

# Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly



Scott A. Cassidy, Ralitzia Dimova, Benjamin Giguère,  
Jeffrey R. Spence<sup>ID</sup>, and David J. Stanley

Department of Psychology, University of Guelph

Advances in Methods and  
Practices in Psychological Science  
2019, Vol. 2(3) 233–239  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2515245919858072  
www.psychologicalscience.org/AMPPS  
 SAGE

## Abstract

Null-hypothesis significance testing (NHST) is commonly used in psychology; however, it is widely acknowledged that NHST is not well understood by either psychology professors or psychology students. In the current study, we investigated whether introduction-to-psychology textbooks accurately define and explain statistical significance. We examined 30 introductory-psychology textbooks, including the best-selling books from the United States and Canada, and found that 89% incorrectly defined or explained statistical significance. Incorrect definitions and explanations were most often consistent with the odds-against-chance fallacy. These results suggest that it is common for introduction-to-psychology students to be taught incorrect interpretations of statistical significance. We hope that our results will create awareness among authors of introductory-psychology books and provide the impetus for corrective action. To help with classroom instruction, we provide slides that correctly describe NHST and may be useful for introductory-psychology instructors.

## Keywords

*p* values, null-hypothesis significance testing, statistical inference, quantitative literacy, quantitative psychology, open data, open materials

Received 8/22/18; Revision accepted 5/21/19

Statistical (adjective): relating to the use of statistics (“Statistical,” n.d.)

Significance (noun): the quality of being worthy of attention; importance (“Significance,” n.d.)

Statistical significance (noun): Assuming that the null hypothesis is true and the study is repeated an infinite number of times by drawing random samples from the same population(s), less than 5% of these results will be more extreme than the current result (based on Kline, 2013, p. 75).

It is difficult to argue that statistical significance is a simple or an intuitive idea. At face value, statistical significance seems straightforward because it combines two relatively common words to form a description.

However, a commonsense interpretation of statistical significance is misleading. As illustrated by the definitions just presented, the term *statistical significance* denotes much greater technical complexity than suggested by the aggregation of the respective definitions of *statistical* and *significance*.

Since its introduction nearly 90 years ago, null-hypothesis significance testing (NHST) has been the most widely used statistical approach to data analysis in psychology (Nickerson, 2000). Yet, despite its ubiquity, the history surrounding significance testing reveals that researchers misunderstand, misinterpret,

## Corresponding Author:

Jeffrey R. Spence, Department of Psychology, University of Guelph,  
50 Stone Rd. East, Guelph, Ontario, N1G 2W1, Canada  
E-mail: spencejr@uoguelph.ca

and misapply the technique with alarming regularity—a situation methodologists have long criticized and attempted to correct (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994; Nickerson, 2000). After close to a century of consistent corrections and explanations regarding how to interpret and use NHST correctly, incorrect interpretations and applications have proven to be rather resilient.

Where do these incorrect interpretations come from? And why do they persist? In this article, we present an exploratory investigation in which we attempted to gain some insight into these questions. Specifically, we examined if and how introductory-psychology textbooks define and explain statistical significance, in order to determine if the earliest pedagogical efforts may be contributing to misinterpretations.

## Understanding and Defining Statistical Significance

In psychology, NHST is typically used to determine if it is reasonable to state that a population-level effect size may not be zero (or another value specified as the null hypothesis; Spence & Stanley, 2018). Such a conclusion is arrived at through use of an index called the  $p$  value. Conceptually, to calculate a  $p$  value, the researcher must calculate an observed test statistic for the sample (e.g., an observed  $t$  value) and then compare this observed test statistic with a distribution of hypothetical test statistics. This distribution of hypothetical test statistics typically assumes that the population-level effect under investigation is zero. This assumption is referred to as the null hypothesis. The distribution of hypothetical test statistics can be thought of as being created by repeating the study under investigation a large number of times, using the same sample size and randomly sampling from the same population, when the population effect size is zero. A  $p$  value is obtained by comparing the observed test statistic with this hypothetical distribution. It specifically indicates the *proportion* of hypothetical test statistics that are equal to or more extreme than the test statistic obtained in the actual study. A particular test statistic is conventionally deemed to reach the status of “statistically significant” when its corresponding  $p$  value is less than .05. This indicates that the test statistic obtained (or a more extreme one) is unlikely when the starting assumption (the null hypothesis) is true. Consequently, researchers use the criterion of statistical significance to reject the null hypothesis. Unfortunately, some researchers incorrectly use a lack of statistical significance (i.e., a nonsignificant  $p$  value) to explicitly or implicitly “accept the null.” This is but one of many common errors researchers make when interpreting  $p$  values (cf. Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019).

## The Current Study and Its Context

NHST has a long history of being misunderstood by the very researchers who rely on it (Nickerson, 2000). In the 1960s, Nunnally (1960) referred to NHST as being “misused and misconceived” (p. 642), and Bakan (1966) stated, “The psychological literature is filled with misinterpretations of the nature of the test of significance” (p. 428). In 1996, the American Psychological Association formed a Task Force on Statistical Inference to “elucidate some of the controversial issues surrounding the applications of statistics including significance testing and its alternatives” (American Psychological Association, 2019). More recently, in 2016, the American Statistical Association issued a statement that noted, “While the  $p$ -value can be a useful statistical measure, it is commonly misused and misinterpreted” (Wasserstein & Lazar, 2016, p. 131).

Given the ubiquity and persistence of misinterpretations, we wanted to investigate how researchers are first exposed to NHST in psychology. Although many of the criticisms have focused on researchers, it is important to note that researchers all start out as students and learn NHST in the process of obtaining their degrees. As students, future researchers are taught by current research psychologists. As a result, it is perhaps not surprising that Haller and Krauss (2002) found that 100% of psychology undergraduates they sampled incorrectly interpreted statistical significance. These results are not too dissimilar from those for research psychologists in the same study, as 80% of methodology instructors and 90% of scientific psychologists made at least one error when identifying the correct meaning of a  $p$  value (Haller & Krauss, 2002). Similarly, Oakes (1986) found that 97% of scientific psychologists made at least one mistake when trying to identify the correct meaning of a  $p$  value. Psychology researchers may be learning to interpret statistical significance incorrectly quite early in their careers and are likely learning incorrect interpretations in their classes, from their instructors. Without knowledge of the pedagogical materials the students in Haller and Krauss’s study were exposed to, at least two explanations of their performance are possible: (a) The students were taught incorrect interpretations of statistical significance, or (b) they were taught correct interpretations but forgot them or supplanted them with incorrect interpretations.

In order to gain some insight into a possible source of misinterpretations of statistical significance in psychology, we decided to examine if and how NHST is introduced to students during their first psychology course. Introductory-psychology courses are where many future psychology researchers are first exposed to psychological research and the statistical approaches used to generate psychological knowledge. One of the

**Table 1.** Coding of the Fallacies in the Textbooks' Presentation of Statistical Significance (Based on Kline, 2009)

Number	Fallacy	Description
1	Odds against chance	Statistical significance means that the likelihood that the result is due to chance is less than 5%.
2	Local Type I error	Statistical significance means that the likelihood that a Type I error was committed is less than 5%.
3	Inverse-probability error	Statistical significance means that the likelihood that the null hypothesis is true is less than 5%.
4	Replicability	Statistical significance means that the probability of finding a statistically significant result in a replication is greater than 95%.
5	Validity	Statistical significance means that the probability that the alternative hypothesis is true is greater than 95%.
6	Meaningfulness	A finding of statistical significance confirms the alternative hypothesis and the research hypothesis.
7	Quality	A finding of statistical significance means the study was of good quality.
8	Other	Unclassifiable fallacy

main ways in which these introductory courses are structured and standardized is through textbook readings (Wandersee, 1988). Consequently, we chose to examine the presentation of NHST in introductory-psychology textbooks. We focused specifically on coding textbooks for the presence of common fallacies in definitions and explanations of statistical significance (e.g., odds-against-chance fallacy, inverse-probability fallacy, validity fallacy; Kline, 2009).

Table 1 provides an overview of some commonly identified fallacies in the interpretation of NHST. These fallacies are useful for classifying common interpretational mistakes and organizing them around key themes. For instance, failure to recognize that the null hypothesis is always assumed to be true is responsible for the odds-against-chance and inverse-probability fallacies. Specifically, because the null is assumed to be true, random sampling (i.e., chance) is the only explanation for any result. Also, if the null is assumed to be true, there is no likelihood of it being true; it is a given. Moreover, several of the fallacies are due to assigning probability to something other than hypothetical data. For example, the odds-against chance, local-Type-I-error, inverse-probability, and validity fallacies all assign probability to something other than hypothetical data (e.g., Type I error, the truth of the null or alternative hypothesis). The replicability, meaningfulness, and quality fallacies are examples of overextending the interpretational meaning of statistical significance to qualities and characteristics outside the scope of  $p$  values. They cannot predict the future, nor can they speak to the truth of a statement or premise or the quality of a study, any more than they can be used draw conclusions about how many outstanding parking tickets the researcher has.

## Disclosures

Our data and a list of the textbooks coded are available at the Open Science Framework (<https://osf.io/z7kq2/>). The files at the Open Science Framework also include the markdown document for the submitted version of this manuscript, which contains the R scripts for our summary statistics and graphs. Finally, we have also provided a PowerPoint file with scans of the coded textbook passages.

## Method

### Textbooks

Our sample consisted of 30 textbooks, including the best-selling textbooks<sup>1</sup> in the United States and Canada from 2017 to 2018. These textbooks were acquired by one of the authors (an introduction-to-psychology instructor), who contacted publishers requesting textbooks for his class and supplemented those textbooks with purchased ones. A full list of the 30 textbooks we coded is available at <https://osf.io/z7kq2/>.

### Content analysis

We coded each textbook for the presence or absence of a definition of statistical significance after examining the main text, sidebars, appendices, and glossary. We then used content analysis to code whether the textbook's definition contained any commonly known fallacies. Many of the textbooks had definitions in multiple locations. In such cases, either exactly the same definition was used in all locations or a version of the same definition was used with slight syntactic variation. For the analyses we report here, we coded one definition

per textbook even if there were multiple definitions. For the purpose of reproducibility, we chose the definition that was coded by prioritizing the definition that was most central to a reader's experience. Specifically, we coded a definition in the main text in preference to one in a sidebar, a definition in a sidebar in preference to one in an appendix, and a definition in an appendix in preference to one in a glossary. For example, if a textbook presented a definition in all four areas, we coded the main text's definition, and if a textbook had definitions in both a sidebar and a glossary, we coded the definition in the sidebar.<sup>2</sup>

It was also common for textbooks to provide an explanation of statistical significance. As a result, in addition to definitions, we coded explanations for the presence of fallacies. It was straightforward to differentiate explanations from definitions because definitions were in boldface, italics, or a different-color font. Our coding of definitions and explanations was based on common and known fallacies (Kline, 2009). Each explanation and definition was coded as "true" or "false" for the presence of each fallacy.

Coding was conducted by three of the authors, all of whom had experience teaching NHST to psychology students. Each coder coded the content of all the books independently, and the three met to discuss coding inconsistencies in an attempt to reach consensus. The coders all coded the same files, which contained electronic images of the content of the textbooks.

## Results

### ***Did the textbooks contain definitions, and did the definitions contain fallacies?***

Figure 1 shows that 25 of the 30 textbooks provided a definition of statistical significance. Of these 25 definitions, 22 contained a coded fallacy, and 3 did not.

As shown in Figure 2, of the 22 definitions that contained a fallacy, 20 contained the odds-against-chance fallacy, 1 contained the validity fallacy, and 1 contained the meaningfulness fallacy. In an additional case, the error was not classifiable according to our list of fallacies. All the coders agreed that the definition was incorrect, and it was coded as "other."<sup>3</sup>

### ***Did the textbooks provide explanations, and did the explanations contain fallacies?***

Figure 1 shows that 28 of the 30 textbooks provided an explanation of statistical significance. Of these explanations, 25 contained a fallacy.

Figure 2 presents the frequency of the types of fallacies in the explanations of statistical significance. The odds-against-chance fallacy was present in 24 of the 28 explanations. The inverse-probability fallacy was in 2 explanations, and the meaningfulness fallacy was in 10 explanations. The same textbook whose definition was not classifiable according to our list of fallacies provided an explanation that all the coders agreed was incorrect and unclassifiable. It was coded as "other."

## Overall results

Of the 28 books that presented a definition, explanation, or both, 25 contained at least one fallacy (Fig. 1). Thus, 89% of the textbooks that presented NHST contained at least one fallacy.

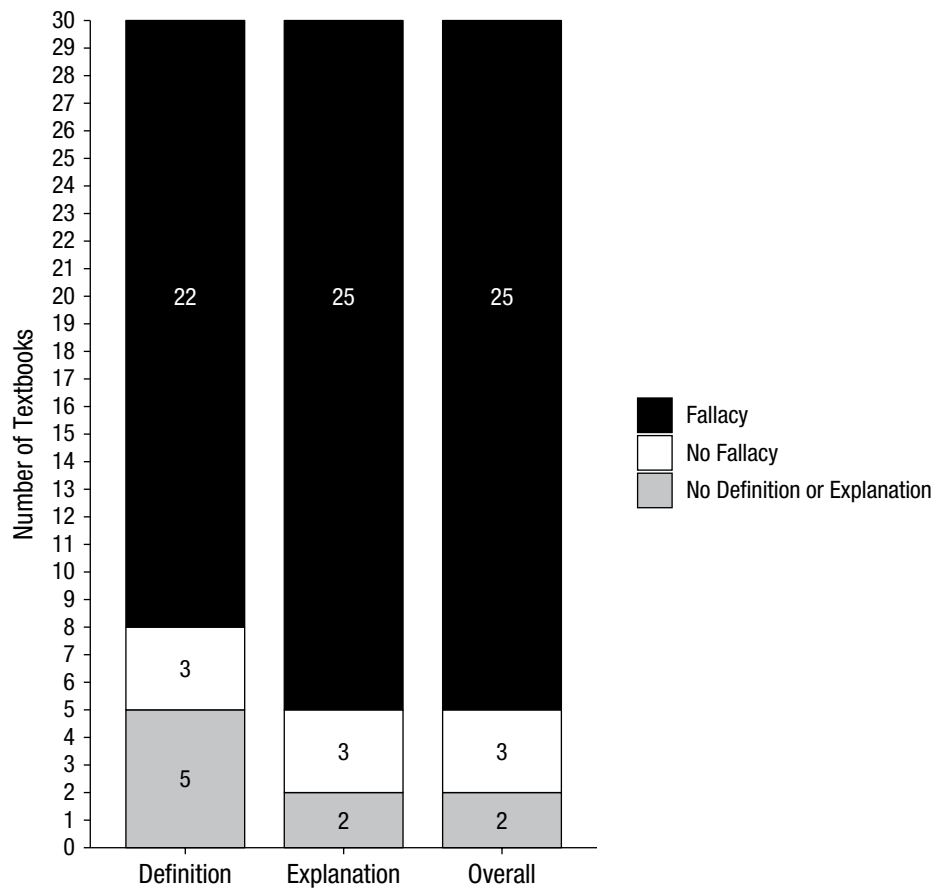
## Discussion

We found that most definitions and explanations of statistical significance in introduction-to-psychology textbooks contained common fallacies. Overall, 89% of the textbooks (i.e., 25 of 28) incorrectly defined or described statistical significance. Of the 25 definitions, 88% (i.e., 22) contained a fallacy. The most common was the odds-against-chance fallacy, which was evident in 80% (i.e., 20 of 25) of the definitions. Of the 25 explanations, 89% (i.e., 25) contained a fallacy. The fallacies included in these explanations were the odds-against-chance, inverse-probability, and meaningfulness fallacies, as well as an unnamed fallacy. The most common fallacy in the explanations (as in the definitions) was the odds-against-chance fallacy, which was found in 86% (i.e., 24 of 28) of the explanations.

Overall, these results suggest that students' misinterpretations of statistical significance may not be the result of their failing to remember the correct interpretation they were taught. Instead, students may be accurately recalling incorrect pedagogy.

Although the introductory-psychology textbooks were found to present a variety of fallacies, they were quite consistent in presenting the odds-against-chance fallacy. This uniformity may suggest that these textbooks' authors drew from similar sources when formulating their definitions. It may also suggest that the odds-against-fallacy is a particularly tempting fallacy in the context of trying to communicate statistical significance to a novice audience.

Our data also suggest that the persistent and ongoing efforts to correct inaccuracies in the interpretation of statistical tests have not been effective in reaching authors of introductory-psychology textbooks. By extension, our investigation provides some insight into



**Fig. 1.** Frequency of definitions and explanations of statistical significance in the 30 textbooks. The “Definition” and “Explanation” bars show the number of textbooks without a definition or explanation, the number with a correct definition or explanation, and the number with a fallacious definition or explanation. The “Overall” bar indicates the number of textbooks lacking either a definition or explanation, the number with a definition or explanation (or both) with no fallacies, and the number with at least one fallacy.

how deeply rooted incorrect interpretations of statistical significance are in psychology. Our results may not be surprising given the previously reported prevalence of misunderstandings among academic psychologists and instructors (Haller & Krauss, 2002; Oakes, 1986).

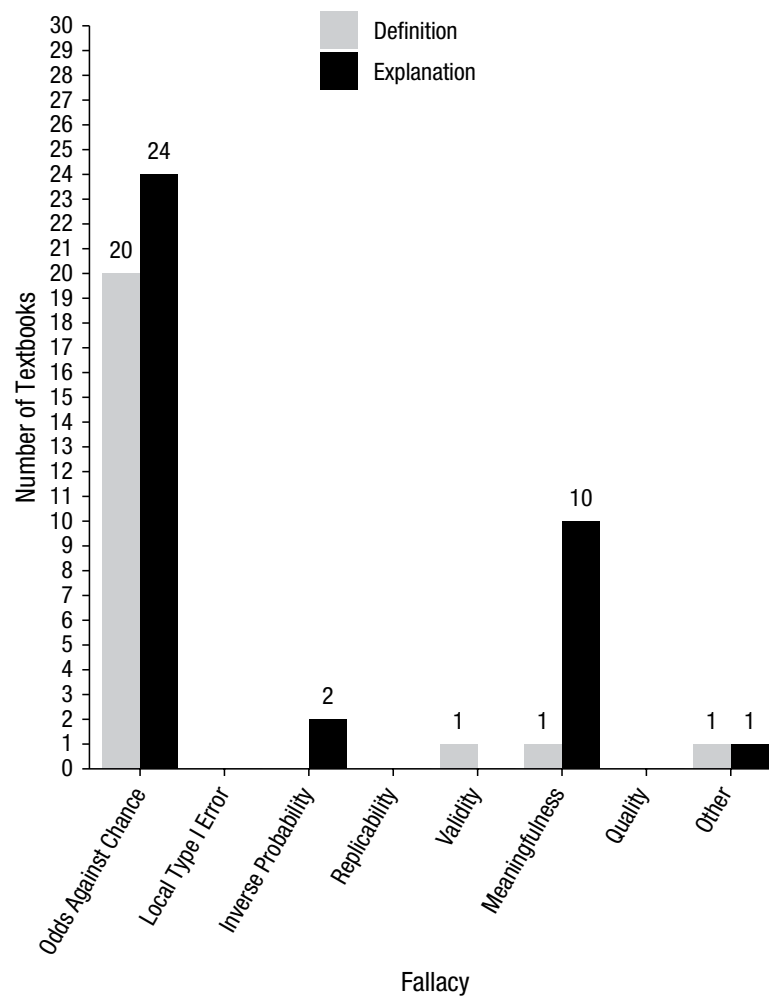
Given the role that textbooks play in early undergraduate learning and pedagogical structure (Wandersee, 1988), our findings point to a possible source of psychologists’ widespread misinterpretation of NHST. One option moving forward is simply to encourage authors to correct the textbook passages that contain fallacies. In the process of coding the textbooks, we noted that in many cases the definitions and explanations could be substantially improved by simply indicating that NHST begins with the assumption that the null hypothesis is true (and then removing passages inconsistent with this fact).

A second option for improving introductory-psychology textbooks is to encourage authors to completely remove discussions of statistical significance from

them. These textbooks could discuss the findings of studies in general terms, without mentioning “statistical significance.” With this approach, an introduction to NHST could be delayed until students’ first statistics course. If publishers are reluctant to remove content regarding statistical significance, we have provided some teaching materials that instructors may find useful to explain statistical significance and avoid fallacies in their presentation of the topic (see <https://osf.io/z7kq2/>).

However, it is possible that students learn misinterpretations from a variety of sources (Murden & Gillespie, 1997). In that case, one potential area for future inquiry would be to investigate the role that other pedagogical material (e.g., lectures) may play in promoting misinterpretations of statistical significance.

We believe that textbooks should provide correct information and hope that our results will create awareness among authors of introductory-psychology books and provide the impetus for corrective action.



**Fig. 2.** Frequency of each of the coded fallacies in the textbooks' definitions and explanations of statistical significance.


### Action Editor

Simine Vazire served as action editor for this article.

### Author Contributions

Order of authorship was determined alphabetically by last name. J. R. Spence generated the idea for this study. B. Giguère acquired the textbooks. B. Giguère, J. R. Spence, and D. J. Stanley coded the data. J. R. Spence and D. J. Stanley wrote the manuscript, and S. A. Cassidy and R. Dimova helped code a previous version of the manuscript.

### ORCID iD

Jeffrey R. Spence  <https://orcid.org/0000-0002-2652-3307>

### Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

### Open Practices



Open Data: <https://osf.io/z7kq2/>

Open Materials: <https://osf.io/z7kq2/>

Preregistration: no

All data and materials have been made publicly available via the Open Science Framework and can be accessed at <https://osf.io/z7kq2/>. The complete Open Practices Disclosure for this article can be found at <http://journals.sagepub.com/doi/suppl/10.1177/2515245919858072>. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.

### Notes

1. We would like to thank Kimberley Veevers, an executive portfolio manager at Pearson, for providing us with this information.

2. One textbook's glossary (Feldman, 2019) contained two separate definitions. One of these definitions was used in the body of the text, whereas the other appeared only in the glossary. In this case, we coded the definition that appeared in the body of the text and ignored the other definition.

3. Of the 25 definitions coded, 18 were presented in the main text, 3 in sidebars, 4 in an appendix, and none in a glossary.

## References

- American Psychological Association. (2019). *Task Force on Statistical Inference*. Retrieved from <http://www.apa.org/science/leadership/bsa/statistical/>
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437.
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378–399.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997–1003.
- Feldman, R. S. (2019). *Understanding psychology* (14th ed.). New York, NY: McGraw-Hill.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1–20.
- Kline, R. B. (2009). *Becoming a behavioural science researcher: A guide to producing research that matters*. New York, NY: Guilford Press.
- Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- Murden, T., & Gillespie, C. S. (1997). The role of textbooks and reading in content area classrooms: What are teachers and students saying? In W. M. Linek & E. G. Sturtevant (Eds.), *Exploring literacy* (pp. 87–96). Pittsburg, KS: College Reading Association.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641–650.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.
- Spence, J. R., & Stanley, D. J. (2018). Simple, concise, and not wrong: In search of a short-hand interpretation of statistical significance. *Frontiers in Psychology*, 9, Article 2185. doi:10.3389/fpsyg.2018.02185
- Statistical. (n.d.). In *Google dictionary online*. Retrieved from <https://www.google.com/search?q=Dictionary#dobs=statistical>
- Significance. (n.d.). In *Google dictionary online*. Retrieved from <https://www.google.com/search?q=Dictionary#dobs=significance>
- Wandersee, J. H. (1988). Ways students read texts. *Journal of Research in Science Teaching*, 25, 69–84.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on  $p$ -values: Context, process, and purpose. *The American Statistician*, 70, 129–133.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond " $p < 0.05$ ." *The American Statistician*, 73(Suppl. 1), 1–19.