Effects of a simulation-based training on students conceptual understanding of the Binomial test

Karin Binder[1], Stephan Schnaitmann[2] and Tim Erickson[3]
[1]Paderborn University, Germany; [2]LMU Munich, Germany; [3]Epistemological Engineering, U.S.A.
Karin.Binder@uni-paderborn.de

Significance tests are used intensively in quantitative empirical research and are also taught at schools and universities. However, even experts and statistics lecturers are subject to misconceptions when interpreting significance tests. Our study focuses on the binomial test, and examines the extent to which a refresher course that focuses on a simulation-based approach (using CODAP) is more conducive to learning than typical education on the binomial test that focuses on calculations. While the conceptual knowledge in the experimental group with simulations improved slightly more than in the control group, the students in the control group showed more improvement in procedural knowledge. However, the pre-test performance was weak overall and only a slight increase was observed in both groups after the 100 minutes of training. A comprehensive development of understanding hypothesis testing is important in teaching, and the results suggest that this cannot be sufficiently improved by our short training session.

INTRODUCTION

Significance tests are a difficult topic for students and are associated with many misconceptions. Even in scientific articles, these tests are often applied like a recipe without sufficient understanding of the underlying concept of significance testing (Gigerenzer, 2004). We conjecture that a simulation-based and software-supported approach can help to develop that conceptual knowledge on significance testing (Chandrakantha, 2020; Podworny, 2018; Rößner, Binder & Ufer, 2025). Therefore, we present a quantitative empirical study with 208 students in which half of the participants had a refresher course on binomial tests with a strong emphasis on calculations and the other half of the participants had a refresher course that focused slightly more on promoting conceptual understanding. Both groups of participants had prior knowledge on binomial tests, because they have already learned this content in school (usually not simulation-based and with a strong focus on procedural knowledge, which in this paper we take to mean the ability to apply the traditional procedure and calculate values like a correct probability of a Type II error).

THEORETICAL BACKGROUND

A p-value smaller than a predetermined significance level—and therefore a significant test result, for example in a binomial test—means that the probability of obtaining these data (or even more extreme data), under the assumption that the null hypothesis is true, is small. However, there are a lot of typical misconceptions in the interpretation of small or large p-values. One famous misconception is the assumption that the results of significant tests can be seen as clear proof (Oakes, 1986; Haller & Krauss, 2002), which is of course not true.

The inverse probability fallacy is also very well known. Here it is assumed that the probability that the null hypothesis (or the alternative) is correct (or incorrect), assuming the data, is now known on the basis of the p-value (Oakes, 1986; Haller & Krauss, 2002). Furthermore, some people commit the replication fallacy: If you repeat the experiment many times, you would obtain a significant result in (1 – p) % of occasions, (Gigerenzer, Krauss & Vitouch, 2004, Herrera-Bennett et al., preprint).

Misinterpretations of significant results in scientific articles have even prompted the American Statistical Association (ASA) to publish an official statement on typical errors in the use of significance tests (Wasserstein & Lazar, 2016). Other journals even prohibit reporting empirical evidence based on p-values (e.g., the journal Basic and Applied Social Psychology, Trafimow & Marks, 2005). They suggest—as already recommended by the American Psychological Association (APA, 2010)—that researchers use confidence intervals as an alternative (Cumming, 2012), even if these are also associated with corresponding misconceptions (Hoekstra et al., 2014, Herrera-Bennett et al., preprint).

Inference methods that are based on simulations are grounded in what Cobb (2007) describes as "randomize, repeat, reject". Frequently it has been suggested to teach various concepts of statistical literacy with the help of simulations (Chance et al., 2022; Estrella, 2025; Jamie, 2002; Zavez & Harel,

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*.

2025). Case et al. (2019) and Chandrakanta (2020) suggest to use physical or computer simulations to estimate p-values for a better understanding.

RESEARCH QUESTION

Does a simulation-based approach in a refresher course on binomial testing with CODAP help to better understand binomial tests compared to a training that focuses slightly more on procedures?

METHOD

Participants and design of the study

371 students participated voluntarily in preparatory refresher course for the German "Abitur," which is the final exam in secondary school. They were randomly assigned into the control group or simulation group. Since some of the students did not participate in the pre-test or the post-test, we include only those students who participated the whole training and in both tests.

Therefore, this contribution presents a pre-test-post-test-study with 208 students (shortly before graduating school) in an experimental-control group design (see Table 1). All students had prior instruction on the binomial test in school. Within our workshop the students refreshed their knowledge of binomial tests. Since the binomial test is frequently part of the final exam, this topic was of special interest to the students.

Table 1. Design of the study with a control and experimental group and a pre- and post-test. Differences in the training courses are **bold.**

Pre-test (10 minutes)	**Training (100 minutes)**	**Post-test (10 minutes)**
8 items on conceptual knowledge (typical misconceptions) 1 item on the calculation of the probability of a Type II error in a binomial test (procedural knowledge)	**Control group (94 participants)** • The principle of hypothesis testing • Calculating binomial tests on **three different tasks** • Type I and Type II error	Identical to the pre-test
	Simulation group (114 participants) • The principle of hypothesis testing • **Simulation of a null hypothesis world with the help of CODAP** • Calculating binomial tests on **two different tasks** • Type I and Type II error	

In the pre- and post-test, we tested typical misconceptions (conceptual knowledge) like in Oakes (1986), and had the students solve a typical calculation task from the final exam (procedural knowledge). One task for conceptual knowledge was (presented in slightly abbreviated form): "True or false? The null hypothesis was rejected. It has thus been clearly proven that the null hypothesis is false." See Figure 1 for an example of a typical procedural task in the final exam.

Some researchers plan to test the null hypothesis: "Less than 90% of 18—30 year-olds have an instagram profile". They plan to collect a simple random sample of 200 people in that age group and use a 5% significance level.

They (correctly) calculate a decision rule and decide to reject the null hypothesis if more than 187 of the people studied have an instagram profile.

Calculate the probability of a type II error, assuming that in reality 95% of 18—30-year-olds have an instagram profile. State your answer in whole percentage points!

Correct Solution: 20 %

Figure 1. Pre-test task, which is similar to a typical task in the final exam.

Simulation-based training with the help of CODAP

Two training methods are compared. Both training courses contain pure calculations and both contain comprehension-based elements. However, the simulation group focuses more on understanding, and students are also allowed to simulate independently using CODAP, to better understand p-values and what a "significant test result" means. The control group has a stronger focus on calculations. For example, whereas the students in the control group see a static picture of a simulated null hypothesis world, the students in the simulation group had the chance to simulate a null hypothesis world on their own with the help of the Binomial simulator in CODAP (which was not available in a German version at that time, see Figure 2; Binder & Erickson, in preparation). The diagram shows 100 experiments in which 10 flips of a coin are considered. In 95 percent of the experiments 7 or fewer of the 10 flips show heads. Figure 3 shows another simulation of a null hypothesis world, in this case the simulation of a specific situation where the probability of a Type II error can be read off.

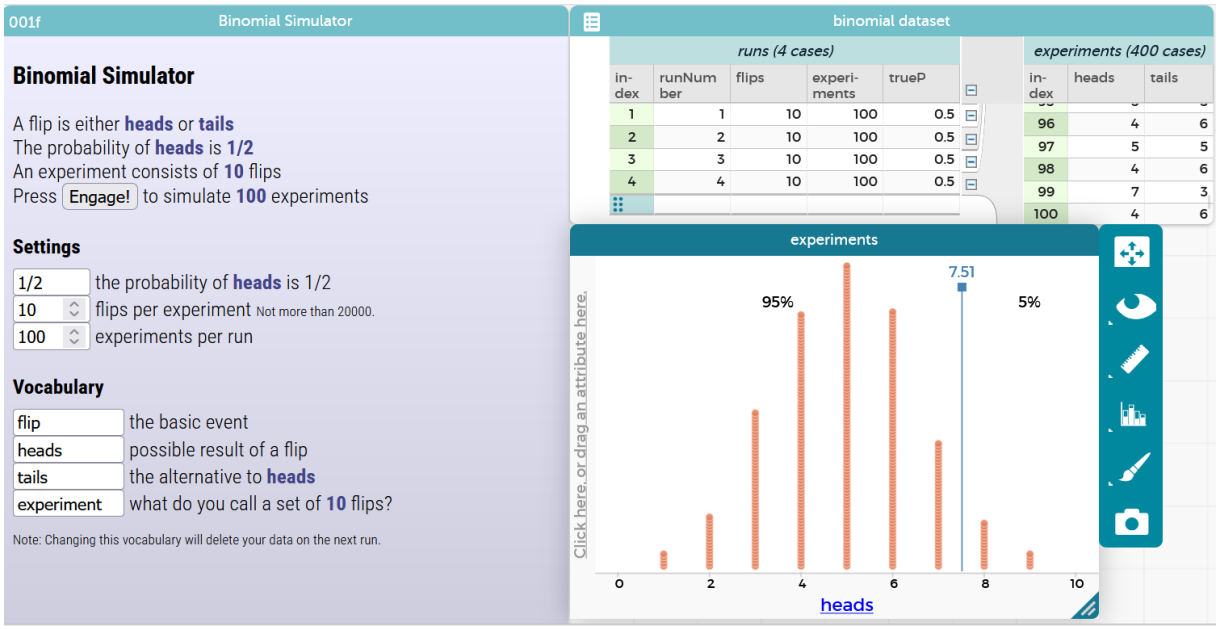

Figure 2. Students created this simulation on their own with the help of CODAP (using the Binomial simulator in the English version; www.codap.xyz).

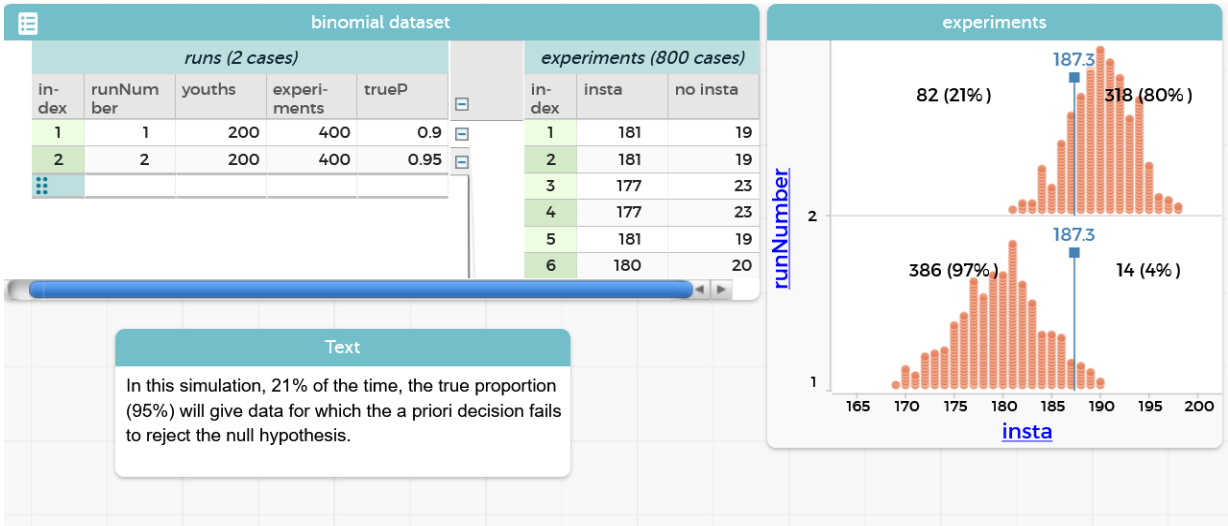

Figure 3. Simulation of a null hypothesis world, and the probability of a Type II error for a specific situation (Instagram context of Figure 1; it's the 21% value in the upper left of the graph).

Students in the simulation group spent about 45 of the 100 intervention minutes working with CODAP, about 15 of those working independently with the Binomial Simulator.

RESULTS

The pre- and post-tests each had a maximum score of 9 points for both sections together—conceptual (8 points) and procedural knowledge (1 point). As the students already had prior knowledge on binomial tests, we used the same 9 items in the pre-test and the post-test. In two of the items for conceptual knowledge, partial credit (i.e., 0.5 points) were possible.

The results of the pre-test show that although the students had already been taught about binomial tests and the exams were imminent, both groups had little knowledge of binomial tests. Mean scores were only 4.49 points out of 9 in total in both the control and simulation groups. This suggests that traditional instruction may not have sufficiently supported the development of a deep conceptual understanding of significance testing.

Regarding conceptual knowledge, the control group improved from 4.37 points in the pre-test to 5.22 points (out of 8 points) in the post-test, whereas the simulation group improved slightly more, from 4.36 points to 5.38 points (out of 8 points). For the "procedural" item where students had to calculate the probability of a Type II error in a binomial test, the proportion of students getting the item right in the control group increased from 12 percent to 35 percent, whereas this proportion increased in the simulation group from 14 percent only to 24 percent, showing that both groups still had problems calculating the correct result.

Figure 4 shows the pre- and post-test results (conceptual and procedural knowledge combined) for the control group and the simulation group. The students could only achieve 4.49 points in the pre-test. This performance increased to 5.63 (simulation group) and 5.57 (control group) points during the rather short training session of 100 minutes. Considering that the students should have already built up all the necessary knowledge on the binomial test from their normal lessons in school and that their school exams were shortly before, this does not speak well for the school training, nor does it represent a major improvement due to the intervention.

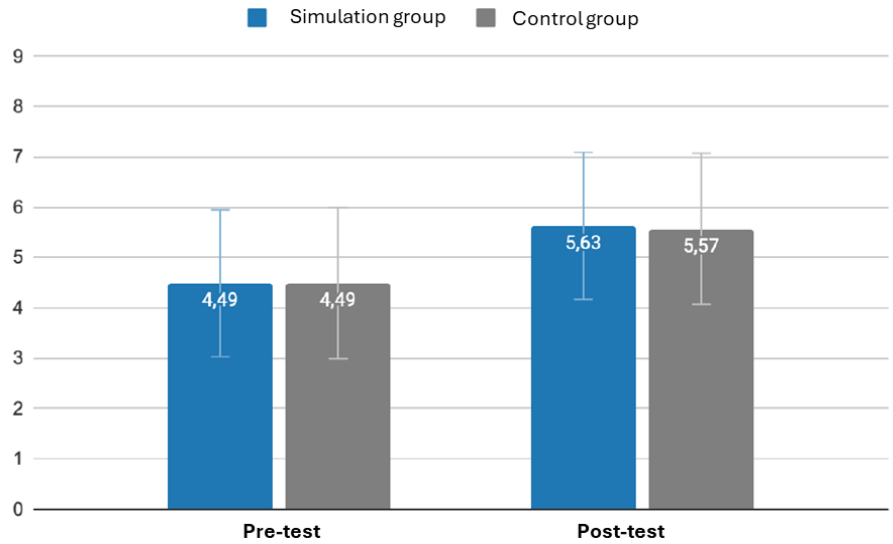

Figure 4. Results students achieved in the pre-test and the post-test, separately for the control group (grey) and the experimental group (blue).

Figure 5 A) shows how individual students' *conceptual knowledge* changed from the pre-test to the post-test. Negative results mean that students performed worse in the post-test compared to the pre-test. The improvement of the simulation group was better, but there is a large variability among the students in both groups.

Figure 5 B) shows that most students did not show any change in *procedural knowledge* as a result of the intervention. 82 percent of the students of the simulation group and 71 percent of the students in the control group received the same result in the post-test as in the pre-test. In the simulation

group only 14 percent of the students performed better in the post-test as in the pre-test, compared to 26 percent in the control group.

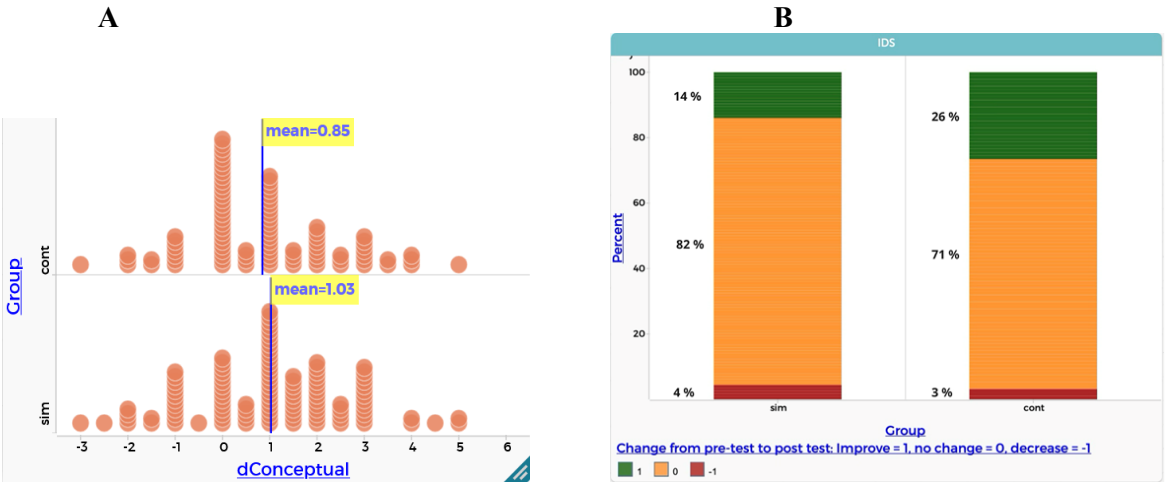

Figure 5. A) Change in *conceptual* knowledge score (dConceptual) from pre-test results to post-test in the simulation group and the control group; B) Proportion of students whose *procedural* knowledge has improved (green), remained the same (orange) or decreased (red).

The results of a linear mixed model indicate what could actually be expected from the descriptive results: For both the conceptual and procedural knowledge, students achieved significantly more points in the post-test compared to the pre-test. In the experimental group, the learning gain for *conceptual knowledge* was descriptive but not significantly higher (p=.41; see Table 2). However, in the control group, the learning gain for *procedural knowledge* was significantly higher (p=.02; see Table 3).

Table 2. Linear mixed model for predicting *conceptual knowledge* in the post-test. The results of the control group in the pre-test serve as the reference group. Dummy-variable PrePost: 0 pre-test, 1 post-test; Dummy-variable Group: 0 control group, 1 simulation group.

Covariates	**Estimate**	***SE***	***z***	***p***
Intercept	4.36	0.14	30.87	<0.001
PrePost	**0.86**	**0.16**	**5.28**	**<0.001**
Group (simulation group)	-0.01	0.19	-0.06	0.95
PrePost × Group	0.18	0.22	0.82	0.41

Table 3. Linear mixed model for predicting *procedural knowledge* in the post-test. The results of the control group in the pre-test serve as the reference group. Dummy-variable PrePost: 0 pre-test, 1 post-test; Dummy-variable Group: 0 control group, 1 simulation group.

Covariates	**Estimate**	***SE***	***z***	***p***
Intercept	0.11	0.04	2.73	<0.01
PrePost	**0.24**	**0.05**	5.15	**<0.01**
Group (simulation group)	0.03	0.06	0.52	0.60
PrePost × Group	**-0.15**	**0.06**	**-2.27**	**0.02**

DISCUSSION

In this experimental control group study, a pre-post design was used to investigate whether training in which students (with prior knowledge on binomial tests) are allowed to independently create simulations with CODAP is superior to training without these independently-created simulations. In the short training period of 100 minutes, although the mean increase was slightly better in the simulation group compared to the control group, there was a large variability among the students and therefore, no significantly higher learning gains in conceptual knowledge were observed in the simulation group. At the same time, a higher proportion of students in the control group improved their procedural knowledge. However, the training was very short and the differences in terms of content between the two training courses were small (large parts of the training sessions were absolutely parallel).

The interpretability of the results is limited by the short intervention and the lack of long-term results. Furthermore, the German students worked with an English version of the Binomial simulator (there was no German version of the Binomial simulator available at that time—but there is now). This could have had a negative impact on the benefits of the tool.

Case et al. (2019) suggest also to use physical simulations. We showed such kind of simulation with a shaker box, compare Figure 6, but students did not have the chance to use this kind of physical simulation on their own (compare Estrella, 2025).

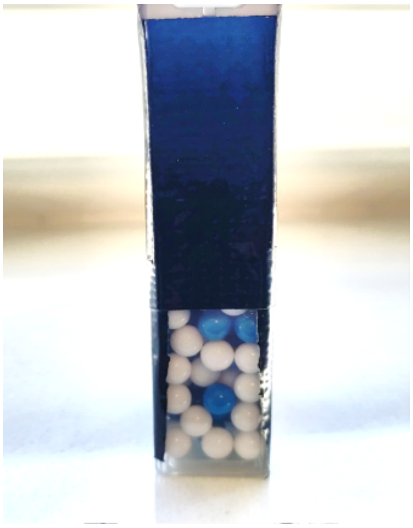

Figure 6. Physical simulation of the binomial test with the help of a shaker box.

Previous work in this area recommends looking at randomization tests to better understand the principle of hypothesis testing (Podworny & Biehler, 2022; Budgett & Wild, 2014). However, since the focus in the final exam is on binomial tests, the present study was limited to this type of test. It would be interesting to investigate whether a prior introduction to the principle of classical hypothesis testing (like binomial tests, t-tests) with the help of randomization tests would have further improved the understanding of the binomial test. Furthermore, it would be interesting to see if students would understand the whole procedure better if they had learned to calculate (empirical) p-values using the simulator (instead of learning the typical procedure to set up a decision rule and calculate a test statistic value and a rejection area). A further limitation of the study is that procedural knowledge was only measured with one task—only focusing on the Type II errors—and the conceptual knowledge only with single and multiple choice items.

The use of simulations in school lessons, including allowing students themselves to do the simulating, is a recurring demand. In this study, we wanted to take a first step and test whether such a simulation-based approach to teaching binomial tests can strengthen students' conceptual knowledge particularly well (compared to a more calculation-based approach). This may be possible with a more extensive intervention and, above all, if simulations are used to a greater extent during the introduction rather than in a crash course shortly before the exam.

REFERENCES

Binder, K. & Erickson, T. (2026). *Traditional tests through a randomization lens: treating p-values as data with CODAP* [Manuscript submitted for publication].

Budgett, S., & Wild, C. J. (2014). Students' visual reasoning and the randomization test. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. ISI/IASE.

Case, C., Battles, M., & Jacobbe, T. (2019). Toward an understanding of p-values: Simulation-based inference in a traditional statistics course. *Investigations in Mathematics Learning*, *11*(3), 195–206. https://doi.org/10.1080/19477503.2018.1438869

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal*, *21*(3), Article 4. https://doi.org/10.52041/serj.v21i3.6

Chandrakantha, L. (2020). Visualizing the p-value and understanding hypothesis testing concepts using simulations in R. *Electronic Journal of Mathematics & Technology*, *14*(3).

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). https://doi.org/10.5070/T511000028

Cumming, G. (2012). *Understanding The New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis*, Routledge. https://doi.org/10.4324/9780203807002

Estrella, Soledad (2025, July 28 – August, 2). *What does it mean to prepare our boys and girls to interact with Artificial Agents? Perspectives from early statistical education in today's classroom* [Plenary session]. 48th Conference of the International Group for the Psychology of Mathematics Education, Santiago, Chile. https://eventos.cmm.uchile.cl/pme48/program/

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 391–408). SAGE Publications.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1–20.

Herrera-Bennett, A. C., Heene, M., Lakens, D., & Ufer, S. (2020). *Improving statistical inferences: Can a MOOC reduce statistical misconceptions about p-values, confidence intervals, and Bayes factors?*. PsyArXiv. https://doi.org/10.31234/osf.io/zt3g9

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Jamie, D. M. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*(1). https://doi.org/10.1080/10691898.2002.11910548

Oakes, M. (1986). *Statistical significance: a commentary for the social and behavioural sciences*. Wiley.

Podworny, S. (2018). Students' Reflections About a Course for Learning Inferential Reasoning Via Simulations. In C. Batanero, & E. Chernoff (Eds.), *Teaching and Learning Stochastics: Advances in Probability Education Research* (pp. 333–349). Springer. https://doi.org/10.1007/978-3-319-72871-1_19

Podworny, S., & Biehler, R. (2022). The process of actively building a model for a randomization test–insights into learners' modeling activities based on a case study. *Mathematical Thinking and Learning*, *24*(4), 291–311. https://doi.org/10.1080/10986065.2021.1922837

Rößner, M., Binder, K., & Ufer, S. (2025). Simulationsbasiert Signifikanztests verstehen. *Mathematica Didactica, 48*(2), 1–22. https://doi.org/10.18716/ojs/md/2025.2296

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1)*,* 1–2. http://dx.doi.org/10.1080/01973533.2015.1012991

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Zavez, K., & Harel, O. (2025). Teaching Statistical Concepts Using Computing Tools: A Review of the Literature. *Journal of Statistics and Data Science Education*. https://doi.org/10.1080/26939169.2024.2445541