An example of rich, real and multivariate survey data for use in school

Susanne Podworny, Yannik Fleischer, Dietlinde Stroop, Rolf Biehler

An example of rich, real and multivariate survey data for use in school

Susanne Podworny[1], Yannik Fleischer[1], Dietlinde Stroop[2] and Rolf Biehler[1]

[1]Paderborn University, Germany; podworny@math.upb.de

[2]District government Detmold, Germany

Data exploration and getting new insights from data is more important today than ever before. Exploring data is an important part of data science and can be a fruitful topic in middle school. New insights can be gained from rich data, but this requires a good data basis. Therefore, we present multivariate and rich data of over 1200 young people who answered an online survey on more than 150 questions. Several examples of possible data explorations of this data are presented. A short glimpse is given to a corresponding teaching unit for grade 8-10. With a suitable data science tool, students can handle and analyze such rich data. Students' presentations at the end of the teaching unit show the insights students get from the data but also, show challenges when exploring multivariate data.

Keywords: Data science, multivariate data, statistics education, CODAP.

Introduction

Statistics and data exploration have a long tradition and is fortunately also getting more attention in the classroom. With the rise of the new field of data science in recent years, its importance is once again increased (Ridgway, 2016) and gets new perspectives from computer science and special domain knowledge. In a data science project (Rubin & Mokros, 2018), various skills can be acquired and used that are helpful for good data exploration. On the one hand, this requires having rich, real data for students to draw real insights and conclusions that are important and motivating to them (Garfield & Ben-Zvi, 2008). On the other hand, a tool is needed that provides easy access to data analysis (Biehler, Ben-Zvi, Bakker, & Makar, 2013). In this paper, we present an example of such data which were collected as part of the ProDaBi project (https://www.prodabi.de/en/) and use the CODAP tool for exemplary explorations.

Background

The Project Data Science and Big Data in Schools (ProDaBi) aims at investigating in which way and with what topics data science can be implemented in the school curriculum. The project was initiated by Deutsche Telekom Stiftung and is conducted by an interdisciplinary team with members from statistics and computer science education. In this context, we collaborated with a media education research association[1] that conducts representative surveys among young people. We received an elaborated questionnaire for telephone interviews about media use of young people (so called "JIM-study"; JIM=Youth, Information, Media) and adapted this as an online survey. Since 2019, many young people participated in this online survey, so that we got rich and real data on young people's media use, which we call *JIM-PB*, based on the official JIM study and our regional reference

[1] https://www.mpfs.de

(PB=Paderborn). Please note, in contrast to the official data from the JIM-study, our data does not claim to be representative.

The data is used in different teaching settings. At first, it is used as an introduction in a project course on data science in grade 12 (Frischemeier, Biehler, Podworny, & Budde, 2021). This project course consists of three modules: (1) basics of data analysis and statistical thinking, (2) algorithmic thinking and machine learning, and (3) application in a comprehensive project.

Second, adapted from these modules 1 and 2, we have developed a teaching unit for grades 8-10, which also introduces statistical reasoning and machine learning. Here, students work entirely with the CODAP tool, both for data exploration and for creating decision trees based on data. This entire unit makes use of the JIM-PB data. Enhancing statistical thinking as proposed by Wild and Pfannkuch (1999) and introducing predictive modelling (Ridgway, Ridgway, & Nicholson, 2018) with the use of decision trees are the two main purposes of the teaching unit.

Rich data that allow multivariate explorations consist of many variables of different types to motivate creative investigations. How the JIM-PB data fulfils this is shown in the next section.

The data "JIM-PB"

The JIM-PB data is based on an online survey that contains various thematic blocks. The first block contains questions on age, sex, grade, type of school. The following blocks contain question in the style "How often do you…"

- … do leisure activities (e.g. meeting friends, playing an instrument, etc.),
- … use classical and digital media (e.g. newspapers, radio, etc.),
- … use media devices (e.g. tablet, gaming console, PC, etc.),
- … use social media (e.g. Facebook, Twitter, WhatsApp, etc.),
- … watch different YouTube videos (e.g. on product tests, letsplays, sports, etc.),
- … play different electronic games (e.g. car racing, shooter, adventure, etc.),
- … use specific apps (e.g. for news, public transport, school, etc.).

Values for all the corresponding questions on the frequency range from "daily", over "several times a week", "once a week", "once in a fortnight", "once a month", "less often" to "never".

Some questions are posed concerning the weekly time in minutes e.g. on watching TV, playing electronic games or using a PC for school at home. Additionally, questions are asked about availability or owning devices for media use (e.g. PC, Laptop, Tablet, WiFi, etc.). Questions are like "Is the device available at your home?" with values "available at home" or "not available at home" and "Do you own such a device?" with values "I own it myself" and "I do not own it".

This leads to 161 different questions, resp. variables. 1287 young people fully answered the survey, so the data contains as many cases from September 2020 until June 2021. A list of variables lists the original questions and the variable's names in the data. Data handling is done in advance. Variables are defined and little data cleaning like correction of German umlauts is necessary.

This results in rich and multivariate data with numerical and many categorical variables. In particular, the large number of variables makes it possible to develop and investigate very creative questions.

The Jim-PB data in a teaching unit for middle school

The idea of using the Jim-PB data in school is mainly to introduce reasoning about data in the frame of a data project. The PPDAC-cycle (Wild & Pfannkuch, 1999) frames the teaching unit and

> [w]orking with SP [statistical projects] thus represents a strategy that can enrich curricula because each phase involved in developing a project entails the use of various statistical concepts and processes that go beyond the topics normally included in curricula (Gómez-Blancarte & Ortega, 2018, p. 5)

For use in school, we created two versions of the data. The "standard" version is didactically reduced and contains 50 variables. A selection was made for this based on content criteria. For example, all questions about "owning" a device and playing different games were abandoned. We recommend using this version in grade 8-10. The "large" version contains all 161 variables.

The selection of a tool is crucial for doing data science in middle school. We selected and applied our JIM-PB data to the online data exploration platform CODAP (https://codap.concord.org), which is a free and online software that allows an easy and quick start of data exploration for beginners. For a detailed description of CODAP see for example Haldar, Wong, Heller, and Konold (2018).

For the teaching unit in middle school, we have chosen personalized advertising as the context. This has a motivation and an everyday relevance for students. The students are asked to explore the data with regard to four topics: user groups of TikTok, online newspapers, Youtube Letsplay videos and use of game consoles. Eight lessons with 45 minutes each can be used to enhance students' statistical skills when exploring Jim-PB data with CODAP in the first part. Additional eight lessons then build on findings from the first part to create and understand the method of decision trees for predictive modelling within CODAP. For the first part, three lessons can be dedicated to introducing basic terms, the data and tool, and concepts like row, column and cell percentages. Posing statistical questions (Arnold, 2013) and preparing group work can happen in lesson 4. The following two lessons can be used for students' own data exploration and preparation of a presentation. That results in a following presentation lesson. Reflections on the data, the PPDAC cycle and the conclusions can happen in the eighth lesson of the first part.

Explorations of the Jim-PB data with CODAP

What's in the Jim-PB data? At the beginning of an analysis, it is worth taking a brief look at one-dimensional distributions to characterize the sample of respondents.

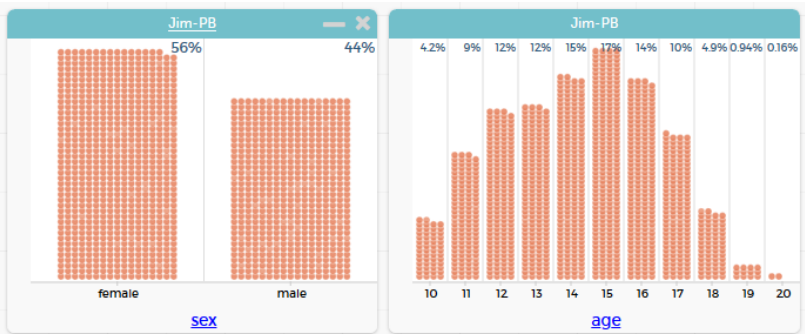

Figure 1: Distribution of sex and age in the Jim-PB data

A little more female (56%) than male (44%) students participated in the survey, aged 10-20 with peak at age 15 (Figure 1). For example, 78% have a game console available at home and a tablet is used by slightly less than half for gaming (Figure 2).

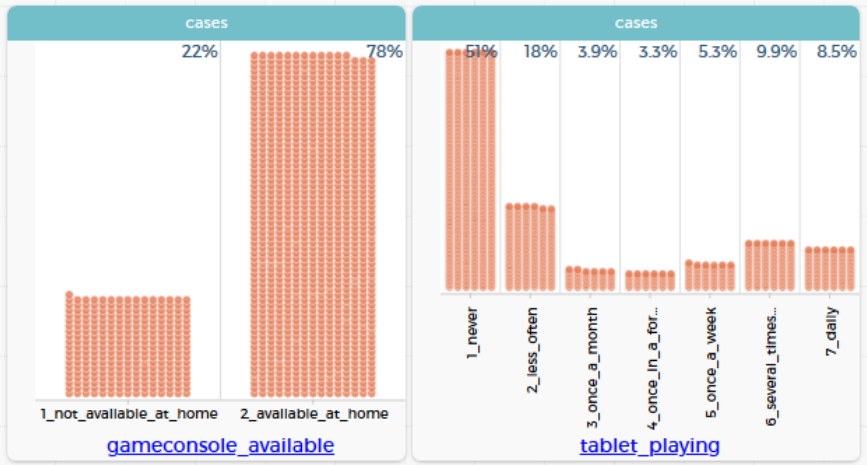

Figure 2: Distribution of game console available at home (left) and playing on a tablet (right)

Exemplarily, we look at a typical distribution of how often the participants use a tablet or Instagram (Figure 3). Such an "u" shape often occurs in these data, which means that the students can be split in two major subgroups ("rarely" and "frequently").

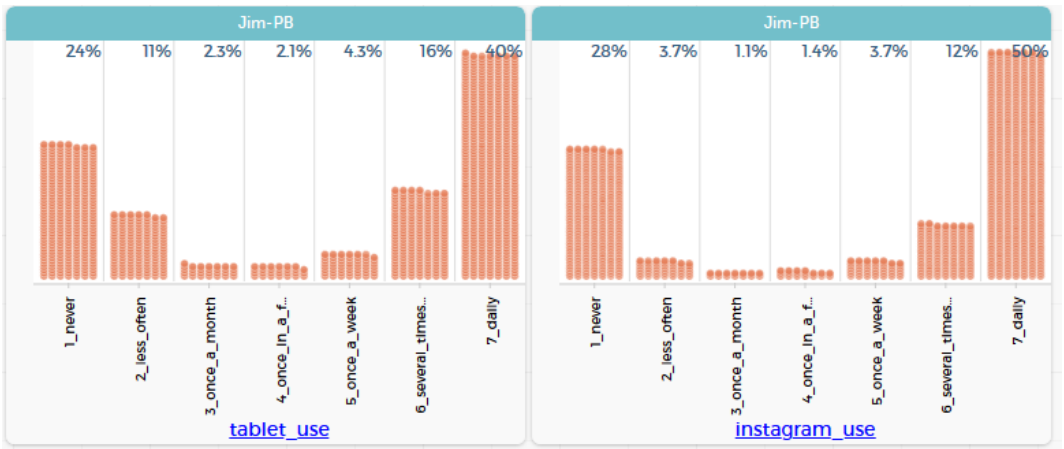

Figure 3: Distribution of tablet use (left) and Instagram use (right) in the Jim-PB data

Who belongs to those who daily use a tablet? Are those, who use daily a tablet those who use Instagram daily? For answering questions like this, two-dimensional diagrams are necessary like in Figure 4.

The challenge of interpreting complex diagrams like the one in Figure 4 is to use the correct percentages. This is often difficult for students (Watson & Callingham, 2014) and must therefore be well addressed in class. In Figure 4, row percentages are shown. A correct interpretation of the value 53 % at the top right corner is: Of those who use a tablet daily, 53 % use Instagram daily. But this is the same for those who never use a tablet (lowest right value in Figure 4)! Now we look at those who use a tablet daily.

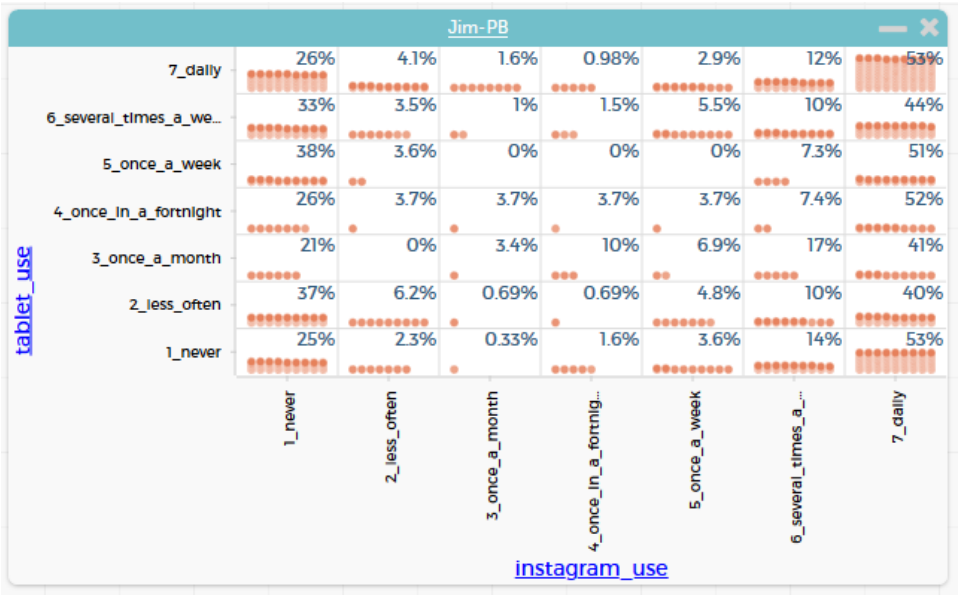

Figure 4: Do those who daily use a tablet those who daily use Instagram? (with row percentages)

7x7 tables like in Figure 4 are quite complex and hard to interpret not only for middle school students. By using data moves (Erickson, Wilkerson, Finzer, & Reichsman, 2019) like filtering with the "eye" function in CODAP, one can concentrate on a subgroup of the participants by selecting the corresponding cases and hiding all other cases (Figure 5).

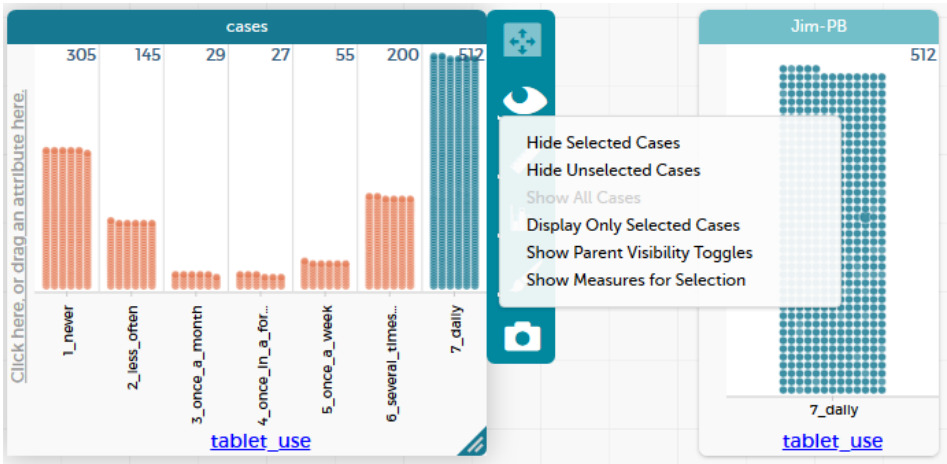

Figure 5: Filtering the data by selecting a subgroup (left) and the subgroup only (right)

Now the analysis can be focused on the subgroup of participants of those who use a tablet daily. For example, in this subgroup two third are female (Figure 6 left) in contrast to all participants (56%, Figure 1); distributed over all grades (Figure 6 middle) like in the whole group. 38% play often with the tablet (Figure 6 right, values for 'several times a week' plus 'daily') in contrast to the all participants, of whom less than 20% play tablet often (Figure 2 right). As a story, daily-tablet-user are more often female than male and use a tablet for playing more often than the whole group.

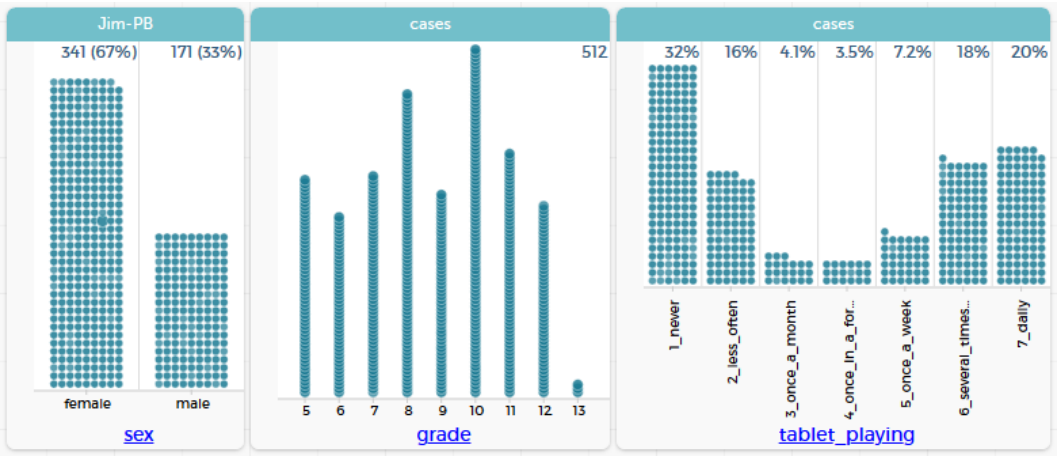

Figure 6: Subgroup displays of those who use a tablet daily

This is a small impression of an analysis of the Jim-PB data that can be done with CODAP.

Students' explorations of the JIM-PB data: some impressions

The first part of the teaching unit was implemented in four different classes in different schools in 2020, once in grade 9 and 10 and twice in grade 11. This includes a total of 77 students aged 13 (grade 9) to 17 (grade 11). All students had nearly no previous knowledge in statistics and data analysis. We collected their final presentations from lesson 7 and analyzed their diagrams and written interpretations. Student groups had between two and six members, we collected n=15 presentations because one group did not hand in their presentation.

Here we give a short overview of the 15 students' presentations, the diagrams used and written interpretations. Major differences between the three grades could not be identified, but the groups differed greatly in their work, independently of the grade. There were great and poor presentations in every grade. So, we analyzed on all presentations without taking the grade into account.

The number of slides and diagrams varied much between the presentations. The number of slides varied between two and 37, with an average of 13. Many slides of a presentation had only little or no text, except one presentation that was like a report. Every group stated their topic at the beginning.

The number of diagrams per presentation varied between two and 16 with an average of six. One group did not include any diagram, they only reported several percentages (that must have been figured out with diagrams in CODAP). The presentations often had no written explanations in the slides but were explained orally during the presentation lesson. Three groups used column percentages, all other used row percentages which are the default setting in CODAP. For these groups it is not clear whether they also considered using column percentages to answer their underlying question. Of course, a rearrangement of the variables would be equivalent to changing row and column percentages. Using and interpreting percentages was often challenging for students. The use of row percentages did not always go along with the interpretations, several times column or cell percentages would have been appropriate for the written statements. This is similar to results found by Watson and Callingham (2014). Two groups used the filtering function to focus on subgroups like in Figure 5.

A nice example is from a group from grade eleven that looked at how the frequency of reading online newspapers is changing with age. They computed the average age of readers in the different groups as shown in Figure 7.

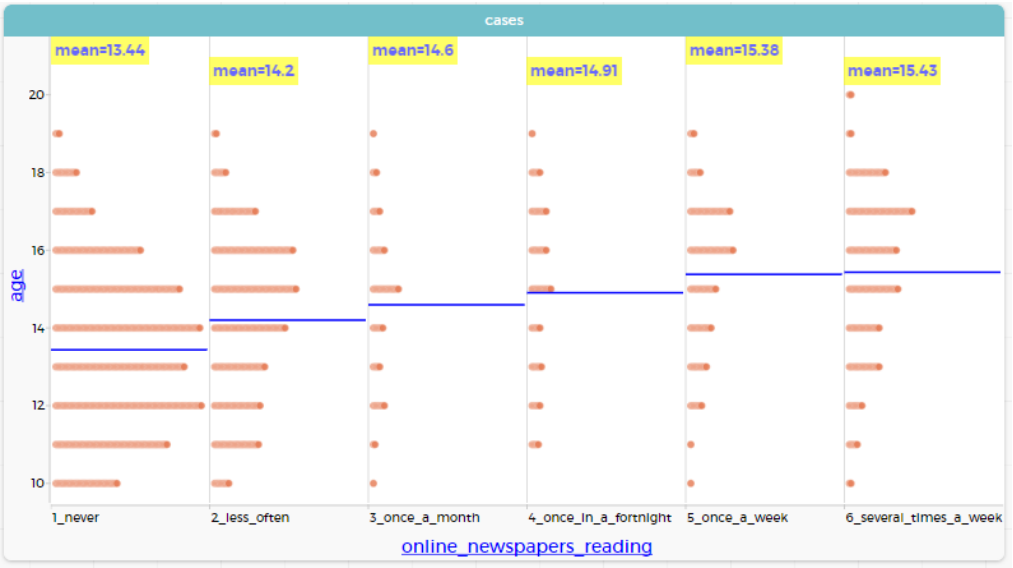

Figure 7: Students' exploration of age and online newspaper reading

As an interpretation they wrote "For each 'reading frequency' you can see the average age indicated by the blue line. Newspapers are read online from the age of about 15, while the average age of non-online newspaper readers is 13. Therefore, it can be said that the target group of online newspapers has the age of more than 15 years, while the largest and most active part of the readers is represented by people aged 16 years and older."

Students' investigations were rich and showed interesting relationships in the data. Diagrams created and used for presentation by the students offered great opportunity to tell many stories about the data. Nevertheless, only few detailed descriptions like the one for Figure 7 occurred and these were not yet completely satisfactory due to inaccuracies in the explanations. Cultivating interpretations turned out to be a challenge.

Conclusion

The JIM-PB data provides rich exploration opportunities and is suitable for students aged 13 and up. For students, the version with 50 variables is sufficient to make interesting discoveries and looking for many relationships in the data. CODAP has proven as an appropriate tool for an easy entrance to explorations.

Such rich data not only brings advantages, but also challenges. The type of the data means that variables with seven values have to be related to each other. This results in 7x7 matrices with a total of 49 entries. In class, it became apparent that the interpretation of 7x7 matrices was challenging for students and patterns were rarely well described and interpreted. As an implication, we added a lesson on data preparation in the teaching unit, where the seven values can easily be merged so that only the values "frequently" and "rarely" remain by using the "hierarchize" function in the CODAP's table. Then only four-field tables occur, which are much easier to interpret than the 7x7 matrices.

Additionally, modelling aspects of such a data preparation with relation to possible interpretations can be discussed in class. As another implication, data moves (Erickson et al., 2019) like filtering (easily done in CODAP with the “eye” function) are a useful concept that can enrich students’ data science projects in middle school.

References

Arnold, P. (2013). *Statistical investigative questions - an enquiry into posing and answering investiative questions from existing data.* (Doctor of Philosophy). The University of Auckland, New Zealand. http://hdl.handle.net/2292/21305

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for Enhancing Statistical Reasoning at the School Level. In M. A. Clements (Ed.), *Third International Handbook of Mathematics Education* (pp. 643-689). New York: Springer. https://doi.org/10.1007/978-1-4614-4684-2_21

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data Moves. *Technology Innovations in Statistics Education, 12*(1). Retrieved from https://escholarship.org/uc/item/0mg8m7g6

Frischemeier, D., Biehler, R., Podworny, S., & Budde, L. (2021). A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. *Teaching Statistics, 43*(S1). doi:10.1111/test.12283

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

Gómez-Blancarte, A., & Ortega, A. S. (2018). Research on statistical projects: Looking for the development of statistical literacy, reasoning and thinking. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute.

Haldar, L. C., Wong, N., Heller, J. I., & Konold, C. (2018). Students making sense of multi-level data. *Technology Innovations in Statistics Education, 11*(1), 2-32. Retrieved from https://escholarship.org/uc/item/7x28z96b

Ridgway, J. (2016). Implications from the data revolution for statistics education. *International Statistical Review, 84*(3), 528-549. https://doi.org/10.1111/insr.12110

Ridgway, J., Ridgway, R., & Nicholson, J. (2018). Data science for all: a stroll in the foothills. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., & Mokros, J. (2018). Data clubs for middle school youth: engaging young people in data science. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10), Kyoto, Japan*. Voorburg, The Netherlands: International Statistical Institute.

Watson, J., & Callingham, R. (2014). Two-way tables: Issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning, 16*(4), 254-284. http://dx.doi.org/10.1080/10986065.2014.953019

Wild, C., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223-265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x