EARLY INTRODUCTION OF HYPOTHESIS TESTING IN INTRODUCTORY STATISTICS: A PILOT STUDY

Heidi HULSIZER
Department of Mathematics & Computer Science
Benedictine College
United States
hhulsizer@benedictine.edu

Aminul HUQ
Center for Learning Innovation
University of Minnesota Rochester
ahuq@umn.edu
United States

Wei WEI
Mathematics Department
Metropolitan State University
United States
wei.wei@metrostate.edu

Abstract: The placement of hypothesis testing in the timeline of an introductory statistics curriculum might have significant effects on learning outcomes. This study aims to investigate whether the introduction of the concept of hypothesis testing early in the semester significantly increased student understanding of the topics surrounding it. Students were assessed on various aspects of hypothesis testing: the inquiry process, formulation, algorithm, and decision making. The data indicated that the introduction of hypothesis testing early in the semester had significant, positive results on student performance, compared to introducing hypothesis testing later in the semester.

INTRODUCTION

Creating an effective, meaningful, and goal oriented introductory statistics curriculum is one of the core curricular problems one can encounter in teaching undergraduates. Several researchers over the years have suggested a major resequencing of introductory statistics courses (Wardrop 1995; Cobb and Moore 1997; Cobb 2007; Malone et al. 2010). One of the learning objectives in an introductory statistics course is the mastery of statistical inference, particularly hypothesis testing. Results from the delMas et al. (2007) study indicate that the introductory statistics course had not significantly improved students' correct interpretation of significance tests. Cobb (2007, pp. 11) recommended that "We need a new curriculum, centered not on the normal distribution, but on the logic of inference." Wardrop (1995) proposed a new sequence of topics that focused on introducing inference much earlier in the course. Garfield et al. (2012) suggested a new curriculum with a simulation-based approach and informally introduce hypothesis testing before making formal procedures. Prodromou (2017) brought up a similar idea of introducing a model-based inference informally before a formal procedure.

With the current push toward evidence and modeling based teaching (ASA 2014; MAA 2015) we see the need to re-evaluate the classical topic sequence in introductory statistics course that is found in most textbooks. The report "Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics" (Pearl et al. 2012) lists researching efficient learning progressions for introductory statistics as a need in current curricula. Many times instructors pick up a textbook and follow the sequence of topics laid out in it, leaving the planning of the course design on the book selected. Attempts should be made, and are being made, to connect the current research findings to our teaching. In this study we hope to focus on one particular topic in a typical introductory statistics learning progression - hypothesis testing. We hope to contribute to the research on when this topic should be discussed in the sequence of course material.

In this study we wanted to determine the effectiveness of introducing the concept of hypothesis testing on the first day of class (e.g. Aliaga and Gunderson 2006) or very early in the semester. Our goal for this rearrangement is to achieve better student understanding of the concept of hypothesis testing by the end of the semester. By shifting the topic of testable hypotheses to the beginning of the semester, the concepts then have the opportunity to be reinforced and corrected throughout the semester. Also, instructors can analyze short term learning objectives at different points in the semester and see if they align well with the long term course goals related to modeling and inference. What we call the hypothesis first approach (or HF for short) is similar to the way

Aliaga and Gunderson (2006) sequenced the topics in the beginning chapters of their textbook. The hypothesis first curriculum introduces the formulation of a testable hypothesis, the concepts of p-value, type I and II error, and the logic of a decision rule at the beginning of a semester in an intuitive way. Therefore, instructors can discuss the topics informally early, and then came back later to discuss the formal mechanics. Hong and O'Neil (1992) found evidence that teaching the ideas behind hypothesis testing before the formal procedure was a beneficial instructional technique. This type of informal introduction has also been recommended and formalized by Zieffler, et al. (2008). Dolor and Noll (2017) have summarized three newest approaches to an introductory statistics curriculum, and one of them was this type of informal inference. This study focuses on comparing student performance using two curricular approaches - an early introduction to hypothesis testing (HF - hypothesis first) and a traditional approach midway through the semester (denoted HL throughout this paper to stand for hypothesis later). Outside of the introduction of hypothesis testing, the rest of the course material is introduced in the same way as the traditional curriculum.

One reason that we chose to emphasize hypothesis testing in this research is because we believe it is an essential topic to introductory statistics and, even more so, it is a threshold concept (Taylor and Meyer 2009). The term threshold concept has been used to describe troublesome topics in many disciplines (Meyer and Land 2003; Land et al. 2008; Meyer and Land 2009; Walker 2013). Meyer and Land (2005) characterized threshold concepts using four criteria: a threshold concept is transformative, irreversible, integrative, and troublesome. Bulmer et al. (2007) also included hypothesis testing as a troublesome concept or, as others may call it, a threshold concept. Studies have shown that the concepts and applications related to hypothesis testing are probably the most misunderstood and misapplied topics of statistics (Brewer 1985; Batanero et al. 1994; Cobb and Moore 1997; Castro Sotos et al. 2009; Aquilonius and Brenner 2015; White 2004).

One benefit to the hypothesis first curriculum is that it includes more opportunities for reinforcement of the material, sometimes called distributed practice. When student learning activities are spaced out over the course of a semester, students may recognize that they have forgotten some of the material in that time period. When this happens they will tend to implement encoding methods that lead to better retention of the material; for example they may begin new study strategies that will lead to less forgetting (Benjamin and Bird 2006). Research also indicates that the knowledge structures of the students became more consistent and correct during the period of the course if it is reinforced (Bude et al. 2011). One question of other statistics researchers has been "how much repeated exposure is necessary for students to develop a deep understanding of statistical significance..."? (Chance, Wong & Tintle, 2017) As other researchers are examining this while using randomization methods, it will be interesting to compare those results to investigations using non-randomization techniques. As one group of authors reported "A key advantage of the randomization-based curriculum may be that students are able to conduct formal and informal inference on data early in the curriculum." (Tintle, Topliff, VanderStoep, Holmes, & Swanson, 2012). We hope this article will help to contribute to this discussion.

We collected data from one campus where all introductory statistics classes introduced parametric hypothesis testing rather than simulation/randomization tests. In our study we assessed the students' ability to formulate and test a hypothesis, make a decision and contextualize the result in both hypothesis first (HF) and hypothesis later (HL) curriculum.

METHODS

OVERVIEW OF INSTITUTION AND CLASSES

Our study compared the hypothesis first and hypothesis later curriculum from one instructor teaching two sections of the same course. The college involved in the study is a private, residential liberal arts college for men with a 95% confidence interval for the math SAT score of (556.45, 562.51). The institution does not offer a Statistics major.

The statistics course is a four credit hour course that meets four times a week in the morning hours and is fourteen weeks long. The statistics classes range in size from 20-30, in particular, the two sections involved in this study have 26 (hypothesis later) and 21 (hypothesis first) students. The classes are service courses for most students and required for Economics majors, which is the predominant major at the college. Students usually take this course their first year or second year.

Students self-selected into the classes without knowledge of the topic sequence. The HF class was told on the first day of class the steps of the scientific method. The instructor wrote on the board two distributions of colored poker chips (about 10 chips per distribution). The students then sampled from two bags to determine which bag corresponded to the appropriate distribution on the board. To introduce p-value and type I/II errors, students were asked to think about the situation of removing only one poker chip and then having to make the decision as to which distribution of chips was in the bag. Because the sampling (n=1) was small calculating the probabilities was not complicated. The steps of a hypothesis test were described, but not with the formal language of statistics (however, the definitions of type I/II errors were given). The HL group was introduced to hypothesis testing after learning descriptive statistics, the normal distribution, z-scores, and sampling distributions (see Table 1). The first hypothesis test introduced to the HL group was using a normal distribution with the logic of a sampling distribution (following the text *The Basic Practice of Statistics* by Moore). The HL group was given a different first example to hypothesis testing because they had just learned the sampling distribution and the hypothesis test was a use of this concept.

ASSESSMENTS

We analyzed different aspects of hypothesis testing: inquiry process, formulation, algorithm, and decision making and contextualizing. The baseline assessment was given on the first day of the class and the final assessment was embedded in the final exam. We assessed the different classes at different times in the semester (see Table 1), however most assessments were given a week or two after the material was introduced. Several of the assessment questions were from the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) (Garfield et al. 2006). A discussion on the formulation of the test is found in an article by delMas et al. (2007). A list of the assessment questions is given in the Appendix.

Table 1. Course Outline. Assessments in Bold.

Hypothesis First Course Topics	**Hypothesis Later Course Topics**
Baseline Assessment given -first day	**Baseline** Assessment given -first day
Hypothesis Tests/Inference, *p*-value, type I/II error (**Formulation - In week one**)	Displaying Distributions/Descriptive Stats
Displaying Distributions/Descriptive Stats	Normal Distribution
Normal Distribution	Linear Regression
Linear Regression	Correct Sampling Designs/Experiments
Correct Sampling Designs/Experiments	Probability
Probability	Binomial Distribution
Binomial Distribution	Law of Large Numbers and Central Limit Theorem
Law of Large Numbers and Central Limit Theorem	Hypothesis Tests/Inference, *p*-value, type I/II error (**Formulation - In week seven**)
Z Confidence Intervals/Tests (**Algorithm**) - Formal introduction to hypothesis tests	Z Confidence Intervals/Tests (**Algorithm**)
T Confidence Intervals/Tests	T Confidence Intervals/Tests
Intervals/Tests for Proportions	Intervals/Tests for Proportions
Chi-Square Test	Chi-Square Test
Inference for Regression	Inference for Regression
Final assessment (**Decision Making/Contextualizing)**	Final assessment (**Decision Making/Contextualizing)**

Four assessments were given during the course (see Table 2). On the baseline assessment there was a question regarding the inquiry process. This assessment on the inquiry process was to get a baseline of how the students would form a statistical study without any teaching from the class. The assessment question asked students to "describe a multi-step process of how" a professor could prove a group of students was significantly shorter in height than the campus population. The second assessment tested formulation, to see if students could correctly create a null and alternative hypothesis in words and symbols; this was given a week after the topic was introduced to students. The third assessment tested students on the hypothesis testing algorithm. Our goal for this assessment was to evaluate a student's ability to go through the complete process of: creating hypotheses, determining the appropriate test, calculating the test statistic, determining the p-value, determination of significance, and creating a conclusion in context. The final assessment was given at the very end of the semester to evaluate students' ability to make a decision using statistical thinking.

Table 2. Assessment Outline.

Assessment Name	**Topic(s) Included**
Baseline (first assessment)	Describe the process of a hypothesis test - do the students have any prior knowledge of the process
Formulation (second assessment)	Create a null and alternative hypothesis in words and symbols
Algorithm (third assessment)	Complete all steps necessary for a hypothesis test
Decision Making / Contextualizing (final assessment)	Complete Baseline question again and make decisions using statistical thinking

All assessments were given during class or exam periods and graded by the instructor according to a predetermined rubric. The questions for each assessment were worth different points according to how many steps were needed for a particular question, and the final grade of each assessment was recorded as percentages.

PROCEDURES

We collected data from one instructor and the instructor used the same textbook for both hypothesis first and hypothesis later curriculum, *The Basic Practice of Statistics* by David Moore (2010), which introduces hypothesis testing near the middle of the text. Texas Instruments (TI) calculators were used to find descriptive statistics, calculate binomial probability, and to conduct formal hypothesis tests. The instructor lectured for three of the four classes weekly and used the remaining one class entirely on group/cooperative activities and had one semester long group project. During the group activities, students were given problems to work using TI calculators. These group activities were graded by the instructor after class. Group activities included 3-4 individuals and were self-selected for all classes.

We compared two distinct time frames (see Table 1) for the introduction of the concept of hypothesis testing, one during the first week of instruction (hypothesis first - HF) and one where hypothesis testing was introduced around the seventh week of instruction (hypothesis later - HL). The instructor gave the baseline assessment at the beginning of the semester. The time in which assessments two and three were given is shown on Table 1, and the final assessment was given at the final exam. The assessments were given one to two weeks after the topics were introduced to the classes.

STATISTICAL ANALYSIS

We calculated the descriptive statistics for each assessment for HF and HL groups, and the 95% confidence intervals of the mean difference between HF and HL for each assessment. To perform the analysis, we firstly tested the normality assumption for the four assessment scores. The histograms of the baseline assessment and assessment two showed that the data were skewed to the right. The data from assessment three and the final assessment showed normal distributions. Thus, we conducted a 2 by 2 ANOVA, with one factor being pedagogy with the two levels (HF and HL), and one factor being assessment with two levels (the third and final assessments). The dependent variable was the average score from each assessment. Both pedagogy and assessment were fixed factors in the analysis. The ANOVA model is $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \epsilon_{ijk}$, where α_i represents the pedagogy effects, β_j represents the assessment effects, and $\alpha_i\beta_j$ is the interaction between pedagogy and assessment.

Given the data did not follow normal distributions for the baseline assessment and assessment two, nonparametric analyses, Mann-Whitney U tests, were conducted to compare the median differences between HF and HL groups.

We used an alpha level of 0.05 for all statistical tests. All the analyses were conducted using the IBM SPSS Statistics 24.0 software (IBM Corp. Armonk, NY).

RESULTS AND DISCUSSION

The descriptive statistics for all four assessments are shown in Table 3. We found that the average assessment scores for the HF section were consistently higher than those for the HL section for the third and final assessments at the end.

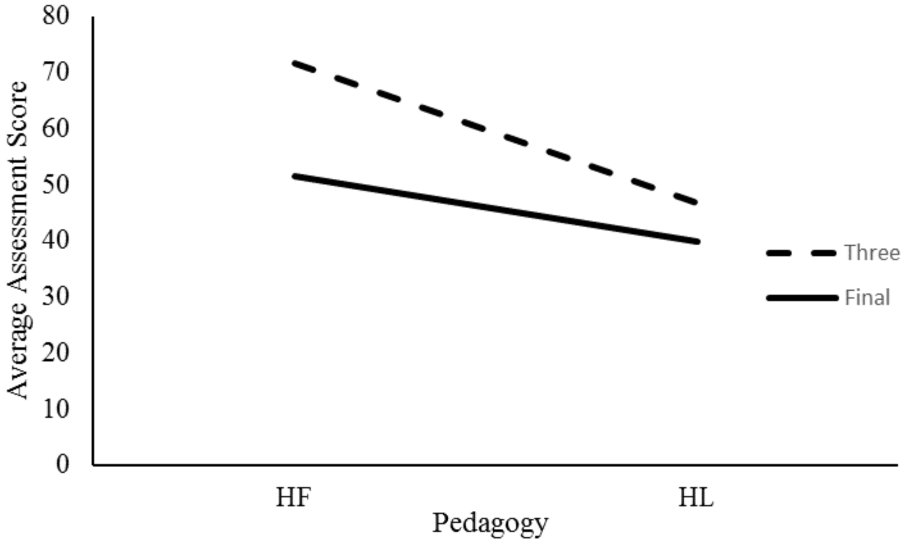
Table 3. Descriptive Statistics for All Assessments for HF and HL Sections.

Assessment	Pedagogy	Mean	SD
Baseline (Inquiry Process)	HF	6.40	9.52
	HL	19.13	15.34
Two (Formulation)	HF	15.91	27.33
	HL	23.44	29.54
Three (Algorithm)	HF	71.59	20.36
	HL	46.76	24.28
Final (Decision making / contextualization)	HF	51.62	15.42
	HL	39.92	19.57

The Mann-Whitney U test of the baseline assessment shows that the median assessment score of HF was significantly lower than the median assessment score of HL (U=155.5, p-value=0.003).
The Mann-Whitney U test of assessment two indicates that there is no significant difference between the median scores of HF and HL (U=146.5, p-value=0.321).

The two-way ANOVA for assessment three and the final assessment shows that there is a significant main effect of pedagogy (F(1,75)=16.34, p-value<0.001), there is a significant main effect of assessment (F(1,75)=8.80, p-value=0.004), but there is no significant interaction between pedagogy and assessment (F(1,75)=2.11, p-value=0.15). Figure 1 shows that the mean assessment score of HF is significantly higher than that of HL.

Figure 1. Average assessment scores for HF and HL for the third and final assessments



The 95% confidence intervals of the mean differences between HF and HL for all the assessments are provided in Table 4. Based on the 95% confidence intervals, we also observed that the average score of HF is significantly higher than that of HL for the third and final assessments.

Table 4. 95% confidence intervals for mean differences.

Assessment	95% confidence intervals (HF-HL)
Baseline (Inquiry Process)	(-20.086, -5.375)
Two (Formulation)	(-26.365, 11.309)
Three (Algorithm)	(10.550,39.112)
Final (Decision making /contextualization)	(0.363,23.051)

We noticed a trend through our assessments: the HF group started out significantly lower on the baseline assessment, gained ground on the second, then obtained a significantly higher score than HL by assessment three, and maintained significant better performance through the final assessment.

LIMITATIONS OF THIS STUDY

We outline some of the major limitations of this pilot study. First, gender of students could be an issue as the data collected was solely from male students. Secondly, the assessment questions are not all validated, even though some of them were from ARTIST questions that are validated. Third, the sample sizes are small. For a pilot study and data collected at a small institution this is to be expected; however, in order to extend these results, more data is needed. Finally, only one instructor from one institution was used to compare the two instructional approaches. In the future it would be essential to gather data from multiple instructors at various types of institutions.

CONCLUSION AND FUTURE GOALS

This study was initiated with the expectation that the hypothesis first curriculum would help students to better understand and internalize the process of hypothesis testing, and related concepts, over the course of the semester. Our statistical analysis revealed that there was a consistent pattern that the average assessment grade of HF was higher than that of HL for the last two assessments. Overall, this pilot study indicates the hypothesis first curriculum is beneficial to student learning in an introductory statistics course at a traditional institution.

A tremendous number of research studies have focused on the teaching and learning of statistics over the past decade (van der Merwe and Wilkinson 2011) and the difficulty encountered by students in understanding the concepts related to distribution, center and variability is one of the focuses of these research studies along with misconceptions about interpretation of p-value (Batanero et al. 1994; delMas et al. 2007; Garfield and Ben-Zvi 2007). There is also research that suggests instructors should develop concepts related to hypothesis testing early in a statistics class in an intuitive/informal way before formally developing the theory (Garfield et al. 2007, Zieffler et al. 2008). Introducing hypothesis testing informally at the beginning of the class can give students an overview of part of the process of statistical analysis before they dive into the individual concepts more rigorously. Formal introduction of the whole process of hypothesis testing later on in the semester for the HF group can then give more opportunity for students to review.

As a future goal it would be interesting to conduct the study with a larger sample size and diverse student population. Another goal could be to analyze the efficacy of other threshold concepts in the introductory statistics course as it is laid out currently; and also, from a course design perspective, to investigate the importance of timing or placement of these topics in the course. With the new curriculum recommendation from the ASA (2014) that focuses on a more data oriented and modeling approach, it may be the appropriate time to look at the sequencing of the topics in introductory statistics courses for non-majors and majors alike.

APPENDIX: ASSESSMENT QUESTIONS

BASELINE

1. (Include this problem as part of the initial test and on the final assessment.) A professor thinks the incoming class of freshmen at his university is significantly shorter than the rest of the student population. Suppose the total student population is around 60,000 students. Describe a detailed multi-step process of how the professor could try to prove his claim. (5pts)

FORMULATION

1. Dr. Smith thinks the students in his mathematics class are smarter than the students in other sections of the same class. He wants to use the students' first exam scores to prove his point. Luckily, each section of the math course takes the exact same exam. He has access to the grades from the six other courses at his university. Dr. Smith's class average was 85% for his 32 students. The class average for the other sections was 80% for 250 students. Using this data, what hypotheses should Dr. Smith test? Write your answer in a complete sentence and then in symbols. (2pts)

2. (ARTIST question) Suppose you want to determine whether students' expected grades at the beginning of an introduction to statistics course are positively related to their final course grade. Write the null and alternative hypothesis in words. (2pts)

ALGORITHM

1. Suppose that the average shoe size of 8-year-olds is 7. In a class of 8-year-olds the teacher thinks the average is not 7. For this situation suppose a test statistic of $z = 1.76$ has been calculated. Assuming shoe sizes are normally distributed, is this test significant for $\alpha = 0.05$? State your hypotheses in symbols and words, perform the appropriate test, make a decision and state your conclusion in context. (8pts)

2. A study followed 89 infants from low-income families from birth to adulthood. At age 20, the mean IQ score for these infants was 98.7. IQ scores follow a Normal distribution with $\sigma = 15$. IQ tests are scaled so that the mean score in a large population should be $\mu = 100$. Researchers suspect that the low-income population has mean less than 100. Does this study give good evidence that this is the

truth? State your hypotheses in symbols and words, perform the appropriate test, make a decision and state your conclusion in context. (8pts)

3. (ARTIST question) A manufacturer of light bulbs claims that its light bulbs have a mean life of 1520 hours. A random sample of 40 such bulbs is selected for testing. If the sample produces a mean value of 1498.3 hours and a standard deviation of 85 hours, is there sufficient evidence to claim that the mean life is significantly less than the manufacturer's claim, using the α= .01 significance level? State the hypotheses, report the test statistic, and draw the appropriate conclusion in context. (8pts)

DECISION MAKING/CONTEXTUALIZING

1. Which of the following is the best completion of the statement? A hypothesis test…: (1pt)
   a. Proves the null hypothesis is true or false.
   b. Proves the probability of the null hypothesis occurring.
   c. Evaluates the evidence in favor (or against) the null hypothesis.

2. (From first assessment.) A professor thinks the incoming class of freshmen at his university is significantly shorter than the rest of the student population. Suppose the total student population is around 60,000 students. Describe a detailed multi-step process of how the professor could try to prove his claim. (5pts)

3. Why do very small p-values indicate that the evidence against the null hypothesis is strong? (Circle one.) (1pt)
   a. Because the p-value is the probability that the null hypothesis is true.
   b. Because the small p-value indicates that the data lie within the confidence interval.
   c. Because the small p-value indicates that data like ours would be very uncommon if the null hypothesis were true.
   d. Because the small p-value indicates that data like ours would be very common if the alternative hypothesis were true.

REFERENCES

Aliaga, M., and Gunderson, B. (2006), *Interactive Statistics*, 3rd Edition, Upper Saddle River, NJ: Pearson Prentice Hall.

Aquilonius, B.C., and Brenner, M.E. (2015), “Students’ Reasoning about p-Values,” *Statistics Education Research Journal*, 14 (2), 7-27.

ASA American Statistical Association Undergraduate Guidelines Workgroup. (2014), “2014 curriculum guidelines for undergraduate programs in statistical science,” Alexandria, VA: American Statistical Association [online]. Available at http://www.amstat.org/education/curriculumguidelines.cfm

Bahrick, H., and Hall, L. (2005), “The importance of retrieval failures to long-term retention; A metacognitive explanation of the spacing effect,” *Journal of Memory and Language*, 52 (4), 566-577.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., and Holmes P. (1994), “Errors and difficulties in understanding elementary statistical concepts,” *International Journal of Mathematical Education in Science and Technology*, 25 (4), 527-547.

Benjamin, A. S., and Bird, R. D. (2006), “Metacognitive control of the spacing of study repetitions,” *Journal of Memory and Language*, 55, 126–137.

Brewer, J. K. (1985), “Behavioral statistics textbooks: Source of myths and misconceptions?” *Journal of Educational Statistics*, 10 (3), 252-268.

Bude, L., Imbos, T., van de Wiel, M. W., and Berger, M. P. (2011), “The effect of distributed practice on students’ conceptual understanding of statistics,” *Higher Education* [online], 62(1), 69-79, DOI: 10.1007/s10734-010-9366-y. Available at http://link.springer.com/article/10.1007/s10734-010-9366-y/fulltext.html

Bulmer, M., O’Brien, M., and Price, S. (2007), “Troublesome concepts in statistics: a student perspective on what they are and how to learn them,” in *UniServe Science Teaching and Learning Research Proceedings* [online], pp. 9-15. Available at http://science.uniserve.edu.au/pubs/procs/2007/06.pdf

Castro Sotos, A. E., Vanhoof S., Van den Noortgate W., and Onghena, P. (2009), “How confident are students in their misconceptions about hypothesis tests?” *Journal of Statistics Education* [online], 17 (2). Available at www.amstat.org/publications/jse/v17n2/castrosotos.html

Chance, B., Wong, J. and Tintle, N. (2016), “Student Performance in Curricula Centered on Simulation-Based Inference: A Preliminary Report,” *Journal of Statistics Education*, 24 (3), 114-126.

Cobb, W. G. (2007), "The introductory statistics course: A Ptolemaic curriculum?" *Technology Innovations in Statistics Education* [online], 1(1), 1-15. Available at http://escholarship.org/uc/item/6hb33k0nz

Cobb, W. G., and Moore, S. D. (1997), "Mathematics, statistics, and teaching," *The American Mathematical Monthly*, 104 (9), 801-823.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing students' conceptual understanding after a first course in statistics," *Statistics Education Research Journal* [online], 6 (2), 28-58. Available at http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

Dolor, J., and Noll, J. (2017), "Using guided reinvention to develop teachers' understanding of hypothesis testing concepts," Statistics Education Research Journal, 14(1), 60-89.

Garfield, G., and Ben-Zvi, D. (2007), "How students learn statistics revisited: a current review of research on teaching and learning statistics," *International Statistics Review*, 75 (3), 372-396. DOI: 10.1111/j.1751-5823.2007.00029.x

Garfield, G., delMas, R. and Chance, B. (2006), *Assessment Resource Tools for Improving Statistical Thinking (ARTIST)* [online]. Available at https://apps3.cehd.umn.edu/artist/index.html

Garfield, J., delMas, R. and Chance, B. (2007), "Using students' informal notions of variability to develop an understanding of formal measures of variability", In M. Lovett and P. Shah (Eds.), *Thinking with Data*, 117–148. Mahwah, NJ: Lawrence Erlbaum.

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

Hong, E., and O'Neil Jr., H. F. (1992), "Instructional strategies to help learners build relevant mental models in inferential statistics," *Journal of Educational Psychology*, 84(2), 150–159.

Land, R., Meyer, J. H. F., and Smith, J. (Eds.). (2008), "Threshold Concepts within the Disciplines," *Educational Futures: Rethinking Theory and Practice* (Vol. 16), Rotterdam, The Netherlands: Sense Publishers.

MAA Mathematical Association of America. (2015), "A common vision for the undergraduate mathematics program in 2025," [White paper] [online]. Available at http://www.maa.org/programs/faculty-and-departments/common-vision

Malone, C., Gabrosek, J. Curtiss, P., and Race, M. (2010), "Resequencing topics in an introductory applied statistics course" *The American Statistician* [online], 64 (1), 52-58. DOI: 10.1198/tast.2009.08090. Available at http://www.tandfonline.com/doi/abs/10.1198/tast/2009.08090

Meyer, J., and Land, R. (2003), "Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practicing in the disciplines", Enhancing Teaching-Learning Environments in Undergraduate Courses Project occasional report 4, University of Edinburgh, School of Education [online]. Available at http://www.tla.ed.ac.uk/etl/docs/ETLreport4.pdf

Meyer, J., and Land, R. (2005), "Threshold concepts and troublesome knowledge (2): epistemological considerations and a conceptual framework for teaching and learning," *Higher Education*, 49, 373-388.

Moore, D. (2010), *The Basic Practice of Statistics*, 5th Edition, New York, NY: W.H. Freeman and Company.

Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., and Lee, H. S. (2012), "Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics," [online] Available at http://www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf

Prodromou, T. (2017), "Model-based Informal Inference," International Journal of Statistics and Probability, Vol. 6, No. 5, 140-147. DOI: 10.5539/ijsp.v6n5p140. Available at http://doi.org/10.5539/ijsp.v6n5p140

Rohrer, D., and Taylor, K. (2006), "The Effects of Overlearning and Distributed Practice on the Retention of Mathematics Knowledge," *Applied Cognitive Psychology*, 20, 1209–1224.

Taylor, C. E. and Meyer, J. H. F., (2009), "The testable hypothesis as a threshold concept for biology students," In J.H.F. Meyer, R. Land & C. Baillie (Eds.), Threshold Concepts and Transformational Learning (pp.179-191), *Educational Futures: Rethinking Theory and Practice*, (Vol. 42). Rotterdam, The Netherlands: Sense Publishers.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V. and Swanson, T. (2012), "Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum," *Statistics Education Research Journal*, 11(1), 21-40.

Van der Merwe, L. and Wilkinson, A. (2011), "Mapping the Field of Statistics Education Research in Search of Scholarship," *International Journal for the Scholarship of Teaching and Learning*, 5(1), Article 29.

Walker, G. (2013), "A cognitive approach to threshold concepts," *Higher Education* [online], 65, 247-263, DOI: 10.1007/s10734-012-9541-4. Available at http://link.springer.com/article/10.1007/s10734-012-9541-4

Wardrop, R. L. (1995), *Statistics: Learning in the Presence of Variation*, Dubuque, IA: Wm. C. Brown Communications, Inc.

White, B. (2004), "Reasoning maps: a generally applicable method for characterizing hypothesis-testing behavior," *International Journal of Science Education*, 26(14), 1715-1731.

Wiggins, G. and McTighe, J. (2006), *Understanding by Design*. Pearson: Merrill Prentice Hall.

Zieffler, A., Garfield, J., DelMas, R., and Reading, C. (2008), "A Framework to Support Research on Informal Inferential Reasoning," *Statistics Education Research Journal*, 7(2), 40-58.