Formalizing students' informal statistical reasoning on real data: Using Gapminder to follow the cycle of inquiry and visual analyses

Martin Andre, Zsolt Lavicza, Theodosia Prodromou

Formalizing students' informal statistical reasoning on real data: Using Gapminder to follow the cycle of inquiry and visual analyses

Martin Andre[1], Zsolt Lavicza[2] and Theodosia Prodromou[3]

[1]Pedagogical University of Tyrol & Johannes Kepler University Linz, Austria; m.andre@tsn.at
[2] Johannes Kepler University Linz, Austria; lavicza@gmail.com
[3]University of New England, Australia; theodosia.prodromou@une.edu.au

Graphical representations of Open Data facilitate students' personal engagement in discovering and exploring data and enable students to intuitively access certain statistical concepts. The aim of this study is to identify important patterns and fundamental limitations in the learning processes when students acquire deeper understanding of "big ideas of statistics" (Garfield & Ben-Zvi, 2008). The objective of the study is to nurture students' intuitive knowledge (Fischbein, 1987) by using the software Gapminder for data visualization in classrooms. In this exploratory study 19 students aged 14-15 were guided through two lessons of 1 hour and 40 minutes, each, using the Gapminder tool and materials to explore and visualize sets of Open Data while working on several worksheets and test items developed by the research team. Findings of this initial study are being used to further conceptualize and design successive investigations.

Keywords: Statistics education, statistical literacy, visual data analysis, statistical reasoning, intuitions.

Theoretical background

This study is primarily based upon the cycle of inquiry and visual analysis (Prodromou, 2014), complemented by works on students' statistical literacy (Prodromou & Dunne, 2017), their development of the big ideas of statistics (Garfield & Ben-Zvi, 2008) and various research studies on the use of technology in the era of Open Data with a special focus on the visualization of context-based data. Acknowledging that graphical representations play a fundamental role in intuitively understanding of abstract mathematical or scientific concepts (Fischbein, 1987), theories on intuition (Fischbein, 1987), heuristics and biases (Tversky & Kahnemann, 1974) are considered.

Tversky and Kahnemann (1974), Fischbein (1987) stress the influence of connate heuristics, biases and misconceptions on making decisions or judgements under uncertainty (Tversky & Kahnemann, 1974). Studies have examined specific heuristics, biases and misconceptions within the learning processes of statistics and probability theory (Garfield & Ben-Zvi, 2008). Thus, it is a central challenge in statistics education, to address students' intuitions in order to identify their biases and misconceptions and support students to overcome their misconceptions. In fact, misconceptions are strongly bound to context because students think differently about stochastic concepts in various contexts.

Makar et al. (2011), as well as other researchers in Statistics Education, emphasize the importance of context and the use of real data in statistical reasoning. Ben-Zvi and Aridor-Berger (2016) demonstrate students' transition between context and data and also report on students' growing understanding of the ways to combine these contexts and data. Although there is little research on

the question what makes data "real", we can presume that students perceive data to be real when they get personally affected by the data. Therefore, Open Data on topics of general interest may be considered as real data easily accessible for teachers and students.

Advances of new technology with numerous opportunities for data visualization prompted Prodromou and Dunne (2017) to argue for the profound importance of using Open Data in statistics education. Analyzing Open Data with the software TinkerPlots in classroom, Watson (2017) describes the chances of learning statistics in a transdisciplinary way, integrating statistics education in non-mathematical topics, such as to Health and Physical Education, Social Science or History.

Garfield and Ben-Zvi (2008) outline various research studies on students' development of different big ideas of statistics, which they identified as crucial for statistical reasoning: a) *data*, including the nature of data and the diverse types and sources of data b) *statistical models*, such as regression and the Normal Distribution as statistical models, c) *distribution*, involving the ideas of shape, centre and spread, d) *centre*, covering the difference between median and mean, e) *variability*, including the measuring of the spread of a distribution, f) *comparing groups* by centre and spread, g) *sampling*, containing, e.g., the effect of sample size, h) *statistical inference* as to testing hypotheses or confidence intervals, and i) *covariation* with scatterplots, correlation and linear regression. The theory-based recommendations given with the presented activities for the development of the big ideas of statistics comprise learning these ideas explicitly, addressed in a constructive way and facilitated by examining real data sets. Furthermore, Garfield and Ben-Zvi (2008) also suggest learning the big ideas of statistics beginning with students' intuitions and preconceptions, taking their informal notions of the single ideas and turning them into formal notions. Teaching must start with introducing students' intuitions and preconceptions, identifying intuitions and preconceptions and rectifying them using different techniques including the use of digital technologies and graphical representations of data to display the summaries and visualizations of data.

Visualizing data is one of the most intuitive ways of understanding underlying concepts from data and "graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space." (Tufte, 1983, p. 51). Moreover, presenting information with diverse graphical representations is a highly important skill for many scientific disciplines aiming to make data meaningful (Tufte, 1983). Therefore, new types and the diversity of data in our age also require new ways of structuring, organizing, and visualizing data. Traditional, static visualization techniques such as bar charts, boxplots or distributions for single variables often fail to meet requirements of current data sources and analyses. Thus, processing and visualizing data with new techniques using specific, sophisticated technology plays an important role in students' developing understanding of statistics.

Using TinkerPlots for exploratory data analysis in the classroom, Ben-Zvi and Ben-Arush (2014) focus on the instrumentation process of transforming an artifact (i.e., a software with no meaning for the students) into an instrument (i.e., a tool, that is meaningful and useful to the learners). Based upon their observations, the authors suggest three types of instrumentation processes: a) *unsystematic*: students playing and experimenting with the software in a non-intentional way, b)

systematic: students being focussed on the tool rather than on the task, but applying the software purposefully, and c) *expanding*: students using the software fluently and focussing on the task.

Prodromou (2014) presents a project concluding that the students were able to flexibly use the software Gapminder to build visual structures that highlight information relevant to their analysis task. In addition, various studies concerned with the implementation of Open Data in school education (e.g., in Prodromou, 2017; Engel et al., 2016) and the technologies used for this purpose (e.g., Forbes et al., 2014) show promising possibilities to beneficially integrate Open Data in statistic courses by operating with visual methods.

To meet the needs of exploring and analysing data particularly by visual methods, Prodromou (2014) presents the cycle of inquiry and visual analysis, especially elaborated for that purpose. The steps of this cycle are: a) identifying the task, b) foraging for data, c) searching for visual structure and implementing visualizations, d) developing insight through interaction with the resulting data visualizations, and e) acting with regards to work further on any step to develop deeper insight or ending the cycle. All these stages are connected to each other and the process of visual analysis includes continuous interactions between the stages. Furthermore, basic statistical literacy is a prior condition for exploring data by visual methods, being deepened with these investigative processes. Therefore, Prodromou and Dunne (2017) have introduced a framework for constructing statistical literacy in schools, incorporating new possibilities arising in the age of Open Data and strongly bound to the diverse graphical representations of data.

In the teaching of basic statistics courses, students' intuitions and preconceptions have to be addressed and we also need to consider their biases and misconceptions. Working with graphical representations assists students to access their *intuitive knowledge* (Fischbein, 1987). According to researchers (e.g., Makar & Ben-Zvi, 2011), it is crucial to work with context-based data that are meaningful and important for students. Thus, meaningful data-processing needs visualizing real data, includes entire cycles of inquiry, and particularly addresses the *big ideas of statistics* (Garfield & Ben-Zvi, 2008). Furthermore, an updated concept of statistical literacy required for our age of data society (Prodromou & Dunne, 2017) should address students' perceptions and assessment of large sets of data. Finally, using technology autonomously and fluently, allowing focus on the task rather than on the software, is a fundamental factor of students' constructing their statistical knowledge and skills (Ben-Zvi & Ben-Arush, 2014). Therefore, the aim of this study is to develop new ways of teaching basic statistics, integrating students' intuitions addressed by the visualizations of Open Data with the software Gapminder. In our exploratory pilot study, we aimed at identifying important patterns and fundamental limitations regarding students' intuitions on statistical concepts, their statistical literacy, and their instrumentation processes to further conceptualize this investigation.

Methodology and implementation

The present pilot study is embedded in a larger study investigating students evolve their intuitions and preconceptions from informal perceptions of statistical concepts to a higher and more formal level by using different visualization tools. The research question concerning the present study is: How do activities related to analyzing real data by visual methods contribute to formalize students'

statistical reasoning? The entire study follows the methodology of design-based research (Cobb et al., 2003), utilizing an iterative design of the implementation of a learning environment and approach based on analysis of variety of data sources at the subsequent stages. We analyze students' conceptualization of the *big ideas of statistics* inherent in the applied worksheets and tests while examining data with the software Gapminder. The worksheets were used to get a broad range of tasks and corresponding answers; the test items included both closed and open-ended questions. Results of the worksheets and tests were anonymized during transcription and translation by using pseudonyms. Transcripts were analyzed qualitatively using MAXQDA qualitative analytics software by categorizing the answers in response to various statistical ideas addressed. Although it was planned to visit classes and conduct interviews with students, researchers were not allowed to collect any additional data from students beyond the worksheets and tests. But, interviews with the teacher before and after the lessons offered further information on the circumstances and settings of lessons and insights into students' work.

In total, 19 students in the 8th grade participated in the study. They had not worked with the software Gapminder previously nor had they received any prior instruction on distributions and variability in their school. However, their pre-knowledge, which could have an influence on solving tasks, includes calculating the mean and graphical representations of functional relationships in non-statistical, linear contents. The participating teacher was introduced by the researcher to the software and the instructional items of the study. Learning activities took place in a computer lab with a PC for each student and had duration of one week with two lessons of 1 hour and 40 minutes. The topic of the lessons was to find out more about poverty in the world.

Using two worksheets, the first lesson aimed at introducing students to the software as well as to ideas of investigating and exploring data sets. In the second lesson, using a third worksheet, students were asked to explore autonomously visualisations of various datasets. Between the different learning sequences, some test items were introduced to find out more about the students' statistical preconceptions and the conceptions they gained. The test questions were partly open-ended and partly closed, a structure familiar to the students from Austrian national standardized tests. The content of the first worksheet is about eight sets of interactive slides on human development trends, provided by the Gapminder foundation as instructional material (https://www.gapminder.org/downloads/human-development-trends-2005/). These slides show

Figure 1: Visualizations of income distributions used in the first test

different visualisations (see figure 3) of the variables *income* and *child mortality rate (health).* Visualisations explained and described the distribution of income of all countries in the world and the relationship between *income* and *health* over time. This worksheet contains questions on statistical ideas that address central value, variety, distribution's shape, and correlation between variables based on the topic. As the tasks were open-ended and students did not get any instruction before, they were required to describe their ideas on the tasks intuitively. Connected to the first worksheet, the first test items were developed to directly address the students' intuitions of centre, distribution, and variability by giving them two diagrams of different income distributions (see figure 1). The students were asked to reason about the shape of the second distribution and to compare it to the first distribution regarding the centres of these distributions. Additionally, students were asked to explain why the median income in 2015 most probably belongs to a person from a middle-income country and to describe the income of Chinese people in 1987 (light green top of the left peak). The second worksheet was a step-by-step guideline to follow the cycle of visual analysis (Prodromou, 2014) with different variables using the Gapminder software. There was no data collected on this second worksheet. The third worksheet contained some general instructions and suggestions, which variables the students may use for their autonomous investigation of the topic, e.g., income and literacy rate. The students participating were asked to document the findings of their investigations.

A second test, held at the end of the second lesson, included questions on a bubble chart (see figure 2) showing the variables *income* and *sanitation*. The focus of the test was on the ideas of correlation, centre and distribution. Students were asked to describe the bubble chart depending on the income levels and got a set of five statements addressing the various statistical concepts (e.g., variety, centre, or spread) to choose the right ones. With exception of the investigational learning sequences, the students were asked to work independently. The teacher supported the students' learning process by assisting them while working with the software and the topic. When the students solved the tests independently, they did not receive any assistance.

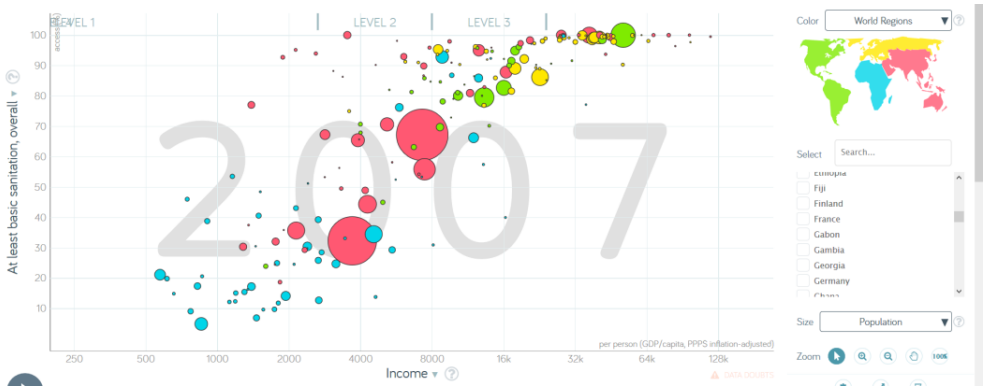

Figure 2: Bubble chart of the variables income and access to improved sanitation used in the second test

Initial results and discussion

In this section we present and discuss some of the main results of the pilot study to pinpoint directions for further investigation. Results presented focus on some of the *big ideas of statistics* (Garfield & Ben-Zvi, 2008) connected to the students' intuitive approaches and their uses of

technology. In the study, the main statistical model addressed was the model of regression. In several tasks students were asked to intuitively reason about the relationship between different variables (e.g., income and health) regarding various countries or a time series of one or more countries. All students recognized a correlation between income and health, building a model of regression. For example, one student, Richard, stated: "The higher the income in a country is, the better is the health rate of children." (Richard, worksheet 1) Students also interpreted the positive slope of the regression correctly: "That means that the countries constantly develop – in other words the GDP per person is growing – and the percentage of surviving children is heightening." (Eva, worksheet 1) Still, students did not compare countries' development by recognizing and building two or more models of regression in one graph (see figure 3) with one exception: "That means that poor people in India are less healthy than the rich ones whereas in Namibia the poor ones have nearly the same rate of health as the rich ones." (Vincent, worksheet 1). Anyway, most students independently used a model of regression during their autonomous investigations: "The poorer the countries are the worse are the numbers [of other variables] or the more children they have." (Paul, Peter and Anna, worksheet 3).

Besides using regressions, some models of probabilities were detected in the data as well. The students were asked to decide whether the person with the median income most probably comes from a low, middle or high-income country. For example, Vincent marked the height of the income of middle-income countries in the graphic of the distribution and connected the height to its probability. But some misconceptions were also observed with students building a model for this

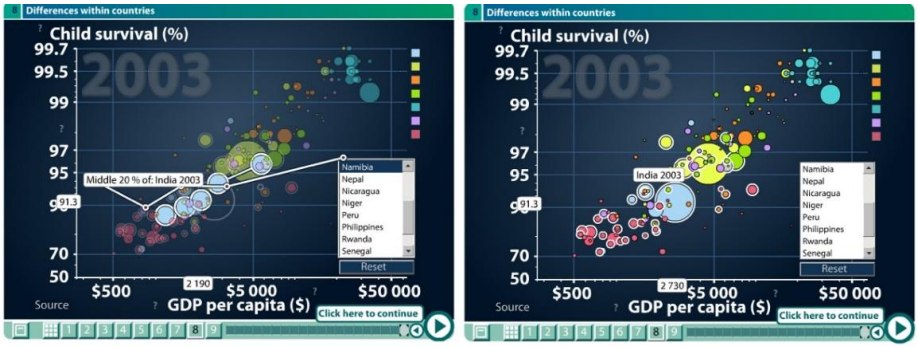

Figure 3: Examples of slides from the Gapminder learning materials

probability while maybe having a linear (regression) model in mind, as exemplified by these remarks: "Poor income country, because 6-7 $ is very little" (Julia, test 1) or "middle income because the number is in the middle" (Ruth, test 1).

These outcomes show that students had an intuitive access to the basic ideas of regression and correlation. The attempts to apply the model of regression to the probability theory may be rooted in the proximity of tasks. Furthermore, we suggest addressing *context, change and causality* as a part of the new framework for statistical literacy (Prodromou & Dunne, 2017) in that connection, as some students intuitively described the relationship between the variables with a focus on causality.

A second statistical idea to be mentioned here is the idea of distribution. With the learning materials on worksheet 1, students were briefly introduced to the idea of statistical distributions. All students recognized a change in the parameters of the income distribution over time and interpreted this change correctly, e.g., "Heavily concerned [with poverty] were South and East Asia (1970). Since

then, Africa got poorer" (Anna, worksheet 1). Moreover, some students gave a correct interpretation of the old-fashioned term "3rd world (in 1987)" regarding the graphic in figure 1. As they did not get any instruction on regular or irregular shapes of distributions, students were not wondering about the different shape of the income distributions of 1987 and 2015.

When asked for the central value of the income distributions in 2015 and 1987, all students gave the right number for the "regularly shaped" distribution of 2015, but around half of the students wrongly identified the higher, more distinctive peak in 1987 (see figure 1) as the centre (in terms of the median) of the distribution, although the poverty line above this peak is marked with "47%". In our data, we have detected two answers of students autonomously differing between median and mean, but most students did not differentiate between these two. In sum, we can conclude that the ideas of distribution and the centre are intuitively accessible, but certainly should be addressed more explicitly to avoid or eliminate the misconceptions. Especially a more detailed focus on the idea of distribution regarding their shape must be explored.

The collected data show that all students adequately applied the idea of variability in different contexts. They intuitively defined and applied the idea of "variability", e.g., in terms of "constant appearing dispensation, spread of something; […] In Namibia there are some people, who earn very little and some who earn very much. Therefore, the spread is as large." (Anna, worksheet 1), "the diverse levels of economy and rate of child mortality in different countries" (Sarah, worksheet 1), or "continuously occurring distribution/spread" (Clemens, worksheet 1). Moreover, all students used the idea of variability in their autonomous investigation to describe different variables. Therefore, we believe that the basic concept of variability is relatively easy to access for students who have adequate pre-conceptions of spread or range.

Relying on the reports of the teacher, the instrumentation processes (Ben-Zvi & Ben-Arush, 2014) of students can be described as unsystematic in the beginning and partly systematic while they are getting familiar with the tool. This does not seem surprising, especially since the students did not spend a long time working with the software. As students were given hints which variables to use in their autonomous investigation, the development of the instrumentation processes from unsystematic to systematic may have been supported. More detailed observation on students' use of the Gapminder tool will be necessary to identify helpful and hindering conditions for their instrumentation processes.

Summary

By addressing students' pre-conceptions and intuitions on various statistical concepts with visual methods, the initial results of the study suggest that some of the basic ideas, such as models of regression, centre or variability, tend to be intuitively accessible for students, whereas especially the idea of distribution seems to demand a more extensive development. In addition, several opportunities to focus on students' statistical literacy, related to explicating *inferences within data* or *context, change and causality* (Prodromou & Dunne, 2017), were detected. Offering a selection of adequate variables to explore, students' instrumentation process using the Gapminder software was apparently supported. Results of this pilot study will lay the basis to redesign this research project and to generalize findings.

References

Ben-Zvi, D., & Aridor-Berger, K. (2016). Children's wonder how to wander between data and context. In D. Ben-Zvi, & K. Makar (Eds.), *The teaching and learning of statistics* (pp. 25–36). Cham, Switzerland: Springer International.

Ben-Zvi, D., & Ben-Arush, T. (2014). EDA instrumented learning with TinkerPlots. 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. 193–208). Wiesbaden, Germany: Springer Spektrum.

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

Fischbein, E. (1987). *Intuition in science and mathematics: An educational approach.* Dortrecht, Netherlands: D. Reidel.

Forbes, S., Chapman, J., Harraway, J., Stirling, D., & Wild, C. (2014). Use of data visualisation in the teaching of statistics: A New Zealand perspective. *Statistics Education Research Journal, 13*(2), 187–201.

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

Engel, J., Gal, I., & Ridgway, J. (Eds.). (2016), *Promoting understanding of statistics about society*. Proceedings of the Round Table Conference of the International Association for Statistics Education. Berlin, Germany. Retrieved from https://iase-web.org/Conference_Proceedings.php?p=Promoting_Understanding_of_Statistics_about_Society_2016

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152–173.

Prodromou, T. (2014). Drawing Inference from Data Visualisations. *International Journal of Secondary Education, 2*(4), 66–72.

Prodromou, T. (Ed.). (2017). *Data Visualization and Statistical Literacy for Open and Big Data*. Hershey, PA: IGI Global.

Prodromou, T., & Dunne, T. (2017). Data Visualisation and Statistics Education in the Future. In T. Prodromou (Ed.), *Data visualization and statistical literacy for Open and Big Data* (pp. 1–28). Hershey, PA: IGI Global.

Tufte, E. R. (1983). *The Visual Display of quantitative information*. Cheshire, CT: Graphics Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science, 185*, 1124–1131.

Watson (2017). Open Data in Australian schools: Taking statistical literacy and the practice of statistics across the curriculum. In T. Prodromou (Ed.), *Data visualization and statistical literacy for Open and Big Data* (pp. 29–54). Hershey, PA: IGI Global.