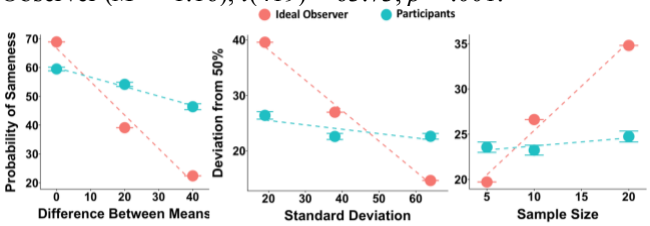

Figure 5. The Ideal Observer and the participants' sensitivity to sample factors. Difference between means was plotted against original estimates while standard deviation and sample size were plotted against deviation from 50%. Error bars represent one standard error of the mean.

Learning & Transfer. Learning was measured by the correlation between the number of trials completed and the absolute difference between the Idea Observer's estimates and the participants' guesses. Intuitively, if learning occurred during the perceptual training, the difference should decrease as the number of trials increased. No learning was observed, $r(142)$ = .056, p = .51. Given that the scenarios in the posttest questions were only distantly related to the factory-based stories we used in training, we treated posttest performance as a measure of far transfer. No transfer was observed, $t(139)$ = 1.08, p = .28.

The findings of Experiment 1 suggested that participants had difficulty reacting in the mathematically warranted way to varying levels of sample size, standard deviation, and difference between means. They were not influenced by any of these factors as was the Ideal Observer. Specifically, the participants gave the most weight to difference between means, less weight to standard deviation, and very little weight to sample size.

More surprisingly, no learning was observed despite 144 repetitions and immediate, trial-by-trial feedback. The participants were told that the relevant settings information should be important for their judgments and encouraged to explore how difference between means, standard deviation, and sample size affected ideal estimates. Thus, we speculate that the absence of learning was due to no explicit instructions were given as to how participants should integrate numerical information in the feedback to adjust their responses. Hence, our findings suggest that explicit descriptions of how learners' estimates deviated from the ideal estimates in accord with each of the three factors should be given, especially at the initial stage of learning when participants did not yet have "internal" guidance.

Experiment 2

Experiment 1 revealed that presenting relevant information without explicit guidance of how to use them in probability judgments was of little to no use to inducing learning. Thus, in Experiment 2, we provided tailored instead of generic feedback. Moreover, different levels of standard deviation, sample size, and difference between means were used to

make more clear-cut trials near 0% and 100% judgments. There were also some aesthetic modifications to avoid encouraging mistakes due to interface layout. The main goal of Experiment 2 was to examine whether explicit guidance of how to use perceptually salient cues could bridge the gap between perceptually grounded experience and learning, and perhaps even transfer.

Participants We recruited 239 undergraduates at Indiana University, Bloomington in exchange for course credit.

Stimuli. The design of Experiment 2 was identical to that of Experiment 1 with a few modifications to feedback: a) an ideal estimate was mapped onto an identical slider immediately below the participants' slider, in the hope of letting participants visually see how far their answers were away from the Ideal Observer's; b) sliders were centered to avoid the tendency to always move the thumb to the center of screen, an act which would lead to higher probabilities of sameness (we had many flipped answers in Experiment 1); c) increased differences between each factor level: mean (100, 125, 150; thus difference between means: 0, 20, 50), standard deviation (20, 36, 54), and sample size (4, 9, 20); d) tailored feedback was given to show how the Ideal Observer reached its estimates (discussed below).

Tailored feedback. Regardless of the direction from 50% (complete uncertainty), no justification was given when the deviation between a participant's guess and the ideal estimate was smaller than 15% ($P_{\text{difference}} < 15\%$). When $P_{\text{difference}} > 15\%$ and judgments were in the opposite directions, difference between means was highlighted to reveal the wrong judgment of which type was more likely. When the directions were the same (both below or above 50%) and $P_{\text{difference}} > 15\%$, in an Ideal-Observer-more-confident-same (different) case, a larger sample size, a smaller (larger) standard deviation, and a more extreme difference between the two means were highlighted whenever applicable. Likewise, in an Ideal-Observer-less-confident-same (different) case, a smaller sample size, a larger (smaller) standard deviation, and a more moderate difference between the two means were highlighted whenever applicable (see Figure 6).

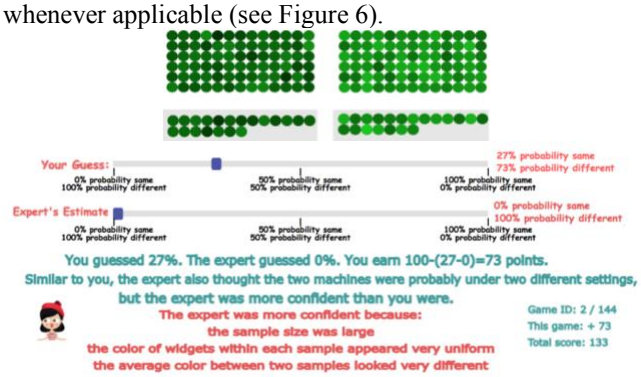

Figure 6. Experiment 2's complete feedback page (color condition). The location condition had an identical layout.

Results & Discussions

We applied the same exclusion criteria used in experiment 1 and included 218 participants for the following analysis ($N_{color} = 113, N_{location} = 105$).

Accuracy analysis. Similar to Experiment 1, we flipped estimates by 100 – estimates when the ground truth was different. Again, participants were consistently more conservative than the Ideal Observer across difference between means, sample size, and standard deviation.

Sensitivity analysis. Identical to Experiment 1's analysis, we conducted a sensitivity analysis for Experiment 2 by analyzing relative steepness of changes in responses. Interactions between factor levels and the estimator were observed at each factor, *p*s < .001. A one-way ANOVA revealed a significant difference between sensitivity to three sample factors, $F(2,651) = 64.18$, $p < .001$. As demonstrated in Figure 7, the pattern was identical to that in Experiment 1: participants were influenced by difference between means the most ($M = .56$, $SD = .21$), followed by standard deviation ($M = .35$, $SD = .29$), $t(434) = 8.42$, $p < .001$, and sample size ($M = .26$, $SD = .33$), $t(434) = 38.79$, $p < .001$. They were also more influenced by standard deviation than sample size, $t(434) = 3.17$, $p = .0016$.

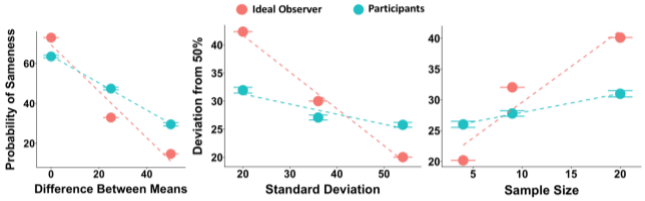

Figure 7. The Ideal Observer and the participants' sensitivity to sample factors. Difference between means was plotted against original estimates while standard deviation and sample size were plotted against deviations from 50%. Error bars represent one standard error of the mean.

Learning & Transfer. Learning was observed as there was a significant correlation between the absolute deviation between participants' guesses and ideal estimates and number of trials, $r(142) = -.18$, $p = .029$ (see Figure 8a). This effect was not simply due to repetition because participants in Experiment 1 went through the exact same procedures without revealing any signs of learning. Transfer of learning was also revealed by a paired sample t-test, with higher posttest accuracy ($M = .58$, $SD = .24$) than pretest accuracy ($M = .61$, $SD = .25$), $t(217) = -2.18$, $p = 0.030$ (Figure 8b). The correlation between learning during training and transfer of learning, however, was only marginally significant, $r(216) = .12$, $p = .068$. Thus, there was some suggestion that participants who established a perceptual grounding of fundamental statistical concepts during training did not necessarily develop the ability to apply that new gain to contextually dissimilar but structurally similar problems.

Overall, the findings of Experiment 2 showed that with tailored feedback, participants in Experiment 2 showed both significant learning and transfer of learning. This suggests the importance of analytic feedback that specifies not just how good a response was, but what factors were probably not influencing judgments sufficiently. Meanwhile, the two

types of improvement seems to indicate that statistical reasoning is not just one thing. While most people think the kind of quantitative and immediate feedback we gave in Experiment 1 is more common, and is perhaps the gold standard, the type of relatively unusual feedback we use in Experiment 2 is applicable to many educationally relevant interventions.

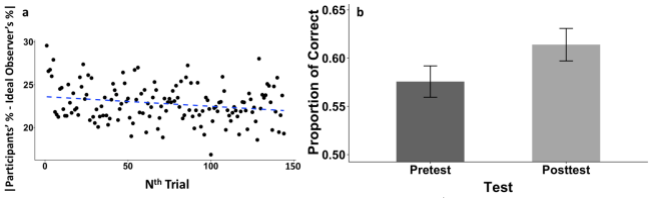

Figure 8a. Learning: Correlation between the n^{th} trial and the absolute difference between the Ideal Observer's estimates and the participants' guesses. *Figure 8b.* Transfer: Error bars represent one standard error of the mean.

General Discussion & Conclusion

The central goal of the present study was to assess the applicability of a grounded cognition approach to learning the relations between basic statistical concepts. Across two experiments, we found that tailored guidance along with perceptually salient properties has the potential to induce both learning (during training) and transfer of learning.

Psychometrically, the present study proposes a new paradigm to quantitatively measure people's sensitivity to three factors underlying statistical inference (difference between means, variance, and sample size). Our task is promising because it 1) allows us to compare learners' relative sensitivity to these three factors so as to make quantitative claims about how much each factor is influencing their judgments, 2) allows us to compare these influences to an Ideal Observer that makes optimal use of the information in a display, 3) allows us to quantitatively measure improvements in the use of these factors over training, 4) allows us to give learners quantitative and objective feedback on their task performance, and 5) gives us a method for quantitatively assessing performance on a statistical inference task that is potentially independent from, but possibly correlated with, other explicit measures of statistical reasoning. This last feature allows us to empirically determine if implicit and explicit measures of statistical reasoning are tapping into the same knowledge.

Experiment 2 produces two kinds of improvement: a) training improvement from beginning to the end and b) improvement from pretest to posttest, but they are not correlated. This suggests that reasoning with statistics is not just one thing. An improved ability to reason with variability, sample size, and difference between means is manifested in both a quantitative and a qualitative way. A quantitative understanding is through improved numerical integration. Some people improve their ability to integrate variability, sample size, and difference between means, and hence improve during training. They learn how to give more appropriate consideration to the relative weight of each factor. A qualitative understanding, on the other hand, is manifested through transfer of learning from a pretest to a posttest. These people pay attention to the detailed, analytic feedback and they do better on posttest (Experiment 1 does not have this kind of feedback). Compared to numerical judgments during training, pretest and posttest questions are more concerned with (at a descriptive level) how changing variability, sample size, or difference between means affects confidence respectively. Hence, learners who demonstrate a qualitative understanding learn directionally how important each factor is in statistical reasoning. More importantly, as suggested by the lack of correlation between the two improvements, they do not occur in an all-or-none fashion. Specifically, integrating statistical information is different from just being able explicitly state if and how sample size, variability, and difference between means should affect judgments.

References

Ahn, J. (2007). Application of experiential learning cycle in learning from a business simulation game. Doctoral Dissertation, New York: Teachers College, Columbia University.

Alvarez, G.A. and Oliva, A. (2008) The representation of ensemble visual features outside the focus of attention. *Psychol. Sci. 19,* 678–685.

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

Black, J. B. (2010). An embodied/grounded perspective on educational technology. In M. S. Khine, & I. M. Saleh (Eds.), *New science of learning: Cognition, computers and collaboration in education*. New York, NY: Springer.

Chin, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63(1),* 1-49.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21,* 1–78.

Dewey, J. (1916) Democracy and Education. New York: Touchstone.

Dewey, J. (1986). How we think. (Rev. ed.). In J. A. Boydston (Ed.), *The collected works of John Dewey:* Vol. 8. Carbondale, IL: Southern Illinois University Press. (Originally published in 1933).

Gardner, H., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education, 7(1).*

Garfield, J., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review, 67,* 1–12.

Humphreys, M. A. The dependence of comfortable temperatures upon indoor and outdoor climates. In K. Cena and J. A. Clark (Eds.), Bioengineering, thermal physiology and comfort. New York: Elsevier Scientific Publishing, 1981.

Rossman, A., and Chance, B. (2004), "The Rossman/Chance Applet Collection," 2009.

Sorto, M. A. (2006). Identifying content knowledge for teaching statistics. In A. Rossman & B. Chance (Eds.), Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil.

Suh, J. M., Moyer, P. S., & Heo, H.-J. (2005). Examining technology uses in the classroom: Developing fraction sense using virtual manipulative concept tutorials. *The Journal of Interactive Online Learning, 3(4),* 1-22.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science, 1974, 184,* 1124-1131.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, The Netherlands: Kluwer.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students-some interesting parallels. *British Journal of Educational Psychology, 61*, 319-328.