Synthesis of elements for conducting a randomization test

Síntesis de los elementos para la realización de un ensayo de aleatorización

Susanne Podworny

Paderborn University, Germany

Abstract

Randomization tests offer an access to inference statistics, which is regarded as particularly simple in the didactic literature. Above all, the logic of the inferential reasoning should become particularly clear. Nevertheless, in order to carry out a randomization test, some elements are needed that must be understood in order to successfully draw statistical conclusions. In this article, various elements from the literature are collected and compiled in order to create a scheme for the hand of learners to carry out a randomization test.

Keywords: Inference, randomization test

Resumen

Las pruebas de aleatorización ofrecen un acceso a las estadísticas de inferencia, lo que se considera particularmente sencillo en la literatura didáctica. Sobre todo, la lógica del razonamiento inferencial debería ser particularmente clara. Sin embargo, para llevar a cabo una prueba de aleatorización, se necesitan algunos elementos que deben ser comprendidos para poder sacar conclusiones estadísticas con éxito. En este artículo, se recogen y compilan varios elementos de la literatura, con el fin de crear un esquema para que ayude a los alumnos a llevar a cabo una prueba de aleatorización.

Keywords: Inferencia, prueba de aleatorización

1. Introduction

Inferential reasoning is a cornerstone on which the practice of statistics is based. Data and conclusions drawn from data play an important role in daily life. Computer-based methods and evaluations are part of the daily repertoire of statistics. Every day we encounter news in the media in which "a study has shown" or it is concluded that "the effect of A is B". Frequently, however, it is not mentioned that these results are subject to a certain degree of uncertainty.

The process of reasoning required for this is seen as an important skill of every adult. "Drawing inferences from data is part of everyday life and critically reviewing results of statistical inferences from research studies is an important capability for all adults" (Garfield & Ben-Zvi, 2008, p. 262). In order to be able to draw such conclusions or to reflect critically on drawn conclusions, it is necessary to understand the logic of inferential reasoning. However, many (learning) difficulties are known about this (e.g. Haller & Krauss, 2002; Thompson, Liu, & Saldanha, 2012).

To provide a simple introduction to inference statistics, Cobb (2007) and others (e. g. Rossman (2008)) pointed out the randomization test method, which should be central in a newly created curriculum "whose center is the core logic of inference" (Cobb, 2007, p. 11). In doing so, he is taking up a demand formulated almost ten years earlier in the fundamental article by Wild and Pfannkuch (1999).

> Statistics education should really be telling students something every scientist knows, 'The quest for causes is the most important game in town.' It should be saying: 'Here is how statistics helps you in that quest'. (Wild & Pfannkuch, 1999, p. 238)

Podworny, S. (2019). Synthesis of elements for conducting a randomization test. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística.* Disponible en www.ugr.es/local/fqm126/civeest.html

Since the suggestions of Cobb (2007) and Rossman (2008) to use randomization tests to introduce inference statistics, some curricula and learning units have emerged that build on them. In addition, there are a few empirical studies that investigate how learners perform a randomization test.

This article synthesizes the various elements found in the literature to be addressed in a randomization test by novices. For this purpose, the diverse literature is reviewed and synthesized in order to comply with the suggestions of Cobb and Rossman.

2. Didactic aspects of the randomization test method

A randomization test is a non-parametric procedure almost without formal calculations because of the use of computer-based simulation. This makes it more flexible than traditional statistical tests and more intuitive to understand for beginners (Pfannkuch & Budgett, 2014; Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012). Randomization testing allows conclusions to be drawn from data, even from small samples or non-random collection methods, as is often the case in empirical research (Zieffler, Harring, & Long, 2011).

An advantage of the randomization test is that the design of an experiment is represented by the inference procedure. This makes it more accessible for beginners.

> One advantage of this procedure [simulation-based randomization tests] for introducing introductory students to the reasoning process of statistical inference is that it makes clear the connection between the random assignment in the design of the study and the inference procedure. It also helps to emphasize the interpretation of a p-value as the long-term proportion of times that a result at least as extreme as in the actual data would have occurred by chance alone under the null model. (Rossman, 2008, p. 10)

An important component of the simplicity of the approach is the use of stochastic simulations (Batanero & Borovcnik, 2016). Simulations do not require formal calculations and therefore can focus on the logic of reasoning. In addition to the content-related advantages mentioned by Rossman (2008), randomization tests offers a further advantage, because they reduce the cognitive load (Chandler & Sweller, 1991) of learners:

> Norm-based and randomization methods have the same reasoning process but the norm-based methods rely on many invisible concepts behind mathematical manipulations. The randomization method will decrease cognitive load by reducing the number of concepts that need to be activated simultaneously. (Pfannkuch et al., 2011, p. 911)

Finally, a third advantage is mentioned at statistical level. In general, statistical significance is a difficult concept for learners (e.g. Batanero, 2000; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002), but some authors also see randomization tests as an easier approach. Chance and Rossman (2006, p. 5) emphasize, for example, that "a randomization test can lead students to a deeper understanding of the concepts of statistical significance and p-value" and Holcomb, Chance, Rossman, Tietjen, and Cobb (2010) even say that there is nearly no need for prior knowledge to carry out a simulation-based randomization test.

While the elaborated perspectives highlight the simplicity of the method and the low entry to statistical inference, for example Batanero (2000) has a critical look at statistical testing in general and difficulties in teaching about it.

3. Towards a scheme for randomization testing

Conducting a randomization test requires certain procedures and the use of certain elements. In this section, I will present the sequence of steps and their description in literature[1]. At the end of the section, a general scheme will be developed that can be used to structure the learners' process of randomization testing.

3.1. Normative approaches to randomization tests

Cobb (2007) has proposed a three-step scheme consisting of the "three Rs of inference: randomize, repeat, reject" (Cobb, 2007, p. 12) to elementarize a randomization test. The first step of *randomize data production* is to describe how the observed data of an experiment may be collected in order to check which conclusions may be drawn later (see also Ramsey & Shafer, 2013). Behind the second step *repeat by simulation to see what's typical*, which appears to be simple with the explanation "randomized data production lets you re-randomize, over and over, to see which outcomes are typical, which are not" (Cobb, 2007, p. 12), stands the entire creation of a simulation and thus of a null model, which must be expressed in a null hypothesis. For the user of a randomization test, this second step therefore involves much more (and more difficulties) than Cobb initially seems to briefly present in this one sentence. Finally, in the third step *reject any model that puts your data in its tail*, a conclusion must be drawn. The most important thing in this formulation is "reject", which implies that a model can be rejected, but not necessarily confirmed. Cobb uses a metaphor commonly used in American English for the *p*-value. In the "tail" or at the edge of a distribution are the results that are just as extreme or even more extreme than the observed value, i.e. the results that the *p*-value includes. As a basic framework for the introduction to hypothesis testing, Cobb's scheme seems to be well suited, the logic of inferential reasoning is well represented, but on the other hand, the whole process is very briefly summarized.

Building on the ideas of Cobb (2007), Rossman, Chance, Cobb, and Holcomb (2008) have developed a number of modules on randomization tests to provide access to inferential reasoning. In order to make it easier for learners to access this way of thinking, they propose a four-step scheme that clarifies the logic. In addition to the four main steps "Observed Data, Null Model, Statistical Test and Scientific Inference" (Rossman et al., 2008, p. 6f), there are between two and five explanatory sub-steps, each of which provides further help in the form of questions or instructions on what exactly to do. These explanatory sub-steps seem to be very helpful for the use of learners, as it becomes very clear which steps are to be carried out. Thus, this scheme seems to be suitable as a direct template for use in a teaching situation.

Tintle, VanderStoep, and Swanson (2009) have developed a complete curriculum that introduces statistical inference through randomization tests and focuses on randomization tests based on Cobb's demand. They first develop a six-step scheme for the general statistical investigation process, which can be seen in Figure 1, and is appropriate for randomization tests as well. This scheme is strongly reminiscent of the PPDAC cycle of Wild and Pfannkuch (1999) and, like the PPDAC cycle, it is cyclical.

[1] The following description is abridged from my dissertation (Podworny, 2018).

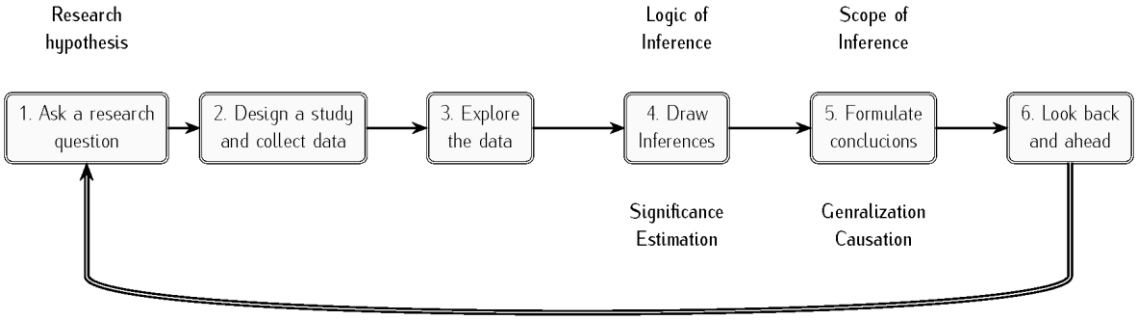

Figure 1. The "six step statistical investigation process" of Tintle et al. (2009, p. 2, own representation)

In contrast to the previous schemes, the authors here begin with a research question from which all further steps result. In the sense of the PPDAC cycle (Wild & Pfannkuch, 1999), the complete run-through of a cycle is stimulated. The schemes considered so far ultimately start with the performance of a randomization test if an experiment has already been carried out. A desirable situation for a learning situation would of course be to start with a research question and to carry out all further steps by oneself. In reality, however, the schemes that start two steps later, i.e. with the exploration of existing data, seem to be much more frequent and can be realized in the classroom with much less effort by using existing data.

Another framework for the realization of a randomization test can be found in Biehler, Frischemeier, and Podworny (2015) where a distinction is made between three "worlds" which are embedded in each other and which should be addressed in the respective step.

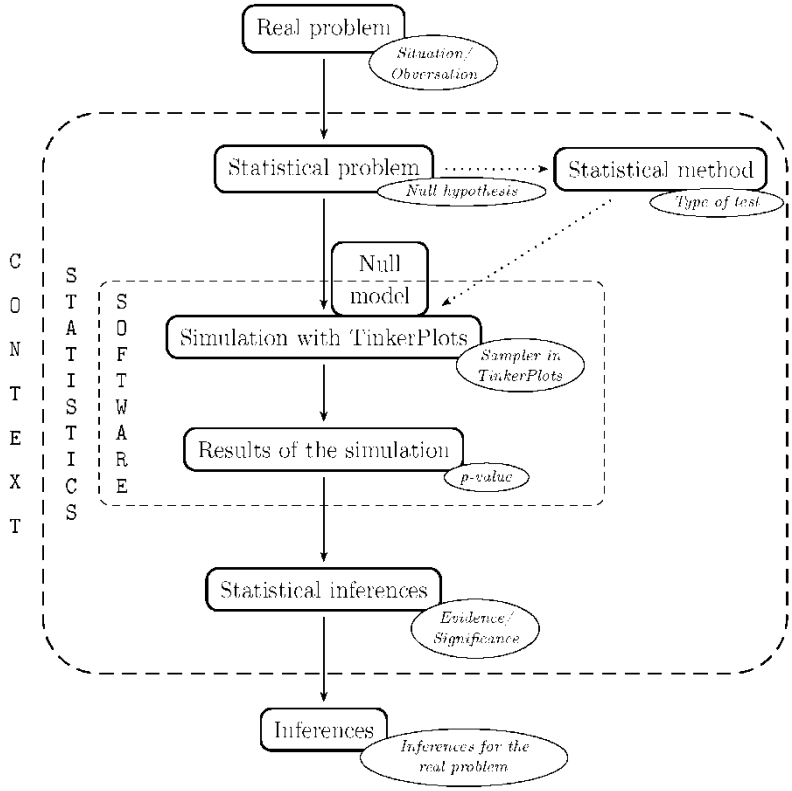

Figure 2. Framework for randomization testing (taken from Biehler et al., 2015, p. 139)

The first step in Figure 2 is the “observation” in the context world, which is followed in the statistics world by the “null hypothesis” and next, within the software world “sampler in tinkerplots” that corresponds to the null model. it remains in the software world with “p-value”, switches back to the statistics world with “evidence/significance” and finally turns to the context world with “Inferences for the real problem”. This embedding in the three worlds can be a useful addition to the previous schemes since the steps are quite alike.

3.2. Selected results of empirical research

Specific analyses of learners performing a randomization test are still rare. Budgett, Pfannkuch, Regan, and Wild (2012) conducted a half-day session on randomization tests with students in the last year of high school and with students in the first year of college. Of these, they conducted interviews with ten selected participants who performed a randomization test with the VIT software (*VIT: Visual Inference Tools*, https://www.stat.auckland.ac.nz/~wild/VIT/). For the interviews they used the example of the “fish oil and blood pressure study” (Budgett et al., 2012, p. 5) and investigated some selected aspects, e.g. which possible explanations for observed differences were given by learners and how the VIT software supported the implementation of a randomization test. They highlight the importance of giving possible explanations for observed differences in the experiment and using these explanations for the further reasoning process. Using the software was not a difficulty for learners in this study.

Biehler et al. (2015) have investigated which steps preservice teachers for mathematics at university courses can take successfully during a randomization test along three worlds (context, statistics, software) in which learners move. Above all, they identified learning difficulties in establishing the null hypothesis, its transfer into a simulation model and in finding the *p*-value. The use of the software (in this case TinkerPlots), though, was not a problem. However, they conclude that the relationship between the statistical world and the context world should be established more strongly in a teaching unit, as this was a hurdle for the participants of their study.

Noll and Kirin (2017) used the framework of Biehler et al. (2015) to investigate how learners link the null hypothesis with the sampler of TinkerPlots and how they reason with it. The authors report that the required random assignment of the experimental units to the new groups represented a difficulty in understanding. For this, there were also problems to transfer the independent assignment of observed values to two groups into the simulation model. In the end, Noll and Kirin (2017) insist that more research is needed on how learners interpret simulation models (here with the TinkerPlots sampler).

Justice, Zieffler, Huberty, and delMas (2018) took up on the research of Noll and Kirin (2017) and examined the argumentation process of four teachers in an AP statistics program regarding different simulation models and the related data generating process. The authors suggested some questions about the data generating process that can promote understanding, such as “‘is it essential that it happens in the same order?’ or, when the sampler devices are swapped, ‘will this affect the resulting distribution of statistics?’” (Justice et al., 2018, p. 10). One of their main outcomes is that the participants of their study regarded decision-making as the primary goal of statistical reasoning. As a further result, they formulated that participants valued the design of the original experiment in order to transfer it into a simulation model. However, if they did not understand that the design was needed to model variation, it hindered the ability to

draw statistical conclusions. Justice et al. (2018) highlighted the need to refer to important elements of the context for a problem in the reasoning process to understand the randomization test method.

3.3. Synthesis of elements for randomization tests

In the schemes and studies presented here, certain elements can be found for conducting a randomization test that seem to be necessary for the reasoning process. Some of them are set up as normative steps, others are used as evaluation categories. In Table 1, the various elements are extracted and arranged for a synopsis.

Table 1. Synopsis of elements for conducting a randomization test

Element	Description	Reference
Ask or reconstruct a *research question*	The question that led to the experiment is named.	Tintle et al. (2009); Wild and Pfannkuch (1999)
Explain the *random assignment* in the experiment	The design of the experiment is determined or explained retrospectively with regard to whether a random allocation of experimental units to groups has actually taken place and what meaning this has with regard to the randomization test	Budgett et al. (2012); Justice et al. (2018)
Analyse *observed data*	The observed data of the experiment are analyzed. For example, a group comparison based on the mean values or a comparison of certain proportions can take place and be noted as $X_{observed}$. According to Biehler et al. (2015) this takes place at the context level	Biehler et al. (2015); Rossman et al. (2008); Tintle et al. (2009)
Give *possible explanations* for observed differences	Two possible explanations are to be found for the observed differences, as these provide the motivation for a randomization test. One possible explanation can be the random assignment of the experimental units to the groups, the other possible explanation can be the effectiveness of a procedure	Budgett et al. (2012); Justice et al. (2018)
Set up the *null modell*	Null and alternative hypotheses are to be formulated. In the null hypothesis, the random assignment to the groups is expressed as an explanation for the observed differences. In the following, the null hypothesis is assumed to be true and modelled. According to Biehler et al. (2015) this connects the context level with the statistics level	Biehler et al. (2015); Noll and Kirin (2017); Rossman et al. (2008)
Set up the *simulation modell*	Based on the null hypothesis, the randomization of the data for the randomization test is explained. This must be transferred to software, where the model is expressed and tested. According to Biehler et al. (2015) this takes place at the software level and it should be strongly connected with the context level according to Noll and Kirin (2017)	Biehler et al. (2015); Cobb (2007); Noll and Kirin (2017);
Visualize *test statistic* and *sampling distribution*	The test statistic X is defined for the simulated data according to the value in the observed data. A frequent repetition of the simulation takes place and the sampling distribution is generated from the collection of the test statistics. According to Biehler et al. (2015) this takes place at the software level	Cobb (2007); Rossman et al. (2008)

Table 2. Synopsis of elements for conducting a randomization test (continuation)

Element	Description	Reference
Identify the *p-value*	The *p*-value $P(X \geq X_{Observed} \mid H_0$ applies) is estimated from the simulation as the probability of obtaining a value such as the observed or an even more extreme one, assuming that the null hypothesis applies. According to Biehler et al. (2015) this takes place at the software level	Biehler et al. (2015); Rossman et al. (2008)
Draw *conclusions*	Conclusions are drawn from the results. Here it is checked whether the *p*-value is small enough, e.g. $p < 5\%$, to reject the null hypothesis. Remaining uncertainties are discussed. Furthermore, reference is made to the design of the experiment and corresponding causal conclusions are drawn. It is also discussed whether the results can be generalized, which is only possible if a random sample was used. According to Biehler et al. (2015) this takes place at the statistics and the context level	Biehler et al. (2015); Cobb (2007); Justice et al. (2018); Rossman et al. (2008); Tintle et al. (2009)

From the synopsis of Table 1, a scheme for teaching purposes can be developed that brings together the elements that are necessary for performing a randomization test (Figure 3). Such a scheme structures the process of a randomization test and can be helpful for learners (Biehler et al., 2015; Rossman et al., 2008).

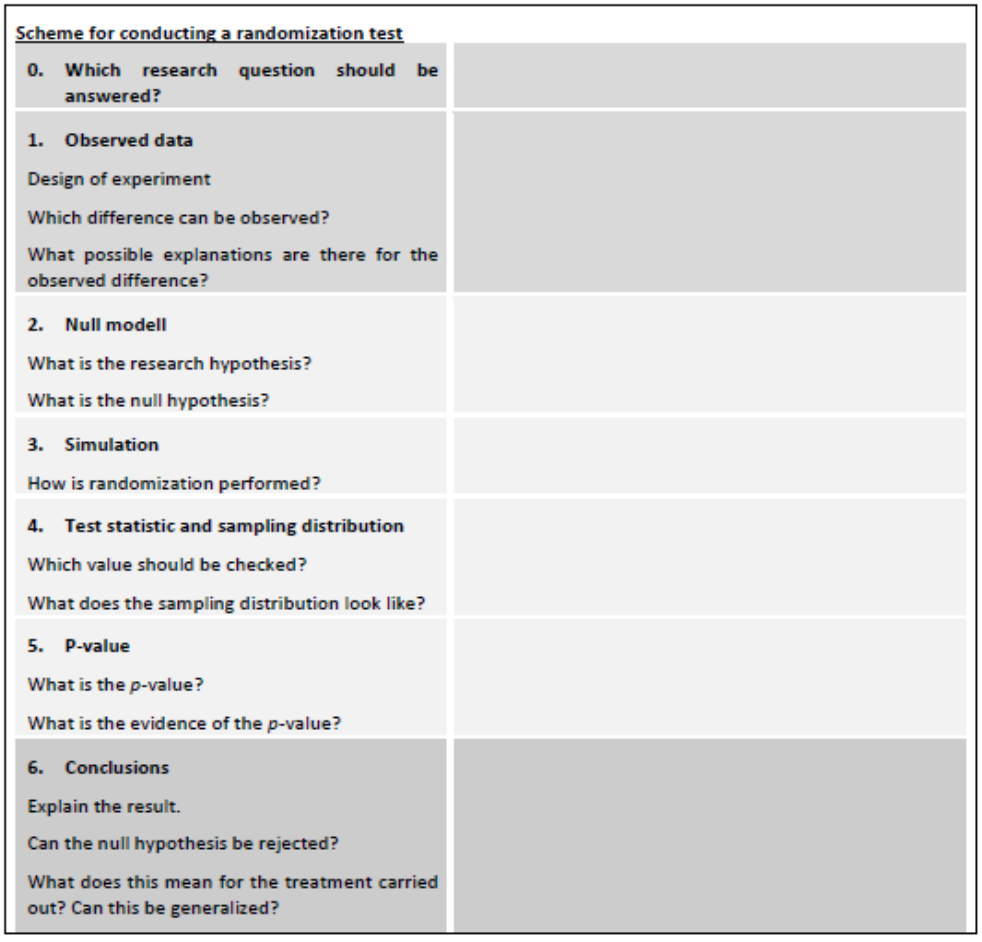

Figure 3. Scheme for conducting a randomization test

4. Conclusion

In this article, theory-driven elements were identified that should be addressed during the implementation of a randomization test in learning contexts. From this synopsis, a scheme was developed that is intended for the hand of learners to structure the reasoning process. An initial evaluation of this scheme took place in the study of Podworny (2018) with promising results. For the scheme itself, these results are in line with other studies in which schemes were also used successfully.

In general, it can be said in the words of Batanero (2000) that "statistics is not a way of doing, but a way of thinking that helps us solve problems in science and everyday life, teaching statistics should begin with real problems" (Batanero, 2000, p. 94). From this perspective, randomization tests offer an access to inference statistics, since they are always assumed to solve a real problem. Moreover, in didactic literature they are regarded as particularly suitable for introducing the way of thinking in inference statistics or to the logic of inference. At the same time, randomization tests play the "most important game in town" (Wild & Pfannkuch, 1999, p. 238) and thus certainly satisfy a need for the question of causality. However, as appears in some articles, randomization testing should not be seen as a panacea by which everyone is now able to understand the logic of inference statistics. As shown above, many elements are included (and must be understood) in the randomization test procedure in order to draw meaningful conclusions from data. An essential element here is the importance of the design of an experiment, which makes or makes not conclusions possible in the first place. However, it should not be overlooked that certain designs and thus possibly entire branches of research (e.g. empirical educational research) are completely questioned by some authors (e.g. Saint-Mont, 2011) or at least viewed very critically (Batanero & Borovcnik, 2016).

Despite all the simplicity, this approach "may be seen as an *intermediate step* before students can learn more formal inference" (Batanero & Borovcnik, 2016, p. 192). This opinion is shared by the author of this article in the sense that randomization tests provide a good approach to introduce the logic of inference statistics, but should not be the endpoint. Randomization tests, as described by most of the authors mentioned here, certainly offer a good first (informal) approach to inference statistics. However, this should not be stopped at, but, as already demanded almost 20 years ago (Batanero, 2000), further statistical methods should be explored and possible shortcomings and difficulties pointed out.

References

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1-2), 75-97.

Batanero, C., & Borovcnik, M. (2016). *Statistics and probability in high school*. Rotterdam, Boston, Taipei: Sense Publishers.

Biehler, R., Frischemeier, D., & Podworny, S. (2015). Preservice teachers' reasoning about uncertainty in the context of randomization tests. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning*. Minneapolis, Minnesota: Catalyst Press.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. (2012). Dynamic visualizations for inference. Paper presented at the T*he International Association for Statistical*

Education Roundtable Conference: Technology in statistics education: Virtualities and Realities, Cebu City, The Philippines.

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In B. Phillips (Ed.), *Proceedings of The Sixth International Conference On teaching of Statistics (CD-ROM)*. Voorburg, The Netherlands: International Statistical Institute.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Cobb, G. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1), 1-15. doi:https://escholarship.org/uc/item/6hb3k0nz

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer Science+Business Media.

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

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute.

Justice, N., Zieffler, A., Huberty, M. D., & delMas, R. (2018). Every rose has its thorn: secondary teachers' reasoning about statistical models. *ZDM*. doi:10.1007/s11858-018-0953-1

Noll, J., & Kirin, D. (2017). TinkerPlots model construction approaches for comparing two groups: Student perspectives. *Statistics Education Research Journal, 16*(2), 213-243.

Pfannkuch, M., & Budgett, S. (2014). Constructing inferential concepts through bootstrap and randomization-test simulations: A case study. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the ninth international conference on teaching statistics, Flagstaff, USA*. Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M., Regan, M., Wild, C., Budgett, S., Forbes, S., Harraway, J., & Parsonage, R. (2011). Inference and the introductory statistics course. *International Journal of Mathematical Education in Science and Technology, 42*(7), 903-913.

Podworny, S. (2018). *Simulationen und Randomisierungstests mit der Software TinkerPlots. Theoretische Werkzeuganalyse zur stochastischen SImulation und explorative Fallstudie zum statistischen Schließen mit Randomisierungstests [Simulations and randomization tests with TinkerPlots]*. Paderborn: Universität Paderborn.

Ramsey, F. L., & Shafer, D. W. (2013). *The Statistical Sleuth. A Course in Methods of Data Analysis*. Boston, Massachusetts: Cengage Learning.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19.

Rossman, A., Chance, B., Cobb, G., & Holcomb, R. (2008). *Concepts of statistical inference: Approach, scope, sequence and format for an elementary*

permutation-based first course. Available from: http://statweb.calpoly.edu/bchance/csi/CSIcurriculumMay08.doc

Saint-Mont, U. (2011). *Statistik im Forschungsprozess*. Berlin, Heidelberg: Springer.

Thompson, P. W., Liu, Y., & Saldanha, L. (2012). Intricacies of Statistical Inference and Teachers' Understandings of Them. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 207-231). New York: Psychology Press.

Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

Tintle, N., VanderStoep, J., & Swanson, T. (2009). *An active approach to statistical inference, preliminary edition*. Holland, Michigan: Hope College Publishing.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-265.

Zieffler, A., Harring, J. R., & Long, J. D. (2011). *Comparing groups. Randomization and bootstrap methods using R*. Hoboken, New Jersy: John Wiley & Sons.