DESIGNING AND IMPLEMENTING AN ALTERNATIVE TEACHING CONCEPT WITHIN A CONTINUOUS PROFESSIONAL DEVELOPMENT COURSE FOR GERMAN SECONDARY SCHOOL TEACHERS

Janina Oesterhaus and Rolf Biehler
University of Paderborn, Germany
janina.oesterhaus@math.upb.de

New standards in mathematics at German high-school level laid a plethora of stress on teaching statistical inference in senior classes, a domain that enjoys little popularity among teachers. Their personal perception not being adequately trained in teaching statistics during their professional careers is considered as one possible explanation for this deplorable state of affairs. In 2013 the German Center for Mathematics Teacher Education (DZLM) implemented a specially designed continuous professional development (CPD) course for teaching statistical inference at high-school level using a new and illustrative teaching approach. This approach was designed by the authors of the paper by adopting main characteristics from new Anglo-American curricula and accommodating them to German syllabi. The paper will focus on presenting the design process of the teaching approach and its evaluation within the CPD course.

TEACHING STATISTICAL INFERENCE IN GERMANY – POINT OF DEPARTURE

Strong fixation on deciding whether to reject or not to reject a null-hypothesis and little attention that is paid to reflection and interpretation of test results in major German textbooks, particularly when hypothesis testing is taught at German secondary high-school level, has long been criticized as being detached from reality and actual scientific practice (see e.g., Buth, 1993). This deplorable state of affairs is considered as a plausible explanation for well-known difficulties learners have when interpreting results of inferential procedures. In particular, the lack of authentic contexts and discussions of subsequent statistical methods and interpretations, are considered to be a paramount factor for inferior connections between statistical skills taught at German secondary high-school level and further education and career (see e.g., Kraus & Wassner, 2001).

However, the decree of new national standards for mathematics education at German high-school level in 2012 entailed that in North Rhine – Westphalian curriculum more emphasis has been laid on teaching probability and statistics in senior classes. But, in particular teaching statistical inference is a domain that enjoys little or no popularity among German teachers. This might be a consequence of their personal perception not being adequately trained in teaching probability and statistics during their own school and university careers (see e.g., Eichler, 2002). With the increased development and evaluation of appropriate learning environments, media, and materials for teaching probability and statistics of secondary and tertiary level (see e.g., Meyfarth, 2009), the mentioned deficiency in Germany has been increasingly taken into account with respect to preservice teacher education in recent years. However, inservice teachers thus far have experienced too little support and, therefore, could barely benefit from these new developments.

The German Center for Mathematics Teacher Education (DZLM) is trying to close this gap. By pooling the expertise of its various partners in schools and educational institutions, it has set itself the goal of developing CPD courses for mathematics teachers, which are available nationwide. In 2013, DZLM implemented under this intent a specially designed CPD course for teaching statistics in grade 12. In this course, secondary school teachers were given a demand-oriented update with respect to content and pedagogical content knowledge. Moreover, concrete assistance for their teaching in statistics was provided by introducing the teachers into an alternative teaching concept for teaching hypothesis testing, which had been developed by the authors.

DEVELOPMENT OF AN ALTERNATIVE TEACHING APPROACH

The aim of the teaching approach "BeSt@Kontext" [student activating teaching of inference statistics by using computer-based simulation methods in authentic contexts] (Oesterhaus & Biehler, 2013), is to foster an early and a continuous development of a conceptual understanding for the core-logic of statistical inference.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

The development of the teaching concept was mostly based on research findings concerning the design of comprehension-oriented statistics curricula as presented in the *Guidelines for Assessment and Instruction in Statistics Education* (Aliaga et al., 2010). Main characteristics of our approach were adopted from new Anglo-American curricula for Introductory Statistics Courses at college level (see e.g., Garfield, delMas & Zieffler, 2012; Tintle, 2011; Watkins, Scheaffer & Cobb, 2011) as well as recently developed German media and materials for teaching probability and statistics at high-school level (see e.g., Meyfarth, 2009). Those were then adapted to German, namely to North Rhine - Westphalian syllabus. These curricula are characterized by focusing on students´ activation, using intuitive approaches to inferential procedures by introducing them via informal hypothesis testing using p-values, use of technology and simulation with real data, and the use of authentic contexts as recommended in Aliaga et al. (2010). Such foci are rarely fully represented in current German textbooks.

"BeSt@Kontext" Version 0.8

The teaching approach "BeSt@Kontext" is modularized. In its first stage of development (version 0.8), which was also the underlying content of the mentioned CPD course in 2013, it comprised of *six basic compulsory modules*: 1) testing with p-values, 2) interpretation of p-values, 3) significance testing, 4) errors of a test and the operating characteristic (power) curve, 5) planning and validating a test design and 6) one-sided tests with composite null hypothesis (see Figure 1). Modules 3), 4) and 6) are the required curriculum in accordance with the new German standards and in conformity with the subject material that is usually found in German textbooks for senior classes, whereas module 4) and 6) are not equally common content in Anglo-American introductory statistics education. This content was supplemented with additional statistical content, methods and applications – modules 1), 2), and 5) – that appear to the authors as a minimum level for discussing statistical inference in senior classes, in the sense of the above mentioned objectives of BeSt@Kontext. Emphasis was laid on introducing formal significance testing via informal testing with p-values because, thus far, p-values have not been part of the German teaching content when teaching statistical inference. Evidence was found in earlier studies that this concept will enrich German traditional teaching by supporting students' understanding by making transparent the basic logic of statistical inference (see e.g., Meyfarth, 2009). Simulation methods and dynamic visualizations are a further ingredient of the approach. Skills acquired in the basic modules could then be expanded by *five optional advanced modules* (see Figure 1): tests with two known alternative point hypotheses and simple Bayesian inference, reflection on hypothesis testing, testing in randomized experimental and control group designs, test of a proportion and confidence intervals, further tests. Whereas the latter advanced modules are common content in Anglo-American introductory statistics education, that was adapted to the German approach as enriching content, the first module is not a common Anglo-American teaching content in introductory statistics education, but in Germany has long been discussed as an alternative or complementary contrast to non-Bayesian hypothesis testing. The advanced modules were meant to deepen and extend the knowledge and give access to further relevant applications of statistics.

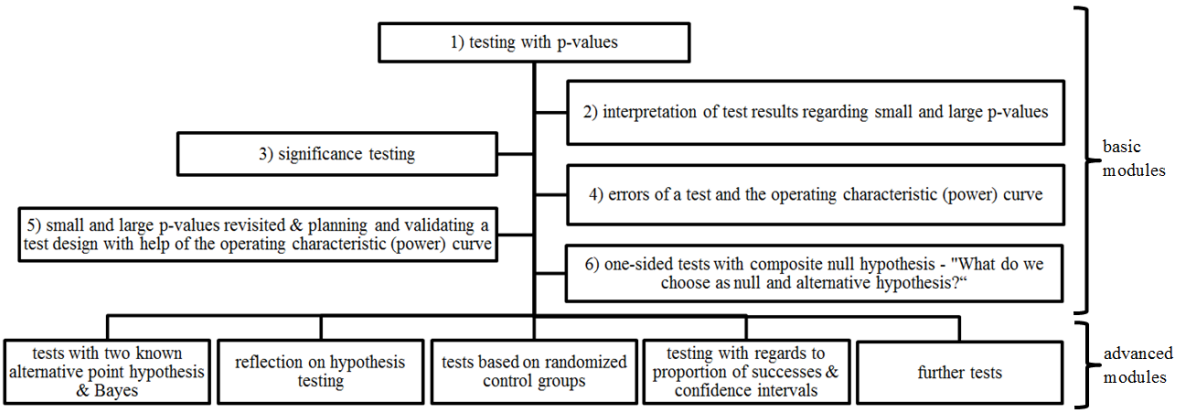

Figure 1. Modules and structure of BeSt@Kontext 0.8

The Basic Modules

In *module 1* the calculation of one-sided and small p-values (< 0.10) and the valuation of hypotheses using evidences are introduced in context of a taste-test called *Guessing or Knowing?* (Biehler et al., 2011, p. 129). Simulation methods are used to answer the central question "How likely is it to get results as extreme or even more extreme as the ones observed if someone is only guessing?" by making transparent which results are more or less likely to appear if someone is just guessing, especially when students themselves have the opportunity of performing the real experiment. Due to the fact that computer-based simulation methods have only recently been made obligatory as teaching and learning methods and, therefore, appropriate simulation tools are not yet widely available in schools, applets from the *Rossman and Chance Applet Collection* (http://www.rossmanchance.com/applets/OneProp/OneProp.htm) were adapted with permission of the respective authors. Subsequently, the acquired skills are extended to contexts where calculation and interpretation of two-sided p-values is required. In *module 2* the setting of taste-tests is extended to numbers of successes that now lead to the issue of how to correctly interpret large p-values (> 0.10) with respect to context. To strengthen their understanding, students are to simulate the taste-test with other probabilities than $p = 0.5$ which also lead to no evidence against the null hypothesis. Thus, students see that no final conclusion about the true probability can be drawn when large p-values are observed. Going more in depth, the interpretation of large p-values is extended to different contexts which include different types of a priori assumptions about the validity of the null hypothesis: *neutral a priori assumptions* – as, for example, given in context of a taste-test where usually no reasonable a priori assumption about the true ability of the participant can be made – or *a priori assumptions promoted by previous evidence* based on further data – given, for example, when testing a roulette table on its fairness where previous evidence about the fairness of the object is promoted by the usually given obligation to regularly check those roulette tables in casinos. Both situations lead to somewhat different valuations of the validity of the underlying null hypotheses; in the latter context, given a large p-value, one can conclude to further stick to the assumption of the validity of the null hypothesis because of the existing further data that also promote this assumption, whereas in context of the taste-test no statement about the true probability p can be made. Working on this module, students learn that the overall validation of the null hypothesis, in absence of evidence, depends on the context.

In *module 3* the need of a predetermined limit for the p-value (significance value) is motivated by modifying the taste-test by adding a final decision whether or not the test person finally gets a price for his or her abilities and the minimum number of correct answers to receive this price has to be fixed in advance. The terms "statistical significant", "significance level" and "(not) rejecting an assumption" are introduced and subsequently reflected against the background of scientific practice by exemplarily discussing authentic research reports where these terms are used to evaluate hypotheses but where no final decision is needed. In *module 4* possible errors of a test are introduced and discussed with respect to contexts such as sensory tests and medical studies. Digital simulation and dynamic visualizations are used to discover and discuss the contrariness of type I and type II errors, as well as the connection between types of errors and sample size. Then the operating characteristic (power) curve is introduced to analyze the type II error in dependence of the true value of p via dynamic visualizations. After having introduced and analyzed type II errors, in *module 5* interpretation of small and large p-values are revisited by arguing with type II errors: in case of absence of evidence against the null hypothesis no positive conclusion about the validity of the null hypothesis can be drawn due to the fact that there are also other values p than p_0 consistent with the null hypothesis and values of p that are close to p_0 will lead to a high type II error. Subsequently, dynamic visualizations are used to reflect on the operating characteristic curve as a powerful instrument for planning as well as for validating the design of a test.

After a reflection on meaning and difference of one-sided hypothesis tests with point null hypotheses versus one-side hypothesis tests with composite null hypotheses, in *module 6* the issue of what to choose as null- and alternative-hypothesis is addressed by contexts of quality control where two one-sided composite hypotheses face each other as possible null hypotheses. Digital visualizations are used to analyze the respective type II error from each perspective. Hence, the case of indifference where none of the two hypotheses can be rejected is introduced and possible

subsequent methods to solve this situation in practice (by means of sequential tests) are discussed exemplarily.

The Advanced Modules

Whereas the former modules dealt with contexts where no preliminary information about the true probability *p* was given, the module *test with two known alternative point hypothesis and Bayesian inference* introduces test procedures for situations where a priori probability assumptions about the true probability *p* can reasonably be made. After discussing tests where a decision between only two possible values for the true probability *p* is required and other values can a priori be excluded, Bayesian testing is introduced with the purpose to clarify which probabilities can be calculated with each of the two methods – significance testing and simple Bayesian inference – by noting down those probabilities as conditional probabilities. Learners are ought to understand that only with simple Bayesian inference it is possible to calculate probabilities of hypotheses. In *reflection on hypothesis testing* typical problems such as the use of data for generating hypothesis versus testing hypothesis, the case of practical relevance versus statistical relevance, the problem of multiple testing and examples for publication bias are discussed.

Sampling methods, experiment design and data collection are topics that are barely discussed explicitly in German textbooks but, in the authors' view, are essential knowledge for a meaningful interpretation of test results with respect to context. The advanced module *tests based on randomized control groups* introduces basic ideas of randomization methods and discusses the different types of conclusions that can be drawn out of tests based on randomized samples in comparison to tests based on simple random sampling. Hence, in the context of clinical studies the idea of randomized controlled experiments as a method to enable causal conclusions drawn out of tests is introduced. As Fisher's exact test and the hypergeometric distribution are no obligatory teaching contents and are hardly to be found in current German textbooks, simulation methods are used to solve this problem by performing the simulation of a permutation test with help of an applet (http://www.rossmanchance.com/applets/Dolphins/Dolphins.html).

Since hypothesis testing in German curricula is mediated with regards to inferences for number of successes (because this assures a direct application of the binomial distribution), in the module *testing with regards to proportion of successes and confidence intervals* inferences for proportion of successes are introduced by tailoring tasks and associated simulations and visualizations of the former modules accordingly. By focusing on expanding questions such as "What are plausible proportions?" when the null hypothesis could not be rejected, informal ideas of confidence intervals are introduced and extended to scientific contexts. In *further tests*, a deeper insight into the calculation, use and typical contexts of those tests, other than the binomial tests, that have already been used in contexts of the former modules, such as Fisher's exact test or permutation tests, is given.

IMPLEMENTATION AND EVALUATION WITHIN A CPD COURSE

The above mentioned CPD course was particularly designed to accomplish the first steps of the dissemination process of this alternative teaching concept. In four all-day face-to-face meetings secondary school teachers were given a demand-oriented update with respect to basic content and pedagogical content knowledge regarding the teaching of probability and statistics. Furthermore, they were urged to critically analyze and reflect on common teaching materials and previous teaching practices with a special focus on teaching statistics with the aim to raise their awareness of the existing shortcomings. Subsequently, they were introduced into the alternative teaching approach BeST@Kontext, its underlying ideas and goals, as well as, associated materials and tools. At the end of the course they reflected on the new materials with respect to their valuation of its appropriateness and practicability for classroom practice. By discussing this, it was intended to gain the teachers' interest and motivation to themselves test and evaluate the materials in their classrooms and, in this way, to get a deeper insight into limitations of the approach with respect to the German school system and, therefore, required changes in content, structure or methods. In consideration of the cognitive load within the four all-day meetings, digital simulations and visualizations were, in a first stage, introduced in form of ready-made learning environments and worksheets by using the same applets that later on were meant to be used by their students.

A further face-to-face meeting took place in November 2013 with teachers who had already tested parts of the new materials and ideas and those who were interested in implementing ideas of the course in their future teaching.

Evaluation and redesign of BeSt@Kontext 0.8

The development of the teaching approach is inspired by paradigm of design-based research (see e.g., The Design-based Research Collective, 2003) with a more complex cycle of redesigning the material through interaction in the CPD course and in mathematics classrooms. First design was based on theories derived from prior research on comprehension-oriented teaching and learning of statistics. Then, first demands for revision and redesign were identified within the face-to-face meetings with the teachers through several group work and discussions on the material. These discussions were videotaped in order to have them backed up for further development work. Moreover, all professional training done by DZLM is – for the purpose of quality assurance – evaluated by pre- and post-questionnaires. These surveys were also used to evaluate the teaching approach. Based on this analysis, the approach was redesigned during summer and fall 2013. Further cycles of formative evaluation (with enactment not only in professional training but also in teaching practice) will follow in spring 2014.

Figure 2. Development process of BeSt@Kontext

Teachers' feedback given within this first phase of evaluation can be summarized as follows: Firstly, the modules have to be connected more explicitly to popular textbooks. Taken that textbooks are used as a central tool for teaching and learning, it has to be made clear where and how contents of the modules can be implemented in classrooms' textbook work. Moreover, terminology and nomenclature need to be in accordance with current textbooks. Secondly, learning environments and worksheets for digital simulation and visualization have to be in accordance with, firstly, tools that are available and commonly used in mathematics classrooms (e.g., Microsoft Excel, GeoGebra, TI-Nspire) and, secondly, should proactively consider the forthcoming obligatory use of graphic calculators which has recently been mandated by an official enactment of the North Rhine – Westphalian government. Hence, materials will have to gradually be adapted for a flexible implementation of different statistical tools.

STATUS QUO AND FUTURE WORK

Based on teachers' feedback, BeSt@Kontext 0.8 has been revised during summer break with regards to the obligatory basic modules (see Figure 2). Contents that are obligatory according to core-curriculum for high-intensity mathematics courses have been set as two obligatory anchor modules *significance testing* and *errors of a test, operating characteristic curve & use of the operating characteristic curve for experiment design* (indicated by the bolded frame). Here, contents and tasks can be replaced or upgraded by the particular textbook used in the classroom. The contents of the former module 1) have been split into two optional introductory modules. These newly constructed modules are now self-constrained and independent from other modules, but differ in depth with regards to content that is discussed. In the same sense, the contents of the former module 2) have been split into three optional modules that can serve teachers to consolidate knowledge and competences acquired with regards to significance testing as it is mediated in German textbooks. This would also allow them to expand and deepen these skills in terms of the objectives of BeSt@Kontext mentioned above. Whereas it was decided to redesign the modules in a way that they can be implemented flexibly and independently into classroom work in accordance to teachers need and feasibility (indicated by the thin arrows), the sequence of the modules in accordance with the original objectives of BeSt@Kontext is indicated by the bolded arrows.

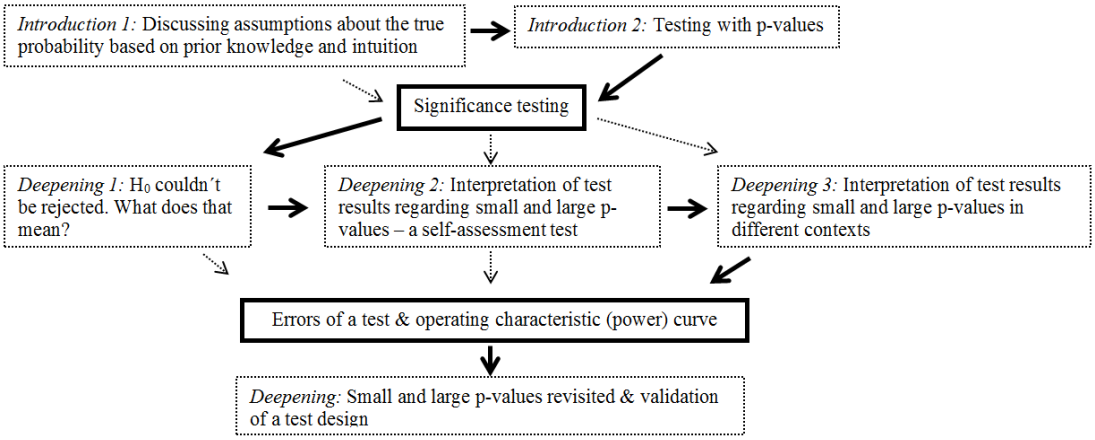

Figure 3. Basic modules of BeSt@Kontext 0.9 after a first phase of revision

These basic modules (version 0.9) were reflected back into the face-to-face meeting in November 2013. In their role as experts for teaching practice, teachers worked on optimizing parts of the learning material in small groups. The original advanced modules of BeSt@Kontext 0.8 are still being reworked. In a next step, learning environments and worksheets will be adapted to a flexible use of digital tools for simulation and visualization. Continuative, these reworked materials will be reflected back into a next upcoming face-to-face meeting in March 2014, which was set up at the request of the participating teachers in November 2013. In this sense, we hope, in the long run, to be able to cultivate professional learning groups and to continue the collaborative work on specializing the teaching approach in accordance with German teachers' needs. BeSt@Kontext 0.9 will again be revised based on teachers' feedback in March 2014 with the aim to be able to implement BeSt@Kontext 1.0 into the reissue of the CPD course in 2014.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., et al. (2010). *Guidelines for assessment and instruction in statistics education (GAISE) college report.* Alexandria, VA: American Statistical Association.

Biehler, R., Hofmann, T., Maxara, C., & Prömmel, A. (2011). *Daten und Zufall mit Fathom: Unterrichtsideen für die SI und SII mit Software-Einführung*. Braunschweig: Schroedel.

Buth, M. (1993). Testen von Hypothesen: Was man aus der Forschungspraxis für die Schule lernen kann. *Stochastik in der Schule, 13*(2), 35–46.

Eichler, A. (2002). Vorstellungen von Lehrerinnen und Lehrern zum Stochastikunterricht. *Der Mathematikunterricht, 4–5*, 26–44.

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

Krauss, S., & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule, 21*(1), 29–34.

Meyfarth, T. (2009). Die Konzeption, Durchführung und Analyse eines simulationsintensiven Einstiegs in das Kurshalbjahr Stochastik der gymnasialen Oberstufe. *Kasseler Online-Schriften zur Didaktik der Stochastik (KaDiSto), 6*. Kassel: Universität Kassel.

Oesterhaus, J., & Biehler, R. (2013). BeSt@Kontext: Ein schüleraktivierendes Unterrichtskonzept für die Beurteilende Statistik mit computergestützter Simulation in authentischen Kontexten. In G. Greefrath, F. Käpnick & M. Stein (Eds.), *Beiträge zum Mathematikunterricht 2013* (Vol. 1, pp. 720–723). Münster: WTM-Verlag.

The Design-based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

Tintle, N. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1), n1.

Watkins, A. E., Scheaffer, R. L., & Cobb, G. W. (2011). *Statistics: From Data to Decision* (2nd ed.). Hoboken, NJ: John Wiley & Sons.