PRESERVICE TEACHERS COMPARING GROUPS WITH TINKERPLOTS—AN EXPLORATORY LABORATORY STUDY

DANIEL FRISCHEMEIER
University of Paderborn
dafr@math.upb.de

ROLF BIEHLER
University of Paderborn
biehler@math.upb.de

ABSTRACT

Group comparisons offer students opportunities to reason about many fundamental statistical concepts like center, variation, or distribution. When doing such activities using large, real datasets, technology becomes an essential tool for exploring the data. With its large variety of features and its user-friendly handling, TinkerPlots^{TM}—as a software for learners and teachers—can facilitate the process of comparing distributions. In this article we focus on eight preservice teachers´ reasoning when comparing groups with TinkerPlots. We present ideas on the design of a course to develop statistical reasoning with TinkerPlots, present a framework to rate learners´ performance when comparing groups with TinkerPlots, and present results of a laboratory study about preservice teachers´ reasoning when comparing groups with TinkerPlots. Findings suggest that the TinkerPlots tool and design of the course supported these preservice teachers' reasoning and that more learning opportunities are needed to increase their group comparison elements' repertoire and interpretation in context.

Keywords: *Statistics education research, Preservice teacher education, Frameworks*

1. INTRODUCTION

In Germany, the teaching and learning of statistics has received increased attention in the primary and secondary mathematics curriculum. This is evident by the emergence of German national recommendations for student learning of statistics at the primary (Hasemann & Mirwald, 2012) and secondary level (Blum, Drüke-Noe, Hartung, & Köller 2006). Additionally, recommendations have emerged in Germany (Sill, 2018) and internationally (e.g., Batanero, Burrill, & Reading, 2011) that explicate the statistical knowledge teachers need to develop to effectively teach school statistics. Salient across these recommendations is that student learning of statistics should be grounded in opportunities to use technology to investigate real datasets while engaged in statistical enquiry cycles (e.g., the PPDAC cycle; Wild & Pfannkuch, 1999). The PPDAC cycle provides learners with opportunities to generate their own statistical problems and questions (first "P" in PPDAC), to plan data collection (second "P" in PPDAC), to collect data ("D" in PPDAC), to analyze data ("A" in PPDAC), and to interpret the findings of their data exploration by drawing conclusions ("C" in PPDAC).

Of course, any efforts to improve student learning of statistics at the primary and secondary level necessarily depends on the teacher's ability to realize these recommendations. Thus providing future and in-service teachers with learning experiences that enable them to effectively enact these recommendations in their classrooms is of growing importance. With this in mind, we designed a statistics course for preservice teachers called *Developing Statistical Reasoning with TinkerPlots* aimed at developing future

Statistics Education Research Journal, 17(1), 35-60, http://www.stat.auckland.ac.nz/serj

primary and secondary teachers' content and technological knowledge in statistics. The course was developed based on a design-based research paradigm (see Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). Because making group comparisons offers learners opportunities to reason about and coordinate many fundamental statistical concepts (e.g., center, variation, distribution, etc.), developing preservice teachers' abilities to draw conclusions on the basis of describing similarities and differences between two samples (without making further inference to a population or process) was a primary goal in the course. The preservice teachers were also provided with opportunities to draw inferences beyond the data at hand by conducting randomization tests with TinkerPlots. Although not a focus of this paper, Frischemeier and Biehler (2014) describe in detail this second component of the course.

As pointed out by Pfannkuch and Ben-Zvi (2011), "teachers can be challenged to explore and learn from data in ways similar to the ways their students will explore data" (p. 328). Thus throughout the course, preservice teachers engaged with activities designed to have them go through the entire PPDAC cycle and use technology to investigate large, real datasets (self-collected or downloaded from statistical bureau websites). We selected the TinkerPlots (Konold & Miller, 2011) software as the primary technology used throughout the course. From our perspective, TinkerPlots serves multiple purposes in teacher education. First, research has shown that TinkerPlots, as educational software, is accessible to students of all ages and can be used as a tool to develop statistical reasoning (see Biehler, Ben-Zvi, Bakker, & Makar, 2013). As such TinkerPlots can be seen as appropriate software to facilitate the learning of data analysis both for students at the school level and for preservice teachers enrolled in university courses. Second, with its wide variety of features, students can also leverage TinkerPlots as a tool for doing data analysis. Third, because the preservice teachers that enroll in our course will become teachers in primary or secondary school, it is important to introduce them to a tool they can use in their future work as teachers. Thus TinkerPlots can be seen as a tool for teachers to learn to implement in their classrooms (e.g., for classroom demonstrations or to facilitate student activities) as well.

The primary goals of this paper are to present a framework developed to assess preservice teachers' performance when making group comparisons in TinkerPlots and to present results of a study where we assessed our preservice teachers' ability to do group comparisons with TinkerPlots after participating in the course *Developing Statistical Reasoning with TinkerPlots*. Prior research on learners comparing groups has described teachers' understanding of fundamental concepts underlying comparing groups (e.g., Batanero, Burrill, & Reading, 2011; Jacobbe & Carvalho, 2011; Reading & Canada, 2011; Sánchez, da Silva, & Coutinho, 2011), has pointed out approaches learners use to make group comparisons (e.g., Ben-Zvi, 2004; Biehler, 2007b; Frischemeier & Biehler, 2011), and provided frameworks for assessing learners´ outcomes when comparing groups (e.g., Makar & Confrey, 2002; Pfannkuch, Budgett, Parsonage & Horring, 2004; Pfannkuch, 2007; Watson & Moritz, 1999). Although technology was a component of many of the aforementioned studies, current frameworks for assessing learners' abilities to conduct group comparisons do not attend to learners' abilities to use technology during this process. As such, we begin this paper with a review of previously proposed frameworks and taxonomies that rate learners' performance when comparing groups or rate learners' performance when using TinkerPlots in an effort to design a framework that will allow us to assess our preservice teachers' reasoning when making group comparisons as well as their ability to implement their approaches in TinkerPlots. We then apply this framework on data collected during a laboratory study conducted with eight preservice teachers at the end of the *Developing Statistical Reasoning with* TinkerPlots course. In the study, the preservice teachers were asked to use TinkerPlots to compare two distributions of a numeric variable from a large, real dataset.

To assess our preservice teachers' performance when making group comparisons in TinkerPlots and to develop a framework for this assessment, the following questions guided our study:

1. What are adequate group comparison elements in the TinkerPlots software?
2. Which of these adequate group comparison elements (mentioned in 1.) are used by the participants in our study when comparing groups in real data sets using TinkerPlots?
3. To what extent do the participants in our study interpret their findings in a group comparison process?

4. In which ways are the participants able to handle the TinkerPlots software competently when comparing groups?

2. LITERATURE REVIEW

Previous efforts to rate learners' performance when comparing two groups have not attempted to assess the influence of their software capabilities on their reasoning processes. In order to answer our research questions, a framework is needed that incorporates the fundamental statistical features learners attend to when making group comparisons as well as learners' software skills when comparing two groups. In the sections that follow we present a review of the literature related to our study. The underlying goal of this literature review is to identify aspects of a framework that could be used to assess learners' performance when making group comparisons with TinkerPlots. Towards this end, we review literature that provides insights into approaches learners use when comparing groups as well as the statistical features they attend to when making group comparisons. Additionally, we review previously proposed frameworks that assess learners' performances when making group comparisons as well as frameworks that rate learners' software skills. Throughout the following section we highlight findings from the literature that we see as necessary components of our framework and offer a rationale for the inclusion of those components.

2.1. ADEQUATE ELEMENTS FOR COMPARING GROUPS

In their text, Rossman, Chance, and Lock (2001) identified several ways to describe and interpret a distribution of a numerical variable. These include "center, variability, shape, peaks, and clusters and outliers" (p. 48). These elements might also be adequate elements to take into account for comparing two distributions.

Pfannkuch et al. (2004) identified different approaches students used when comparing groups using boxplots. They observed that the students in their study employed various strategies including comparing equivalent summary statistics, comparing non-equivalent summary statistics, comparing variability, and comparing distributions. In a subsequent study, Pfannkuch (2007) investigated student responses on a comparing boxplot distributions task. Her analysis of students' responses revealed additional statistical concepts that students used to describe and interpret similarities and differences they identified when comparing two groups. More specifically, she distinguished (among others) comparison elements like summary, spread, shift, and signal. Whereas Pfannkuch (2007) concentrated on comparing distributions displayed by boxplots, we aim to cover a broader spectrum of representations. Therefore, we decided to attend to categories that could be used when making group comparisons across a variety of representations and named these categories as "center," "spread" and "shift."

Biehler (2001, 2007a) gives a normative point of view on comparing groups, and emphasizes additional approaches learners might use when making group comparisons. Amongst others, Biehler suggests that learners might attend to differences in the skewness of the distributions being compared. Additionally, Biehler (2001) suggests that when comparing two numerical distributions learners may rely on p-based and q-based comparisons. The following definitions are translated from Biehler (2001):

> Comparisons of two distributions of numerical variables (say V and W) are called p-based, if for x the relative frequencies $h(\mathrm{V} \leq x)$ and $h(\mathrm{W} \leq x)$ are compared. So in p-based comparisons a specific point x can be given (for example, 10 hours) and the proportion of cases which are less than or equal to 10 hours is compared in both groups. Comparisons of two distributions of numerical variables are called q-based, if for a proportion p between 0 and 1 the matching quantiles of the variables V and W (written $q_{\mathrm{V}(p)}$ and $q_{\mathrm{W}(p)}$) are compared. With $q(p)$ we mean the quantile associated with p. For $p = 0.5$ this is a comparison of medians. (p. 110)

While observing preservice and in-service teachers comparing groups with TinkerPlots, Hammerman and Rubin (2004) identified two strategies learners use when comparing groups with TinkerPlots. In particular, they observed that the teachers in their study based their comparisons on comparing categorized numerical data (i.e., data that had been categorized into bins) or using proportional reasoning to compare the relative frequencies of given intervals in the distributions. Hammerman and Rubin also distinguished between system-generated cut-points via binning and user-generated cut points via dividers in TinkerPlots. These user-generated cut points serve as a basis for the realization of p-based comparisons in TinkerPlots. So in addition to including center, spread, and shift in our framework, we have added skewness, p-based comparisons, and q-based comparisons as these three comparison elements are suggested in the literature as possible approaches learners may use when comparing groups. Table 1 provides an overview and a description for each of the adequate group comparison elements in our framework.

Table 1. Adequate elements when comparing groups with TinkerPlots

Group comparison element	Description
Center	The centers (mean or median) are compared between the distributions of a numerical variable.
Spread	The components of spread (e.g., interquartile range, standard deviation) are compared between the distributions of a numerical variable.
Skewness	The skewnesses (e.g., left-skewed, symmetrical, right-skewed) are compared between the distributions of a numerical variable.
Shift	The shift between the distributions of a numerical variable is compared.
p-based	Two distributions of a numerical variable are compared p-based.
q-based	Two distributions of a numerical variable are compared q-based.

2.2. FRAMEWORKS TO RATE LEARNERS´ PERFORMANCE WHEN COMPARING GROUPS

Makar and Confrey (2002), Pfannkuch et al. (2004), Pfannkuch (2006, 2007), and Watson and Moritz (1999) all use taxonomies or frameworks in their work to rate the reasoning of learners when comparing groups. In the section that follows we describe these research studies and comment on them with respect to our purpose.

Watson and Moritz (1999) observed school students in Australia (grades 4 through 8) as they compared two distributions in two different settings (equal size groups vs. non-equal size groups). The students were given two distributions in the form of stacked dot plots that displayed the distribution of test scores of two school classes. Students participated in four interviews in which they were asked to compare the two distributions of test scores from the two school classes and asked to decide which class performed better on the test. In the first two interviews, students were asked to compare two groups where each group had the same number of cases. In the third interview, these students were asked to compare two distributions that were equal sized and differed in relation to spread only. Lastly, in the fourth interview, students were given two unequal-sized distributions and were asked to compare them. The students' responses were transcribed and coded based on the hierarchical levels (e.g., unistructural, multistructural, and relational) of the SOLO taxonomy (Biggs & Collis, 1982). In this categorization, the outcome of a learner was rated higher when taking more features into account when making the comparison. Additionally, Watson and Moritz observed that students used both visual and numeric approaches when comparing groups. Students employing a visual approach attempted to compare groups by attending to the skewness or shape of the distributions. In contrast, when using a numeric approach students were observed making comparisons by calculating and comparing statistical measures (e.g., the mean of both distributions). Although the SOLO taxonomy used by Watson and Moritz offers a good basis to rate learners' reasoning in comparing groups in two datasets, it does not seem appropriate for our purposes because it distinguishes non-equal-sized and equal-sized groups for group comparisons, focuses on proportional reasoning of learners, and also deals with small data sets without software.

Makar and Confrey (2002) conducted a course with teachers in a professional development setting and observed how preservice teachers compared two groups with Fathom^{TM} (Finzer, 2001). The preservice teachers were given two distributions in the form of stacked dot plots of test scores of two schools in Fathom. To rate the participants' reasoning, Makar and Confrey developed a five-tier framework, which they described as a taxonomy for classifying levels of reasoning when comparing two groups. The five levels in their taxonomy were pre-descriptive, descriptive, emerging distributional, transitional, and emerging statistical. Whereas learners with a pre-descriptive view show "no recognition of relationships between datasets except based on individual data points or anecdotal evidence" (p. 3), learners characterized as having an emerging distributional view establish "a first holistic view of the data … where informal qualitative descriptors of the data, along with basic summary statistics are used to describe two datasets" (p. 3). Furthermore, at this level "teachers begin to understand the difficulty in creating measurable conjectures, but are unable to successfully resolve the conflict and show frustration in attempting to write an appropriate conjecture. Variability, while acknowledged, is not understood beyond a descriptive level" (p. 3). At the highest level of their taxonomy teachers

> gain confidence in using standard descriptive statistics to compare data sets, taking into consideration the differences between measures of center in light of the variability in the data and the sample size of the datasets. Conjectures demonstrate some ability to frame questions that balance data constraints with the problem at hand. Context and quantified descriptions are well integrated into conclusions and inferences may attempt to draw on statistical models, if relevant (Makar & Confrey, 2002, p. 3).

Further details about the other levels (e.g., level 2 and level 4) can be found in Makar and Confrey (2002).

Makar and Confrey's (2002) taxonomy concentrates on inferences when comparing samples of populations and focuses on comparing the distributions with respect to statistical terms like *evidence* or *significance*. In addition, the levels of their taxonomy show how variability was used to identify and explain differences between the groups. The framework of Makar and Confrey does not seem to be adequate for our purposes because the framework does not reveal which concrete elements learners use when comparing groups with software. A more appropriate application of this framework can be the comparison of pre- and post-results of learners' reasoning when comparing groups with software (see for example Madden, 2008).

Pfannkuch presents another framework for rating learners' abilities when comparing groups (Pfannkuch et al., 2004; Pfannkuch, 2006, 2007). Pfannkuch et al. (2004) observed students (15 years old) comparing boxplot distributions of the temperatures in Wellington and Napier (both cities in New Zealand). Based on an analysis of the students' written work on this comparison task, Pfannkuch et al. distinguished between different types of strategies these students used. In particular, they observed that students made comparisons by comparing equivalent summary statistics, comparing non-equivalent summary statistics, comparing variability, and comparing distributions. Additionally, they rated each kind of response using a SOLO taxonomy. The hierarchal levels of their taxonomy included no response, prestructural, unistructural, multistructural, and relational levels. So at first they structure the students' responses in form of the comparison element, then they rate their quality. One result of the study was that whereas these students seemed to prefer comparing the boxplot distributions using summary statistics and range, they did not directly refer to other measures of variability (like interquartile range) or attempt to make comparisons by attending to a possible shift between the distributions.

Pfannkuch (2007) continued to observe students (Year 10) when comparing boxplots. In this study, the participants were given two boxplot distributions and they were asked to make three comparison statements to explain differences and similarities between the distributions. Student responses were analyzed in two ways. First, Pfannkuch distinguished between the structural components students attended to when making their comparisons. From the analysis she was able to distinguish (amongst others) elements like summary, spread, shift, and signal. Pfannkuch also rated each of the comparison statements students gave, using different levels to describe the quality of the comparison. In particular, Pfannkuch distinguished between the following four hierarchical levels: point decoder (level 0), shape comparison describer (level 1), shape comparison decoder (level 2) and shape comparison assessor (level 3). In point decoder responses, students identified values or points of the distribution. A shape comparison describer made statements on a descriptive level (e.g., with the picture of the plot). The highest level of responses was classified as shape comparison assessor. These students moved beyond describing what they saw to comparing, interpreting, and paraphrasing the differences between the two groups in context. To summarize, Pfannkuch associated lower quality responses with identification, medium quality responses with description, and high quality responses with interpretation. A primary finding of this study was that participants preferred to compare the distributions using summary and spread elements as opposed to shift and signal elements. Additionally, when describing differences and similarities between two boxplot distributions, the learners often relied on identifying and describing what they saw rather than interpreting what they observed.

Pfannkuch's work (Pfannkuch, 2006, 2007; Pfannkuch et al., 2004) provides structural and evaluative elements that allow us to distinguish several approaches learners use when comparing groups. In particular, the framework used by Pfannkuch (2007) allows the identification of the group comparison elements used and also offers details to what extent the difference between two groups was interpreted.

2.3. FRAMEWORKS TO RATE LEARNERS SOFTWARE SKILLS WHEN COMPARING GROUPS

As previously mentioned, something that is missing in the studies and research reports mentioned above is an explicit focus on the extent to which learners are able to use software in the group comparison process. Currently, there are no frameworks that assess software competence for comparing groups with TinkerPlots. However, there are available frameworks that rate learners' software competencies when conducting chance experiments using Fathom.

Maxara (2009, p. 293) identified different facets of Fathom competences when conducting simulations of chance experiments with Fathom: "General Fathom competence, Formula competence in Fathom, Simulation competence in Fathom, and Strategic competence in Fathom." For further details, see Maxara (2014, p. 327). Maxara (2014) mentioned that this framework could also be adapted for assessing the software competence of learners using other tools.

Biehler's (1997) research focused on software use during data analysis tasks. Biehler identified four phases that learners must reason through to solve statistical problems with software: *Statistical problem*, *problem for the software*, *results of software use*, and *interpretation of results in statistics.* Biehler points out that "…we can often reconstruct in our students a direct jump from a real problem to a problem for the software without an awareness of possible changes" and also that "… students are satisfied with producing computer results that are neither interpreted in statistical nor subject matter terms" (p. 175). Additionally, Biehler mentions a "degenerate use of software for problem solving, where it only counts that the computer does it" (p. 175).

A primary purpose of this literature review was to identify components of a framework that would allow us to assess learners' performance when comparing groups with TinkerPlots. The frameworks we highlighted in our literature review supported our development of six adequate group comparison elements (see Table 1) as well as two "dimensions" for assessing learners' performance when comparing groups with TinkerPlots:

- Dimension 1: Assesses a learner's ability to use TinkerPlots when conducting group comparisons. It focuses exclusively on the software and rates users based on their ability to use TinkerPlots to execute planned actions that arise during the group comparison process.
- Dimension 2: Assesses a learner's statistical reasoning when comparing groups with the software. It covers how (center, spread, skewness, shift, p-based comparison or q-based comparison) and in what way (descriptive or interpretative) two groups are compared on the base of the six adequate elements.

We elaborate on the dimensions of this framework in Section 3.4.

3. CONTEXT, DATA, AND METHODS

3.1. THE COURSE: DEVELOPING STATISTICAL REASONING WITH TINKERPLOTS

We report findings from data collected as part of a larger project (Frischemeier, 2017) aimed at developing a course for preservice primary and secondary teachers of statistics. The course, named *Developing Statistical Reasoning with TinkerPlots,* was developed based on the Design-based research paradigm (Cobb et al., 2003) for the purpose of deepening the statistical and technological content knowledge of preservice teachers at the University of Paderborn. As part of the course design, we integrated many of the components of a Statistical Reasoning Learning Environment (Garfield & Ben-Zvi, 2008), including "focusing on central statistical ideas," "using real and motivating data sets," "using classroom activities," "integrating the use of appropriate technological tools," "promoting classroom discourse," and "using assessment." Additionally, our course relied on the PPDAC cycle (Wild & Pfannkuch, 1999) in its chronological sequence. This means that the participants in our study were asked to generate a problem and plan the data collection needed to approach the problem. After collecting data, participants were asked to analyze their data with TinkerPlots and interpret the results of their analysis. Overall, the course consisted of 14 sessions, each lasting approximately 90 minutes.

Group comparisons played a fundamental role throughout the course and students frequently engaged in activities in which they were asked to make group comparisons in real datasets. For each activity, participants generated their own statistical questions leading to group comparisons, explored real datasets with TinkerPlots, and wrote down their findings in a statistical report. Within the TinkerPlots environment, data for group comparisons was often displayed using stacked dot plots, histograms, and boxplots. When making group comparisons, students were instructed to decide between producing and analyzing all the displays at once or generating and analyzing single displays successively.

As part of our course design we drew on insights gleaned from Pfannkuch (2007), Biehler (2001), and Biehler (2007a) to foster preservice teachers' abilities to compare groups by attending to various features between the distributions. In particular, in this course students were taught to compare groups using center, spread, shift, and skewness, as well as to conduct p-based and q-based comparisons. Throughout the course, students were expected to identify and describe as many similarities and differences as they could between the groups under consideration (e.g., comparing center, spread, skewness, shift, p-based, and q-based). Additionally, students were expected to interpret the various differences they identified in the sense of paraphrasing the differences between two groups in context, rather than just describing these differences.

Although students in this course did receive instruction on conducting randomization tests with TinkerPlots towards the end of the course (for more detail see Frischemeier & Biehler, 2014), this study reports on efforts to develop students' abilities to identify, describe, and interpret group differences.

3.2. PARTICIPANTS, TASK, AND DATA COLLECTION

The data presented here were drawn from a laboratory study conducted at the end of one implementation of the *Developing Statistical Reasoning with TinkerPlots* course. There were 22 preservice teachers enrolled in the course, 15 of whom were pursuing degrees to teach at the primary level and 7 of whom were pursuing degrees to teach at the secondary level. Of the 22 preservice teachers enrolled in this course, eight agreed to participate in the laboratory study (6 primary and 2 secondary preservice teachers). In addition to enrolling in our course, all participants had previously completed a course in elementary statistics (data analysis, combinatorics, and probability theory) as part of their basic studies.

These eight participants were randomly placed into one of four pairs (All names are pseudonyms): Hilde and Irene (preservice teachers for primary school), Conrad and Maria (preservice teachers for secondary school), Ricarda and Laura (preservice teachers for primary school), and Sandra and Luzie (preservice teachers for primary school).

At the conclusion of the course, each of the four pairs of students were asked, in a laboratory setting, to solve a comparing two groups task using the TinkerPlots software. Given the prevalence of group comparisons in real datasets throughout the course, we selected a large, multivariate dataset for participants to use when making group comparisons. In particular, the dataset used throughout this study consisted of a random sample of German employees taken from a dataset available at the German Bureau of Statistics. The original dataset (called VSE), available at the German Bureau of Statistics, contained 60,552 cases selected from the population of all German employees using stratified random sampling techniques. The dataset contained variables such as monthly income, gender, region, status of employment, etc. To guide participants' exploration of this dataset, each pair of students was given a task (named the VSE task) which asked students to consider the ways in which male and female German employees differed with respect to their monthly income. Our goal was for participants to see this task as a comparison of two samples rather than to make further inferences in regard to the population.

As each group of students worked through the VSE task they were asked to think aloud, describing their intentions and planned activities throughout the group comparison process. Their verbal communications and computer work were recorded using Camtasia. Following the completion of the task, students written notes and TinkerPlots files were also collected. Prior to elaborating on the methods used for data analysis in this study, we first present a possible way to work through the VSE task in order to familiarize the reader with the task and our expectations for student work.

Anticipated procedure of our participants on the VSE task In the following we will point out possible ways to work on the VSE task from the perspective of the knowledge our participants may have ideally acquired in our course.

As students primarily generated graphs like boxplots, histograms, and stacked dot plots in our course, we expected our participants to produce these kinds of graphs in TinkerPlots for the VSE task. Figure 1 shows the typical displays for group comparisons in TinkerPlots: boxplots (Figure 1a), histograms (Figure 1b), and stacked dot plots (with dividers, Figure 1c). As students worked through the task we expected them to work out differences in center (mean and median), to identify the shift between the two distributions, to make p-based and q-based comparisons, and also to work out differences in regard to the skewness of both distributions.

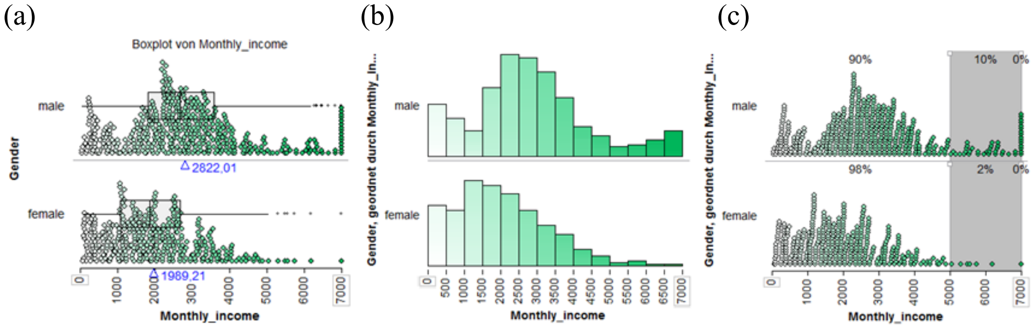


Figure 1. Possible TinkerPlots graphs for comparing groups (a) boxplots, (b) histograms, (c) stacked dot plots

As mentioned in Section 3.1, throughout the course students were asked to decide between producing and analyzing all the displays at once or generating and analyzing single displays successively. We also expected our students to go beyond a descriptive level when comparing two groups and interpret the differences between two groups in the sense of paraphrasing the differences between two groups in the context. Thus when starting to work on the VSE task, our participants could, for example, produce a

TinkerPlots graph like the one in Figure 1a, calculate the means of both distributions, and identify that the mean of the distribution of monthly income of male employees is larger than the mean of the distribution of monthly income of female employees. A more sophisticated approach would be to paraphrase the difference between the means in context (interpret) and to state that the male employees in this dataset earn 833€ more than female employees on average. Similar comparisons could also be done on the difference between the medians of both distributions. However, because both distributions are skewed, the participants might choose to only concentrate on differences in the medians of both distributions rather than on the difference in the mean monthly income of male and female employees.

Learners could also attend to differences with respect to spread when comparing the distributions in Figure 1a. For example, students might notice the interquartile ranges of both distributions and state that the interquartile range of the distribution of monthly income of male employees is larger than the interquartile range of the distribution of monthly income of female employees. A more sophisticated approach would be to state that the distribution of monthly income of male employees seems to be more heterogeneous when compared to the distribution of monthly income of female employees. The TinkerPlots graph in Figure 1a also enables learners to identify a shift between the distributions. For example, the learners might compare non-equivalent summary statistics (see Pfannkuch et al., 2004) and might say that the first quartile of the distribution of monthly income of male employees equals the median of the distribution of monthly income of female employees. Attending to this shift might lead the learner to state that the male employees tend to earn more than the female employees. So we see that a TinkerPlots graph like in Figure 1a enables learners to identify several differences between both distributions.

Students could also produce histograms (see Figure 1b), changing the bin width in a flexible way, and comparing the skewness of both distributions. Here one might identify that the distribution of the female employees is more skewed to the right than the distribution of the male employees. This observation might also lead to the assumption that the male employees tend to earn more than the female employees in this dataset. Another possibility is to use the divider feature in TinkerPlots (see Figure 1c) and to conduct p-based and q-based comparisons. With the TinkerPlots graph in Figure 1c, for example, one can make a p-based comparison by stating that 10% of the male employees earn 5000€ or more per month, but only 2% of the female employees earn 5000€ or more per month. Students could also use the divider feature in TinkerPlots to make q-based comparisons. Here they could compare the upper 10% of the distribution of monthly income of the male employees with the upper 10% of the distribution of monthly income of the female employees.

Based on the activities and norms of the course, these are the types of comparisons we would expect from our participants. As we see there are different qualities of comparisons. For example, sometimes differences between both distributions are just described, whereas other times these differences are interpreted in the sense of paraphrasing the differences in context. We will discuss this in further detail in Section 3.4.

3.3. DATA ANALYSIS

Transcription standards for computer supported data analysis (Kuckartz, 2012) were used to transcribe the students' communications and actions with the software as they worked through the VSE task. Based on these standards, transcriptions should include students' communications, their actions with the software, time stamps, and the plots students produced with the software. Figure 2 provides an example of a short excerpt of the transcript from Hilde's and Iris's work.

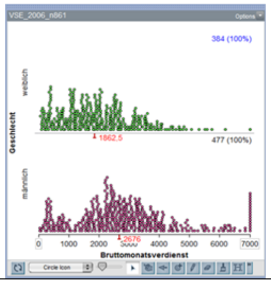
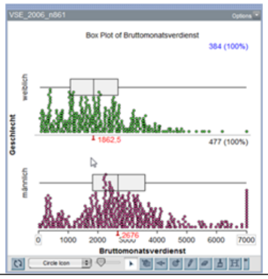
44	I:	Yes. Shall I display the median? Maybe we can find differences… #00:06:12#
45	H:	Yes. #00:06:13# *(Median (with numeric value) is displayed in TinkerPlots.)* 
46	H:	Then we could draw boxplots, (5 sec) Hats. #00:06:28# *(They choose „Hats“, and then „Boxplot“.)* 

Figure 2. Excerpt of transcript of Hilde and Iris

The transcripts were analyzed using qualitative content analysis (Mayring, 2010, 2015). In particular, we employed a structured-scaling approach (Mayring, 2015). To provide a detailed account of our analysis, we outline the steps of this approach below.

(1) We began by considering our research questions with respect to the current field of research (existing theory) by identifying and reviewing relevant literature.

(2) From our review of the literature, we derived two dimensions related to evaluating learners' performance when comparing groups with software. The focus of Dimension 1 is on the learner's ability to use TinkerPlots when conducting group comparisons. This dimension focuses exclusively on the software and rates users based on their ability to use TinkerPlots to execute planned actions that arise during the group comparison process. Dimension 2 covers how (center, spread, skewness, shift, p-based comparison, or q-based comparison) and in what way (descriptive or interpretative) two groups are compared.

(3) The basis of the analysis is a coding agenda for each of the research dimensions that emerged in step (2). The coding agenda consists of categories, definitions, anchor examples, and coding rules. In the definition of the category "it is precisely determined which text components belong in a given category" (Mayring, 2015, p. 377). As anchor examples, "concrete passages belonging in particular categories are cited as typical examples to illustrate the character of those categories" (Mayring, 2015, p. 377). Finally "where there are problems of delineation between categories, [coding] rules are formulated for the purpose of unambiguous assignment to a particular category" (Mayring, 2015, p. 377). The categories can arise "deductively, inductively and mixed" (Kuckartz, 2012, p. 69). We describe the generation of our coding agenda and elaborate on the frameworks used during analysis in Section 3.4.

(4) After developing a coding agenda, we selected the coding unit we would use. Although a coding unit can range in size (e.g., from a word to a phrase), a common choice of a coding unit is a unit of meaning. We describe our choice of the units and the definition of a unit of meaning in Section 3.4.

(5) Using the coding agenda, the materials (mostly transcripts) were analyzed according to structuring and scaling. Following this analysis, the first author and an independent researcher conducted a test coding. Discrepancies in coding were discussed between both researchers and the coding agenda was modified to reflect the discussion and the decision made by the two coders. For example, if the coders determined that the discrepancy was a result of an unclear definition, the definition of a category was modified to make it clearer.

(6) Alongside step (5) checks of reliability (e.g., inter-coder reliability) were conducted. Cohen's Kappa was used as a measure for inter-coder reliability. Cohens´s Kappa can be calculated with the following formula (see Mayring 2010, p. 120):

$$\kappa = \frac{\frac{x}{n} - \frac{1}{k}}{1 - \frac{1}{k}}$$

where x is the number of codes which match between researcher and independent coder, n is the overall number of codes, and k is the number of categories. According to Mayring (2001), $\kappa \geqslant 0.7$ is adequate for demonstrating inter-coder reliability. We present the findings related to reliability in the Results section.

(7) Lastly, to answer our research questions we conducted a frequency analysis of occurrence of the categories. A frequency analysis of the categories associated with Dimension 1 was used to provide an overview of our participants' abilities to enact their planned group comparisons for the VSE task in TinkerPlots. Similarly, a frequency analysis of the categories associated with Dimension 2 was used to provide an overview of the group comparison elements used by our participants (structural aspect) as well as the quality (scaling aspect) of how the group comparison elements were used by our participants (i.e., to describe an observed difference or to interpret an observed difference in context). Lastly, a frequency analysis was used to identify possible relationships between Dimension 1 and Dimension 2. We present the results of the frequency analysis for this study in the Results section.

3.4. THE FRAMEWORKS FOR THE DATA ANALYSIS

In this section, we outline our frameworks and expand on the qualitative content analysis methods mentioned in Section 3.3. In the literature review we identified two dimensions, Dimensions 1 and 2 (see Section 2.3). We will now outline the construction of the frameworks for both dimensions.

For Dimension 1 we developed a framework for TinkerPlots competence (see Table 2). We chose Mayring's (2010) structural-scaling content analysis approach and used Maxara (2009) as a basis for this framework. Maxara's framework rated learners' Fathom competence when simulating chance experiments with Fathom. We adapted the categories of this framework for TinkerPlots skills using a

Table 2. Framework for rating students´ TinkerPlots skills

Step	Definition	Example
TinkerPlots skill high	Learners have a concrete plan in mind and can fulfill it with TinkerPlots.	Conrad & Maria: "Let's do a boxplot." *Conrad and Maria produce a boxplot in TinkerPlots.*
TinkerPlots skill medium	Learners have a concrete plan in mind and can fulfill it with TinkerPlots after a trial-and-error approach,.	Hilde & Iris are unsure which button is for displaying the mean and which button is for displaying the median.
TinkerPlots skill low	Learners have a concrete plan in mind and cannot fulfill it with TinkerPlots.	Laura & Ricarda want to plot a boxplot in TinkerPlots. After some time Laura said: "I do not know how."

deductive approach. As the participants of the laboratory study were asked to articulate aloud their thinking processes when working through the task, we rated their performance with the software as "TinkerPlots skill high" if the pair was able to complete their idea (which occurred when thinking aloud) in TinkerPlots successfully. If students experienced slight problems using TinkerPlots (e.g., a trial-and-error approach) to complete their planned approach to the task, we rated their TinkerPlots skill as "medium." Lastly, if students were unable to use TinkerPlots to execute their intended plan for analyzing the task, we rate their TinkerPlots skill as "low." Table 2 presents the framework used for rating students' TinkerPlots skills and provides illustrative examples of each rating level.

For Dimension 2 we developed a framework for evaluating learners' performance when comparing groups (see Table 3). As with the framework developed for Dimension 1, we leveraged Mayring's (2010) structural-scaling content analysis approach. Additionally, the findings of our literature review were used to construct the codes. First, we considered *Which elements are used by the learners?* And second, *To what extent were learners able to interpret their findings in a group comparison process with TinkerPlots?* Using a deductive approach we identified and named the categories in our framework: "center," "spread," "shift," "skewness," "p-based," and "q-based" (see Table 1). We sought to refine these categories using an inductive approach (Kuckartz, 2012, p. 69). In particular, when analyzing the transcribed data of the group comparison processes of the four pairs, several comparison elements different from those listed in Table 1 arose inductively. In deciding whether to add these additional categories to the framework we decided to add them only if they occurred within the work of at least two pairs. As we did not find any additional element that occurred during the comparison process of at least two pairs, we did not add any additional categories.

Table 3. Framework for rating students´ statistical reasoning when comparing groups with TinkerPlots

Item	High quality	Medium quality	Low quality
Center	Measures of center (mean, median) are compared in a quantitative way and are interpreted.	Measures of center (mean, median) are compared in a qualitative way and are not interpreted.	Measures of center (mean, median) are compared inappropriately.
Spread	Measures of spread (IQR) or informal descriptions of spread (such as "density," "close") are compared and interpreted.	Measures of spread (IQR) or informal descriptions of spread (such as "density," "close") are compared and not interpreted.	Spread is compared using inadequate measures (like range) and/or are interpreted incorrectly.
Shift	Shift between both distributions is quantified correctly (with comparing the position of the middle 50% or with using the "Shift model")	Shift between both distributions is described in a qualitative way.	Shift between both distributions is worked out incorrectly.
Skewness	Skewness of both distributions is described correctly and the differences between the distributions are interpreted correctly.	Skewness of both distributions is described correctly but not interpreted.	Differences in skewness are worked out incorrectly.
p-based	p-based differences are identified and interpreted.	p-based differences are identified but not interpreted.	p-based differences are worked out incorrectly.
q-based	q-based differences are identified and interpreted.	q-based differences are identified but not interpreted.	q-based differences are worked out incorrectly.

Learners' discussions of their findings were rated using a scale of high-medium-low. This approach is consistent with Mayring's (2010) procedure of a structuring and scaling content analysis. In general, we rated a phrase as "high" when participants interpreted the difference by paraphrasing the differences in context. For example, if the difference between the means was interpreted in a way such as "the men tend to earn more money than women" (paraphrasing in context), we rated this aspect as "high" because the difference between the two means is interpreted in context ("…tend to earn more money than…"). This is similar to the assessor level identified by Pfannkuch (2007). We rated a phrase "medium" if a difference between the distributions was described but not interpreted, and "low" if the difference is worked out incorrectly. For example, when comparing the means of two distributions, the phrase "the mean of A is larger than the mean of B" would be rated as "medium" because in this instance the difference between the means is stated but not interpreted. Table 3 provides definitions of each coding element (see also Frischemeier & Biehler, 2016, p. 646).

Mayring (2010) points out that key examples are necessary for coding. Therefore, Table 4 (also see Frischemeier & Biehler, 2016, p. 647) provides key examples of all of the components of the framework for statistical reasoning when comparing groups with TinkerPlots.

Table 4. Key examples for coding "Statistical reasoning when comparing groups with TinkerPlots"

Item	High quality	Medium quality	Low quality
Center	The men earn 29.5% more than women on average. (Laura & Ricarda)	The mean of men is higher than the mean of women (Hilde & Iris)	No example.
Spread	The middle 50% of men spreads more than the middle 50% of the women. (Hilde & Iris)	The Interquartile Ranges of the distributions are almost identical. (Conrad & Maria)	No example.
Shift	The first quartile of the distribution of monthly income of male employees equals the median of the distribution of female employees, so the male employees tend to have a larger monthly income then the female employees. (No example from our data)	The middle 50% of men are shifted right compared to the middle 50% of women. (Hilde & Iris)	No example.
Skewness	The distribution of men seems to have some peaks but the distribution of women seems to be right skewed, so there might be more women earning little money compared to men. (Laura & Ricarda)	Here [distribution of salary of women] we can find a peak at 400€…the men [distribution of salary of men]…okay there is also a peak, but it is not so high. (Laura & Ricarda)	No example.
p-based	10% of the men earn more than 5000€, only 2% of the women earn more than 5000€, so the male employees tend to have a larger monthly income than female employees. (No example from our data)	10% of the men earn more than 5000€, only 2% of the women earn more than 5000€. (Sandra & Luzie)	No example.
q-based	The lower 25% of the women is smaller than the lower 25% of the men, so the male employees tend to have a larger monthly income than female employees. (Laura & Ricarda)	The lower 25% of the women earn between 71 € and 1076.50 €. The lower 25% of the men earn between 47 € and 1825 €. (Conrad & Maria)	No example.

The transcribed data were coded using these two frameworks (Table 2 and Table 3). The analysis unit included the transcripts and the written notes of the four pairs of students: Hilde & Iris, Conrad & Maria, Laura & Ricarda und Sandra & Luzie. The coding unit was a unit of meaning. When coding for TinkerPlots skills, a unit of meaning consisted of the articulation of the participants' intended action with TinkerPlots and their actual TinkerPlots action (see for example Table 2). When coding for statistical reasoning when comparing groups, a unit of meaning consisted of a comparison statement (e.g., "In 2006 the men earn 29.5% more on average than the women"). In general, we avoided multiple coding in both dimensions. Thus, the codes produced by the two frameworks can be seen as disjoint. We did however, allow for multiple codes within one framework. For example, if a learner articulated two comparisons in one statement (e.g., "the men have a larger mean and there is a shift to the right"), that statement received multiple codes (i.e., one coding in regard to center and one coding in regard to shift). Coding was done with support of the qualitative computer coding software MAXQDA (Kuckartz, 2012) by the first author.

4. RESULTS

In this section, we present the results of our analysis. We begin by providing an overview of all four pairs' performance on the VSE task, with particular attention paid to assessing the quality of each group's software skills and statistical reasoning. Then we provide a closer look at each of the four pairs of students' work. In particular, for each pair we present details of the statistical features (e.g., center, spread, etc.) attended to when comparing the distribution of monthly incomes of men and women in the VSE task and provide illustrative examples of the pair's verbal, written, and TinkerPlots work.

4.1. OVERVIEW: TINKERPLOTS SKILLS AND STATISTICAL REASONING

TinkerPlots skills Students' TinkerPlots skills were coded by the first author using the framework presented in Table 2. The transcript of Hilde's and Iris's activity during the VSE task was used to establish inter-coder reliability. This double coding process resulted in a value of $\kappa = 0.8558$, surpassing the requirements set forth by Mayring ($\kappa \geqslant 0.7$). Table 5 presents an overview of the ratings for each pair of students' TinkerPlots work throughout the VSE task.

Table 5. Percentage of codes "TinkerPlots skill" (absolute frequencies in brackets)

	TinkerPlots skill		
Pair	High	Medium	Low
Conrad & Maria	75.9% (22)	3.4% (1)	20.7% (6)
Hilde & Iris	80.5% (33)	17.1% (7)	2.4% (1)
Laura & Ricarda	81.5% (22)	7.4% (2)	11.1% (3)
Sandra & Luzie	45.5% (5)	9.0% (1)	45.5% (5)
Overall	75.9% (82)	10.2% (11)	13.9% (15)

Throughout the activity, three out of the four pairs of students consistently were able to use TinkerPlots to perform their intended plans for making group comparisons. In particular, apart from Sandra and Luzie, each pair of students exhibited high TinkerPlots skills at least 75% of the time. Overall, approximately 76% of the codes related to students' TinkerPlots skills were coded as "high."

Both Hilde and Iris and Laura and Ricarda were able to accomplish nearly every statistical activity they planned in TinkerPlots. In particular, Hilde and Iris were able to execute their intended approaches for answering the VSE task in TinkerPlots with little to no problems 98% of the time. Similarly, Laura and Ricarda were able to execute their intended approaches in TinkerPlots with little to no problems 89% of the time. Conrad and Maria were also relatively successful using TinkerPlots throughout the activity, experiencing little to no problems using the technology to execute 79.3% of their plans for comparing the distributions of monthly income of male and female; however, they also were unable to utilize

TinkerPlots for their intended analysis 20.7% of the time. Throughout their work, Sandra and Luzie experienced consistent difficulties when working with TinkerPlots. That is, although Sandra and Luzie were able to effectively use TinkerPlots to implement their plans 45.5% of the time, they were unable to execute their plans using TinkerPlots 45.5% of the time. In particular, we observed that Sandra and Luzie were unable to use the divider tool on their own and were unable to generate a boxplot in TinkerPlots.

Statistical reasoning when comparing groups Students' statistical reasoning components were coded by the first author using the framework presented in Table 3. As before, inter-coder reliability was established by coding the transcripts of Hilde's and Iris's activity during the VSE task and resulted in a value of κ = 1.000. Again, this value exceeds the required value (κ ⩾ 0.7) proposed by Mayring. Table 6 provides insight into the level of statistical reasoning the students demonstrated while working on the VSE task.

Table 6. Frequency of codes "Statistical reasoning when comparing groups" (absolute frequencies in brackets)

	Statistical reasoning		
Pair	High	Medium	Low
Conrad & Maria	0% (0)	100% (4)	0% (0)
Hilde & Iris	20% (2)	80% (8)	0% (0)
Laura & Ricarda	60% (6)	40% (4)	0% (0)
Sandra & Luzie	0% (0)	100% (4)	0% (0)
Overall	28.6% (8)	71.4% (20)	0% (0)

As can be seen in Table 6, none of the pairs of students demonstrated low statistical reasoning while working on the VSE-task. This implies that these pairs of students were able to select appropriate approaches or use important statistical features in an appropriate manner to describe or interpret differences between the distributions of monthly income of men and women. Whereas each pair of students demonstrated the ability to appropriately describe differences in these distributions, only two groups of students were able to adequately interpret their findings using the context of the problem. In particular, Hilde and Iris moved beyond simply describing group differences to interpreting their findings (and showed high statistical reasoning) 20% of the time, while Laura and Ricarda interpreted the results of their group comparisons (and showed high statistical reasoning) 60% of the time. Although both Conrad and Maria and Sandra and Luzie were able to adequately describe group differences using a variety of statistical measures and approaches, none of the differences they described were interpreted using the context of the problem. It is possible that both these pairs of students thought the goal of the VSE task was to simply identify and describe differences between the two groups rather than to interpret the differences they observed within the problem context. However, this explanation does not seem particularly viable because throughout our course students were instructed to and expected to not only describe the differences they observed when making group comparisons, but also to interpret those differences in context.

Overall, these four pairs of students demonstrated high statistical reasoning 28.6% of the time, and medium statistical reasoning 71.4% of the time. These results provide some evidence that the preservice teachers from our course are capable of working out differences when comparing groups in real datasets with TinkerPlots.

In the sections that follow we provide details about the type of approaches and statistical features students attended to when making group comparisons during their work on the VSE task. We present findings for each pair of students independently, assessing the quality of their statistical reasoning, and providing some exemplary excerpts from their work on the task.

4.2. CONRAD AND MARIA

Table 7 presents the distribution of codings for Conrad and Maria.

Table 7. Frequency of codes "Statistical reasoning when comparing groups" – The case of Conrad & Maria

	Statistical Reasoning		
Conrad & Maria	High	Medium	Low
Center	0	0	0
Spread	0	2	0
Skewness	0	0	0
Shift	0	1	0
p-based	0	0	0
q-based	0	1	0
Overall	0 (0%)	4 (100%)	0 (0%)

Conrad and Maria used three comparison elements when working out differences between the monthly incomes of male and female German employees. In particular, they attended to the spread and shift of the two distributions and made a q-based comparison when working out differences between the two distributions. All of their comparison elements were rated as medium quality. With respect to spread, Conrad and Maria observed that "the interquartile ranges [of the distributions of incomes of men and women] are almost identical." This assertion demonstrates a medium quality of statistical reasoning about spread because the group mentioned that the interquartile range of both distributions were identical but did not interpret this finding.

When attending to the shift between the distributions and making q-based comparisons, this group also relied on describing their findings rather than describing and interpreting their findings in context. For example, when making the comparison based on the shift between the distributions of monthly income of male and female employees, Conrad and Maria created boxplots in TinkerPlots (see Figure 3) and stated, "the box of men is shifted more to the right than the box of women." This comparison via shift is rated as medium comparison, because (as with their previous comparisons) we see that Conrad and Maria only state their observation (shift of the boxes) but do not interpret it (e.g., that the male employees tend to have higher monthly incomes than the female employees). Conrad and Maria also made a q-based comparison (see Figure 4).

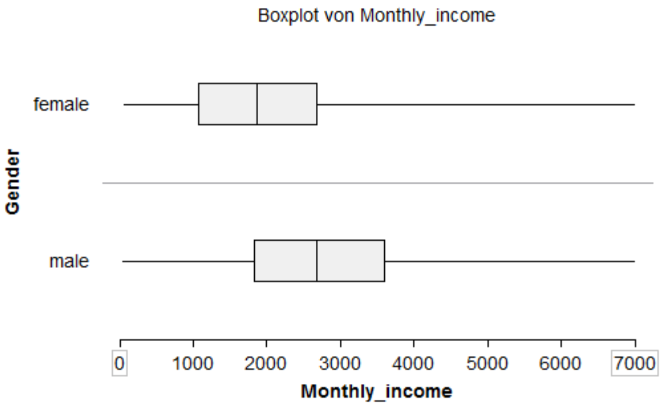
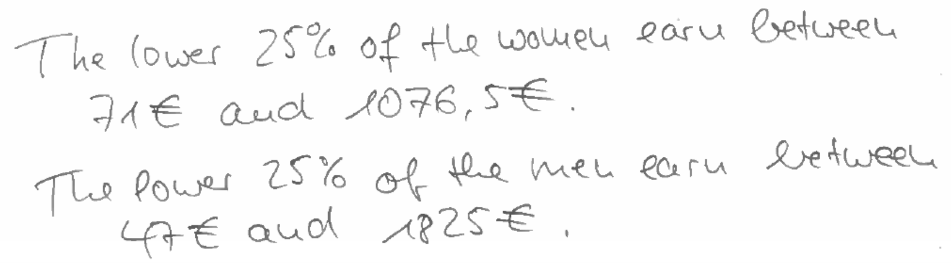


Figure 3. One of the TinkerPlots graphs produced by Conrad & Maria



The lower 25% of the women earn between
71€ and 1076,5€.
The lower 25% of the men earn between
47€ and 1825€.

Figure 4. q-based comparison done by Conrad & Maria (written note)

This q-based comparison was rated medium, because the components of the q-based comparison are only described but not interpreted and not compared.

4.3. HILDE AND IRIS

Table 8 presents the distribution of codes for Hilde and Iris.

Table 8. Frequency of codes "Statistical reasoning when comparing groups" – The case of Hilde & Iris

	Statistical resasoning		
Hilde & Iris	High	Medium	Low
Center	0	2	0
Spread	2	0	0
Skewness	0	0	0
Shift	0	4	0
p-based	0	2	0
q-based	0	0	0
Overall	2 (20%)	8 (80%)	0 (0%)

Hilde and Iris attended to center, spread, and shift, and also made p-based comparisons when comparing the monthly incomes of male and female German employees. Although they used various approaches to make group comparisons, they appeared to prefer to work out differences by attending to shift (four times). For example, after generating the TinkerPlots graph displayed in Figure 5, they stated, "This is interesting, here, the first quartile of male employees starts here, where the median of the women

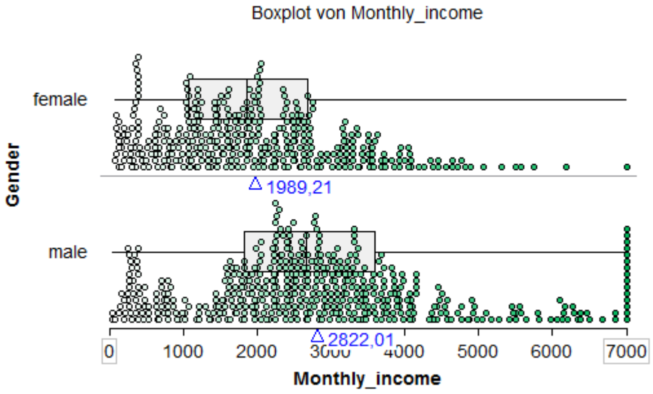


Figure 5. One of the TinkerPlots graphs produced by Hilde & Iris

is." They also compared measures of center for the two distributions to work out differences. For example, they noticed that, "The mean of men is higher than the mean of women" (see Figure 5). In both of these statements, Hilde and Iris describe the difference between the statistical measures for each distribution without attempting to interpret it. When comparing the mean monthly income of males and the mean monthly income of females, a more sophisticated statement would be that the male employees have a larger income than the female employees on average. When attending to center and shift and when making p-based comparisons, Hilde and Iris relied solely on describing differences between the statistical features they attended to for each group rather than interpreting them. Thus, Hilde and Iris demonstrated medium statistical reasoning when working out the majority of their group comparisons.

Hilde and Iris also used the TinkerPlots graph displayed in Figure 5 when comparing the spread of the distributions. For example, they stated that, "The middle 50% of men spreads more than the middle 50% of the women." Because the spread was measured and interpreted using the middle 50% of the data (interquartile range), Hilde and Iris were classified as demonstrating high statistical reasoning about spread when comparing distributions. This statement is representative of the other assertion they made when describing and interpreting differences in spread between the two distributions.

4.4. LAURA AND RICARDA

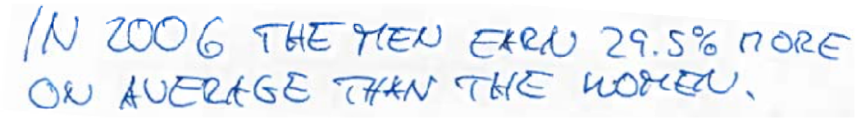
Laura and Ricarda were the only pair of students that employed all group comparison concepts when attempting to work out similarities and differences between the two distributions (for details see Table 9).

Table 9. Frequency of codes "Statistical reasoning when comparing groups" – The case of Laura & Ricarda

	Statistical reasoning		
Laura & Ricarda	High	Medium	Low
Center	2	0	0
Spread	1	0	0
Skewness	2	1	0
Shift	0	1	0
p-based	0	2	0
q-based	1	0	0
Overall	6 (60%)	4 (40%)	0 (0%)

Overall, Laura and Ricarda discussed similarities and differences between the distributions of monthly income between male and female German employees ten times. Of these, the pair made group comparisons based on skewness most frequently (3 times), followed by p-based comparisons and comparisons using measures of center (2 times each). Laura and Ricarda also leveraged q-based comparisons and comparisons using measures of spread and shift, though these group comparison concepts were used less frequently than the others (1 time each). As can be seen in Table 9, in 60% of the comparisons Laura and Ricarda demonstrated high statistical reasoning; and in 40% of the comparisons they relied on describing rather than interpreting, thus demonstrating medium statistical reasoning. For example, when comparing the skewness of the two distributions Laura and Ricarda stated, "The distribution of men seems to have some peaks but the distribution of women seems to be right skewed, so there might be more women earning little money compared to men." As can be seen from this excerpt, Laura and Ricarda go beyond describing to interpreting what they notice. In particular, they determined that "there might be more women earning little money compared to men" by recognizing that the distribution of income for women "seems to be right skewed" whereas for men it is not (or rather it is more symmetric than the distribution of incomes for females).

When discussing differences between the "center" of the two distributions, Laura and Ricarda also demonstrated high statistical reasoning. Figure 6 presents an example of their written work when comparing the average monthly income of male and female German employees. Laura and Ricarda compare the means of both distributions in a multiplicative way and interpret this difference ("on average").



IN 2006 THE MEN EARN 29.5% MORE
ON AVERAGE THAN THE WOMEN.

Figure 6. Comparison of center done by Laura & Ricarda (written note)

When making p-based comparisons and group comparisons based on shifts between the distributions of monthly income, Laura and Ricarda only showed a medium level of statistical reasoning. Figure 7 presents the TinkerPlots graph this pair generated and discussed when making their comparison based on shifts. This particular graph is known as a hat plot in TinkerPlots where the "hat" estimates the location of the central "clump" of data and the "brim" extends to all data points on either side of the hat. Using the divider tool in TinkerPlots, Laura and Ricarda were able to adjust the divider to match the hat and calculate the relative frequency of cases contained in the center clump. Although they are able to calculate these relative frequencies, they do not refer to them in their comparison process. However, they do describe that the location of the hat is shifted when describing differences in the plots of the two distributions.

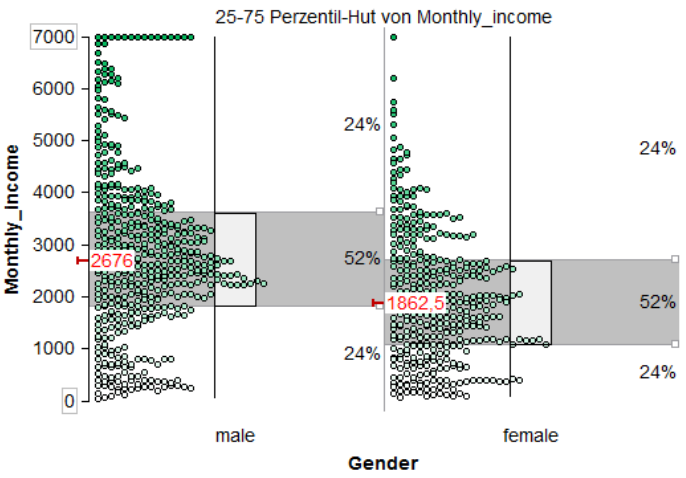


Figure 7. One of the TinkerPlots graphs produced by Laura and Ricarda

4.5. SANDRA AND LUZIE

While working through the VSE task, Sandra and Luzie only made p-based comparisons (see Table 10). Figure 9 presents one example of a p-based comparison that Sandra and Luzie wrote down when investigating the graph in Figure 8. In the TinkerPlots graph (Figure 8) we see that the pair used the divider tool in TinkerPlots to calculate the relative frequencies of the intervals [0€, 2000€], [2000€, 5000€], and [5000€, 7000€] for both distributions.

Table 10. Percentage of codes "Statistical reasoning when comparing groups" – The case of Sandra & Luzie

	Statistical reasoning		
Sandra & Luzie	High	Medium	Low
Center	0	0	0
Spread	0	0	0
Skewness	0	0	0
Shift	0	0	0
p-based	0	4	0
q-based	0	0	0
Overall	0(0%)	4(100%)	0(0%)

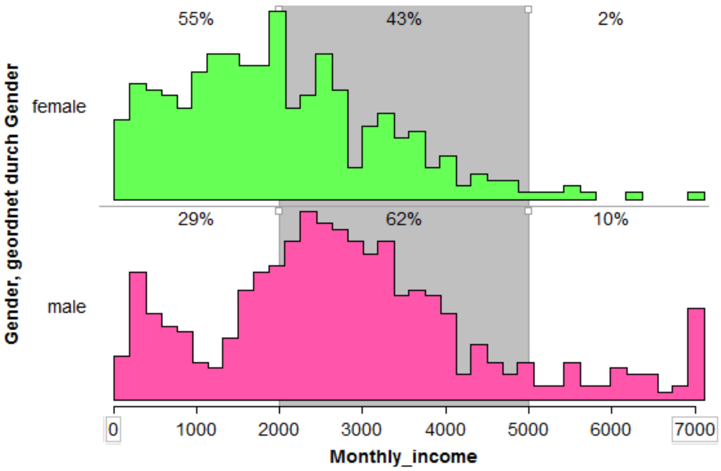
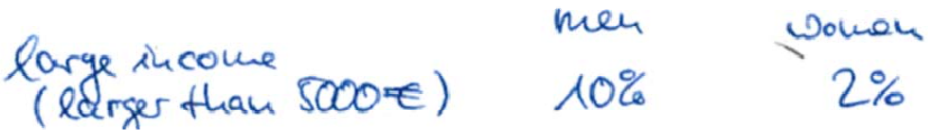


Figure 8. One of the TinkerPlots graphs produced by Sandra and Luzie

For Figure 9, Sandra and Luzie make a p-based comparison by comparing the percentage of male and female employees that have monthly incomes exceeding 5000€. Although their comparison reflects that there is a difference in the percentage of male employees and female employees that make more than 5000€ per month (i.e., 10% of male and 2% of female), they neither compare nor interpret these percentages.



	men	women
large income (larger than 5000€)	10%	2%

Figure 9. p-based comparison done by Sandra & Luzie (written note)

5. DISCUSSION

Table 11 provides insight into some of the successes and challenges students experienced by presenting an overview of the percentage of codes for all four pairs of students in both dimensions.

Table 11. Percentage of codes "TinkerPlots Skill" and "Statistical reasoning when comparing groups" – an overview of the distribution of codings over all four pairs

	Conrad & Maria	Hilde & Iris	Laura & Ricarda	Sandra & Luzie
TP skill high	75.9%	80.5%	81.5%	45.5%
TP skill medium	3.4%	17.1%	7.4%	9.0%
TP skill low	20.7%	2.4%	11.1%	45.5%
Stat.reasoning high	0%	20%	60%	0%
Stat.reasoning med	100%	80%	40%	100%
Stat.reasoning low	0%	0%	0%	0%

In general, each pair of students experienced both successes and challenges using TinkerPlots to enact their plans for making group comparisons. When compared to the other pairs of students, Sandra and Luzie experienced the greatest difficulty using TinkerPlots. In particular, they were unable to execute their plans using TinkerPlots 45.5% of the time. Apart from some difficulties observed in Sandra's and Luzie's work, the majority of students' were able to consistently use TinkerPlots to make group comparisons throughout the VSE task. Additionally, participants in this study were consistently able to identify appropriate ways to compare the two distributions and to describe differences between the distributions of monthly income for male and female employees. Only Hilde and Iris and Laura and Ricarda moved beyond describing the group comparisons to interpreting the differences they observed in the context of the given problem. Given that interpreting group differences (in the sense of paraphrasing the differences between groups in context) was a central focus of the course, we expected that all participants would make efforts to interpret their worked out differences. However, this apparent lack of interpretation is not entirely surprising, as other studies have noted difficulties students' experience when transitioning from describing to interpreting when making group comparisons (e.g., Biehler, 1997; Pfannkuch et al., 2004, Pfannkuch, 2007).

Although the frameworks we developed were designed to measure students' performance of group comparisons across two different dimensions (i.e., TinkerPlots skills and Quality of statistical reasoning when comparing groups), some insight can be gleaned by comparing students' performance across these two dimensions. For example, when comparing the percentage of codes each group received across both dimensions, three types of pairs emerge. In the first type, the pair of students was able to consistently identify and describe fundamental group comparison concepts, made some attempts to interpret their findings, and consistently utilized TinkerPlots to enact their planned comparisons (either immediately or using trial and error). Laura's and Ricarda's performance as well as Hilde's and Iris's performance on the VSE task is consistent with this description. In the second type, the pair of students was able to consistently use TinkerPlots as a tool to perform their planned group comparisons but never attempted to move beyond describing differences between the two groups. Conrad's and Maria's performance throughout the VSE task is consistent with this description. Lastly, in the third type, the pair of students (Sandra and Luzie) was able to identify and describe fundamental group comparison concepts but showed inconsistency using TinkerPlots to make group comparisons. Although it cannot be said that students with high TinkerPlots skills will necessarily exhibit high statistical reasoning skills when making group comparisons (see for example, Conrad and Maria), our participants provide some evidence that students' with high statistical reasoning skills also exhibit high TinkerPlots skills (see for example, Hilde and Iris, and Laura and Ricarda).Nevertheless, TinkerPlots allowed the participants to use their knowledge ("statistical reasoning") and their working methods for comparing groups. Additionally, as most participants were able to consistently use TinkerPlots to enact their plans throughout the task, TinkerPlots allowed our preservice teachers' to carry out their plans in the process of comparing groups in real datasets.

6. CONCLUSION

Providing future teachers with experiences to develop their statistical content knowledge and technological skills has become increasingly important in our data driven society. The course, *Developing Statistical Reasoning with TinkerPlots*, was designed to provide preservice primary and secondary teachers in Germany with just such experiences. Throughout the course, a fundamental activity the preservice teachers engaged in was making group comparisons using TinkerPlots. As a result, we wished to assess the performance of our preservice teachers in making group comparisons with TinkerPlots after taking this course. To aid in our analysis we leveraged prior research to develop a two dimensional framework (see Table 2 and Table 3) for analyzing qualitative video data of students' reasoning processes when comparing groups using TinkerPlots. The results of our study provide insight both at a local level and at a broader, more global level. We begin by describing local insights gleaned from the results of our study by focusing on implications with respect to the design of our course. We then take a more global perspective, pointing out implications for teaching and future research in statistics education.

Developing the ability to use TinkerPlots effectively to perform data analysis is particularly important for future teachers, especially if they plan to use technology as a tool in their future work as teachers. Our findings suggest that the design of the course assisted the majority of these students in developing the ability to use TinkerPlots to enact their planned approaches for making group comparisons. In particular, although one pair of students did experience some difficulties using TinkerPlots to make group comparisons in real data, the majority of the pairs of our study were able to successfully utilize TinkerPlots throughout the VSE task. That is, after creating a plan for the group comparison, most pairs were able to use TinkerPlots to perform that plan.

Our findings also suggest that the design of the course and the use of TinkerPlots as a tool for learning supported these students in learning to make group comparisons. In general, the participants in our study were able to use appropriate group comparisons concepts (like center, spread, skewness, shift, p-based comparisons, q-based comparisons) when working out differences between two distributions in real data. Prior research suggests that teaching students multiple strategies for making group comparisons is often not enough. For example, Biehler (2007b) and Frischemeier and Biehler (2011) found that, despite receiving instruction on a variety of group comparison concepts, students often relied almost exclusively on averages when making group comparisons. However, we observed that the majority of participants in our study were able to attend to a variety of group comparisons concepts after taking our course.

We did observe that our participants often relied on describing the comparisons they made rather than making interpretive statements. This is consistent with findings from other studies (e.g., Biehler, 1997). Given the importance interpretation plays in statistical reasoning and the apparent challenge students and teachers experience in interpreting their findings, future research should continue to study how to support learners' abilities to interpret statistical findings in context.

Although our findings suggest that the current design of the course and the use of TinkerPlots have the potential to support preservice teachers' abilities to make group comparisons in real data using TinkerPlots, they also suggest some aspects of the course that may need to be reconceived. Such implications might also assist others when designing different courses focused on developing teachers' tatistical reasoning, especially with respect to group comparisons. For instance, our analysis suggests that after taking our course these preservice teachers made relatively few q-based comparisons and seldom used skewness as a means to compare groups. Additionally, we observed one pair of students (Sandra and Luzie) focusing entirely on p-based comparisons. Thus when re-designing our course (or designing similar courses with the aim of developing students' statistical reasoning regarding group comparisons) teachers and researchers have to set the focus on taking into account all different elements when comparing groups. Especially for the case of Sandra and Luzie, who only used p-based comparisons to compare groups, teachers have to introduce students to other global features which allow them to compare two groups in a sustainable way (like center, shift, spread). Our analysis also suggests that efforts to improve our course could focus on better supporting students' abilities to interpret the differences they observe when making group comparisons within the problem context. One such approach for supporting this development might be to provide students with activities that focus on describing and interpreting the differences between group comparison concepts. For example, asking students to consider whether a particular group comparison concept would be “adequate” or “not-adequate” for supporting a given interpretation may assist students in recognizing the utility of various group comparison concepts. Additionally, providing students with more opportunities to determine and discuss appropriate interpretations for each of the group comparison concepts may support students’ abilities to move beyond describing the differences they see to interpreting what those differences mean in context. A group comparison scheme in the form of a process working sheet might guide learners through their group comparison process in regard to two aspects: (1) the scheme might structure the group comparison process, showing the learners the variety of different group comparison elements; and (2) the scheme can emphasize the interpretation of the group differences in context.

From a more global perspective, the framework developed and applied in this study (see Table 2 and Table 3) may assist teachers and researchers interested in scaffolding students’ comparative reasoning or in assessing students’ abilities when making group comparisons with TinkerPlots. For example, the framework for rating students’ statistical reasoning when comparing groups (Table 3) may assist teachers and researchers in identifying which group comparison elements students take into account during the group comparison process and reveal whether students interpret, describe, or incorrectly work out the differences they observe. In this way, the framework can support a teacher’s ability to identify challenges learners experience when comparing groups (like focusing on specific group comparison elements only or neglecting interpretation in context, etc.), providing them with the opportunity to intervene and better support their students through the learning process. Additionally, the framework for rating students’ statistical reasoning when comparing groups (Table 3) may assist teachers in establishing norms for making group comparisons in their classroom by providing them with a resource they can share with their students that shows the different quality levels of comparisons a learner can make during the group comparison process. The framework for rating TinkerPlots skills (Table 2) can also be adapted to rate skills of learners using other software tools (like Fathom^{TM} or Excel) in a data analysis process. The data analysis method (Mayring’s qualitative content analysis) we have used in our study can also be applied in other research studies in statistics education to generate frameworks and to assess learners’ statistical reasoning. The qualitative content analysis method offers a traceable approach to generate frameworks (in a deductive, inductive, or mixed way) for structuring and evaluating qualitative video data as we have seen in this study. Finally, we do wish to acknowledge several limitations of this study. First, given the small number of participants in our study (8 participants) and the fact that the participants volunteered to be a part of our study, we cannot generalize our findings to all students who were enrolled in the *Developing Statistical Reasoning with TinkerPlots* course. Additionally, given that these students engaged with specific curricular materials and have a certain background (preservice teachers at the University of Paderborn) we are not able to generalize our findings for other samples of preservice teachers.

ACKNOWLEDGEMENTS

Many thanks to Dana Kirin for proofreading and providing helpful comments on previous versions of this paper. We also thank Maxine Pfannkuch (as *SERJ* co-editor), the *SERJ* associate editor, and the two anonymous reviewers for providing constructive feedback and suggestions.

REFERENCES

Batanero, C., Burrill, G., & Reading, C. (Eds.) (2011). *Teaching statistics in school mathematics–Challenges for teaching and teacher education. A Joint ICMI/IASE Study: The 18th ICMI Study.* New York: Springer.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42–63.
[Online: https://iase-web.org/documents/SERJ/SERJ3(2)_BenZvi.pdf]

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Role of Technology in Teaching and Learning Statistics* (pp. 169–190). Voorburg, The Netherlands: International Statistical Institute.
[Online: https://www.dartmouth.edu/~chance/teaching_aids/IASE/14.Biehler.pdf]

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern - Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens empirischer Verteilungen. In M. Borovcnik, J. Engel, & D. Wickmann (Eds.), *Anregungen zum Stochastikunterricht* (pp. 97–114). Hildesheim, Germany: Franz Becker.

Biehler, R. (2007a). Denken in Verteilungen - Vergleichen von Verteilungen. *Der Mathematikunterricht, 53*(3), 3–11.

Biehler, R. (2007b, August). *Students' strategies of comparing distributions in an exploratory data analysis context*. Paper presented at the 56th session of the International Statistical Institute, Lisbon, Portugal.
[Online: https://www.stat.auckland.ac.nz/~iase/publications/isi56/IPM37_Biehler.pdf]

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In K. Clements, A. Bishop, C. Keitel, J. Kilpatrick & F. Leung (Eds.). *Third International Handbook of Mathematics Educ.* (pp. 643–689). New York: Springer

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Blum, W., Drüke-Noe, C., Hartung, R., Köller, O. (2006). *Bildungsstandards Mathematik: Konkret*. Berlin, Germany: Cornelsen Verlag.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Finzer, W. (2001). *Fathom Dynamic Statistics* (version 1.0) [Computer software]. Emeryville, CA: Key Curriculum Press.

Frischemeier, D. (2014). Comparing groups by using TinkerPlots as part of a data analysis task—Tertiary students' strategies and difficulties. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://icots.info/9/proceedings/pdfs/ICOTS9_8J3_FRISCHEMEIER.pdf]

Frischemeier, D. (2017). *Statistisch denken und forschen lernen mit der Software TinkerPlots*. Wiesbaden, Germany: Springer.

Frischemeier, D., & Biehler, R. (2011). Spielerisches Erlernen von Datenanalyse mit der Software TinkerPlots—Ergebnisse einer Pilotstudie. In R. Haug & L. Holzäpfel (Eds.), *Beiträge zum Mathematikunterricht 2011* (pp. 275–278). Münster, Germany: WTM.

Frischemeier, D., & Biehler, R. (2014). Design and exploratory evaluation of a learning trajectory leading to do randomization tests facilitated by TinkerPlots. In B. Ubuz, C. Haser & M. A. Mariotti (Eds.). *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (pp. 799–809). Ankara, Turkey: Middle East Technical University and ERME.

Frischemeier, D., & Biehler, R. (2016). Preservice teachers´ statistical reasoning when comparing groups facilitated by software. In K. Krainer & N. Vondrova (Eds.). *Proceedings of the 9th Congress of the European Society for Research in Mathematics Education* (pp. 643–650). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.

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

Hammerman, J. K. L., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistical Education Research Journal, 3*(2), 17–41.
[Online: https://iase-web.org/documents/SERJ/SERJ3(2)_Hammerman_Rubin.pdf]

Hasemann, K., & Mirwald, E. (2012). Daten, Häufigkeit und Wahrscheinlichkeit. In G. Walther, M. van den Heuvel-Panhuizen, D. Granzer, & O. Köller (Eds.). *Bildungsstandards für die Grundschule: Mathematik konkret* (pp. 141–161). Berlin, Germany: Cornelsen Verlag Scriptor.

Jacobbe, T., & Carvalho, C. (2011). Teachers' understanding of averages. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education* (pp. 199–209): New York: Springer.

Konold, C., & Miller, C. (2011). *TinkerPlots*: *Dynamic data exploration* (Version 2) [Computer software]. Emeryville, CA: Key Curriculum Press.

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung*. Weinheim, Germany: Beltz Juventa.

Madden, S. R. (2008). *High school mathematics teachers' evolving understanding of comparing distributions* (Unpublished doctoral dissertation). Western Michigan University, Michigan, USA.

Makar, K., & Confrey, J. (2002). Comparing two distributions: Investigating secondary teachers' statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*, Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://www.stat.auckland.ac.nz/~iase/publications/1/10_18_ma.pdf]

Maxara, C. (2009). *Stochastische Simulation von Zufallsexperimenten mit Fathom—Eine theoretische Werkzeuganalyse und explorative Fallstudie*. Hildesheim, Germany: Franz Becker.

Maxara, C. (2014). Konzeptualisierung unterschiedlicher Kompetenzen und ihrer Wechselwirkungen, wie sie bei der Bearbeitung von stochastischen Simulationsaufgaben mit dem Computer auftreten. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen—Using Tools for Learning Mathematics and Statistics* (pp. 321–336). Wiesbaden, Germany: Springer Spektrum.

Mayring, P. (2001). Combination and integration of qualitative and quantitative analysis. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, *2*(1).
[Online: http://www.qualitative-research.net/index.php/fqs/article/view/967]

Mayring, P. (2010). *Qualitative inhaltsanalyse: Grundlagen und techniken*. Wiesbaden, Germ: Beltz.

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to Qualitative Research in Mathematics Education* (pp. 365–380). Dordrecht, The Netherlands: Springer.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27–45.
[Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Pfannkuch.pdf]

Pfannkuch, M. (2007). Year 11 students' informal inferential reasoning: A case study about the interpretation of box plots. *International Electronic Journal of Mathematics Education*, *2*(3), 149–167.
[Online: http://www.iejme.com/makale_indir/330]

Pfannkuch, M., & Ben-Zvi, D. (2011). Developing teachers' statistical thinking. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 323–333). Dordrecht/Heidelberg/London/New York: Springer.

Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004, July). *Comparison of data plots: Building a pedagogical framework*. Paper presented at the Tenth International Congress on Mathematics Education (ICME-10), Copenhagen, Denmark.
[Online: http://iase-web.org/documents/papers/icme10/Pfannkuch.pdf]

Reading, C., & Canada, D. (2011). Teachers' knowledge of distribution. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education* (pp. 223–234). New York: Springer.

Rossman, A., Chance, B. L., & Lock, R. (2001). *Workshop statistics: Discovery with data and Fathom* (2nd Edition). Emeryville, CA: Key College Publishing.

Sánchez, E., da Silva, C. B., & Coutinho, C. (2011). Teachers' understanding of variation. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education* (pp. 211–221). New York: Springer.

Sill, H.-D. (2018). Zur Stochastikausbildung im Primarstufenlehramt. In R. Möller & R. Vogel (Eds.), *Innovative Konzepte für die Grundschullehrerausbildung im Fach Mathematik, Konzepte und Studien zur Hochschuldidaktik und Lehrerbildung Mathematik* (pp. 71–93). Wiesbaden, Germany: Springer Spektrum.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168.

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

DANIEL FRISCHEMEIER
University of Paderborn
Institute of Mathematics
Warburger Straße 100
33098 Paderborn
Germany