Content Audit for P-value Principles in Introductory Statistics

Karsten Maurer, Lynette Hudiburgh, Lisa Werwinski, John Bailer
Miami University

Revision Submitted: 10 September 2018

Abstract

Longstanding concerns with the role and interpretation of p-values in statistical practice prompted the American Statistical Association (ASA) to make a statement on p-values (Wasserstein & Lazar 2016). The ASA statement spurred a flurry of responses and discussions by statisticians, with many wondering about the steps necessary to expand the adoption of these principles. Introductory statistics classrooms are key locations to introduce and emphasize the nuance related to p-values; in part because they engrain appropriate analysis choices at the earliest stages of statistics education, and also because they reach the broadest group of students. We propose a framework for statistics departments to conduct a content audit for p-value principles in their introductory curriculum. We then discuss the process and results from applying this course audit framework within our own statistics department. We also recommend meeting with client departments as a complement to the course audit. Discussions about analyses and practices common to particular fields can help to evaluate if our service courses are meeting the needs of client departments and to identify what is needed in our introductory courses to combat the misunderstanding and future misuse of p-values.

1. Introduction

The publication of this journal's special issue reflects a growing consensus: p-values are often misused, and that misuse often leads to bad science. Many argue that the main challenges are to understand the logic of testing scientific hypotheses, and to get away from mechanical rules like $p < .05$ as a substitute for contextual reasoning. The logic of testing, we argue, is best taught in a first statistics course. However, research in statistics education makes clear that this logic is far harder to teach and to learn than is the simple $p < .05$. This poses a particular challenge for those who teach introductory statistics courses. In this article we propose a process for auditing the coverage of p-value principles in an introductory statistics course.

The misuse of p-values is frequent, well-documented, and potentially leads to bad science; most notably, the "reproducibility crisis." Ioannidis sounded an alarm with his paper, "Why most published research findings are false" (Ionnidis, 2005). A decade later, the Open Science Collaboration (2015) repeated 100 experiments taken from the psychology literature. Of these hundred, only 39 produced results that replicated the original findings. In the health sciences, Greenland et al. (2016) offered readers a catalogue of misinterpretations of p-values. As noted by Berry (2016), "(o)ur collective credibility in the science community is at risk". These and other articles led the American Statistical Association (ASA) to issue a statement on proper use of p-values (Wasserstein and Lazar, 2016). The ASA statement spurred a flurry of responses and discussions by statisticians, along with the 2017 American Statistical Association Symposium on Statistical Inference. We now look toward the next steps necessary to expand the adoption of these principles.

Although there is a growing consensus about the nature of the problem, there is little consensus on simple remedies like banning p-values altogether or reducing the threshold for "significance" to $p < 0.005$. Most agree that the heart of the problem is reliance on mechanical rules like $p < 0.05$. Such rules cannot substitute for the logic of hypothesis testing applied in the scientific context. Unfortunately, decades of research have shown that this logic is not easy to learn (Falk and Greenbaum, 1995; Williams, 1999; Batanero, 2000; Garfield and Ben Zvi, 2003; Harradine et al., 2011). Del Mas et al. (2007) gave a multiple choice test to students who had completed an introductory

statistics course and found that only 54.5% could identify the correct interpretation for p-values; only 58.6% could identify incorrect interpretations. Rossman (2008) cites the work of Nickerson (2004) in cognitive psychology to argue that one explanation “surely rests in all of the research that has shown how difficult probabilistic reasoning is for people.”

Stangl (2016) argues that we have a responsibility in statistics education to preempt and end the perpetual misuse of p-values. Cobb (2016) responded to the ASA statement by saying, “(w)hat ASA has done here should spur a reshaping of the way we teach – both p-values in particular and statistics generally.” There are many reasons to focus on the introductory course in statistics. For many students it is their first encounter with the logic of statistical inference. For most of those students it is also their last formal encounter with that logic in an academic setting. Moreover, courses that introduce statistical thinking and methods reach a very large percentage of the students who will become practicing scientists and evidence-based decision makers.

The ASA-sponsored report, *Guidelines for Assessment and Instruction in Statistics Education* (GAISE), outlines goals and methods for teaching introductory statistics (Carver et al., 2016). The GAISE report sets as a goal “(s)tudents should demonstrate an understanding of, and be able to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings.” Millar (2016) argues that “(s)tudents of other disciplines will be in our service courses, and while we should not advocate for hypothesis tests as the monolithic statistical inference method, they do need to know what it is and what its shortcomings are because they will encounter it.” Goodman (2016) also points out the responsibility our courses have to future scientists, saying “(t)he fact that statisticians do not all accept at face value what most scientists are routinely taught as uncontroversial truisms will be a shock to many. But if we are to move science forward, we must speak to scientists.” Berry (2016) is more direct: “We must communicate better even if we have to scream from the rooftops.” But as statisticians, before we scream, we should gather data.

In the remainder of this article we describe and illustrate the use of a rubric that can be used to assess how well a course addresses the ASA’s principles for sound use of p-values and the use of focus group discussions between teachers of statistics and

their colleagues in the sciences. We hope these will serve as tools to help frame systematic conversations within and across departmental lines. In what follows, Section 2 describes the rubric and focus group questions. Section 3 illustrates an example of conducting the course audit with these methods. Section 4 concludes with a discussion.

2. Methods

For a department to perform a comprehensive audit of p-values concepts in their introductory statistics courses, we propose a framework that elicits both intra-departmental and inter-departmental feedback on the current curriculum. The combination of introspection and external suggestions can then drive targeted curricular adaptation to better cover the challenging topics related to p-values. The key idea for this assessment is that each of the six principles articulated by the ASA Statement provide a target that can be directly evaluated with a common rubric as we discuss next.

2.1 Intra-Departmental Evaluation

For a department with several instructors involved in teaching introductory statistics courses, it may be helpful to use a common rubric as a unified starting point for identifying strengths and/or weaknesses in teaching about p-values. We propose the rubric found in Table 1. The rubric is structured to be applied to each of the six principles from the ASA statement:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

The rubric evaluates how each principle is formally introduced by the instructor, reinforced with classroom activities, assessed for comprehension and is supported with appropriate supplementary learning resources. These broad categories of course components are used because they are likely to be present, in some form, in most introductory courses.

Rubric Instructions for Evaluating ASA P-Value Principle:

Repeat the following steps for each principle i=1,...,6

1. Reread the i^{th} ASA p-value principle.
2. Reflect on how completely, correctly and consistently each of the four broad curricular components (instruction, class activity, assessment and support materials) reflect the i^{th} ASA p-value principle.
3. Score the corresponding item based on the statement in Table 1 that you feel most accurately describes how substantially the curricular component reflects the i^{th} ASA p-value principle. Enter this in the curriculum audit report card in Table 2.

The curricular component does not reflect the ASA Principle.	The curricular component barely reflects the ASA Principle.	The curricular component mostly reflects the ASA Principle.	The curricular component fully reflects the ASA Principle.
Score=0.0	Score=0.3	Score=0.7	Score=1.0

Table 1: Curricular Component Rubric.

4. Sum the score from the four curricular components. The sum will range from 0-4. Treat this sum as a GPA value and assign a letter grade. (1.0=D, 1.3=D+, 1.7=C-, 2.0=C, 2.3=C+, 2.7=B-, 3.0=B 3.3=B+, 3.7=A-, 4.0=A)

With such a general grading guide, the scores and letter grades are clearly subjective in nature but may provide an informative shorthand familiar to most educators. It is important to highlight the rubric is primarily designed as a support tool to help frame discussion and help target areas for improvement in an internally administered content audit. Creating an overall score is secondary, as it is *not* a nationally-normed instrument for comparing quality of curriculum across all universities. The rubric can be completed collectively or by individual instructors and then used to drive a discussion on steps to improve the curriculum coverage of the ASA principles. Alternative rubrics could be devised by moving to simple dichotomous responses for the inclusion or exclusion of components covering particular ASA principles, or by including

weights for components that a department finds more essential, or by breaking from the letter-grade theme altogether. In the end, we employed a relatively interpretable scale to provide a basis for reflection and discussion. If several department members identify a deficiency in teaching one of the principles from the ASA statement, a collective curricular remediation can be planned. In the case that a principle is covered well by one instructor but not by another, class materials and teaching advice can be shared to help patch the gap. A result of this audit conducted by a subgroup of a department might provide the basis for discussion at a departmental retreat or meeting.

		ASA P-Value Principles					
		1	2	3	4	5	6
Curricular Component	*Instruction:* lecture, discussion, video lecture, etc.						
	Activity: lab, worksheet, case studies, etc.						
	Assessment: homework, quiz, exam, etc.						
	Support Materials: readings, tutorials, apps, etc.						
	Totals						
	Overall P-value Principle Grades						

Table 2: Curriculum Audit Report Card

We conducted the content audit for p-value principles within our department at Miami University, a mid-sized public Midwestern university. Multiple introductory statistics courses are offered through our department, each geared toward a different sub-population of students. We applied the rubric to our algebra-based introductory statistics service course for undergraduate students, STA 261. This course serves an assortment of majors. The course begins with examples that challenge students to

begin inferential thinking by using simulation-based methods to evaluate likelihood of observed data under assumed conditions. The curriculum then proceeds through a unit on probability and sampling distributions before introducing p-values in a probability-based framework.

This class is taught to 600 students each semester using a hybrid model with online introduction of concepts, just-in-time teaching of problematic concepts in a large lecture meeting, and a smaller lab section where statistical principles are explored. In a typical semester, two large lectures are taught by a continuing lecturer and four other large lectures are taught by term-limited faculty, namely Instructors or Visiting Assistant Professors. Graduate students facilitate lab sections. Course materials and labs are all centrally constructed by the course coordinator. In our case, the rubric was completed collectively by members of the author group and the results can be found in Section 3.1.

2.2 Inter-Departmental Evaluation

Next, we recommend running a small focus group discussion with analytically savvy members from client departments in order to gain an interdisciplinary perspective on the statistics curriculum at your institution. A great starting place is to identify departments with large numbers of undergraduate majors who are required take your introductory service course in statistics. These will often be psychology, political science and biology departments, but this will vary from campus to campus. The goals would be to advocate for good analytical practice in the scientific community, in this case pertaining to p-values, and to ask for candid feedback on how the statistics service classes meet the needs of their respective fields based on their general observations. Our focus group included faculty members from psychology, biology, geology, and kinesiology.

Along with an invitation to the meeting, we suggest sending a brief description with a web-link to the ASA p-value statement and a short list of questions you plan to discuss. We encourage you to consider a set of questions that can be used as discussion prompts. For example, when we conducted this focus group, we asked a number of questions to prompt additional discussion including:

- Teaching in our introductory statistics service courses:
    - Do your students show understanding of p-values and hypothesis testing?

    - What methods should our introductory statistics course include?
- Teaching in our advanced statistics service courses:
    - What do our advanced service courses do well to prepare your students?
    - Do your advanced students show understanding of p-values and hypothesis testing?
    - What is missing and what would you like to see us address in more detail?
- Why is a fundamental understanding of statistics important in your field?
    - What attracted and engaged you personally with statistics?

3. Results

In Fall 2017, we conducted a content audit for the coverage of p-values in our algebra-based introductory statistics course and held a focus group meeting with representatives from client departments. The results of the content audit and focus group are summarized in subsections 3.1 and 3.2, respectively.

3.1 Application of Course Content Audit Rubric

Table 3 contains the rubric that was collectively completed by members of the author team to evaluate the curriculum of the introductory course discussed in Section 2.1. The audit for our algebra-based introductory course found that we avoided improper probabilistic interpretations and clearly articulate the difference between statistical and practical significance as highlighted in principles two and five, respectively. However, the results of the audit suggested that we had ample room to improve our coverage of some principles. As a result, we developed a set of action items to address these areas' curricular weaknesses. A natural follow up to the assessment is to document what features were highlighted in this evaluation as a justification of the assigned grade. Table 4 above provides this reflection.

The most urgent action item was to replace the rote procedural approach of running down the checklist of model assumptions taught to accompany each hypothesis test, and instead begin to frame hypothesis tests more comprehensively with respect to clearly specified statistical models. For example, we can discuss a hypothesis test for a population proportion of success equal to 0.5, where a binomial exact test runs from a

		ASA P-Value Principles					
		1	2	3	4	5	6
Curricular Component	*Instruction:* lecture, discussion, video lecture, etc.	0.3	0.7	0.7	0.3	0.7	0.3
	Activity: lab, worksheet, case studies, etc.	0.3	1.0	0.7	0.7	1.0	0.3
	Assessment: homework, quiz, exam, etc.	0.3	1.0	0.7	0.7	1.0	0.7
	Support Materials: readings, tutorials, apps, etc.	0.3	1.0	0.7	0.7	1.0	0.7
	Totals	1.2	3.7	2.8	2.4	3.7	2.0
	Grades	D	A-	B-	C+	A-	C

Table 3: Completed Curriculum Audit Report Card

simple and familiar binomial model. Here, we can note that the p-value calculated could reflect the true proportion differing from 0.5 or that the observations were not from a binomial experiment. Another action item from the audit is to include a case study to deconstruct a real-world analysis involving hypothesis tests, where we can highlight study design, dangers of "p-hacking," effect sizes, integration of additional evidence, and resulting policy decisions. The last action item is to expand our discussions of reproducibility beyond documenting data handling to include all stages of a study, from design through analysis. The Reproducibility Project led by the Center for Open Science can provide a good launching point for this discussion (Weir, 2016).

We have shared results of the audit of this introductory course curriculum with our colleagues at a recent departmental retreat. The discussion encouraged change to modify content and exercises to better address the ASA p-value principles throughout our curriculum. We also asked our colleagues to consider how we can similarly focus the curriculum of our upper-level classes to reinforce areas that are challenging to impart with students in an introductory course.

1. P-values can indicate how incompatible the data are with a specified statistical model. **Grade: D** Reason: Hypothesis tests (HT) are presented and assessed with statements of assumptions, but these assumptions are often made without consideration of why these assumptions are important. May not be until the first regression / modeling class before intuition about the importance of assumptions on validity of HT. Even then, the p-value is more often emphasized as linked to a parameter vs. a model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that data were produced by random chance alone. **Grade: A-** Reason: Conditional probability is strongly emphasized and p-values are consistently discussed and assessed in the context of a particular hypothesis being true. Minor problems remain where we may still see the second interpretation sneak into instructor or graduate assistant descriptions.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. **Grade: B-** Reason: We regularly and thoroughly challenge the bright line rules (P<0.05 significant) with examples, readings and discussions; for instance questioning what the difference between P=0.049 and P=0.051 really means. However, context factors -- design of study, quality of measurements, and data analysis assumptions -- are often given little emphasis or done as a topic of discussion separate from the p-value calculation and interpretation.
4. Proper inference requires full reporting and transparency. **Grade: C+** Reason: Our curricular components currently discuss p-values in the context of a single study. The class does a lab exercise on data cleaning and ethical / transparent data handling procedures, but largely omits the discussion about transparent reporting of analysis choices. Topics like multiplicity of testing and “p-hacking” are rarely discussed in our introductory classes, which don’t tend to emerge until a second course in statistics.
5. P-value does not measure the size of an effect or the importance of a result. **Grade: A-** Reason: Exercises and class activities address practical significance vs. statistical significance. Emphasize interval estimation and reporting effect sizes across curricular components. Need to be careful that we do not implicitly suggest that larger p-values imply “lack of importance or even lack of effect” or that smaller p-values necessarily imply “larger or more important effects.”
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. **Grade: C** Reason: We do not suggest that data analysis ends with the construction of a p-value; but rarely do we discuss or evaluate other evidence in support of our models/hypotheses.

Table 4: Results of our content audit with a short explanation specific to each principle.

3.2 Summary of Focus Group Discussion

Looking externally for feedback, our focus group was comprised of five colleagues from the fields of biology, kinesiology, psychology and environmental science. During the ninety minute session with these representatives of our client departments, we asked the questions listed in Section 2.2. The questions were posed as prompts to discussion periods where a member of the statistics faculty primarily acted as a facilitator: taking notes, asking for additional details and prompting for points of clarification. The discussion yielded fruitful conversation; the detailed summary is found in Table 5 below.

The focus group provided strong insights for us to consider when we engage students from many scientific disciplines in our service courses. We are not proposing that we completely restructure our course based this input, only that we acknowledge what is viewed as important by client departments when evaluating service course curricula. A major theme that came out of the conversation was that hypothesis testing and p-values are used widely in their fields, so students need a better understanding of the ASA principles. The focus group would also like to see a more thorough coverage of additional concepts such as the fundamentals of probability, model fitting, effect size estimation, and Bayesian methods in our curricula for future STEM students. Lastly, the personal feelings toward statistics by the faculty members from other disciplines were illuminating and encouraging. Several of them divulged an animosity or an ambivalence toward statistics in their undergraduate educations, but later came to recognize the value in their careers during graduate study. This revealed the urgent need to provide stronger motivation for STEM students in an undergraduate statistics course. We should strive to motivate *why* the statistical methods they are learning are necessary within their respective fields, and demonstrate the value that statistics provides within science. A final benefit of this meeting with representatives from other departments is that this strengthens the connections between statistics departments and client departments.

4. Discussion

We encourage statisticians in academia to evaluate the coverage of p-values and hypothesis testing in their curriculum, in light of the ASA p-value statement. Within our

Question Prompts	**Summary of Focus Group Discussion Responses**
Teaching in our *introductory* statistics service courses: - Do your students show understanding of p-values and hypothesis testing? - What methods should an introductory statistics course include?	*P-value understanding:* ● Students don't adequately understand p-values after intro courses ● Need more theory and intuition about p-values, less procedural mechanics and thresholds, show impacts of assumptions *Desired Coverage:* ● Want students to be good data consumers – need 'soft touch' in intro ● Considering alternative sources of uncertainty ● Students don't seem to understand discrete probability distributions ● More fundamental probability, including conditional probability ● More exposure and practice with software and packages ● Emphasize effect sizes and interval estimates ● May benefit from resampling / randomization – illustrate null model ● Open question: How can we get more advanced methods from intro?
Teaching in our *advanced* statistics service courses: - What does it do well to prepare your students? - Do your students show understanding of p-values and hypothesis testing? - What is missing or needs more detail?	*Coverage of Methods:* ● P-value good starting point, necessary for reading literature ● Many students don't understand p-values even after a second course ● Becoming practitioners, need to know principles and adv. methods *What is needed:* ● More modeling vs. HT – focus on fit (R^2) and selection vs. tests ● Behind in teaching Bayesian methods *Challenges to extending beyond HT:* ● May go against adviser's training and level of expertise ● Books/literature still have heavy focus on HT ● "P<0.05 will continue to be taught b/c it is so much easier to teach" ● "What do we reinforce? We implicitly support this bright line" ● Grad students too focused on application, ignoring important nuance ● Simplicity of "yes/no" type of outcome is attractive
Why is a fundamental understanding of statistics important in your field? What attracted and engaged you personally with statistics?	*Big issues for stats in STEM education:* ● Too much focus on "what of statistics", need more "why of statistics" ● Motivation gap for learning statistics vs. science specialty ● Fighting student expectations (and sometimes their adviser's too) ● Challenging idea: something could be wrong in a scientific model *Individual responses for personal experience with learning statistics:* ● Tools for addressing the idea of significance / expressing confidence ● Not interested as an undergraduate – took off as a graduate student when encountered Markov models to explain behavior ● "Hated undergraduate stats class but in graduate school it made sense and I love it now" ● Interest and passion emerged in grad school but stat classes not directly relevant to their field – 90% of statistics background was learned as a researcher. Need to start motivating the learning earlier.

Table 5: Highlights from focus group meeting with colleagues from client departments.

statistics department, we conducted an audit of our algebra-based introductory course using our broad rubric to evaluate how effective the curriculum is at conveying the six principles from the ASA p-value statement. We then looked outside the department, through a focus group conversation with colleagues across several scientific disciplines. Through these activities, we have formed a better understanding of the coverage of p-values and hypothesis tests within our introductory curriculum and have developed a set of action items to remediate areas where we can improve student learning outcomes. Following the research of John Dewey (1933), we feel that the critical reflection on the state of our inference curriculum for our introductory service courses will make us more aware of how we approach these concepts in our other courses. We propose the framework that we used in our reflective process as a basic template that can be adapted and implemented at other universities.

Several challenges remain for fully improving the quality of p-value understanding in introductory courses. At many institutions, introductory statistics courses are staffed by temporary instructors or visiting faculty, who may lack experience in the teaching or study of statistics. In addition, lab sections are often facilitated by graduate assistants, many of whom are new to statistics. Continuing research methodology courses in other departments are also often taught by non-statistician methodological instructors who may hold misconceptions about significance and p-value interpretations (Harradine et al., 2011). This suggests that instructor preparation is a necessary component for improving the teaching of p-value concepts. We need to have a continuing process of training the instructors for our introductory classes. A great set of resources for staying current with instructional practices can be found through the ASA Section on Statistics Education (community.amstat.org/statisticaleducationsection), the Consortium for the Advancement of Undergraduate Statistics Education (causeweb.org) and the Statement on Qualifications for Teaching an Introductory Statistics Course by the ASA/MAA Joint Committee on Undergraduate Statistics (2014).

Another challenge for introductory statistics is that the ASA statement encourages that p-values be framed within a wider range of methods. We cannot expect students new to statistics to appreciate the value of robust analyses and reporting without some real exposure to those methods. For this, we need to find space in the

curriculum -- a scarce commodity in any statistics course -- to cover new analysis methods, such as model selection, bootstrapping, generalized linear models, or Bayesian methods. Including a broader context of p-values and/or reinforcing the introductory curriculum will almost certainly come at the cost of another topic, and clearly not all of these topics can be added to an introductory statistics class. While the particular weight given in the curriculum is subjective, it should be heavily considered when striking the balance, given the general consensus in statistics education that a proper understanding of inference is a foundational learning objective in introductory statistics (Cobb, 2007; Garfield and Ben-Zvi, 2008; Rossman, 2008; Carver et al., 2016). Realistically, we concede that a truly robust understanding of statistical analyses is developed through a continued statistical education beyond the introductory course. While this is a given for the future statisticians in our classrooms, our discussions with colleagues from the focus group made it clear that we need to provide consistent motivation, starting in the earliest service courses, on how statistical foundations -- such as a correct understanding of p-values -- are valuable across all scientific fields. We agree with Harradine, Batanero, and Rossman (2011) who argue that with students understanding of inference, “the underpinning ideas need to be developed over years, not weeks."

As statistics educators, we need to recognize the potential of our service courses to inform the next wave of scientists and scholars about the correct interpretation and fundamental use of p-values. Statisticians should also strive to reinforce the value that statistical analyses bring to general scientific inquiry by actively engaging in conversation with our peers in other scientific professions.

5. Acknowledgements

We thank our collaborators -- Drs. Tom Crist, Joe Johnson, Jonathan Levy, Hank Stevens, and Rose Marie Ward -- for representing their respective scientific disciplines and providing insights in the curricular focus group discussed above. We also thank the anonymous referees and associate editor for the helpful suggestions they provided on the first version of this paper.

6. References

1. ASA/MAA JCUS (2014). “Qualifications for Teaching an Introductory Statistics Course”. American Statistical Association and Mathematical Association of America Joint Committee on Undergraduate Statistics. [amstat.org/asa/files/pdfs/EDU-TeachingIntroStats-Qualifications.pdf].
2. Batanero, C. (2000). “Controversies around the role of statistical tests in experimental research”. *Mathematical thinking and learning*, *2*(1-2), 75-97.
3. Berry, D. A. (2016). “P-Values Are Not What They’re Cracked Up to Be”. Online Discussion: ASA Statement on Statistical Significance and P-values. *The American Statistician*, 70(2), 1-2.
4. Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G.H., Velleman, P., Witmer, J. and Wood, B (2016). “Guidelines for assessment and instruction in statistics education (GAISE) college report 2016”. American Statistical Association. [amstat.org/education/gaise].
5. Cobb, G.. (2016). “ASA Statement on P-Values: Two Consequences We Can Hope For”. Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values. *The American Statistician*, 70(2), 1.
6. Cobb, G. W. (2007). “The introductory statistics course: A Ptolemaic curriculum?”. *Technology Innovations in Statistics Education*, 1(1). [escholarship.org/uc/item/6hb3k0nz]
7. DelMas, G., Joan, G., Ooms, A., & Chance, B. (2007). “Assessing Students’ Conceptual Understanding After a First Course in Statistics”. *Statistics Education Research Journal*, *6*(2).
8. Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process.* New York: D.C. Heath.
9. Falk, R., & Greenbaum, C. W. (1995). “Significance tests die hard: The amazing persistence of a probabilistic misconception.” *Theory & Psychology*, 5(1), 75-98.
10. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations”. *European journal of epidemiology*, 31(4), 337-350.
11. Goodman (2016). “The Next Questions: Who, What, When, Where, and Why?”. Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values. *The American Statistician*, *70*(2), 1-2.
12. Garfield, J., & Ben-Zvi, D. (2008). *Developing students’ statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.
13. Harradine, A., Batanero, C., & Rossman, A. (2011). *Students and Teachers Knowledge of Sampling and Inference in Teaching Statistics in School*

Mathematics - Challenges for Teaching and Teacher Education. A Joint ICMI/IASE Study: The 18th ICMI Study. New York: Springer
14. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.
15. Millar (2016). “ASA Statement on P-values: Some Implications for Education.” Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values. *The American Statistician*, *70*(2), 1.
16. Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Psychology Press.
17. Open Science Collaboration. (2015). “Estimating the reproducibility of psychological science.” *Science*, *349*(6251), aac4716.
18. Rossman, A. J. (2008). “Reasoning about Informal Statistical Inference: One Statistician's View.” *Statistics Education Research Journal*, *7*(2).
19. Stangl, D. (2016). “Comment”. Online Discussion: Official Supplement to ASA Statement on Statistical Significance and P-values. *The American Statistician*, *70*(2), 1.
20. Thompson, P. W., Saldanha, L. A., & Liu, Y. (2004). “Why statistical inference is hard to understand.” In *American Educational Research Association*.
21. Wasserstein, R. L., & Lazar, N. A. (2016). “ASA Statement on P-values: Context, Process, and Purpose.” *The American Statistician* 70:2, pages 129-133.
22. Weir, K. (2015). “A reproducibility crisis? The headlines were hard to miss: Psychology, they proclaimed, is in crisis.” *Monitor on Psychology*, *46*, 39.
23. Williams, A. M. (1999). “Novice students' conceptual knowledge of statistical hypothesis testing.” *Making the difference: Proceedings of the twenty-second annual conference of the mathematics education research group of Australasia*. 554-560.