Preservice teachers' statistical reasoning when comparing groups facilitated by software

Daniel Frischemeier, Rolf Biehler

Preservice teachers' statistical reasoning when comparing groups facilitated by software

Daniel Frischemeier and Rolf Biehler

University of Paderborn, Paderborn, Germany, dafr@math.upb.de, biehler@math.upb.de

Comparing groups is a fundamental skill preservice teachers are supposed to gain after attending a statistics course at university level. Preferably these activities are embedded in the well-known PPDAC-Cycle and contain the exploration of real and motivating data. Adequate software such as TinkerPlots may support learners when exploring data and carving out differences between distributions of numerical variables. In this article we want to present first results of a study on statistical reasoning of preservice teachers while doing group comparisons with TinkerPlots.

Keywords: *Preservice teacher education, statistical education, group comparisons, software use, tinkerplots.*

INTRODUCTION

Comparing distributions of numerical variables lies in "the heart of statistics" (Konold & Higgins, 2003, p. 206). A question which may motivate a group comparison is for example "In which respect do men and women differ regarding their income?". This could be also posed in the daily-life media, so we see the importance of such an activity, not only for students but also for upcoming teachers. Preferably questions and activities are embedded in a cycle (like PPDAC, Wild & Pfannkuch, 1999) where students are enabled to generate their own statistical questions and hypotheses, to design a questionnaire for collecting data, to analyse their data and to draw conclusions from it. When analyzing large datasets, the use of adequate statistical software becomes inevitable. TinkerPlots (Konold & Miller, 2011) yields many requirements for statistical software for use at school and university level. At the one hand TinkerPlots can be seen as educational software supporting learners to learn data analysis, on the other hand it can be seen as a tool for exploring multivariate data. Additionally, the teacher can also use it as demonstration medium in the classroom. From our research point of view, we are primarily interested how preservice teachers compare groups with TinkerPlots. In this article we want to introduce a framework to rate preservice teachers´ skills when comparing groups in large datasets with TinkerPlots. Therefore we designed a course to deepen preservice teachers´ statistical knowledge and conducted an interview-video-study to evaluate in which way the participants are able to compare distributions with TinkerPlots in large datasets. First results of this study will be reported at the end of this article.

LITERATURE REVIEW

Comparing groups is an important domain in statistics education research: There are several empirical studies which concentrate on rating learners' statistical reasoning when comparing groups. We can differentiate between at least five different ideas in these studies (not in order).

A first idea is set by Watson and Moritz (1999) who investigated Australian 3–8 graders when comparing two data sets. The participants were given two data sets with test scores of two classes and were asked which class has done better in the test? The distributions of the variable test score was displayed as a stacked dot plot and the interview protocol offered different types of group comparisons, e.g. distributions which differed in the number of cases, in variation or in skew (Watson & Moritz, 1999, p. 151). The learners responses were rated via SOLO taxonomy where the responses where rated "unistructural", "multistructural" and "relational" and distinguished between the comparison of equal and unequal sizes. One major result of their study was, that students in higher grades tend to reason proportionally rather than younger students (for further details see Watson & Moritz, 1999, p. 153). A second main idea is displayed by Makar & Confrey (2002). They conducted a one

semester professional course for preservice teachers including group comparison tasks with Fathom and developed a "taxonomy for classifying levels of reasoning when comparing two groups" to evaluate their participants reasoning from an interview task (concluding the course regarding to comparing groups). The participants were asked to compare two distributions of test scores of two schools given as a stacked dot plot in Fathom (with no use of software itself). A crucial point in their analysis was in which way inferential terms (like "evidence" and "significance") were used in the comparison process of the participants and in which way the participants draw conclusions from samples. A third idea of research on comparing groups is given by the work of Biehler (2001) and (2007). Biehler gives a normative view on comparing groups and expresses which elements shall be included in a "good" group comparison. He mentions that *p*-based [1] and *q*-based [2] comparisons (Biehler, 2001, p. 110) might offer intuitive strategies for students and he also emphasizes an interpretation of the difference in the skewness of distributions as a possible comparison element (Biehler, 2001, p. 101). There is also an idea (fourth) which covers the use of software when comparing groups. Biehler (1997, p. 175) has set up a cycle and gives an overview on four phases "statistical problem", "problem for the software", "results of software use" and "interpretation of results in statistics" which are run when doing data analysis with software. Maxara (2009) designed a framework for evaluating learner´s software competencies when simulating chance experiments with Fathom. This is not directly related to group comparisons but nonetheless adaptable for evaluating leaner´s competences when using software (such as TinkerPlots) when comparing groups. Overarching for the fifth idea might be the work of Pfannkuch and colleagues (2004; 2006; 2007). In these research papers a framework for evaluating learner's competencies when comparing two distributions was developed. Since Pfannkuch (2007) is a succeeding study of Pfannkuch and colleagues (2004) and Pfannkuch (2006), we want to refer to Pfannkuch (2007) only. In this empirical study she gave a boxplot comparison task (see Pfannkuch, 2007, p. 157) to "Year 10"- students. They were given two boxplots, asked to compare them in the sense of making three statements to explain differences or similarities between the distributions. Pfannkuch (2007, p. 159) had a look on different statistical aspects that were used by the participants and set up categories for the evaluation of statistical reasoning elements when comparing two distributions by boxplots. On a structural level she distinguished between "summary", "spread", "shift" and "signal". Then she rated each statement regarding to its quality: "point decoder" (level 0), "Shape comparison describer" (level 1), "shape comparison decoder" (level 2) and "shape comparison assessor" (level 3). Main results of this study were that the students mostly refer to summary and spread elements, but neglected elements on shift and signal. Furthermore they tended to stay mostly on the describing and decoding but not on the assessing level when pointing out differences and similarities between both distributions.

All in all, we can derive three dimensions having an influence on the group comparison process from the literature review: *Software cycle when comparing groups* (Biehler, 1997), *competence of using software (TinkerPlots) when comparing groups* (Maxara, 2009) and *"statistical reasoning" when comparing groups* (Watson & Moritz, 1999; Biehler, 2001; Makar & Confrey, 2002; Pfannkuch et al., 2004; Pfannkuch, 2006; Biehler, 2007 and Pfannkuch, 2007). In this study the work of Pfannkuch seems to be the most interesting aspect: Watson and Moritz (1999) deal with given data and given distributions (but in datasets with a small amount of cases) and a focus on counting strategies and proportional reasoning. Makar & Confrey (2002) have had their crucial research point of interest on how learners draw conclusions from samples to a population while comparing samples. Pfannkuch offers an open framework which firstly structures learners' outcome in regard to the statistical element used and secondly rates this in form of quality. Since working with software offers a broad spectrum of statistical elements (e.g. center, spread, shift, etc.) which can be used in group comparisons even in large datasets, the framework of Pfannkuch with enrichment of Biehler´s (2001) suggestions (skewness, *p*- and *q*-based comparisons) seems to offer possible and adequate comparison elements when comparing groups and a solid basis for evaluating the outcomes of learners when comparing groups. These ideas and aspects motivated us to design a course for the education of preservice teachers in statistics in which we want to teach the comparison of groups (with TinkerPlots) with the elements (such as center, spread, etc.) described above.

COURSE "DEVELOPING STATISTICAL REASONING WITH TP"

The authors of this article have designed a course for preservice teachers called "Developing statistical reasoning with using the software TinkerPlots" (Frischemeier & Biehler, 2012) in the sense of the design based research paradigm (Cobb et al., 2003). In this course, which goal is the development of statistical content (but not pedagogical) knowledge, the participants go through the whole PPDAC-cycle (Wild & Pfannkuch, 1999) which includes analysing self-collected data with TinkerPlots and writing down findings in statistical reports. In the analysing section the participants got to know about how they could compare distributions via different aspects (such as center, spread, shift (see Pfannkuch, 2007) and skewness, *p*-based- and *q*-based-comparisons (see Biehler, 2001)) with TinkerPlots. At first they were taught to identify differences between distributions (regarding to center, spread, etc.), then they were told to interpret these differences. A norm set by us was to work out as many differences regarding to center, spread, etc. between both distributions as possible. At this stage we firstly want our participants to work out as many differences as possible and to interpret them. In a next step, not reported in this paper, the participants are asked to synthesize their findings (e.g., in form of writing a statistical report). For further details see (Frischemeier & Biehler, 2012).

RESEACH QUESTIONS

Since the ability of handling a large set of real and multivariate data is important for upcoming teachers in statistics, we want to investigate how preservice teachers explore large datasets and compare distributions with TinkerPlots. In this article we want to concentrate on the "statistical reasoning" component of preservice teachers when comparing groups with TinkerPlots only, so two research questions arise: Which group comparison elements (which were taught in our course - such as center, spread, shift, etc.) are used by the participants when comparing groups? How is the quality level of these group comparison elements used by the preservice teachers?

DESIGN OF THE STUDY

As part of the Ph.D. study of the first author, an interview-video study was designed in which the participants were asked to compare two distributions with TinkerPlots in pairs of two. For the selection of task we chose a task which deals with the exploration of the income distributions of male and female employees, which has got a lot of publicity in Germany with regard to gender biases in monthly incomes. The dataset taken from the German Bureau of Statistics was imported in TinkerPlots and contains 861 cases and more than 20 variables (such as gender, monthly income, region, kind of employment, etc.). This was drawn as a random sample out of 60,552 which itself was sampled at random (stratified) from the population of all German employees. Furthermore we handed out a TinkerPlots file containing the dataset and an exercise sheet where the participants were asked to make notes on it. After motivating the problem of a "gender difference income" with a newspaper article, the task for the participants was: "*In which way do the men and women differ regarding their income? Carve out differences in both distributions!*" Some impulses which differences can be carved out between the distributions can be found in (Biehler & Frischemeier, 2015). The video study was two-phased adapted from the design of Busse & Borreomeo-Ferri (2003). In phase 1, the "working phase", the participants work on the task in pairs and were forced to communicate to each other while doing the task. Figure 1 shows a TinkerPlots graph which displays the differences between both distributions using boxplots (and the mean) produced in TinkerPlots by participants during the working phase.

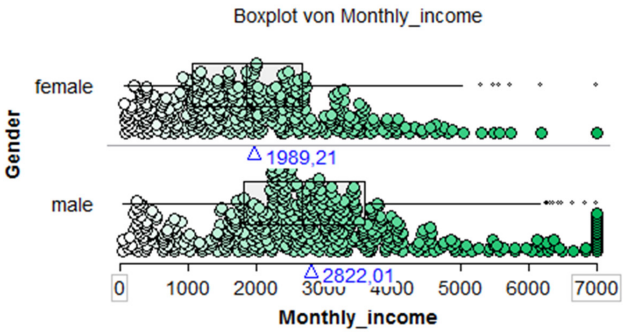

Figure 1: TinkerPlots Graph with boxplots and mean

In this phase there was no intervention by the interviewer (first author of this article). In phase 2 cognitive processes of the pairs should be revealed via "stimulated recall".

DATA, PARTICIPANTS AND METHODOLOGY

All in all 14 participants (7 pairs) took part in the study. All of them were preservice teachers for mathematics for primary and secondary school at the University of Paderborn and all of the participants attended the course "Developing statistical reasoning with TinkerPlots". The interviews were done 4–6 weeks after the last session of the course "Developing statistical reasoning with TinkerPlots". The participants were asked to work on the task in teams of two while they and their screen activities were video recorded. TinkerPlots files, exercise sheets and video recordings were also collected. The communication and action with TinkerPlots was transcribed. Our goal is to evaluate the whole communication of the participants regarding to their statistical reasoning elements when working on the task. So a large amount of data is to handle and there is also a need for an evaluation procedure which is comprehensible. The qualitative content analysis in the sense of Mayring (2010) "can be well applied in the case of having the intention to analyse a huge amount of transcribed data" and follows the "reduction of the huge amount of data in form of an analysis via category systems" (Kohlbacher, 2006). One main goal of this procedure is "to filter out a particular structure from the material" (Kohlbacher, 2006). Mayring (2010, p. 63) has pointed out different streams of the qualitative content analysis, such as "structural", "explicating" or "summarising". We want to structure the group comparison elements of the participants first and then scale (evaluate) them. The sequence of a structured-scaled qualitative content analysis (Mayring, 2010, p. 101) starts with choosing the analysis units (Step 1). In this case the analysis units are the transcribed data from the communication of the pairs when the participants were working on the task. As a second step the dimension of analysis is set up (step 2) – in our case this is *statistical reasoning when comparing groups*. In a further and third step we determine the characteristics of "dimensions of analysis" (Step 3) in form of setting up a category system, which is seen as the basis of this method where coding rules and key examples are given for an exact assignment between coding and data material and therefore define categories for evaluating the transcribed communication of the participants (Step 4). These categories can be defined deductively, inductively or mixed (Kuckartz, 2012, p. 62).

FRAMEWORK FOR COMPARING GROUPS

We used this construction of categories in the sense of Kuckartz (2012, p. 62) for our purpose. The aim was firstly to structure the transcribed data regarding the statistical elements used for group comparison and secondly to evaluate the quality of each element. So we took into account the elements "summary", "spread", "signal" and "shift" of the framework of Pfannkuch (2007) in the sense of a deductive approach. Since our participants were asked to find differences between both distributions and since they have had the possibility of a free data exploration with TinkerPlots, we have had to modify the categories of Pfannkuch (2007) for our purpose. Having a look on Biehler's (2001) normative point about group comparisons, where the comparison of skewness and *p*/*q*-based-comparisons also plays a huge part in the comparison process, we decide to add the elements "skewness", "*p*-based" and "*q*-based" to our framework. The further step sees an inductive refinement of the categories (Step 5) in the sense of Kuckartz (2012, p. 69). In this step 5 we have

	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 in a wrong way.
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 with inadequate measures (like range) and is wrongly interpreted.
Shift	Shift between both distributions is quantified correctly (with comparing the position of the middle 50% or with comparing non-corresponding numbers)	Shift between both distributions is described in a qualitative way.	Shift between both distributions is worked out in a wrong way.
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 of skewness are worked out in a wrong way.
p-based	p-based differences are identified and interpreted	p-based differences are identified but not interpreted	p-based differences are worked out in a wrong way
q-based	q-based differences are identified and interpreted	q-based differences are identified but not interpreted	q-based differences are worked out in a wrong way

Figure 2: Definitions of codings

gone through the data with our deductively developed categories and refined them inductively. After this process, we added the elements "center" instead of "summary" because we wanted to concentrate on the comparison of mean and median and not on the comparison of all summary statistics and we have left out the category "signal" which is a special element for boxplot comparisons but not necessarily for group comparisons in general. We did not focus on the construction of plots in TinkerPlots, we just focus on working out as many differences with TinkerPlots as possible. Mostly standard displays, also primarily used in our course, like stacked dot plots, histograms and boxplots were used. All in all we finally have the following elements for our analysis: "center", "spread", "shift", "skewness", "*p*-based" and "*q*-based". We see these elements as our categories (see Figure 2).

As we see in Figure 2 we also generated for each of these categories the ratings "high quality", "medium quality" and "low quality" to evaluate the quality of the use of the elements in the comparison process. Generally we coded a group comparison element used by our participants with a "high" quality, if the difference of the distributions was worked out quantitatively and was also interpreted (in the idea of Pfannkuch's category "assessor"). An element is coded in the sense of a "medium" quality, if the difference is at least worked out on a qualitative level ("X is higher than Y") but not interpreted. Finally a "low"-quality code is given if the difference is worked out in a wrong or in an inadequate way. For illustrating the definition of codings we want to give examples (see Figure 3) arisen from our data.

When having set up the category system, we chose (in a sixth step) a word as minimal coding unit and a phrase as maximal coding unit. With this agenda we have coded the transcripts of four of the seven pairs so far. In this paper we refer only to codings belonging to the working phase, the codings of the transcripts of the stimulated recall phase are not reported here. If codings of passages were unclear, we had a look in the video to clarify the situation. A further step (step 7: Revision of codings) included the discussion of the codings with an independent researcher and in the following the revision of categories and definitions of categories. Finally a frequency analysis of the occurrence of the several categories was made (step 8: frequency-analysis of occurrence of steps).

RESULTS

Let us have a look which group comparison elements were used by the teams and how well they did in using them when working on the task with TinkerPlots.

At first we can say that we have 23 codings in total. The codings of the elements center, spread, skew and shift are at least at a medium quality level. All *p*-based comparions were rated with medium quality. There has been no team using *q*-based comparisons. All in all

	High quality	Medium quality	Low quality
Center	The men earn 29,5% more than women on average. (L & R)	The mean of men is higher than the mean of women (H & I)	No example.
Spread	The middle 50% of men spreads more than the middle 50% of the women. (H & I)	The Interquartile Ranges of the distributions are almost identical. (C & M)	No example.
Shift	The median of the distribution of men is almost equal to the first quartile of the distribution of women. (H & I)	The middle 50% of men are shifted right compared to the middle 50% of women. (H & I)	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. (L & R)	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. (L & R)	No example.
p-based	No example.	10% of the men earn more than 5000€, only 2% of the women earn more than 5000%. (S & L)	No example.
q-based	No example.	No example.	No example.

Figure 3: Key examples of codings of the group comparison elements

Overall	High quality	Medium quality	Low quality	Overall
Center	1 (33%)	2 (67%)	0 (0%)	3 (100%)
Spread	2 (40%)	3 (60%)	0 (0%)	5 (100%)
Skewness	2 (50%)	2 (50%)	0 (0%)	4 (100%)
Shift	3 (60%)	2 (40%)	0 (0%)	5 (100%)
p-based	0 (0%)	6 (100%)	0 (0%)	6 (100%)
q-based	0 (0%)	0 (0%)	0 (0%)	0 (100%)
Overall	8 (35%)	15 (65%)	0 (0%)	23 (100%)

Figure 4: Overwiew of all codings related to group comparison elements

Hilde & Iris	High quality	Medium quality	Low quality	Overall
Center	0	2	0	2
Spread	1	1	0	2
Skewness	0	0	0	0
Shift	3	0	0	3
p-based	0	1	0	1
q-based	0	0	0	0
Overall	4	4	0	8

Conrad & Maria	High quality	Medium quality	Low quality	Overall
Center	0	0	0	0
Spread	0	2	0	2
Skewness	0	0	0	0
Shift	0	1	0	1
p-based	0	0	0	0
q-based	0	0	0	0
Overall	0	3	0	3

Ricarda & Laura	High quality	Medium quality	Low quality	Overall
Center	1	0	0	1
Spread	1	0	0	1
Skewness	1	1	0	2
Shift	0	1	0	1
p-based	0	2	0	2
q-based	0	0	0	0
Overall	3	4	0	7

Sandra & Luzie	High quality	Medium quality	Low quality	Overall
Center	0	0	0	0
Spread	0	0	0	0
Skewness	1	1	0	2
Shift	0	0	0	0
p-based	0	3	0	3
q-based	0	0	0	0
Overall	1	4	0	5

Figure 5: Codings (group comparison elements) distinguished by pairs

we can say that all of the statements and conclusions which were done by the pairs are at a high (35% of the codings) or medium (65% of the codings) quality level. Let us now have a look on the codings distinguished by pairs.

Hilde & Iris use amongst others center and spread elements - both of their statements using the group comparison element center were on a medium quality, one of their two elements regarding spread are on a high, the other on a medium quality. The conclusions regarding the comparison of shift of the both distributions are all on "high" quality. They also offer a medium quality *p*-based comparison, but they do not use a comparison of skewness or a *q*-based comparison. Conrad & Maria do not show any high quality statements in the whole solving process of the task. They offer at least three statements at medium quality. Conrad & Maria only use spread and shift elements but no center elements. They also do not use any skewness element neither do they use a *p*-based or a *q*-based comparison. Having a look on the codings of Laura & Ricarda we can say that they use every comparison element except *q*-based comparisons. They use center and spread (both in high quality) to compare the distributions and also shift and *p*-based comparisons (all in medium quality). Sandra & Luzie do not use center, spread or shift elements at all and work out differences from both distributions using skewness elements (one in high quality, one in medium quality) and *p*-based comparisons (all (3) in medium quality).

CONCLUSIONS

The statements of the pairs offer a broad variety of the use of comparison elements and none of the teams show low quality group comparison elements when working on the task. Most of the statements relating to the codings of "medium" quality could have been improved with the addition of an interpretation of the differences. Nevertheless we have to report some shortcomings which occurred: The amount of codings (overall: 23) is low. That means that all four teams made 23 comparison statements in total. On a first view this aspect is not necessarily negative, but in the course we have set up the norm that a group comparison should include as many investigations as possible. In this task there could have been found several differences along all aspects (center, spread, skewness, shift, *p*-based and *q*-based), so we finally expected some more codings relating to the comparison of both distributions. Whereas Hilde & Iris and Laura & Ricarda made eight respectively seven group comparison statements, Conrad & Maria only made three of them. Q-based comparisons were not used at all by the teams, although they played a big role when e.g. comparing boxplots in our course. Comparisons of skewness of both distributions were only done by two teams and apart from Hilde and Iris, the shift between both distributions was not worked out adequately. *P*-based comparisons were all done without interpreting the differences and are therefore all rated on a medium quality. As a further step in research we will take into account our findings from all three dimensions and search for relations between them. With these findings and the re-design of the course

in mind we might conclude that our norm "to work out as many differences between two distributions as possible" should be made more explicit. A data analysis scheme, which structures the data analysis process and gives hints of possible comparison elements, might support learners when comparing groups. Furthermore we might conclude that there should be a closer focus on interpreting differences between the distributions regarding center, spread, shift, etc. This may be done with contrasting adequate and non-adequate examples in regard to comparisons via center, spread, etc. Additionally we might reemphasize comparing groups with *q*-based comparisons in an upcoming course.

REFERENCES

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.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 169–190). Voorburg: ISI.

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern - Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens empirischer Verteilungen [Statistical competence of students – Concepts and results of empirical studies on the example of comparing groups]. In M. Borovcnik et al., *Anregungen zum Stochastikunterricht* (pp. 97–114). Hildesheim: Franzbecker.

Biehler, R. (2007). Denken in Verteilungen - Vergleichen von Verteilungen [Thinking in distributions – Comparing distributions]. *Der Mathematikunterricht, 53(3)*, 3–11.

Biehler, R., & Frischemeier, D. (2015). „Verdienen Männer mehr als Frauen?" – Reale Daten im Stochastikunterricht mit der Software TinkerPlots erforschen [Do men have larger income than women? Exploring real data in school with TinkerPlots]. *Stochastik in der Schule,* 35*(1), 7–18*.

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

Frischemeier, D., & Biehler, R. (2012). Statistisch denken und forschen lernen mit der Software TinkerPlots [Learning statistical thinking and exploring with TinkerPlots]. In Kleine, M. & Ludwig, M. (Eds.), *Beiträge zum Mathematikunterricht 2012* (pp. 257–260). WTM: Münster.

Kohlbacher, F. (2006). The Use of Qualitative Content Analysis in Case Study Research. *Forum: Qualitative Social Research*, 7(1), Art. 21. (http://www.qualitative-research.net/index.php/fqs/article/view/75)

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W.G. Martin, & D. E. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: NCTM.

Konold, C., & Miller, C. (2011). *TinkerPlots TM Version 2* [computer software]. Emeryville, CA: Key Curriculum Press.

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung [Qualitative content analysis, Methods, practice, support of computer]*. 1. Auflage. Weinheim und Basel: Beltz.

Makar, K., & Confrey, J. (2002). *Comparing two Distributions: Investigating Secondary Teachers´ Statistical Thinking.* Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Maxara, C. (2009). *Stochastische Simulation von Zufallsexperimenten mit Fathom. Eine theoretische Werkzeuganalyse und explorative Fallstudie [Stochastic simulation of chance experiments – A theoretical tool analysis and an exploratory case study].* Hildesheim: Franzbecker.

Mayring, P. (2010). *Qualitative Inhaltsanalyse [Qualitative content analysis]*. Weinheim und Basel: Beltz.

Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004). *Comparison of data plots: building a pedagogical framework.* Paper presented at ICME-10, TSG11: Research and development in the teaching and learning of probability and statistics.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher´s reasoning, *Statistics Education Research Journal, 5(2)*, 27–45.

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.

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

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

ENDNOTES

1. Comparisons of two distributions of numerical variables are called *p*-based, if for x the relative frequencies $h(V \leq x)$ are $h(W \leq x)$compared. So in *p*-based comparisons a specific argument can be given (for example: 10 hours) und the proportion of cases which are equal or larger than 10 hours is compared in both groups. (see Biehler, 2001, p. 110)

2. 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 und W, $q_V(p)$ with $q_W(p)$, are compared. With $q(p)$ we mean the quantile regarding to p. For $p = 0.5$ this is a comparison of medians. (Since the comparisons of medians is also included in the category "center" we do not want to include this special case for $p = 0.5$ here). (see Biehler, 2001, p. 110)