THE DEVELOPMENT OF INFORMAL INFERENTIAL REASONING VIA RESAMPLING: ELICITING BOOTSTRAPPING METHODS

Jeffrey A. McLean
Syracuse University
jamcle01@syr.edu

Helen M. Doerr
Syracuse University
hmdoerr@syr.edu

This study focuses on the development of four tertiary introductory statistics students' informal inferential reasoning while engaging in data driven repeated sampling and resampling activities. Through the use of hands-on manipulatives and simulations with technology, the participants constructed empirical sampling distributions in order to investigate the inferences that can be drawn from the data. Students' developing reasoning of sampling and informal inference is reported as they move from repeated sampling methods to resampling methods, along with their reasoning of bootstrapping methods and how this reasoning was applied to make informal inferential claims.

Keywords: Data Analysis and Statistics; Modeling

Introduction

Over the past few decades statistics education has become an integral part of the mathematics curriculum at all levels. Influential documents such as the National Council of Teachers of Mathematics standards documents (NCTM, 1989, 2000), the *Guidelines for Assessment and Instruction in Statistics Education College Report* (Aliaga et al., 2005), and The *Common Core State Standards for Mathematics* (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010*)* have emphasized the importance of statistics education at all levels. Prior to these documents, statistics at the K-12 level was often "the mere frosting on any mathematics program if there was time at the end of the school year" (Shaughnessy, 2007, p. 957).

This relatively new emphasis on the learning of statistics brings with it a new emphasis on how statistics is taught. A trend in statistics education is the shift from a focus on theoretical distributions and numerical approximations to an emphasis on data analysis (Cobb, 2007). Cobb asserted that many statistics curricula are outdated and based on how statistics could be learned prior to the computing power of modern times. The use of probability distributions, such as the normal distribution, were once needed since the conceptually simpler approach of simulations by hand was far too tedious to perform. Technology now allows these simulations to be performed nearly instantaneously. New curricula for introductory statistics courses should emphasize the ideas of data creation, exploration and simulation.

This study investigates students' developing informal inferential reasoning while engaging in a data driven instructional unit. Activities in the unit use both hands-on manipulatives and computer simulations to construct empirical sampling distributions from which students made informal inferences.

Related Literature

Informal inferential reasoning has been defined as "the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data" (Pfannkuch, 2007, p. 149), "the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected" (Makar & Rubin, 2007, p.1), and "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (Zieffler, Garfield, delMas, & Reading, 2008, p. 44). Synthesizing these definitions, this study examined the claims that students made about populations of data when examining empirical sampling distributions, and how the students used the distributions of data to support these claims. Researchers suggest the use of informal inference before the use of formal inferential procedures (Zieffler, Garfield, Delmas, & Reading, 2008), such as employing the "three R's: randomize data, repeat by simulation, and reject any model that puts your data in its tail" (Cobb, 2007, p.12). This use of simulation to teach informal inferential reasoning can help students build a deep understanding of the abstract statistical concepts (Burrill, 2002; Maxara & Biehler, 2006). College curricula using simulations have indicated modest improvement in students' understanding of inference (Garfield, delMas, & Zieffler, 2012; Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012).

There are two forms of simulations in this study, simulations that construct an empirical sampling distribution by: 1) repeatedly sampling from an available population; and 2) resampling from a sample with an unavailable population. Research involving repeated sampling activities have indicated that *some* students develop "a multi-tiered scheme of conceptual operations centered around the images of repeatedly sampling from a population, recording a statistic, and tracking the accumulation of statistics as they distribute themselves along a range of possibilities" (Saldanha & Thompson, 2002, p. 261). However, many students do not focus on empirical sampling distributions for inference and instead compare a single sample statistic with a population parameter. Students have also shown difficulty distinguishing the difference in strength of conclusions made from a small number of samples versus those made with large amounts of samples (Pratt, Johnston-Wilder, Ainley, & Mason, 2008).

The second form of simulation activities aimed to elicit and develop students' ideas about the resampling method of bootstrapping. Efron (1979) introduced the method of bootstrapping and asserted that the bootstrap was more widely applicable and dependable than earlier resampling methods, while also using a simpler procedure. Bootstrapping begins with drawing one sample of data from a population. Bootstrap samples are then constructed by choosing elements from this one sample, with replacement, and creating resamples which are equal in size to the original sample. A statistic from these bootstrap samples is then aggregated to form an empirical bootstrap sampling distribution. If done with hands-on manipulatives, this sampling process using replacement can potentially provide insight into the approach, but it is also very time consuming. However, technology can be used to simulate this procedure in a short period of time, but the use of technology may obscure the underlying sampling process.

While limited research has been done on student learning of statistics with bootstrapping methods (Garfield, delMas, Zieffler, 2012; Pfannkuch & Budgett, 2014; Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013), researchers have asserted that bootstrapping may promote student learning of the logic of inference (Cobb, 2007; Engel, 2010, Hesterberg, 2006)). The bootstrapping method has already become part of introductory statistics coursework such as the CATALYST curriculum (Catalyst for Change, 2012). Some textbooks (e.g., Lock, Lock, Lock-Morgan, Lock, & Lock, 2013) introduce the method of bootstrapping to define confidence intervals well before discussing confidence intervals with normal approximation methods. Lock et al. claim that the bootstrapping method has become an important tool for statisticians and that it is also intuitive and accessible for introductory statistics students. The authors further state that bootstrapping capitalizes on students' visual learning skills and helps to build students' conceptual understanding of key statistics ideas. There is not yet research evidence to support these claims, which will be explored in this study.

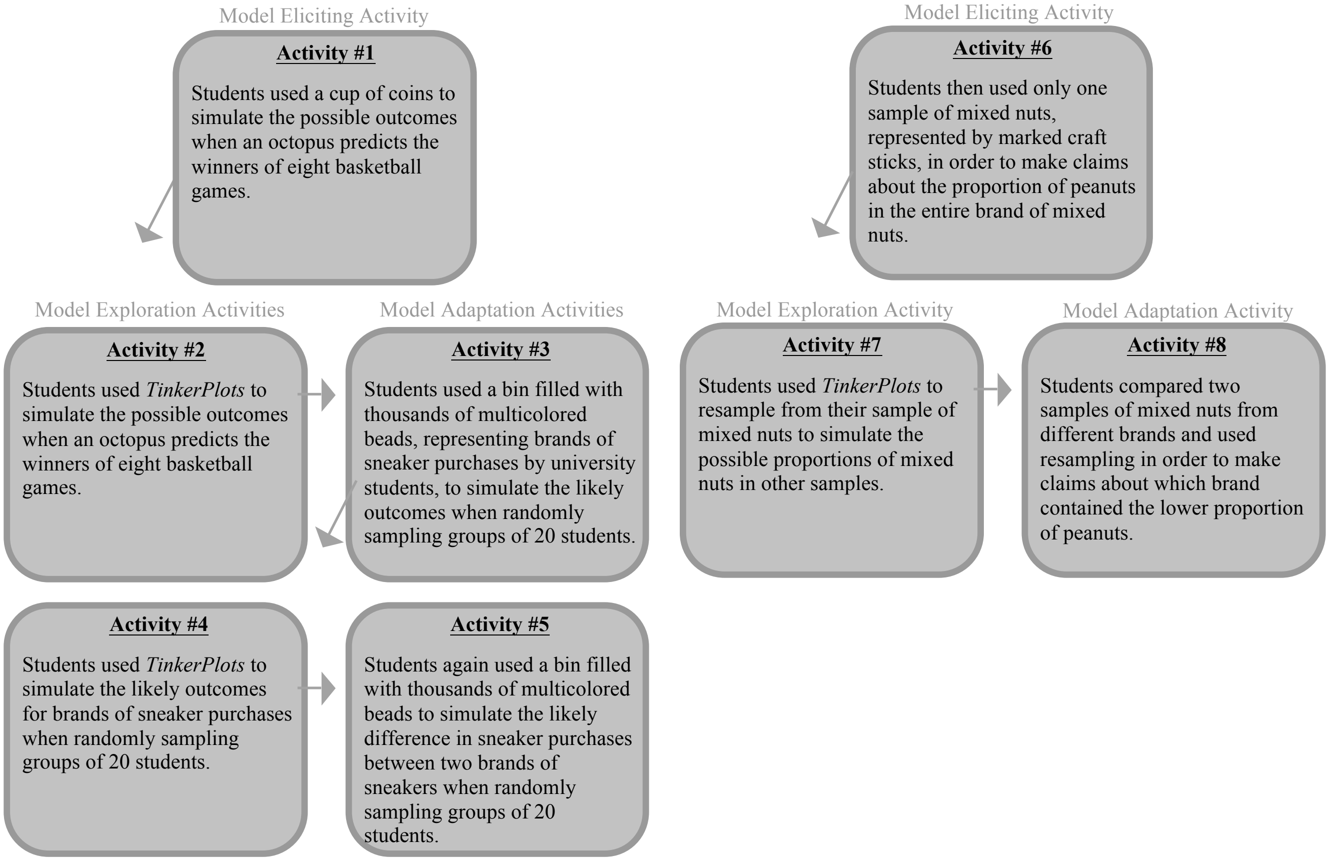
The research questions guiding this study were: 1) What student reasoning develops as they move from repeated sampling methods to resampling methods? 2) How do students develop their reasoning of bootstrapping methods and apply this reasoning to make informal inferential claims.

Theoretical Framework

The focus of analysis for this study was the models of sampling that the students created while engaged in a model development sequence (Lesh, Cramer, Doerr, Post, & Zawojewski, 2003). Drawing on Lesh and colleagues, in this study, we define models as "conceptual systems … that are expressed using external notation systems, and that are used to construct, describe, or explain the behaviors of other system(s)" (Lesh & Doerr, 2003, p. 10). Teaching and learning from a modeling approach shifts the focus of an activity from finding an answer to one particular problem to constructing a system of relationships that is generalized and can be extended to other situations (Doerr & English, 2003). Students' mathematical models are useful for research since they provide a means for investigating students' developing knowledge (Lesh, Hoover, Hole, Kelly, & Post, 2000). Model development sequences consist of three forms of activities: model-eliciting activities that encourage students to generate descriptions, explanations, and constructions in order to reveal how they were interpreting situations; model-exploration activities that focus on the mathematical structure of their models and often use technology in order to develop a powerful representation system; and model-adaptation activities that transform the models created in model-eliciting activities in order to investigate more complex problems (Lesh et. al, 2003). By using a modelling approach to examine student reasoning, we can view reasoning as dynamic and developing over the course of instruction. Student reasoning may not only change from activity to activity, but many times during an activity. This framework allows us to examine the impact of the activities on the development of students' reasoning.

Design and Methodology

In this qualitative case study, the first author collaborated with two introductory statistics instructors to create an instructional unit that consisted of two model development sequences (Figure 1). This study is part of a larger study that examined the development of informal inferential reasoning through simulation activities and the role of hands-on manipulatives versus technological


Model Eliciting Activity
Model Exploration Activities
Model Adaptation Activities
Model Eliciting Activity
Model Exploration Activity
Model Adaptation Activity

Activity #1

Students used a cup of coins to simulate the possible outcomes when an octopus predicts the winners of eight basketball games.

Activity #2

Students used *TinkerPlots* to simulate the possible outcomes when an octopus predicts the winners of eight basketball games.

Activity #3

Students used a bin filled with thousands of multicolored beads, representing brands of sneaker purchases by university students, to simulate the likely outcomes when randomly sampling groups of 20 students.

Activity #4

Students used *TinkerPlots* to simulate the likely outcomes for brands of sneaker purchases when randomly sampling groups of 20 students.

Activity #5

Students again used a bin filled with thousands of multicolored beads to simulate the likely difference in sneaker purchases between two brands of sneakers when randomly sampling groups of 20 students.

Activity #6

Students then used only one sample of mixed nuts, represented by marked craft sticks, in order to make claims about the proportion of peanuts in the entire brand of mixed nuts.

Activity #7

Students used *TinkerPlots* to resample from their sample of mixed nuts to simulate the possible proportions of mixed nuts in other samples.

Activity #8

Students compared two samples of mixed nuts from different brands and used resampling in order to make claims about which brand contained the lower proportion of peanuts.

***Figure 1:* Overview of the instructional unit consisting of two model development sequences**

tools with four classes of students at the secondary and tertiary levels. For this study we focus on one group of four students at the tertiary level engaging in the instructional unit. During the unit, the group of students was videotaped and their written work was collected. One student from the group participated in three interviews to discuss her thinking during the instructional unit. The videos, student work, and interview were analyzed with qualitative methods in order to construct the development of the models of sampling used by the participants, and how they were applied to make inferential claims.

The first model development sequence was intended for students to create models that allowed them to draw inferential claims from empirical sampling distributions constructed from repeated sampling from a known population. The second model development sequence no longer had an available population to repeatedly sample from. This put the students in a situation where they needed to extend their models for drawing conclusions by constructing resampling methods to use with the one available sample. Figure 1 provides an overview of the two model development sequences.

Results

We will report the main reasoning, and changes in reasoning, related to the ideas of sampling and informal inferential claims as demonstrated by the four students during the instructional unit. Changes in reasoning occurred both as students progressed through the activities in the unit and also during group and class discussion within the activities.

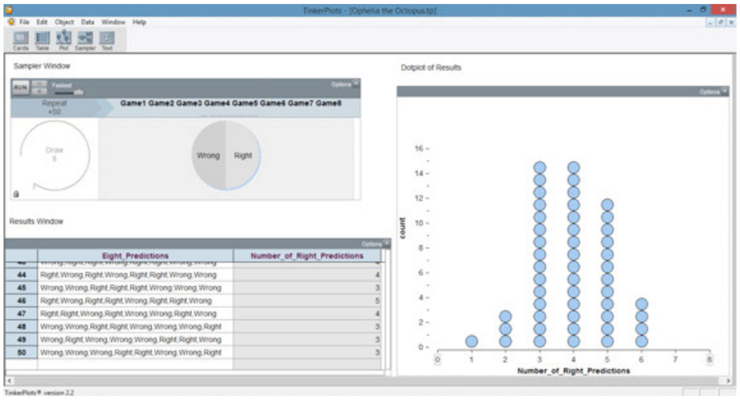
The first activity in the instructional unit asked the students to determine the likely range of correct predictions made when guessing the winner of eight basketball games. A cup of coins was given to the group as an option to use to draw their conclusions. The group showed an aversion to using the coins to simulate guessing and attempted to calculate the probabilities of each outcome. A class discussion encouraged the group to use eight coins to simulate possible outcomes for the number of correct predictions.

Initial Model of Sampling and Inference

The group simulated the outcomes by flipping eight coins five times. For each group of eight coins they counted the number of heads, which represented a correct prediction. They concluded that because of varying values in their simulations, there is “no definitive answer as to the range of possible outcomes”. This view of one definitive and correct range of possible outcomes was in line with their initial attempts to calculate the probability of each outcome occurring. One student in the group, Megan, was later interviewed to discuss her reasoning in the activity and was still not convinced on the value of using simulations to answer the original question. She had read ahead in the course’s textbook to determine a way of answering the activity through calculations, and constructed a 95% confidence interval with a normal approximation to the binomial distribution.

Second Model of Sampling and Inference

The second activity continued in the same context as the first, but moved on from using coins to simulate outcomes, to the use of *TinkerPlots* (Konold & Miller, 2014). *TinkerPlots* was set up for the students using a spinner with half of the area marked ‘Right’, the other half ‘Wrong’, a window that collected the outcomes simulated by the spinner, and a dotplot that collected the number of correct predictions in each sample. A screenshot of the *TinkerPlots* setup is shown below in Figure 2.



	Eight_Predictions	Number_of_Right_Predictions
44	Right,Wrong,Right,Wrong,Right,Right,Wrong,Wrong	4
45	Wrong,Wrong,Right,Right,Right,Wrong,Wrong,Wrong	3
46	Right,Wrong,Right,Right,Wrong,Right,Right,Wrong	5
47	Right,Right,Wrong,Right,Wrong,Wrong,Right,Wrong	4
48	Wrong,Wrong,Right,Right,Wrong,Wrong,Wrong,Right	3
49	Wrong,Right,Wrong,Wrong,Wrong,Right,Right,Wrong	3
50	Wrong,Wrong,Wrong,Right,Right,Wrong,Wrong,Right	3

***Figure 2: TinkerPlots* setup for activity #2**

Each group member first used *TinkerPlots* to simulate 10 samples. Together they determined that the outcomes that occurred the most often could constitute a range of likely values. The group discussed how each of their dotplots were slightly different and led them to have varying predicted ranges. After simulating 1000 samples, they discussed how more samples led to each of their dotplots looking very similar and yielding the same intervals of likely values. The group concluded that more samples lead to more accurate predictions. From their dotplots with 1000 samples they determined that between 2 and 7 correct guesses seemed likely. Megan continued to add samples and came to the conclusion that at some point, adding more samples did not change the look of the dotplot. She concluded that the plot was saturated with data.

Third Model of Sampling and Inference

The next two activities involved sampling from a population of university students to determine what number of sneaker purchases out of groups of 20 students were Nikes. This was first done with a bin of thousands of multicolored beads, each color representing the purchase of a certain brand, and then continued in the next activity using *TinkerPlots* to simulate the outcomes. A similar *TinkerPlots* setup was given to the students as before. The spinner was replaced with a mixer containing the same distribution of balls as the beads used as hands-on manipulatives. The students were asked if it was reasonable for Nike to claim that 7 out of 20 student sneaker purchases by the university students were Nikes.

The group used the same methods to construct an empirical sampling distribution as the previous model but the change came with how they drew their conclusions. When using the hands-on manipulatives, all samples except for one showed more than 7 of 20 sneaker purchases were Nikes. A simulation with *TinkerPlots* also showed a majority of the samples falling above 7 out of 20. The students determined an interval of likely outcomes for Nike sneaker purchases and found 7 of 20 to be below the interval. They concluded that Nike should claim that more than 7 out of 20 sneaker purchases at the college were Nikes. They recommended Nike to claim that 10 in 20 purchases were Nikes, which was approximately the center of the distributions.

Fourth Model of Sampling and Inference

The fifth activity was the first time that the students had to deal with the comparison of populations. The context was similar to the previous two activities. Students were asked to investigate if the difference in Nike and Adidas sneaker purchases at a university was the same as the difference in the global sales of 15%, or 3 in 20 students. The same bin of beads was used as in the previous activity, with one color representing Nike purchases and another color, Adidas purchases. The students decided that since they already constructed an empirical sampling distribution of Nike sneaker purchases, they would conduct the same number of samples and count Adidas sneaker purchases. The group chose to calculate the mean values of each distribution and compare them. The means were approximately three purchases apart, which led the students to conclude that the 15% difference in Nike and Adidas global share held true for the college students.

Initial Model of Resampling and Inference

The next change in the students' models occurred during the beginning of the second model development sequence. The first activity put the students in a situation where they needed to extend their models for drawing conclusions by constructing resampling methods to use with the one available sample. The students were told that the manager of bulk food in a grocery store ordered a sample of a new brand's mixed nuts. She plans to order a large shipment of mixed nuts, but has determined that her customers prefer mixed nuts with fewer peanuts. Before she orders, the manager wants to know more information about the percentage of peanuts in this new brand. The students were given a bag of 25 sticks to represent the sample of mixed nuts. Seven sticks were marked with a 'P' to represent peanuts. The remaining sticks were not marked and represented other types of nuts.

The group began by applying a method similar to the previous activities by taking seven samples of five mixed nuts and calculating the average percentage of peanuts. They chose the size of five since it would be easy to take many samples and also to calculate the percentage of peanuts. After some class discussion, they decided to try and take larger samples. Through interactions with the instructor, they found that if they took samples of size 10 in a similar manner to their previous samples, all outcomes were not possible. Since there were only seven peanuts in their original sample, the largest percentage of peanuts in their sample of size 10 would be 70%. After discussing this issue with the class, they decided to take samples of size 10 from their original sample of 25, with replacement. The process of sampling with replacement was more time demanding than sampling without replacement, so the group decided to collect only three samples. They found the average percentage of peanuts in the three samples and concluded that a likely interval for the percentage of peanuts in the population was that average plus or minus an arbitrarily chosen 4%. During an interview with Megan after the activity, we discussed her group's choice of sample sizes of five and 10. She said that larger samples may have provided more accurate results, but would have taken too long to sample with the sticks.

The next activity worked with *TinkerPlots* and gave each group member a different sample of mixed nuts. From these new samples, each member used *TinkerPlots* to collect resamples of 25 mixed nuts (as set up in *TinkerPlots*) and applied similar methods from the first model development sequence to construct a interval of likely values based on the height of outcomes on the dotplot. The group was told the true percentage of peanuts in the brand of mixed nuts, which was a value not included in all of their predicted ranges. Megan chose a much wider range of values than the other students. She concluded that if others did this as well, more of their likely ranges would capture the true percentage of peanuts.

Second Model of Resampling and Inference

The final activity gave the students two samples of mixed nuts from two brands and asked them to conclude which brand had the lower proportion of peanuts. The samples were each of size 25 and contained six and 10 peanuts. Unlike the previous activity comparing sneaker purchases, *TinkerPlots* was available for the students to construct sampling distributions. The group constructed two empirical sampling distributions with 200 samples in each and determined that the first brand likely had 16%-32% peanuts and the second brand had 32%-48% peanuts. These ranges were chosen based on the height of the dotplot for each outcomes. Since the likely interval for the second brand was higher than the other brand's interval (except for the endpoint) they concluded that the second brand is likely to have more peanuts.

During an interview after the activity, Megan and the first author discussed the possibility that both brands had 32% peanuts. Megan concluded that it was unlikely for both brands to have 32% peanuts, but since that was the only value that overlapped between the two likely ranges, she still believed that the second brand was more likely to have more peanuts.

Discussion

The group's model of inference from simulation developed from not drawing a conclusion from simulated data, to using data as evidence to make informal inferences. The first model development sequence was developed as a means for the students to build the necessary tools to draw these informal inferential claims and attempt to apply them to situations in the second model development sequence with an unavailable population. The group constructed some notion of the bootstrapping process by resampling with replacement from their original sample, but did not take resamples that were equal in size to the original sample. The time demanding nature of resampling by hand was noted as one reason for taking smaller sized samples. Immediately after these topics were covered in the class, the course instructor began topics in formal inference. She believed that the students were better prepared for the concepts of confidence intervals and hypothesis testing by participating in the instructional unit. Further research is needed to indicate the connections between the informal inferential models constructed in this study and students' reasoning of formal inference. This study also has implications with the content and design of introductory statistics curricula, and the role of resampling activities on the development of informal inferential reasoning.

References

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education: College Report.* American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Burrill, G. (2002). *Simulation as a tool to develop statistical understanding*. Paper Presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa

Catalysts for Change. (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Minnesota, MN: Catalyst Press.

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/6hb3k0nz

Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' reasoning about data. *Journal for Research in Mathematics Education, 34*(2), 110-136.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *The Annuls of Statistics. 7*(1), 1-26.

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*. DOI 10.1007/s11858-012-0447-5. Springer.

Hesterburg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. Burrill (Ed.), *Thinking and reasoning with data and chance. Sixty-eighth National Council of Teachers of Mathematics Yearbook* (pp. 391-416). Reston, VA:NCTM

Konold, C., & Miller, C. D. (2014). Tinkerplots: *Dynamic data visualization*.[Computer Software] Amherst, MA: University of Massachusetts Amherst.

Lesh, R., Hoover, M., Hole, B., Kelly, A. & Post, T. (2000). Principles for developing thought revealing activities for students and teachers. In Kelly, A., & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 591-645). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., Cramer, K., Doerr, H. M., Post, T., & Zawojewski, J. (2003). Model development sequences. In R. Lesh, & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on problem solving, learning, and teaching* (pp. 35–58). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Lesh, R., & Doerr, H. M. (2003). Foundations of models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H.M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on problem solving, learning, and teaching* (pp. 3-33). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Lock, R., Lock, P.F., Lock-Morgan, K., Lock, E.F., & Lock, D.F. (2013). *Statistics: Unlocking the Power of Data*. Wiley.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105.

Maxara, C., & Biehler, R. (2006). *Students' probabilistic simulation and modeling competence after a computer intensive elementary course in statistics and probability.* Paper presented at the Seventh International Conference on Teaching Statistics. Salvador da Bahia, Brazil

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for K–12 mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, DC: Authors. http://www.corestandards.org/Math/

Pfannkuch, M. (2007). Year 11 students' informal inferential reasoning: A case study about the interpretation of box plots. *International Electronic Journal of Mathematics Education, 2*(3).

Pfannkuch, M., Budgett, S. (2014). *Constructing inferential concepts through bootstrap and randomization-test simulations: A case study*. Paper presented at the Ninth International Conference on Teaching Statistics: Sustainability in statistics education. Flagstaff, Arizona.

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. (2013). Bootstrapping students' understanding of statistical inference. Summary research report for the Teaching and Learning Research Initiative, www.tlri.org.nz

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal, 7*(2), 107-129. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pratt.pdf

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

Shaughnessy J. M. (2007). Research on statistics learning and reasoning. In F.K. Lester (Ed.), *The second handbook of research on mathematics* (pp. 957–1010). Reston, VA: National Council of Teachers of Mathematics (NCTM).

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), 21-40

Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58.