DEVELOPING STUDENTS' CAUSAL UNDERSTANDING OF SAMPLING VARIABILITY: A DESIGN RESEARCH STUDY

Ethan C. Brown and Robert delMas
Department of Educational Psychology
University of Minnesota—Twin Cities, USA
brow3821@umn.edu

Students struggle with understanding sampling variability. Recent transfer and conceptual change literature suggests a new pedagogical approach: supporting a causal explanation of why sampling variability decreases as sample size increases. We designed technological tools, new representations, and a hypothetical learning trajectory to help students understand causal mechanisms of sampling variability. Two causal explanations were targeted: "swamping", extreme deviations from the mean are less influential in larger samples, and "heaping", values near the population mean become more probable as sample size increases. Preliminary pilot testing of five students who recently completed an introductory statistics course demonstrated the potential of the activities for deepening students' thinking, as well as the challenges of managing and articulating the abstraction of swamping and heaping.

BACKGROUND

Sample size and population variability fundamentally affect the precision and power of a study. Past American Statistical Association president Jessica Utts stated that understanding of the influence of sample size is part of "What Educated Citizens Should Know about Statistics and Probability" (Utts, 2003). However, introductory statistics students struggle to understand the dynamics of sample size and sampling variability (Chance, delMas, & Garfield, 2004), and poor understanding even among empirical researchers (Tversky & Kahneman, 1971) has played a part in the current reproducibility crisis, leading to calls for a "new statistics" (Cumming, 2014). Low-powered studies may be leading to countless spurious results entering the scientific canon (Fiedler, 2011). Without understanding and accounting for sample size, population variability, and their impact on sampling variability, researchers and the general public will continue to have great difficulty making the best decisions based on evidence, whether the statistical methods are old or "new".

RESEARCH ON SAMPLE SIZE & SAMPLING DISTRIBUTIONS

Despite extensive study of sample size reasoning, there is no cogent theoretical framework of why, how, and when people reason about sample size (Lem, Van Dooren, Gillard, & Verschaffel, 2011). When Kahneman & Tversky (1972) provided a classic demonstration of sample size neglect, they theorized that people did not view sample size as relevant to decisions. Yet people correctly attend to sample size in certain contexts (Bar-Hillel, 1979; Obrecht, Chapman, & Suárez, 2010). Some sample size neglect may result from confusion about the difference between the distribution of individual values and the sampling distribution of the means (Sedlmeier, 1998), but confusions and inconsistencies are present even when people appear to have mastered this distinction (delMas, Garfield, & Chance, 1999; Well, Pollatsek, & Boyce, 1990).

Current interventions for sample size neglect have only demonstrated success for limited domains and types of problems. Nisbett, Fong, and Lehman (1987) gave their participants a passage describing the Law of Large Numbers and provided several example applications, and found some improvements in statistical reasoning relative to a control group. However, the authors found that participants often did not successfully transfer sample size reasoning beyond the teaching context, and other researchers have criticized the authors' arbitrary scoring scheme for statistical reasoning (Sedlmeier, 1999). Sedlmeier (1999) created activities based on instructional design principles that used rich computer simulations where participants could experience the process of sampling, viewing individual samples, and viewing the sampling distribution for different problems. Sedlmeier's participants showed a sustained improvement in sample size reasoning across multiple tasks; however, the training and transfer tasks were all isomorphic to the Hospital problem so it is not clear how well thinking would transfer to other tasks.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

In a similar series of simulation-based intervention studies targeting introductory statistics students' understanding of sampling distributions by delMas and colleagues (Chance et al., 2004; delMas et al., 1999), students created empirical sampling distribution histograms based on a given population histogram. When combined with a classical conceptual change approach (Posner, Strike, Hewson, & Gertzog, 1982) that confronted students with their misconceptions after the pretest, this intervention appeared to be quite successful at encouraging most participants to choose histograms with smaller variability for larger sample sizes (delMas et al., 1999). However, qualitative follow-up interviews revealed that students' reasoning was often confused even when they chose correct histograms (Chance et al., 2004). Additionally, an unpublished follow-up study revealed that students still performed poorly on several classic sample size tasks after the histogram-based intervention (delMas, Garfield, & Chance, 2006).

How can people learn to change their incorrect reasoning to correct reasoning? The Knowledge Revisions Components (KReC) Framework (Kendeou & O'Brien, 2014), drawn from extensive experiments on reading comprehension and refutation texts (e.g., Kendeou, Smith, & O'Brien, 2013), suggests that activating old and new information and strengthening the new information with *causal* explanations is an effective way of revising knowledge. Since the effect of sample size on sampling variability is a phenomenon that results from taking the mean of a random process, the causal explanation is not straightforward. Chi (2013) terms these processes *emergent* and her research suggests that understanding the *inter-level* process of summing the individual values at the local level (e.g. the individual cases in the sample) to produce the effect at the aggregate level (e.g. the sampling variability of the mean) is particularly crucial for being able to generate correct causal explanations. No interventions were found in our review that provided a causal explanation, inter-level or otherwise, of the effect of sample size on sampling variability.

CAUSALLY EXPLAINING THE EFFECT OF SAMPLE SIZE

In contrast to prior interventions that only supported students in noticing and applying the effect of sample size on sampling variability, we developed a new hypothetical learning trajectory (HLT) to scaffold two inter-level causal explanations of the reduction in sampling variability as sample size grows:

- *swamping*, the decreasing influence of extreme values on the sample mean
- *heaping*, the increasing concentration of possible sample means near the population mean

Swamping, demonstrated in Figure 1, explains how a particular sample mean will move less and less as more values are added to the sample. This property follows directly from the formula for the mean and does not depend on stochastic processes: means are the sum of the values divided by the number of observations, so any given observation will be divided by a larger number for larger samples and will therefore have less influence. The term *swamping* was coined in Well et al. (1990) based on student observations of this property. However, by itself swamping only explains the decreasing size of movements of the mean relative to a deviation and does not directly explain either the decrease in variability or the increasing concentration around the population mean. For example, it is quite possible for a process to have decreasing possible movements without converging to any particular value, such as the harmonic series ($1 + ½ + 1/3 + \ldots$), which never converges.

Heaping, demonstrated in Figure 2, provides a more complete explanation of the effect of sample size. Viewed from the Laplacian perspective of a sample space with equally likely outcomes, there is an increasing proportion of possible outcomes near μ in the sampling distribution as sample size increases (Kazak & Konold, 2010). Figure 2 applies this logic to the sampling distribution of a Bernoulli random variable with $\pi = 0.5$, which models a coin flip game where heads are scored as 1 and tails scored as 0, and the average score is recorded. When the sample size is 2, there are two ways to get an average score of 0.5. As the sample size increases, *heaping* can be observed: there are increasingly more ways of getting an average score close to population mean, and therefore sample means increasingly "heap" around the true mean.

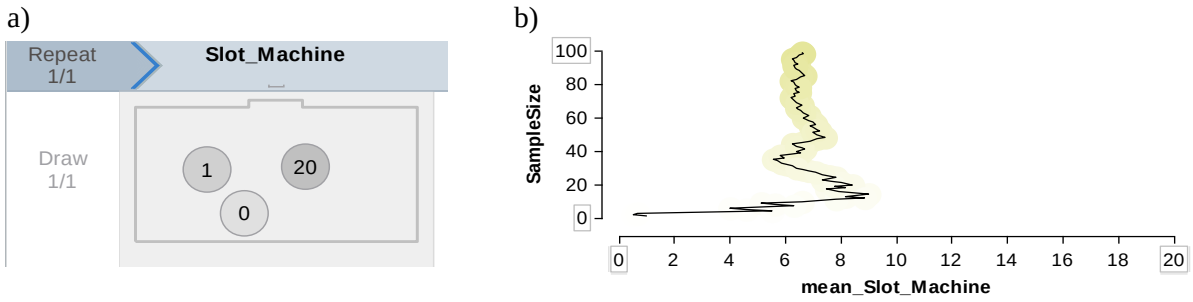

Figure 1. Demonstration of *swamping* (Well et al., 1990), showing how large samples are less influenced by extreme values than small samples, created by a pilot participant using TinkerPlots (Konold & Miller, 2017).

With support, the participant a) created a model that randomly picks one of several balls repeatedly, then (b) viewed a mean against sample size (MASS) plot to show that the mean changes rapidly at first but moves less and less as the sample size grows.

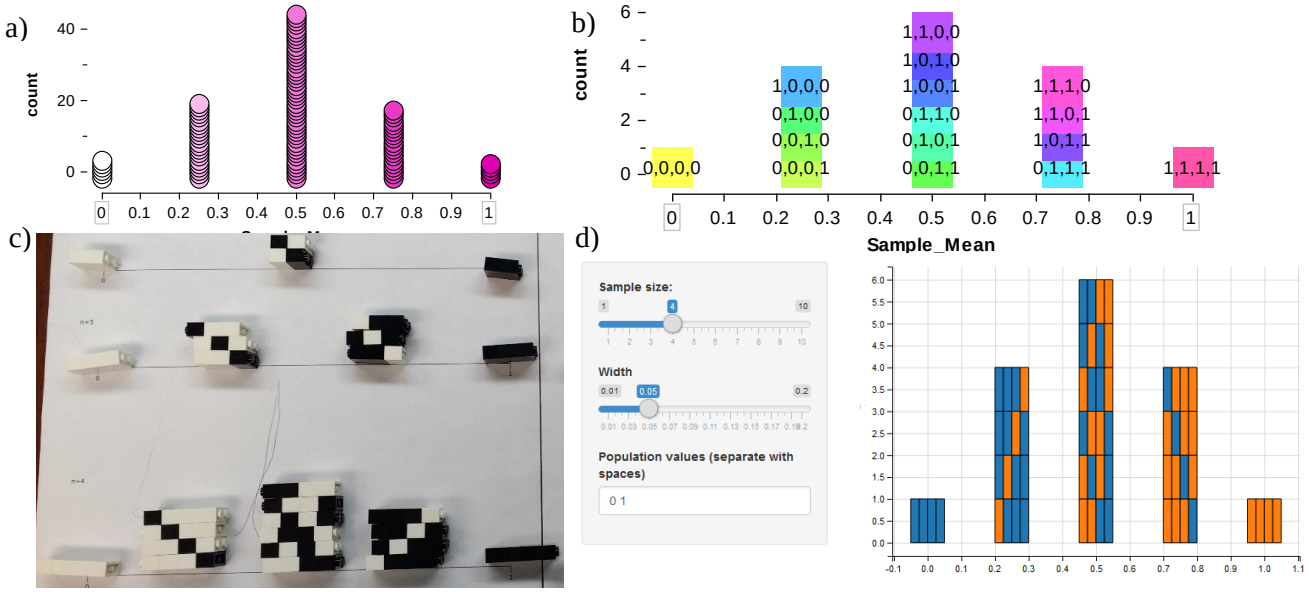

Figure 2. Demonstration of *heaping* (e.g., Konold & Kazak, 2008), showing how means of large samples have more opportunities to be near the true mean than small samples.

The examples in Figure 2 depict the results of repeatedly drawing 0 or 1 with equal probability four times and then recording the mean. Participants first create a model that a) generates 100 means. Participants then b) examine unique combinations of outcomes and note that there are more ways to get an outcome near 0.5 because there are more possible samples near that value. Participants further explore the underlying combinations using c) hands on manipulatives and d) a custom application.

The HLT to support swamping and heaping consists of a sequence of tasks that provide increasingly rich conceptual support for causal thinking about sample size. Students first explore swamping in the context of a sample proportion or mean as sample size grows (Bakker, 2004). These activities support students in growing a sample and describing the effect of sample size on sampling variability. Students visually create random models (Figure 1a) in TinkerPlots (Konold & Miller, 2017) that incrementally add values onto a sample and witness how the mean, initially quite unstable, begins to settle down near an expected value as viewed in the mean against sample size (MASS) plot (Figure 1b) to support their understanding of the mechanism of swamping.

Students explore heaping in the last three activities by growing multiple samples simultaneously and plotting the means (Figure 2a). They then explore the underlying mechanism by examining the unique outcomes that arise in their simulations, and noting that there are more possible outcomes near the mean (Figure 2b). Finally, students connect the simulations with the underlying structure, first by a hands-on exploration constructing the different combinations using building blocks (Figure 2c) and then with a custom-built application for displaying all possible combinations of outcomes (Figure 2d).

Although the basic structure of the HLT is compatible with the literature on sample size and sampling distributions, we were not able to find much previous literature on how to support student understanding of swamping or heaping. To support comparisons with existing research, these activities are bracketed by pre- and post-interviews with students as they solve classic sample size problems to examine the nature of changes in their thinking while working through well-studied problems.

PILOT TESTING OF THE HLT

We conducted an initial pilot test of five students who had recently completed the CATALST course (Garfield, delMas, & Zieffler, 2012), a randomization-based introductory statistics course which uses TinkerPlots extensively. The purpose of the pilot test was 1) to assess how richly the HLT revealed students' thinking about sample size, 2) to assess whether the HLT appeared to be successful in encouraging *causal* thinking about sampling variability, and 3) to improve the tasks, representations, and prompts. Since our focus was on revealing and exploring student thinking, we engaged pilot test participants in semi-structured clinical interviews with a focus on understanding participants' sense-making in the contexts and representations we provided (diSessa, 2007). There were no explicit instructional materials provided, and the interview protocol took pains to ensure that that the interviewer's role was to explore student thinking, not to teach or evaluate.

We were interested in how students causally reasoned and how students used representations to support their reasoning. For instance, after students had spent a half-hour with the MASS plot, four of the five pilot participants described swamping in some form without much prompting. Although students initially varied on whether they thought sample size would have any effect on sampling variability in the pre-interview, after seeing the MASS plot they were usually able to articulate swamping, e.g. "the smaller the sample is, the greater it can be skewed in proportion with just a few … repeated results." However, one participant (P5) used non-inter-level reasoning to explain why the changes in the mean were smaller at larger sample sizes. Here is the participant's reasoning about the MASS plot for a sampler with one 0 and two 1s:

P5: We're looking at changes of, like, 0.3, 0.2, 0.2, so, like, the change, like, keeps getting bigger and bigger than when we go here, we have changes of 0.01 and as we get lower we get a change of 0.001 and as you go lower we have a change of 0.001 and it just keeps getting. So, like, the change keeps getting smaller and smaller as the sample size keeps getting bigger and bigger.

I: Okay, and so why—why do you think that would be?

P5: Um, I think it's because, like, since we're dealing with a huge sample size and that the number is gonna be centered more closer to 0.67, it's not gonna change that much from it. Okay, then that's why our number is like very, very, very low.

Instead of arguing from swamping, P5 argues teleologically that large samples should match the theoretical probability. They mentioned earlier that "we were to do like, like, like do those drawings and have like a huge sample size then our proportion should match like what it should theoretically be." Only after a lot of focused prompting about the effect of extreme values did P5 eventually offer swamping-like reasoning.

Many students were able to express heaping after looking at the combination plots and constructing combinations themselves. However, students' explanations of heaping tended to be more tentative. Here is one participant explaining heaping for three draws of a 0/1 variable:

P3: To equal—to get the average of this probability, this one—this one—this one in this, okay. So there are two up there, three options—there are three different orders that resulted in the same average, 1, 2, 3, all in this average. So, so, every combination is equally unlikely, but some of

the combinations result in the same mean, which means that this is statistically higher chance of getting certain means given the true—or given the different combinations that result in the median.

In the post-interview questions, participants rarely used heaping-style reasoning during problem solving. However, they frequently referred to the building block exploration of heaping as striking and memorable when asked to reflect on what they learned and what surprised them about the HLT:

P3: I guess I'm still, like, processing, like, with the combination thing. I'm thinking about that, that's still on my—on my mind. Um, I guess what we saw there is the—the more number the you drew, the more combinations and the more stacks you have... there's only... [trails off]

P4: I think the [building block] activity definitely surprised me, just because seeing how the graph changed and taking into account different ways that you could get the same average. So, like, having like two blacks and one white, or, like, black black white, or, like, black white black, and how that affected the numbers was surprising to me.

At the end of the HLT, these two participants (whom we judged, out of the five participants, to have the strongest understanding of the tasks) were still struggling with articulating and thinking about heaping.

DISCUSSION AND CONCLUSION

Giving students an opportunity to causally explore the effect of sample size on sampling variability may be a powerful teaching tool. The HLT presented here provides a first attempt to explore the pedagogical possibilities of highlighting swamping, the decreasing influence of extreme values on larger samples, and heaping, the piling up of possible samples near the true mean. Pilot testing has shown the potential depth of student thinking about swamping and heaping, but also demonstrated the difficulty of designing effective interventions. Plotting means of samples against the sample size in rich, well-motivated contexts appears to support understanding of swamping, but understanding of heaping was very tentative even in the strongest students. In both cases, however, the tasks and representations appeared to be successful in revealing complex and previously undocumented student reasoning about sample size.

It is not entirely surprising that swamping came more easily to students, given that it was one of the major correct reasoning types used by college students in Well et al. (1990). Swamping is a relatively simple property of means. Nevertheless, the MASS plot appeared to successfully highlight the decreasing influence of extreme values. Students' expressions of swamping had a variety of nuances that can be examined in follow-up research. For instance, some students seemed to focus more on extreme values, while others talked about how a "streak" of a certain value would affect the mean less at a larger sample size, how it would take a smaller absolute number to change the mean at a lower sample size, or how the equation for the mean implies swamping.

Heaping may be easier to understand in some contexts than others. One student initially gave heaping-like reasoning about the extremes of the sampling distribution of coin flips, arguing that it was much easier to get all heads for a smaller number of flips rather than for a larger number. Students were asked in the post-interview how heaping connected to a problem about a continuous distribution and a problem about a coin flip, and one student who saw no connection to the continuous distribution easily drew a correct connection to the coin flip situation. Further research will explore the potential of these activities by iterating the HLT over a larger sample of students and continuing to refine and improve the materials and prompts to isolate and clarify students' thinking.

REFERENCES

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools* (Doctoral dissertation).

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*(2), 245–257. http://dx.doi.org/10.1016/0030-5073(79)90028-X

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about Sampling Distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295–323). Springer Netherlands.

Chi, M. T. H. (2013). *Two Kinds and Four Sub-Types of Misconceived Knowledge, Ways to Change it, and the Learning Outcomes*. Routledge Handbooks Online.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3).

delMas, R., Garfield, J., & Chance, B. (2006). *Using assessment to study students' reasoning about sampling distributions*. Unpublished manuscript.

diSessa, A. A. (2007). An Interactional Analysis of Clinical Interviewing. *Cognition and Instruction*, *25*(4), 523–565. https://doi.org/10.1080/07370000701632413

Fiedler, K. (2011). Voodoo Correlations Are Everywhere—Not Only in Neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

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

Kazak, S., & Konold, C. (2010). Development of ideas in data and chance through the use of tools provided by computer-based technology. In *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*.

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) Framework: Processes and Mechanisms. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (p. 353).

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 854–865. https://doi.org/10.1037/a0029468

Konold, C., & Harradine, A. (2014). Contexts for Highlighting Signal and Noise. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics* (pp. 237–250). Springer Fachmedien Wiesbaden. Retrieved from http://dx.doi.org/10.1007/978-3-658-03104-6_18

Konold, C., & Kazak, S. (2008). Reconnecting Data and Chance. *Technology Innovations in Statistics Education*, *2*(1). Retrieved from http://www.escholarship.org/uc/item/38p7c94v

Konold, C., & Miller, C. D. (2017). TinkerPlots: Dynamic data exploration (Version 2.3.1). Emeryville, CA: Learn Troop.

Lem, S., Van Dooren, W., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica*, *53*(2), 123–135.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching Reasoning. *Science*, *238*(4827), 625–631.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, *16*(1), 26–44.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227. https://doi.org/10.1002/sce.3730660207

Sedlmeier, P. (1998). The distribution matters: two types of sample-size tasks. *Journal of Behavioral Decision Making*, *11*(4), 281–301.Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, New Jersey: Lawrence Erlbaum.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Utts, J. (2003). What Educated Citizens Should Know About Statistics and Probability. *The American Statistician*, *57*(2), 74–79. https://doi.org/10.1198/0003130031630

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289–312. http://dx.doi.org/10.1016/0749-5978(90)90040-G