The Lowly Worm Climbs Up a Winding Stair: Teaching the Roles of Randomness in Scientific Inference

Ethan C. Brown, Elizabeth B. Fry, Alicia Hofelich-Mohr
University of Minnesota—Twin Cities

Cognitive Challenges of Scientific Inference

Scientists must incorporate both the **reliability of evidence** and the **scope of inference**, yet both are **complex phenomena** that can be difficult to understand.

Difficulties with reliability of evidence

- Tversky & Kahneman (1971) found quantitative psychologists **vastly overestimated replicability** of findings from small studies
- Underpowered studies lead to many **"voodoo correlations"** (Fiedler, 2011) being reported in the literature
- Adults **frequently neglect sample size** (see Lem et al., 2011), which statistical training reduces but does not eliminate (e.g., Fong, Krantz, & Nisbett, 1986)

Difficulties with scope of inference

- Introductory statistics students struggle to understand when **study design** warrants generalizing to a population or concluding causation (e.g., Derry et al., 2000)
- News media reporting on research routinely **overgeneralizes** and adds **inappropriate causal attributions** (Schield, 2010)
- Scientists face many **cognitive biases** (Nuzzo, 2015) and **overstate their own findings** (see Utts, 2003)

Statistics education research can contribute to metascience's investigation of how scientists **interpret and treat evidence**:

- Students face **unique challenges** in learning statistics that require specialized pedagogical research (Garfield & Ben-Zvi, 2007)
- Statistics education research provides perspectives on **students' understanding at all levels**, from kindergarten to the workforce
- Our studies of introductory college statistics' students understanding can illuminate the challenges in **conceptualizing and adequately addressing threats to internal and external validity**.

Teaching Inference with Simulation

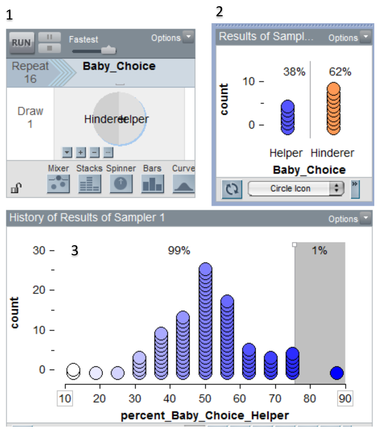
Statistics educators increasingly use **simulation and resampling techniques** to teach frequentist inference (Tintle et al., 2015):

- Simulation, bootstrap resampling, and rerandomization may help students focus on **the unified logic** of the sampling distribution (Cobb, 2007)
- Displaying uncertainty via concrete realizations of a stochastic process may **map on to human cognitive faculties** more naturally than abstract distributions (Gigerenzer & Hoffrage, 1995; Hullman et al., 2015)
- Simulation-based introductory statistics curricula have **some evidence of providing advantages** in student understanding and outcomes (Tintle et al., 2012)

Example simulation-based activity (from Zieffler et al., 2017)

In Hamline & Bloom (2007), 14 out of the 16 (87.5%) babies chose the "Helper" toy. *Does this study provide evidence that infants notice and prefer the "Helper"?*

1. **Students set up null model visually** in TinkerPlots: e.g., a spinner with 50% helper and 50% hinderer, or could be two balls "helper" and "hinderer" chosen with replacement.
2. **Students plot the result of a single trial** under null hypothesis.
3. **Students collect the percentage of means of many trials**, and evaluate how rare results at or above 87.5% would be under the null model.

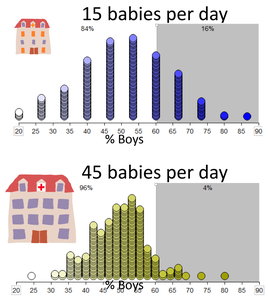


Study 1: Reliability of Evidence (Brown, 2019)

Background

*In one year, which hospital do you expect to have **more days with more than 60% boys**?*

- *The larger hospital*
- *The smaller hospital*
- *About the same* (Kahneman & Tversky, 1972)



As **sample size** increases, we **grow more certain** that a statistic is close to the population value. This fundamental principle is called the **Empirical Law of Large Numbers** (Freudenthal, 1972)

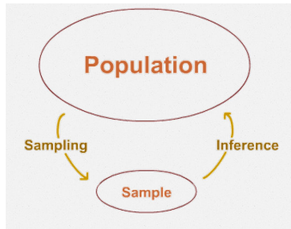
Common confusions include:

- Ignoring sample size, believing that similarity to population is only factor
- Confusing frequency and sampling distributions
- Attending more to the ratio of the sample to the population size

Prior interventions:

- Demonstrate that sample size is responsive to training
- Rely on learning a rule (Fong, Krantz, & Nisbett, 1986) and/or inferring a principle from demonstrations (Chance, delMas & Garfield, 2004; Sedlmeier, 1999)
- Do not support students in exploring the *mechanisms* of power and precision

Study 2: Scope of Inference (Fry, 2017)



According to statistics education recommendations (e.g., Utts, 2003), students should understand:

- **Random sampling** tends to produce representative samples, allowing for **generalization** to a population.
- **Random assignment** tends to balance out confounding variables between groups, helping to enable **cause-and-effect** conclusions.

Students have difficulties such as:

- Confusion between random sampling and random assignment (Derry et al., 2000)
- Disbelief that random assignment can help enable causal claims (Sawilowsky, 2004)
- Believing that larger samples are always better than smaller samples, regardless of method (Wagler & Wagler, 2013)

Methods

***Growing Certain*: a series of activities to provide support for understanding mechanisms of power and precision**

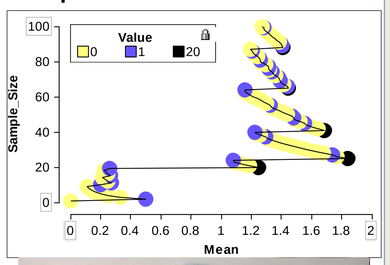
Primary conceptual targets:

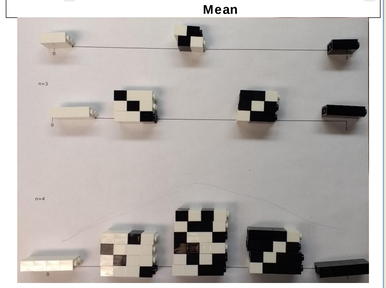
Swamping:

- means of large samples are less influenced by extreme values than small samples (Well et al., 1990)
- Students explore through randomly drawing additional values and watching decreasing impact of a value on mean (right)

Heaping

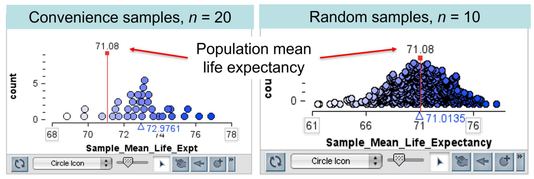
- increasing % of ways to get a sample mean near true mean (inspired by Abrahamson, 2006)
- Participants create both empirical and theoretical sampling distributions and examine permutations





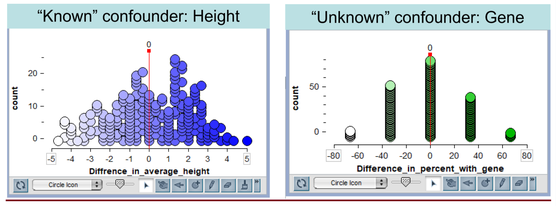
Participants were 5 volunteer introductory statistics students, who participated in 6 videorecorded hours of clinical interviews. One focal participant was coded for mechanistic reasoning (Russ et al. 2008)

Two-and-a-half week *Study Design* unit implemented in undergraduate introductory statistics course



Sampling and bias: Students compare sample means from larger convenience samples (biased method), to sample means from smaller random samples (unbiased method) .

Random assignment and balancing confounding variables: Students contrast purposeful assignment with random assignment, observing how random assignment tends to balance differences in confounding variables

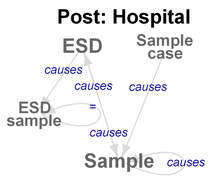


- One activity specifically designed to have students carry out both random sampling and random assignment, and distinguish between purposes of each
- Pretest/posttest taken by n = 125 students, qualitative data analysis of group quizzes and individual homework assignments

Results

Nearly all **students successfully used swamping reasoning**, but application depended on context and student

- Students were able to recognize that having more variability, or a mean closer to hypothesized value, meant a larger sample size was needed
- **Students had difficulty reasoning about many samples and often argued based on results from a single sample**
- Students **successfully drew on their everyday experiences** (getting an F at the end of a course as opposed to beginning) to explain Empirical Law of Large Numbers



Coded mechanistic relationships between entities for the hospital problem for one participant.

ESD = Empirical Sampling Distribution.

Limitations

- Sample size of 1!
- Mechanistic coding had 1 coder
- Attribution to activities unclear
- Low ecological validity
- Not classroom-scalable

"But looking at it **individually** [...] I would say still **the smaller hospital, just because it's more likely to deviate**, while the bigger hospital [...] doesn't go up and down quite as drastically. "

Overall, **evidence of gains** in learning goals related to study design and conclusions, especially those related to understanding:

- The purpose of random assignment
- Correlation does not imply causation

A small, but noticeable portion of students experience difficulties such as:

- **Confusion** between random sampling and random assignment
- **Giving sample size more importance** than sampling method

Limitations

- Neither randomly assigned nor sampled!
- Pretest given just before unit and posttest given just after unit: Did not measure student knowledge at beginning or end of course
- Instrument had evidence of low reliability

Sample item showing some evidence of lingering confusion between random sampling and random assignment:

Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.

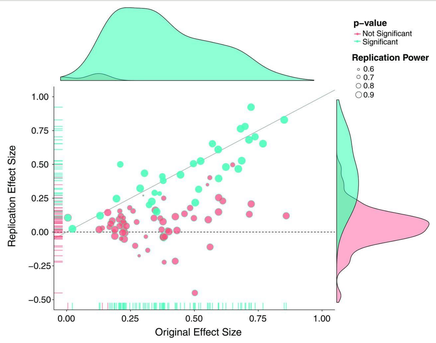
Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

Response	Pretest (%)	Posttest (%)
No, the sample size is too small to allow causation to be inferred.	35.2	7.2
No, the lack of random assignment does not allow causation to be inferred.	**28.0**	**77.6**
Yes, the statistically significant result allows causation to be inferred.	12.0	3.2
Yes, the sample was randomly selected, so causation can be inferred.	**24.8**	**12.0**

Future Directions in Teaching Scientific and Statistical Inference

Statistical Understanding and the Reproducibility Crisis

- **Better instruction** is needed to support scientists and non-scientists in statistical reasoning. Simulation-based representation may be one useful tool.
- **Much more research is needed** into understanding how these crucial types of reasoning can be improved.
- Another crucial topic is people's understanding of **the implications of multiple testing and selective inference** (e.g. Benjamini & Hochberg, 1995), another major contributor to the reproducibility crisis



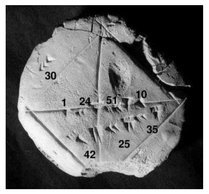
(Open Science Collaboration, 2015)

Topics for Future Research

- What **aspects of a curriculum** are most helpful for students' learning about reliability and scope of inference?
- How well do students **retain conceptual understanding** at the end of the course, or after the course?
- How can **single and multiple samples** be bridged effectively? Would Hypothetical Outcome Plots (Hullman et al., 2015) help?
- What are *researchers*' mechanistic reasoning about reliability and scope of inference?
- How could these concepts be taught in a **shorter unit**, or to researchers outside of a classroom setting?
- Does mechanistic reasoning support decision-making?

Building more experience with statistical processes

- These complex concepts may need more experience beyond standard statistical courses
- **Earlier, deeper, and widely accessible** introductions to statistical reasoning can start in early primary school (Franklin et al., 2005)
- Reliability and scope of inference can be **integrated into science education** to give students more experience
- Arithmetic has moved from an **elite** to a **basic literacy skill**; it may be time for the same transition for reasoning with data and uncertainty



Babylonian clay tablet YBC 7289
https://commons.wikimedia.org/wiki/File:Ybc7289-bw.jpg
CC BY-SA 3.0

References

Abrahamson, D. (2006). The Shape of Things to Come: The Computational Pictograph as a Bridge From Combinatorial Space to Outcome Distribution. *International Journal of Computers for Mathematical Learning, 11*(1), 137–146. https://doi.org/10.1007/s10758-006-9102-y
Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
Brown, E. C. (2019). *Growing Certain: Students' Mechanistic Reasoning about the Empirical Law of Large Numbers* (Ph.D., University of Minnesota). Retrieved from Dissertations & Theses @ CIC Institutions; ProQuest Dissertations & Theses A&I.
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). Retrieved from http://dx.doi.org/10.1007/1-4020-2278-6_13
Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, *1*(1). Retrieved from http://www.escholarship.org/uc/item/6xh34j0q
Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering Students' Statistical and Scientific Thinking: Lessons Learned From an Innovative College Course. *American Educational Research Journal, 37*(3), 747–773. https://doi.org/10.3102/00028312037003747
Fiedler, K. (2011). Voodoo Correlations Are Everywhere—Not Only in Neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. https://doi.org/10.1177/1745691611400237
Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*. Retrieved from American Statistical Association website: http://www.amstat.org/education/gaise
Freudenthal, H. (1972). The 'empirical law of large numbers' or 'The stability of frequencies.' *Educational Studies in Mathematics*, *4*(4), 484–490. https://doi.org/10.1007/BF00567002
Fry, E. B. (2017). *Introductory Statistics Students' Conceptual Understanding of Study Design and Conclusions* (Ph.D., University of Minnesota). Retrieved from http://login.ezproxy.lib.umn.edu/login?url=https://search.proquest.com/docview/2025494905?accountid=14586
Garfield, J., & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review*, *75*(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x
Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684
Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE*, *10*(11), e0142444. https://doi.org/10.1371/journal.pone.0142444
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. *Data and Context in Statistics Education: Towards an Evidence-Based Society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8*.
Lem, S., Van Dooren, W., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica*, *53*(2), 123–135.
Nuzzo, R. (2015). How scientists fool themselves–and how they can stop. *Nature News*, *526*(7572), 182.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716
Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499–525. https://doi.org/10.1002/sce.20264
Sawilowsky, S. S. (2004). Teaching random assignment: do you believe it works? *Journal of Modern Applied Statistical Methods*, *3*(1), 23.
Schield, M. (2010). Association-causation problems in news stories. *Data and Context in Statistics Education: Towards an Evidence-Based Society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*, 6.
Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, New Jersey: Lawrence Erlbaum.
Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum. *The American Statistician*, *69*(4), 362–370. https://doi.org/10.1080/00031305.2015.1081619
Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum. *Statistics Education Research Journal*, *11*(1).
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
Wagler, A., & Wagler, R. (2014). Randomizing Roaches: Exploring the 'Bugs' of Randomization in Experimental Design. *Teaching Statistics*, *36*(1), 13–20. https://doi.org/10.1111/test.12029
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
Zieffler, A., & Catalysts for Change. (2017). *Statistical Thinking: A Simulation Approach to Modeling Uncertainty* (4.0). Minneapolis: Catalyst Press.

UNIVERSITY OF MINNESOTA
Driven to Discover℠