THE ROLE OF CONCEPT IMAGES IN DEVELOPING STATISTICAL UNDERSTANDING: STATISTICAL SIGNIFICANCE

Gail Burrill
Michigan State University
burrill@msu.edu

Instruction often glosses over core concepts, leaving students with fragile understandings and limited ability to use those concepts. Interacting with carefully developed dynamic interactive technology apps can help students build a "movie clip" of features of the concept that can become the basis for understanding. This paper discusses the use of such apps in a simulation based, formula-light statistics and probability course for elementary preservice teachers The discussion includes a description of the course content leading to inference and the results of analyzing student thinking with respect to statistical significance using an adaptation of the SOLO taxonomy. Recommendations are given for instruction and additional research.

INTRODUCTION

Conceptual knowledge can be thought of as implicit or explicit understanding of principles governing a content area and how the units of knowledge in that area are connected (Rittle-Johnson et al., 2001). Tall and Vinner (1981) label such knowledge a "concept image": the total cognitive structure including the mental pictures and processes associated with a concept that accumulate through different experiences associated with the idea. When students have robust concept images, they know more than isolated facts and methods and have organized their knowledge into a coherent whole that enables them to connect new features to what they already know (Bransford et al., 1999). Without careful development of a concept, students are left to construct an understanding based on ill formed and often misguided connections and images (Oehrtman, 2008). The work of understanding subsequent topics is then built on isolated understandings specific to each topic (e.g., center as separate from spread). Somewhat surprisingly, formal definitions play a very small role in the development of concept images but rather intuition or specific elements of examples used in instruction become key in characterizing a concept (Dreyfus, 2014). Both visual and verbal cues contribute to concept formation. Sometimes these cues may interfere with developing accurate concept images: both obvious (visual: an app interface with cognitive overload; verbal: overly formal language with many technical terms) and more nuanced (visual: examples of skewed distributions are skewed right; verbal: using words with meanings different from everyday usage). This paper describes using interactive digital technology to develop preservice teachers' concept images of statistical significance.

BACKGROUND

A number of studies suggest that strategic use of digital tools can help students connect mental images of concepts to visual interactive representations, leading to a robust understanding of the concept (e.g., Guin & Trouche, 1998). Interactive dynamic technology provides students with mental "movie clips" of a concept that can enable them to build images of the properties, processes, and relationships connected to the concept (Drijvers, 2015). Interactive dynamic technology also allows students to link multiple representations of data—graphical, symbolic, numeric, and verbal—and to make connections among these representations to support understanding (Ainsworth, 2006). Using interactive dynamic technology enables students to build representations of sample variability by, for example, generating simulated distributions of sample statistics, comparing random samples, and observing the effect of sample size on sampling distributions (delMas et al., 1999; Prodromou, 2015). Sacristan (2021) argued that digital technologies are "platforms where learners can find and create something personally meaningful, … and thus help them learn best" (p. 3). The applet-like documents used in this project provide such platforms, leveraging dynamic linkages among representations to address specific learning outcomes for statistical concepts.

In their literature review, Nilsson and Schindler (2018) found little evidence of researchers using theories to support statistics learning with technology. The work described here has roots in Peirce's (1931) three stages of diagrammatic reasoning and the notion that in the experimenting stage, some action should be carried out. Extending this to include technology leads to an "action/consequence"

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

principle (Conference Board of the Mathematical Sciences, 2012, p. 34) where students use technology to take an action, observe the consequences, and reflect on what those consequences mean for understanding. The next sections describe the course and the learning trajectory leading to inference.

A PRESERVICE COURSE

The study involved elementary preservice teachers at a large midwestern public university. These preservice teachers (students) had selected a mathematics emphasis for their certification, but their backgrounds in statistics were minimal. They were enrolled in a course related to teaching probability and statistics to elementary/middle school students. The class met twice a week in 110-minute sessions for 15 weeks. Students had their own computers and used TI-Nspire© software to access files from Building Concepts: Statistics and Probability (https://education.ti.com/en/building-concepts) and later StatKey (www.lock5stat.com/StatKey/). At the beginning of the course, students filled out a consent form to take part in the study, but the forms were not shared with the author until final grades for the course had been posted. One student out of the class of 13 did not consent. The course goals were to enable students to interpret and make sense of data, in particular data related to education, such as achievement scores, and to give them tools and strategies they could use in their own teaching. The design of the course was built around two fundamental ideas: (a) a major portion of instruction was devoted to developing conceptual understanding of core statistical ideas and relied heavily on interactive dynamic technology to do so and (b) the approach deemphasized formulas, with inference based on identifying how chance outcomes behave in known situations.

A typical class began with a brief overview of key ideas from the prior class, followed by a short activity that introduced the focus of the day. The activity was followed by small group work on a series of carefully designed questions using the applets to investigate the concepts. The questions were designed to be mindful of the caution that students can go wrong with technology without guidance from the instructor (Drijvers, 2015). The whole class frequently discussed points raised by individual groups. Each class ended with a brief summary of the important ideas covered during the session and a description of how these connected to the bigger conceptual picture of statistics. Weekly homework assignments consisted of short tasks or questions that focused on the key ideas from the prior week. Assignments were ungraded, with students receiving full marks for a completed assignment, but work was returned to the students with problematic answers highlighted by the instructor. Issues that surfaced across the assignments were discussed, usually in a format that engaged groups in considering "why the instructor was worried" about a certain answer or reflecting on how they might improve a response. Students were given frequent formative assessments such as quick quizzes, thumbs up/thumbs down on answers, and exit tickets. Ideas were deliberately revisited to support the development of robust concept images (Oehrtman 2008), as described in the next section.

BUILDING CONCEPT IMAGES

The design of the course provides opportunities for students to develop concept images over time from different experiences. For example, the idea of random is introduced in the probability unit when students observe how the proportion of blue chips in random samples of chips from a bag stabilizes over time (Burrill, 2021). Random is revisited in the Choosing Random Samples activity that involves simulating the random selection of three names from 30 names (in a class of 30) to submit homework assignments each day. The concept of random is connected to chance by considering what "likely" might mean (once every 20 times? Every 100 times?) and how a particular context might affect that decision. Simulations involving spinners are used to estimate probabilities and answer questions such as, "Is it likely a basketball player with a 40% shooting average will make at least 16 shots in the next 30 attempts? Why or why not? (see Figure 1). The next step is to investigate random samples from known populations and the "noise" or variability around an expected outcome.

To smooth the eventual transition to StatKey, the activities move between simulated distributions of outcomes as counts (Figure 2) and as proportions (Figure 3) and include discussion about why proportions are valuable. Simulated distributions of sample means and sample proportions are used to answer questions such as, "If 40% of the students in grade 9 were at the proficient level in reading, would you be surprised that 25 of the 30 students in a certain class were at that level?" or "Estimate the probability that the maximum speed of land animals is greater than 50 mph." Generating simulations provides visual clues about sampling distributions; for example, as the sample size increases, students

can see the sequence of ever tightening distributions as an indicator of less and less variability. Cognizant that words have an impact on concept images, an important part of the process is to have students describe verbally what they see as a simulated distribution stabilizes.

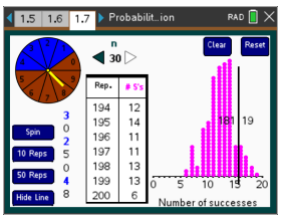

Figure 1. Estimating the probability of at least 16 baskets

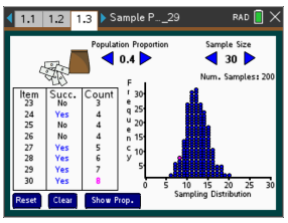

Figure 2. Outcomes as counts

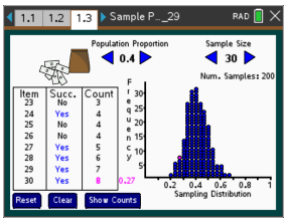

Figure 3. Outcomes as proportions

Apps are then used to develop understanding of margin of error (Burrill, 2021) and normal distribution. In the last part of the course, students transition to StatKey to simulate means and proportions and use randomization tests for significance assuming the null hypothesis is true. StatKey allows more freedom in inputs than the apps, and the area of interest is represented graphically. Note that in the Probability and Simulation app (Figure 1), the point of interest is marked by a moveable vertical line. In StatKey (Figure 4), the area of interest is colored. The expectation is that the concept images of key ideas students have developed can be generalized to other situations.

Figure 4. StatKey simulated distribution of sample proportions

Statistical significance is formally introduced after confidence intervals. Building from earlier work with simulations for probability and sampling distributions, the emphasis is on developing the understanding that a statistically significant outcome is one that is unlikely to occur by chance. Because of this emphasis and because students often confuse concepts that are presented together (e.g., confidence level and confidence interval, Etz, 2015), at the outset the focus is not whether to "reject the null," and the idea of a "*p*-value" is not yet important. The activities consider the status quo or what might be expected given no real difference from the hypothesized parameter, using the notion of a null hypothesis to set up the status quo. Students then consider whether an observed value or one more extreme is likely to occur by chance due just to the variability inherent in the process. From earlier discussions, the decision is made to call something that happens less than 5% of the time by chance statistically significant, noting another level might be appropriate depending on the context. Next, we describe the methods to analyze students' responses to items related to statistical significance.

METHOD

The data come from student responses to final examination questions for which students could use TI Nspire software. The exam was administered online using the university learning management system. Student responses were analyzed to answer the research question: "How do students reason about statistical significance after experiencing a non-formula, applet-based approach to the concept?" To understand students' thinking, their responses were classified using a hierarchical performance level based on the SOLO taxonomy (Structure of Observed Learning Outcomes, Biggs & Collis, 1982). Analysis was done in four parts. The author and a colleague: (a) identified features associated with developing a concept image for statistical significance; (b) associated these features with SOLO levels

(see Table 1) and linked them to possible misconceptions; (c) categorized student responses with respect to the elements in the taxonomy; and (d) summarized the results.

Table 1. Features associated with concept image for statistical significance

SOLO level	Reasoning about statistical significance
Prestructural (P)	Thinks of significance as "important"; associates chance with the number of repetitions of a simulation; uses absolute difference in given values
Unistructural (U)	Uses vague language without clear reference to distribution (i.e., "it" is likely to be significant…"); makes a statement about significance without interpretation; reasons almost correctly but adds an unacceptable comment
Multistructural (M)	Recognizes significance is about the relationship of a sample outcome to what might be expected given the status quo; selects an appropriate method for finding statistical significance; associates a probability with an observed outcome but not a distribution or significance level (5%)
Relational (R)	Links significance to chance; connects significance to a visual image from a sampling distribution; correctly interprets significance in a context; applies concept to both means and proportions; connects non-significance to insufficient evidence that the assumption of the status quo is false

The final examination had four questions related to statistical significance:

- Question 1 gave students the mean scores for large districts from the National Assessment of Educational Progress (NAEP), the scores for a sample of large districts, and a StatKey plot set up for the situation. They were asked to use the plot to answer the question, "Is the mean score for these districts significantly different from those for all large districts? Explain your reasoning."
- Question 2 gave students a simulated distribution of the number of baskets a player with a 60% free throw average would make in 24 free throw attempts. The task was to decide whether making an increased number of shots (two thirds in 24 attempts) after a change in shooting stance was convincing evidence that the player has significantly improved his shooting (is two thirds statistically significant?) with an explanation.
- Question 3 involved a situation where 52 of 69 males sampled said yes, and 120 of 131 females sampled said yes in response to a question. The task was to decide whether the proportions that said yes were significantly different for males and females and to justify the claim.
- Question 4 was, "When you think of an outcome as statistically significant, what image comes to mind?"

RESULTS

Table 2 displays the SOLO taxonomy level classifications for student responses to the questions. A typical prestructural response was to argue using only given numbers such as in response to Question 3, "52/69 ≈ .75 or 75%; 120/131 ≈ .92 or 92%; Yes, I would say there is a significant difference since they vary by 17%." Responses at the unistructural level often did not refer to a graph, making the source of values unclear, such as in response to Question 2, "This is not statistically significant because at random he is likely to make two thirds of his shot with this improvement. 0.077 is greater than 0.05 which indicates that it is not significant." Multistructural responses were the most common for all questions. Students typically entered correct values in StatKey and made right decisions, but very few correctly identified the critical region on their graphs. Others indicated the cutoff point in relation to a critical value (usually 0.05) but did not attend to the notion that the probability refers to a result as or more extreme given that the null hypothesis is true. For example, in response to Question 2, "… I ran 5,000 samples, and then found the probability of the basketball player *getting* [italics added] two thirds of the baskets. The chance of this happening is 19.7%, making it not significant." Relational responses included this connection, such as in response to Question 1:

> This Statkey plot represents sample means of 5000 different samples of all the scores given to the large districts on the 2017 National Assessment of Educational Progress (NAEP), which is centered around their mean of 213. The particular sample we are looking at of the 27 large districts in the list above the graph has a mean of score of 209.926. This value is equal to or

greater than a 0.031 proportion or 3.1% of the other sample means. Because any sample mean that aligns with less than 5% of the rest of the samples means is not likely to occur by chance, the mean score of these large districts is significant.

Table 2. Student responses to questions on final examination (n = 12)

	Reasoning about statistical significance; * indicates a no response			
SOLO taxonomy level	Question 1	Question 2*	Question 3	Question 4*
Prestructural (P)	0%	18%	7%	0%
Unistructural (U)	8%	9%	17%	8%
Multistructural (M)	75%	64%	58%	58%
Relational (R)	17%	9%	17%	33%

DISCUSSION

Several instructional issues emerged from the analysis. Few students reached the relational level on the first three questions. Many did not read an interval from the StatKey graphs, which might suggest the language and examples used in class did not support making the connection between the concept of significance to the chance of an outcome as great or *greater than* an observed value. A possible explanation might be the shift from the vertical line used as a cut point in the Probability and Simulation app (Figure 1) to an area in StatKey (Figure 4); a possible modification of the app could show both the cut point and a shaded area. Despite attention at the beginning of the course to communication, students continued to struggle with telling the story of a data set. In some cases, they misused statistical terms ("Because the correlation is less than 2 standard deviations away from the mean"), had problems attending to the context ("the data"; "the distribution is …"), or tripped over words ("using the left tail, we're looking at what the probability is that one of the people"; "they are 5% different."; "… 3.1% chance of occurring by chance").

Overall, the students seemed to understand that statistical significance is associated with chance and likely outcomes. For Questions 1 and 3, 91% and 83% of the responses, respectively, referred to chance, and in Question 2, two thirds did so. However, students did not always offer enough information to explain their thinking. More than 90% of the student responses were at least at the multistructural level when describing their image of statistical significance, with some at the relational level: "Significant means it is surprising and not likely due to chance. Will naturally be variability around what you expect. Ask yourself is the observed number beyond what might occur at least 95% of the time by chance."

CONCLUSION

The goal of this paper was to describe a formula light, applet-based course in statistics and probability and consider how this approach might support students' reasoning about statistical significance. In three of the questions, nearly three fourths of the students reasoned at least at the multistructural level. This was slightly better than the number of students who seemed to have this level of understanding with respect to other statistical concepts (Burrill, 2019, 2021) and overall encouraging in terms of student learning. However, the findings are observational and limited by sample size. The results could be confounded by the instruction, approach, activities, and other classroom factors.

More research is needed to tease out nuances in the errors students made. In particular, in both this analysis and in earlier work (Burrill, 2021), students struggled with the idea that estimating a probability from a sampling distribution is about something that might occur, not that will happen. This could be because students are asked to shift their thinking about probability from a Bayesian perspective in initially simulating sampling distributions to a frequentist approach that underlies StatKey. This idea needs much more examination if students' concept images built during the simulations will allow students to transition smoothly as they move to more formal procedures. The examples and words used in instruction should be examined to determine if they were sending misleading cues. Further research should also consider practical significance and the role of sample size in determining significant outcomes, which are often misunderstood or misused. (Andrade, 2019).

Thanks to Anna Fergusson for her contributions to the thinking about the ideas in this paper.

REFERENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Andrade C. (2019). The *p* value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, *41*(3), 210–215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* National Academy Press. https://doi.org/10.17226/9853

Burrill, G. (2019). Understanding sampling distributions: The role of dynamic interactive technology. In S. Budgett (Ed.), *Decision making based on data: Proceedings of the Satellite conference of the International Association for Statistical Education*. IASE https://iase-web.org/documents/papers/sat2019/IASE2019%20Satellite%20116_BURRILL.pdf?1569666564

Burrill, G. (2021). *Margin of error: Connecting chance to plausible* [Paper presentation]. The 14th International Congress on Mathematical Education (ICME), Shanghai, China.

Conference Board of the Mathematical Sciences. (2012). *The mathematical education of teachers II.* American Mathematical Society; Mathematical Association of America.

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). https://doi.org/10.1080/10691898.1999.12131279

Dreyfus, T. (2014). Solid findings: Concept images in students' mathematical reasoning. *European Mathematical Society Newsletter*, *93*, 50–52.

Drijvers, P. (2015). Digital technology in mathematics education: Why it works (or doesn't). In S. J. Cho (Ed.), *Selected regular lectures from the 12th International Congress on Mathematical Education* (pp. 135–151). Springer. https://doi.org/10.1007/978-3-319-17187-6_8

Etz, A. (2015, December 3). *Confidence intervals? More like confusion intervals.* Psychometric Society. https://featuredcontent.psychonomic.org/confidence-intervals-more-like-confusion-intervals/

Guin, D., & Trouche, L. (1998). The complex process of converting tools into mathematical instruments: The case of calculators. *International Journal of Computers for Mathematical Learning*, *3*, 195–227. https://doi.org/10.1023/A:1009892720043

Nilsson, P., & Schindler, M. (2018). The nature and use of theories in statistics education—looking back, looking forward. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward: Proceedings of Tenth International Conference on Teaching Statistics (ICOTS 10, July, 2018) Kyoto, Japan.* ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_1G1.pdf

Oehrtman, M. (2008). Layers of abstraction: Theory and design for the instruction of limit concepts. In M. Carlson & C. Rasmussen (Eds.). *Making the connection: Research and teaching in undergraduate mathematics education* (pp. 65–80). Mathematical Association of America. https://doi.org/10.5948/UPO9780883859759.007

Peirce, C. S. (1931). *Collected papers of Charles Sanders Peirce 1931–1958* (C. Hartshorne, P. Weiss, & A. W. Burks, Eds.). Harvard University Press.

Prodromou, T. (2015). Teaching statistics with technology. *Australian Mathematics Teacher*, *71*(3), 32–40.

Rittle-Johnson, B., Siegler, R., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346

Sacristan, A. (2021). *Digital technologies, cultures and mathematics education* [Invited lecture]. The 14th International Congress on Mathematical Education (ICME), Shanghai, China.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*, 151–169. https://doi.org/10.1007/BF00305619