USING TECHNOLOGY IN THE TEACHING AND LEARNING OF BOX PLOTS

Ulrich Kortenkamp, Katrin Rolka

University of Education Schwäbisch Gmünd, University of Cologne

Box plots (or box-and-whisker-plots) can be used as a powerful tool for visualising sets of data values. Nevertheless, the information conveyed in the representation of a box plot is restricted to certain aspects. In this paper, we discuss both the potential and limitations of box plots. We also present a design for an empirical study in which the use of a variety of tasks explicitly addresses this duality. The activities used in the study are based on an interactive box plot applet that surpasses the currently available tools and offers new ways of experiencing box plots.

MOTIVATION

Recently, the mathematics curricula of many parts of the world were revised in order to include more statistics and data analysis. In the literature, one can find an extensive discussion about this idea under the notion of “statistical literacy” (Wallman, 1993; Watson & Callingham, 2003). This reflects the growing importance of the ability to understand and interpret data that has been collected or is being presented by others. The NCTM (2000) standards, for example, state, “To reason statistically--which is essential to be an informed citizen, employee, and consumer--students need to learn about data analysis and related aspects of probability.” The global availability of data through the Internet makes it easy to access and process huge data sets. For these, it is important that students have the skills and tools to summarise and compare the data, also by using the computer.

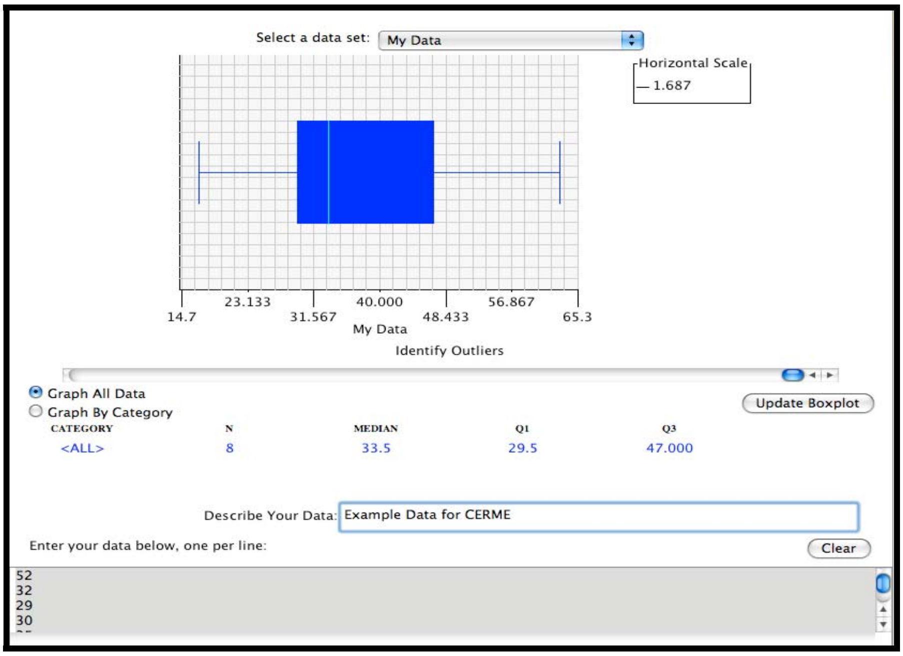
In this paper, we focus on *box plots* as a means to visualize statistical data. Box plots are used not only in textbooks, but are also available in graphing calculators. In order to use statistical information properly, the students have to develop a clear concept of what the information means, no matter whether it is given numerically or, in this case, visually.

The situation described also applies to Germany where some states have incorporated a larger amount of statistics and data analysis into the mathematics curriculum. Our personal experience with teacher students teaching in 8th grade (14-year-olds) has shown that both teachers and learners tend to ignore the mathematical concepts behind the statistical analysis and fall back to recipes that enable them to solve the standard exercises from the text books. In a similar way, Bakker, Biehler and Konold (2004) point out that some of the features inherent to box plots raise difficulties in young students’ understanding and use of them. As a remedy, we developed a series of activities that should enable students to develop a clear understanding of the statistical terms. The ultimate goal of the activities is that students can not only draw box plots for given data, but also interpret box plots that describe real world situations.

THEORETICAL BACKGROUND

Box plots are part of the field of Exploratory Data Analysis where data is explored with graphical techniques. Exploratory Data Analysis is concerned with uncovering patterns in all kinds of data. A box plot (or box-and-whisker-plot) is a relatively simple way of organizing and displaying numerical data using the following five values: the minimum value, lower quartile[1], median[2], upper quartile, and maximum value. Considering a set of data values like, for example, 52, 32, 29, 30, 35, 17, 42, 63, these five values are easy to calculate: minimum value = 17, lower quartile = 29.5, median = 33.5, upper quartile = 47, and maximum value = 63.

Using these five numbers, the related box plot can be constructed on a vertical (which we use in the following description) or horizontal scale (which is used in Fig. 1) by (a) drawing a box that reaches from the lower quartile to the upper quartile, (b) drawing a horizontal line through the box where the median is located, (c) drawing a vertical line from the lower quartile (the lower end of the box) to the minimum value, (d) drawing a vertical line from the upper quartile (the upper end of the box) to the maximum value, and finally (e) marking minimum and maximum with horizontal lines. Figure 1 shows the box plot corresponding to the data above, created with a box plot applet provided by CSERD.



Select a data set: My Data

Horizontal Scale
— 1.687

CATEGORY	N	MEDIAN	Q1	Q3
<ALL>	8	33.5	29.5	47.000

Describe Your Data: Example Data for CERME

Enter your data below, one per line:

```
52
32
29
30
```

Figure 1: Box plot created online for the sample data in this article

At the same time, box plots contain *more* and *less* information. On the one hand, the representation of a box plot communicates certain information at a glance: The median and the quartiles can easily be recognized which is not the case for the

[1] As there is no universal definition of a quartile, we dedicated a whole subsection of this article to this issue. Also, the original box plot uses the lower and upper hinge instead of the quartiles.

[2] The median can be defined as the number separating the lower half of a data set from the higher half in the sense that at least 50% of the values are smaller than or equal to the median.

original set of data values. Moreover, the line indicating the median illustrates the centre of the data, the width of the box demonstrates the spread of the central half of the data, and the length of the two lines next to the box show the spread of the lower and upper quarters of the data. This enables skilled people to interpret the box plot and draw conclusions about the underlying distribution. Various authors have declared that box plots are particularly useful for easily comparing two or more sets of data values (e.g. Kader & Perry, 1996; Mullenex, 1990). In order to illustrate this idea, compare two data sets where the minimum and maximum values as well as the arithmetic mean are equal and reveal no hint of how to draw conclusions about the values as shown in Figure 2.

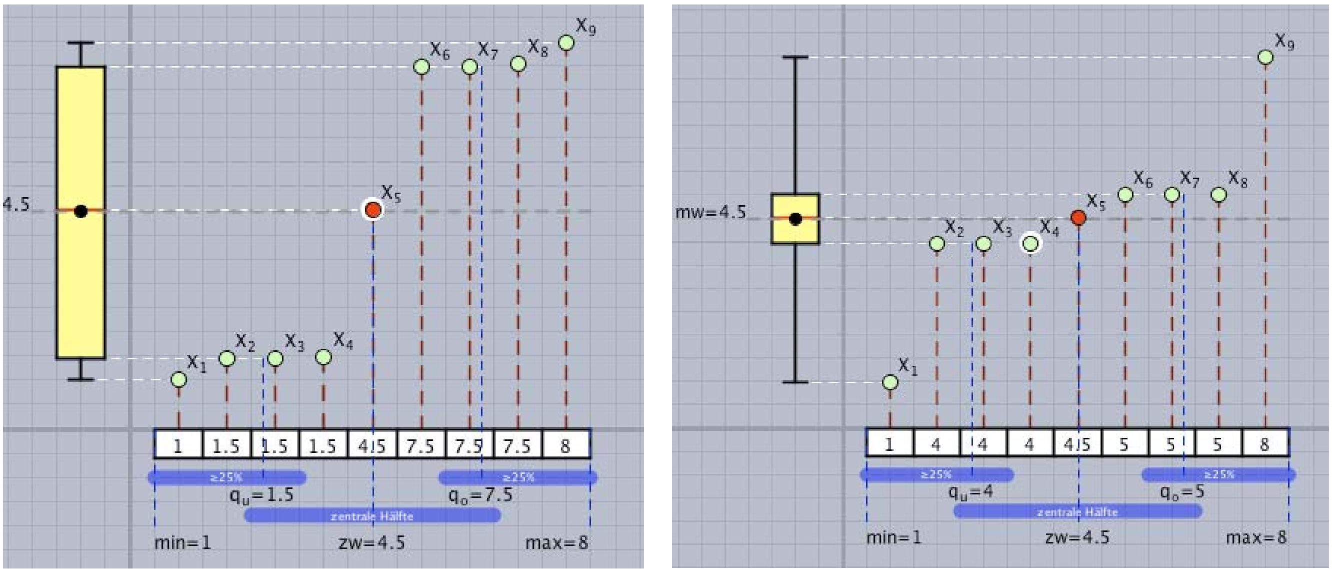


Figure 2: Two box plots with different interquartile ranges

It is obvious that in the second case, the box is much smaller than in the first one, indicating that the spread of the central half of the data is lesser. We use this technique extensively in the exercises that are part of the teaching unit.

On the other hand, the box plot representation is reduced to just five key values and the underlying individual values are not apparent any more – one considerable reason for students' difficulties with this kind of graphical representation (Bakker, Biehler & Konold, 2004). In addition, box plots – compared to many other graphical representations like, for example, histograms – do not display frequencies but rather densities (Bakker, Biehler & Konold, 2004). This means, the smaller a particular area is, the more values are contained in it.

A Useful Quartile Definition

There is no universal definition of a quartile; actually, there are at least five different definitions in use (Weisstein 2008). The situation is even worse for software packages. According to Hyndman and Fan (1996) even within a single software package several definitions might be used concurrently. A visualization sometimes uses a different definition than a numerical calculation. One reason for this is that the original concept of box plots as introduced by Tukey (1977) used the *hinges* of a data set instead of the quartiles, which are different in one of four cases. Unsurprisingly, the concept of a quartile is obscure to most students and even teachers.

School textbooks in Germany usually do not give an exact definition of quartiles, but combine a colloquial description with a recipe to calculate the quartiles. All definitions are not based on the desired result (i.e., "the first quartile is a value such that at least 25% of the values are less or equal, and at least 75% of the values are greater or equal"), but on a specified way to calculate them (i.e. "the first quartile is the value that is placed at position (n+1)/4 if this is an integer, else…" or similar). Unfortunately, these recipes are incompatible with the QUARTILE function as provided by Excel, which is the most common tool for data analysis in German schools, besides the availability of special purpose educational tools for statistical analysis like, e.g., Fathom (Key Curriculum Press, 2008). The documentation of the QUARTILE function in Excel[3] is similar to the text book definitions of quartiles: it lacks a formal definition or explanation of the desired properties, and focuses on examples instead. It is not possible to explain the results of Excel on that basis.[4]

Most of the critique above only applies to small data sets. With larger amounts of data the actual definition used is not as significant as with less than, say, 20 values. Still, these data sets are the ones that are accessible to hands-on manipulation in the classroom.

For our study, we chose a definition that is both easy to understand and easy to use. A lower quartile[5] of a set of values is a number q_u such that at least 25% of all values are less than or equal to q_u, and at least 75% of all values are larger than or equal to q_u. In many cases, this number is a value of the data set, but we do not restrict quartiles to be chosen from the values. The definition for the upper quartile q_o is analogous. Using 50% instead of 25% and 75% we can also use it to define the median. All definitions are valid even if some values occur several times.

Finding the Median and Quartiles

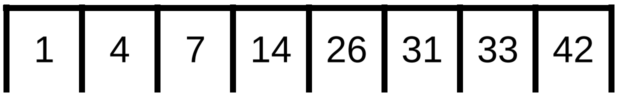
A very useful and action-oriented way to *find* the median and quartiles is the following one:[6] Order all values in increasing order, and write them down in a row of equal-sized boxes. The strip of ordered values may look like this (for 8 values):

[3] We used the German version of Excel 2004 on Mac OS X. There are explanations of the formulas used available, for example, in learn:line NRW at http://www.learn-line.nrw.de/angebote/eda/medio/tipps/excel-quartile.htm. Excel uses a weighted arithmetic mean for the quartiles.

[4] Büchter and Henn (2005) provide a definition of quartiles that is precise and matches the expectation that the lower and upper quartile are the smallest values that cut off at least 25% of the values.

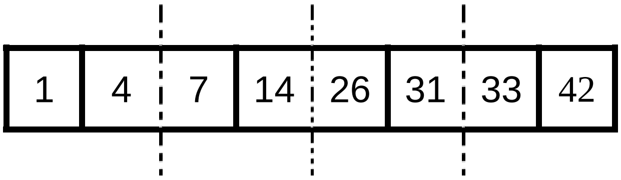
[5] We are using the standard German notation here, instead of Q_1 and Q_3 for lower and upper quartile.

[6] A student teacher, Simone Seibold, came up with this method during her traineeship in school.



Now, fold the strip in the middle by lining up the left and right border. The crease will be between 14 and 26, in this example, as is the median. We may use any number between 14 and 26 (not including them), for example the arithmetic mean, 20.

Finding the quartiles works by iterating the procedure described above. Folding the left and right half of the strip will create creases between 4 and 7, yielding a suitable lower quartile of 5.5, and between 31 and 33, which suggests choosing 32 as upper quartile.



The appeal of this method is that it also applies to situations where the creases pass through the boxes instead of separating two of them (i.e., for odd numbers of values, or if the number is not zero (modulo 4)). In that case, the (only) suitable value for the quartile (resp. median) is the value in that box. The conditions of our definition above are fulfilled automatically.

Of course, the method is not suitable for real computations with data sets of significant size, but only for the proper conceptualisation. It can easily be transferred to a formula for the quartile and medians, however.

Advantages of Using Technology

Computers are a major reason for the increasing importance of statistics, and vice versa. The whole field of *data mining* became feasible only through the computing power to analyse large sets of data easily. Actually, the first applications of mechanized computing were of statistical natures, for example in the 1890 United States census (Hollerith 1894). In general, multimedia learning bears advantages, in particular if several representations of a situation have to be connected mentally (see Schnotz & Lowe 2003; Cuoco & Curcio 2001). Relating to suitable design for multimedia learning, we refer to the book of Mayer (2003) that details some of the guiding principles. This being said, the existing online tools for creating box plots disregard these principles. Even the online tool that is officially endorsed by the NCTM (see Fig. 1) violates most of these rules. For example, the distant placement of the data entry and the box plot is in clear contradiction to the Spatial Contiguity Principle of Mayer. The quality of interaction is another measure for multimedia learning. The direct interaction with a simulation with *immediate* feedback supports the learner (Raskin 2000). Even if there is no such concept of a "level of interactivity," as it is not a one-dimensional scale, such interaction is considered a key ingredient of good software (Niegemann et al. 2003, Schulmeister 2007). Sedig and Sumner (2006) categorized the possible types of interaction in mathematics software. Again, the activities found on the web so far do not obey these rules.

Data Cycle

Biehler (1997) suggests a “Cycle of solving real problems with statistics”, similar to the typical modelling cycle (Fig. 3 left). However, we suggest that in our case another model is more suited. The typical way to work with data and data analysis in school can be described in a “data cycle” (Fig. 3 right), where data is created by, e.g. measurements in the real world, this data is processed to create a representation of it, the representation can be used for interpretation, and this should be connected to the original data. From top to bottom there is less information (in the information-theoretic sense), but more structure. On the left we work with the real world, that is concretely, on the right we work with a mathematized version of it, that is abstractly.

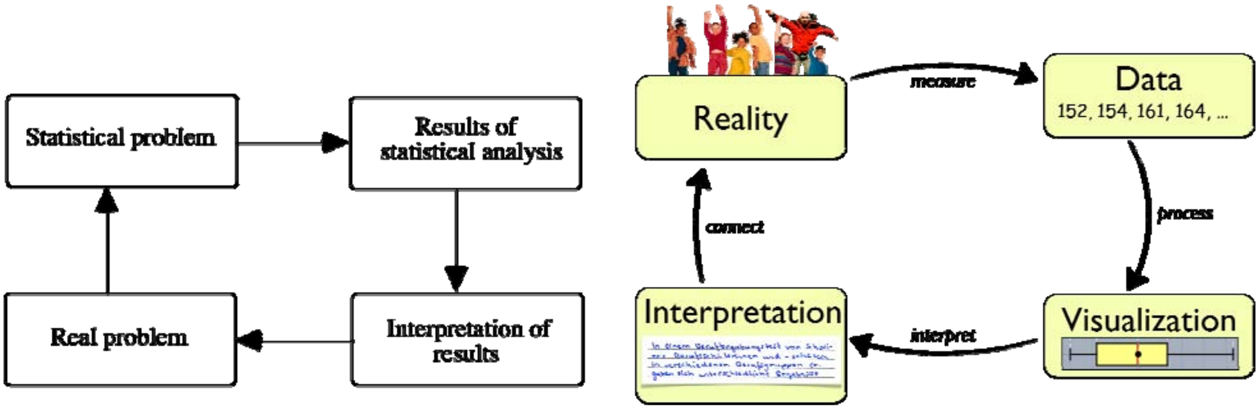


Figure 3: Problem solving cycle by Biehler (1997) on the left, and our proposed data cycle on the right

DESIGN OF THE ACTIVITIES

The design of the study is used in order to answer our main research question: *To what extent are students able to interpret box plots related to real world situations if they work with them interactively on abstract data sets?* Based on the theoretical analysis given above we therefore designed a set of exercises that enables the students to experience both the power and the restrictions of box plots. In all exercises students use the same interactive applet.[7] The applet is embedded into a plain web page and can be used without prior installations using a standard Internet browser. Using this applet, students can view and manipulate data with up to 22 values (the limit is not due to technical reasons, but given by the screen size). They can add or remove data, change data by dragging the associated data point with the

[7] See http://kortenkamps.net/material/stochastik/Quartile.html. The applet is based on Cinderella (Richter-Gebert & Kortenkamp 2006). In our box plot visualization we do not use outliers, as these are not used in the standard textbooks, either.

mouse vertically, and re-order values by dragging the points in direction of the *x*-axis. Points that have been added by the students are shown in red, others that were given are depicted in green.

According to Bakker, Biehler and Konold (2004), it is helpful for students if individual cases can be recognized within the box plot representation. This is granted in the applet that we use in our study. All data is visible at all times. While the students are manipulating the data, the current mean value is displayed both numerically and by a dashed horizontal line. The values that correspond to the data points are shown numerically in a white box below each point (Fig. 4 left).

If the values are ordered ascending the applet adds more statistical information to the visualization. To the left of the values the corresponding *box plot* showing the minimum, maximum, quartiles and median, is drawn. Those are connected through dashed lines with the corresponding "creases" and the values that are shown below the data. The blue bars mark the lower and upper quarters of the values as well as the central half (Fig. 4 right).

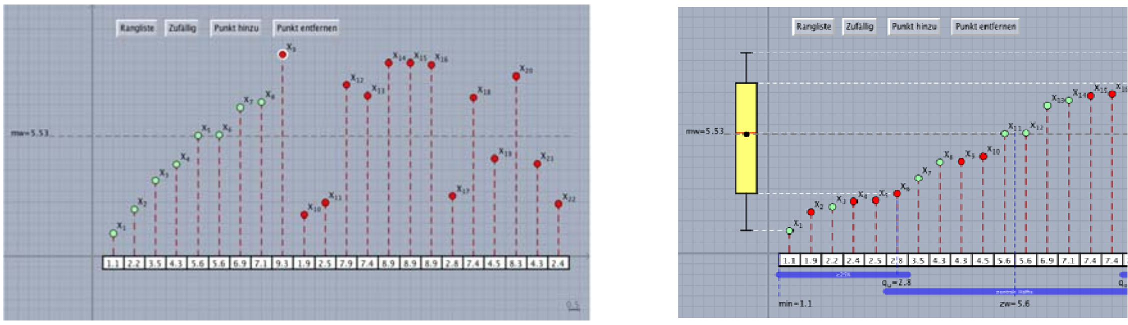


Figure 4: Applet with unordered values on the left, and ordered values on the right

Exploratory Exercises

Assuming that the students cannot master the interpretation step if they already fail at processing the data, we designed a set of exercises that aim at connecting the visualized data and the concepts behind them with the original data. Using the applet, students can easily process data dynamically, while modifying it, with an immediate update of the visualization. The exercises focus on modifying data sets in order to change or preserve the measures of variation: (a) Change *only* the arithmetic mean by changing values, (b) Change *only* the minimum or maximum by changing values, (c) Change *only* the length of the whiskers, (d) Change *only* the size of the box (the interquartile range), (e) Add values without changing the box plot, (f) Remove values without changing the box plot, (g) Try to move the arithmetic mean outside of the box, and (h) Try to move the median outside of the box.

Our primary goal is that students understand that box plots are a compact visualization of five (or six, depending on the plot) statistical measures, which in turn describe the distribution of values in a data set. Based on these measures it is possible to draw conclusion about the original set. Students should be able to find as many conclusions as possible, while not over-interpreting the measures. The activities force the students to create data sets that differ only in certain aspects, while showing an interactive visualization of the data and the measures.

For example, while experimenting with (d) students will see that for a distribution with smaller box (i.e. a smaller interquartile range) the values in the central half are more densely distributed than for a distribution with a larger box. Also, common misconceptions like a correspondence between the size of the box and the number of values in the data set are addressed. Adding or removing values does not necessarily change any of the measures of variation.

SUBJECTS AND METHODS

In line with the recommendations formulated by a group of stochastic educators in Germany (Arbeitskreis Stochastik, 2003), the participants in our study are aged at least 15 years. We conducted preliminary tests with the material in schools in two German states, Baden-Württemberg and North Rhine-Westphalia.

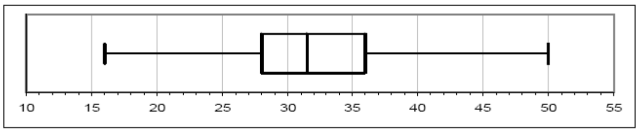
In Baden-Württemberg, we worked with 28 students in grade 9 at the "Realschule" level. They already received some training with box plots, but not with interpretation, in grade 8. In order to let them recall the basics they all received a hand-out about medians, quartiles, and box plots. First, they worked for 20 minutes in pairs with the applet and were asked to answer the exploratory exercises as given in the last paragraph in writing. Next, they were asked to analyze a series of box plots on another (paper) work sheet and interpret them in writing. Their answers were collected for further analysis.

In North Rhine-Westphalia, three students of grade 11 were involved in an interview-like situation where they had the possibility to explore the applet and work on the above presented exercises related to box plots. Beforehand, they had also received a hand-out providing an overview of medians, quartiles, and box plots. Subsequent to the exploration of the applet, they were given two interpretation tasks that they answered in written form.

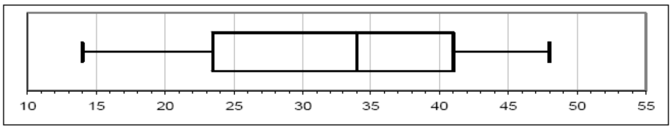
EXAMPLE OF AN INTERPRETATION TASK

In class 10a, there are 30 students, in class 10b 29. In both classes, the same test was written. The two box plots are based on the scores achieved by the students:

Class 10a



Class 10b



a) Describe as detailed as possible which information you can extract from the two box plots and compare them with each other.

b) Which class wrote the better test? Justify your answer.

c) Give examples for scores of the 30 students from class 10a that fit the given box plot and explain your procedure.

FIRST RESULTS

We only report on the results from one of the three students who took part in the interview-based exploration of the applet and then answered the interpretation task presented above. At first, the student describes the two box plots by simply listing the five key values respectively. This observation is in line with results reported on in the literature, and also our observations with the other student group in Baden-Württemberg. However, he does not remain at this merely descriptive level and formulates the following statement:

> In class 10a, a good portion of the students are located in the centre, whereas the points in class 10b are more distributed. However, here the higher points are more pronounced.

Being sympathetic to the student's answer, one could conclude that he has understood some basic principles of the box plot representation. However, in order to get more information about his competencies without construing too much, he was later asked by e-mail to clarify this answer. These are his additional explanations:

> The set of students is divided into four parts by the median and the two quartiles. In class 10a, the two middle areas are particularly small. This means that particularly many students are located there. In class 10b, the four areas are about the same size. This means that the students are distributed equally regarding to the score. The rightmost area in class 10b is considerably smaller than the one in class 10a. This means that the students in this area have achieved particularly high scores.

The additional explanations illustrate that the student has mastered some of the difficulties and challenges related to box plots that are described in the literature (Bakker, Biehler & Konold, 2004). He realizes that a box plot consists of four areas that approximately contain 25% of the data respectively. Moreover, he is able to formulate the relationship between the size of the particular areas and the density of the values contained in them.

CONCLUSION

We agree with the NCTM (2000) standards that students should also be able to create and use graphical representations of data in form of box plots as well as discuss and understand the correspondence between data sets and their graphical representations. The applet presented in this paper and employed in our study does not need any further software packages and therefore provides a basic but powerful tool for students in order to explore the potential and limitations of box plots. The applet is definitely easy to implement in the classroom. However, at the moment we cannot say too much about the effects on the interpretation competencies of the students who worked with the applet in a classroom situation. For the interview-like individual exploration our results show that the work with the applet can support the ability of students to analyze and interpret box plots. Currently, we are concerned with using the promising experiences based on the interview-like situations in order to make the applet also accessible to the work in the classroom.

References

Arbeitskreis Stochastik der GDM (2003). Empfehlungen zu Zielen und zur Gestaltung des Stochastikunterrichts. *Stochastik in der Schule. 23(3), pp. 21-26.* Online at http://www.mathematik.uni-kassel.de/stochastik.schule/sisonline/struktur/jahrgang23-2003/heft3/Langfassungen/2003-3_ak-empfehl.pdf

Biehler, R. (1997). Students' difficulties in practising computer supported data analysis - Some hypothetical generalizations from results of two exploratory studies. In: J. Garfield & G. Bunill (Eds.), *Research on the role of technology in teaching and learning statistics*, pp. 169-190, Voorburg, the Netherlands: Intemational Statistical Institute. Available online at: http://www.dartmouth.edu/~chance/teaching aids/IASE/14.Biehler.pdf.

Bakker, A., Biehler, R. & Konold, C. (2004). Should young students learn about box plots? *Curricular Development in Statistics Education, Sweden*, 163-173.

Büchter, A. & Henn, H.-W. (2005). *Elementare Stochastik. Eine Einführung in die Mathematik der Daten und des Zufalls*. Berlin: Springer Verlag.

Cuoco, A. A. &Curcio, F. R. (2001). *The roles of representation in school mathematics: 2001 Yearbook*. National Council of Teachers of Mathematics, Reston, Va., 2001.

Hyndman, R. J. & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistican, 50*(4), 361-365.

Hollerith, H. (1894). The Electrical Tabulating Machine. *Journal of the Royal Statistical Society*, Vol. 57, No. 4 (Dec., 1894), pp. 678-689. doi:10.2307/2979610.

Kader, G. & Perry, M. (1996). To boxplot or not to boxplot? *Teaching Statistics 18*(2), 39-41.

Key Curriculum Press (2008). *Fathom*. Version 2.1. Computer Software for Windows and Mac OS X. http://www.keypress.com/fathom

Mayer, R. E. (2001). *Multimedia Learning*. Cambridge University Press.

Mullenex, J. L. (1990). Box plots: Basic and advanced. *Mathematics Teacher 83*(2), 108-112.

National Council of Teachers of Mathematics (2000). *Curriculum and evaluation standards for school mathematics*. Reston, VA. Online at http://standards.nctm.org.

National Council of Teachers of Mathematics (2008). *Illuminations*. Web site at http://illuminations.nctm.org

Niegemann, H., Hessel, S., Hochscheid-Mauel, D., Aslanski, K., Deimann,M. & Kreuzberger,G. (2003). *Kompendium E-Learning*. Springer, Berlin.

Raskin, J. (2000). *The Humane Interface. New Directions for Designing Interactive Systems*. Addison-Wesley Longman, Amsterdam

Richter-Gebert, J. & Kortenkamp, U. (2006). *The Interactive Geometry Software Cinderella. Version 2*. http://cinderella.de

Schnotz, W. & Lowe, R. (eds.) (2003). External and Internal Representations in Multimedia Learning. Special issue of *Learning and Instruction. 13(2),* pp. 117-254.

Schulmeister, R. (2007). *Grundlagen hypermedialer Lernsysteme : Theorie - Didaktik – Design*. Oldenbourg, München, Wien.

Sedig, K. & Sumner, M. (2006). Characterizing interaction with visual mathematical representations. *International Journal of Computers for Mathematical Learning, 11(1)1–55.*

Tukey, J. W. (1977). "Box-and-Whisker Plots." §2C in: *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 39-43.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), 1-8.

Watson, J. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46, http://fehps.une.edu.au/serj.

Weisstein, Eric W. (2008). Quartile. From: *MathWorld – A Wolfram Web Resource*. http://mathworld.wolfram.com/Quartile.html