Preservice teachers' exploration of multivariate data based on personal interests

Susanne Podworny[1], Lisa Birk[2], Sibel Kazak[3] and Aisling Leavy[4]
[1]Paderborn University, Germany;
[2]University of Münster, Germany;
[3]Middle East Technical University, Türkiye;
[4]Mary Immaculate College Limerick, Ireland
podworny@math.upb.de

The DataSETUP project addresses the critical need for data science competencies in teacher education by developing short, modular courses for preservice teachers. Based on the DataSETUP framework, these modules guide future teachers through key data science processes, including data exploration, problem formulation, modeling, and results communication. In a pilot study, preservice primary teachers used a real-world dataset to explore digital gaming habits of young people. Findings show that students effectively formulated statistical questions and created visualizations to answer them. However, they tended to select their own topics and struggled to connect their analyses to the module's thematic context, highlighting both the motivational potential and the challenges of working with large, multivariate datasets.

INTRODUCTION

As outlined in the GAISE II report, it is highly relevant to develop students' data literacy and foundational understanding of data science. It emphasizes the importance of engaging all students in the statistical problem-solving process and integrating real-world contexts, technology, and multivariable thinking (Bargagliotti et al., 2021). Teachers play a crucial role in fostering this data literacy in K-12 students, but many studies focus on students and exclude the investigation of preservice teacher preparation (Friedrich et al., 2024). Additionally, teachers and preservice teachers show substantial knowledge gaps in fundamental statistical concepts and reasoning with data–two central components of data literacy and therefore also data science (Schreiter et al., 2024).

The project "Promoting Data Science Education for Teacher Education at the University Level" (DataSETUP, https://datasetup.euc.ac.cy), an Erasmus+ Cooperation partnership, aims to enhance data science education in university-level teacher education programs. The DataSETUP project introduces preservice teachers to data science content in order to promote data-based reasoning. The goal is to foster an attitude that encourages constructive engagement with data science, thereby preparing teachers—and, in turn, their students—for the challenges of an increasingly data-driven world. Often, STEM teacher education programs lack sufficient data science content, which is crucial for preparing preservice teachers for a data-driven world. For example, in selected STEM teacher education programs of the DataSETUP partner universities, the Middle East Technical University (Türkiye), Universities of Münster and Paderborn (Germany), University of Limerick (Ireland), European University Cyprus, and National and Kapodistrian University of Athens (Greece), the connection to data, data science, or statistical literacy is very limited. This poses a challenge, given that data literacy is widely regarded as essential for equipping young people to navigate a data-driven world (Ridgway, 2022; Ridsdale et al., 2015).

The DataSETUP project addresses this by developing and implementing short data science modules for preservice teachers. These modules are developed within a simplified data science and data science education framework (Fig. 1) to foster different data science processes and practices that the framework outlines. Our research explores how preservice teachers with limited prior knowledge engage with these data science processes and practices within their work on a DataSETUP module.

A FRAMEWORK FOR DATA SCIENCE EDUCATION

The DataSETUP Framework (Fig. 1) is a conceptual model designed to support preservice teachers in engaging with data science both as learners and future educators during their university education. It draws on foundational work in statistics and data science education, integrating key ideas from data inquiry cycles like PPDAC (Wild & Pfannkuch, 1999), the GAISE framework (Franklin et

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*.

al., 2007), and recent curriculum innovations such as the IDS curriculum (Gould et al., 2016) and the International Data Science in Schools Project (IDSSP Curriculum Team, 2019).

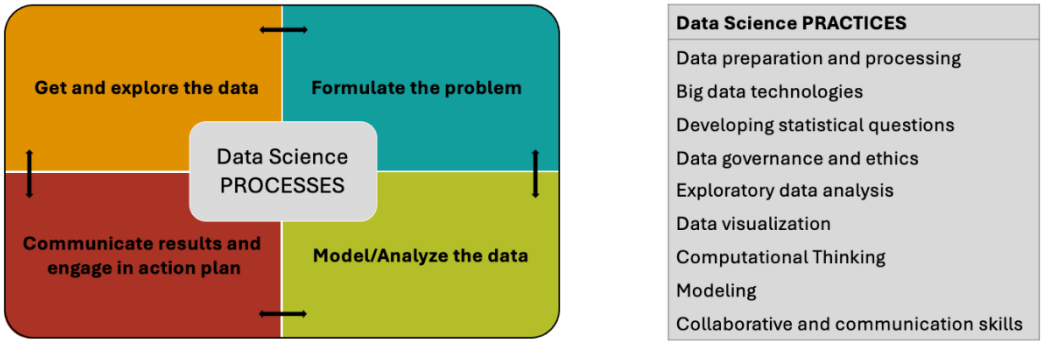

Figure 1. Conceptual model of the DataSETUP framework – dimension 1.

At the core of dimension 1, the framework organizes data science into four flexible, interconnected processes (Fig. 1, left): *Get and explore the data*, *Formulate the problem*, *Model/Analyze the data*, and *Communicate results and engage in an action plan*. These processes are not strictly linear, but reflect the authentic, iterative nature of data investigations. Aligned with these processes are a set of critical data science practices (Fig. 1, right), including data preparation and wrangling, exploratory data analysis, visualization, modeling, computational thinking, ethical reflection, and collaboration—elements commonly emphasized in frameworks by Lee et al. (2022), Ow-Yeong et al. (2023), and Keller et al. (2020).

In addition, the framework introduces a second dimension that highlights pedagogical considerations for teaching data science: selecting relevant, real-world contexts (e.g., social justice, environment), aligning it with key K–12 statistical ideas, choosing suitable tools, and designing meaningful assessments. It responds directly to the challenges identified in the literature, including teachers' lack of confidence, limited curriculum resources, and the complexity of integrating data science concepts like machine learning and ethics into classrooms (Fry & Makar, 2021; Horton & Hardin, 2021; Oliver & McNeil, 2021).

The DataSETUP framework is designed to simplify and adapt existing, often complex data science frameworks in order to provide a focused and practical introduction to data science within a limited time frame in a university-level teacher education course. It distills the content down to core data science processes and selected practices that are particularly relevant for future teachers, while taking into account the specific demands of school-based learning contexts. Through this pedagogical distillation, the framework aims to make data science accessible even for those without a strong background in statistics, computer science, or mathematics, and to promote a reflective, interdisciplinary engagement with data-driven questions. In essence, the DataSETUP framework aims to make data science accessible and actionable for preservice teachers, providing a structured yet flexible guide for developing data-literate, ethically aware, and critically engaged educators.

A STUDY ON A DATASETUP MODULE

Using a design-based research approach, the DataSETUP project develops modular, standalone short courses to familiarize preservice teachers with data science processes and practices. As an outcome of the project, a collection of tested and refined modules will be available to support data science learning across different subject areas and teacher education contexts.

Research Question

Concurrently, from a statistics education perspective, we aim to analyze which data science processes and practices (dimension 1 of the framework) preservice teachers engage with and how they do so. We are particularly interested in identifying aspects of the learning process that work well and where challenges arise. This leads to the guiding research question: *How do preservice teachers with limited prior knowledge engage with data science processes and practices outlined in the DataSETUP framework?* Given that many future teachers have little or no background in statistics or computer

science, there is a need for accessible and pedagogically grounded approaches to introduce them to the field. The modular format of DataSETUP is specifically intended to lower the threshold for engagement, enabling preservice teachers to explore authentic data science activities in a supportive learning environment.

Design of the Study

The study investigates how preservice primary teachers engage with data science processes through the implementation of an exploratory DataSETUP module and which practices they employ. This first practice-oriented module invites participants to explore the digital gaming habits of adolescents and young adults by analyzing a real, multivariate dataset. The data originate from the project Data Science and Big Data at School (www.prodabi.de/en) and include survey responses from over 1,200 young people regarding their leisure and media habits (Podworny et al., 2022).

Each DataSETUP module includes an instructor version and a student version for preservice teachers to work with. The student version is several pages long and starts with a motivational instruction on the data context and some background information. The next section, *Get & explore the data*, asks preservice teachers to begin working with real data, encouraging initial explorations. They are given background information on the origin of the data, including its collection process and limitations. Next, they are guided to review the list of variables to understand the data structure. In *Formulate the problem*, preservice teachers are encouraged to develop their own hypotheses and formulate statistical questions based on their initial data exploration. The section *Analyze data* provides detailed instructions for conducting data analysis (in this study with the free and easy-to-learn software CODAP, codap.concord.org), including creating visualizations, applying statistical methods, and interpreting results. The section *Communicate results & engage in an action plan* guides preservice teachers to summarize findings, interpret results, and suggest practical recommendations or actions based on the analysis. A final section on *Module reflection* encourages reflection from both student and teacher perspectives to evaluate learning outcomes and the integration into educational practice. Students document their process on a specifically designed worksheet.

The module "Gaming habits of young people" was piloted by the second author in a 90-minute session in a university seminar with 19 preservice primary school mathematics teachers (16 female, three male; mean age of 21,05 years; mostly third-year students) at the University of Münster in Germany. Before the pilot session, participants had received a brief introduction to statistics in primary education and were introduced to the CODAP tool in a separate 90-minute session. During the seminar session of the pilot, the students completed a pre-survey, worked through the module in eight small groups of two or three (for approximately 65 minutes), and participated in a post-survey focusing on their attitudes toward the module and their experiences engaging with data science processes and practices. After class, they additionally completed a module review to inform the iterative development of the DataSETUP modules within the design research approach of the project.

Data Sources and Data Analysis

The written work of the eight student groups serves as the basis for analysis. These documents include both written answers to the tasks, especially data interpretations, and screenshots of their explorations and analyses in CODAP. To analyze preservice teacher engagement with the data science processes and practices, in a first step, the documents are analyzed with respect to the four data science processes (Fig. 1), using a qualitative content analysis approach (Mayring, 2015). We developed a coding scheme for our qualitative content analysis (Table 1). For each data science process, we assessed either how many aspects (statistical questions, visualizations, etc.) the preservice teachers documented, the extent to which these aspects were supported by graphs, if the questions were connected to the module's topic, and if there was a description for each graph. In a second step, we considered the data science practices that were addressed in the different data science processes. The coding was undertaken by two authors independently and discussed until an agreement was reached.

Table 1. Coding manual.

Category	Code	Description & *Coding*
Get and explore the data	Number of written aspects documented	Number of written aspects that preservice teachers document in bullet points or sentences *(absolute frequency)*
	Number of graphs used	Number of documented graphs *(absolute frequency)*
Formulate the problem	Number of statistical questions/ hypotheses	Number of statistical questions/hypotheses formulated by students *(absolute frequency)*
	Related to gaming	Relation to the module topic of "gaming" *(yes/no)*
Model/Analyze the data	Number of graphs	Number of documented graphs for exploration *(absolute frequency)*
	Description for each graph	Description for each graph *(yes/no)*
Communicate results and engage in an action plan	Interpretation related to the statistical question	Interpretation related to at least one statistical question/hypothesis posed when formulating a question *(yes/partly/no)*
	Action plan addressed	Action plan addressed *(yes/no)*

RESULTS

All groups worked on all four data science processes outlined in the DataSETUP framework and responded to the different tasks posed in the worksheet section. In the following, we will illustrate the results of our qualitative content analysis, supported by examples from prototypical groups.

Category 1: Get and Explore the Data

All eight groups engaged with the dataset and provided at least some description. Most groups identified specific variables and commented on data characteristics (e.g., types of games, frequencies). Example: Two (of 7) aspects of group 3: "How many females and males participated? 563 (44%) male, 709 (56%) female", "What is the age distribution of the respondents? Most are between 14 and 16 years old", both aspects (distribution of gender and of age) were supported by CODAP graphs.

Table 2. Results for category 1: "Get and explore the data".

Group	1	2	3	4	5	6	7	8
Number of aspects documented	7	3	7	4	4	5	2	3
Number of graphs used	0	5	2	1	0	0	0	0

Graphs were used by three groups to support their first exploration of the data. Focusing on the employed data science practices, we interpreted that the groups used *data processing, big data technology* by using CODAP, *exploratory data analysis* to understand what the data are about, *data visualization* by some groups to support their findings, and *communication skills* to document some of their findings while getting and exploring the data.

Category 2: Formulate the Problem

Each group formulated at least one statistical question or hypothesis involving at least two variables. These included statistical questions and hypotheses on group comparisons (example: group 3) and relationships between different variables (example: group 1).

Examples: "Female users tend to use communicative and creative platforms such as WhatsApp or Pinterest. Male users, on the other hand, prefer platforms such as Twitch or YouTube.", "Men prefer sports and women prefer music.", "Overall, men spend more time online than in the real world." (group 3); "What is the connection between sports activities and the use of social media, using Instagram as an example? Possibly also with regards to music.", "How do attitudes towards social media vary between different age groups?", "Do students at a 'Gymnasium' use ebooks more often than students at other types of school?" (group 1).

Table 3. Results for category 2: "Formulate the problem".

Group	1	2	3	4	5	6	7	8
Number of questions/ hypotheses	3	2	3	2	2	2	2	1
Related to the topic of "gaming"	No	No	No	No	No	Yes	No	No

Regarding the data science practices, we interpreted that the groups used *developing statistical questions* while formulating the problem.

Category 3: Model/Analyze the Data

Each group created and described at least one graph; however, the depth of description varied. Most groups created graphs that were helpful for answering their statistical questions or supporting their hypotheses, which they developed when working on the section *Formulate the problem*. One exception was group 1, which formulated and answered a new question in this section. Still, their graph was helpful to work on their question.

Examples: "What is the connection between internet use and the feeling of missing out (FOMO)?" and the corresponding graph in Fig. 2 (group 1); "Most rate the fear of missing out as 'tend to not agree' and 'do not agree' → in the middle range are more than two thirds of respondents. Most respondents use the internet daily, and of this group, most chose 'tend to not agree' (340) and 'do not agree' (323) as their answer → no trend apparent" as graph descriptions (group 3).

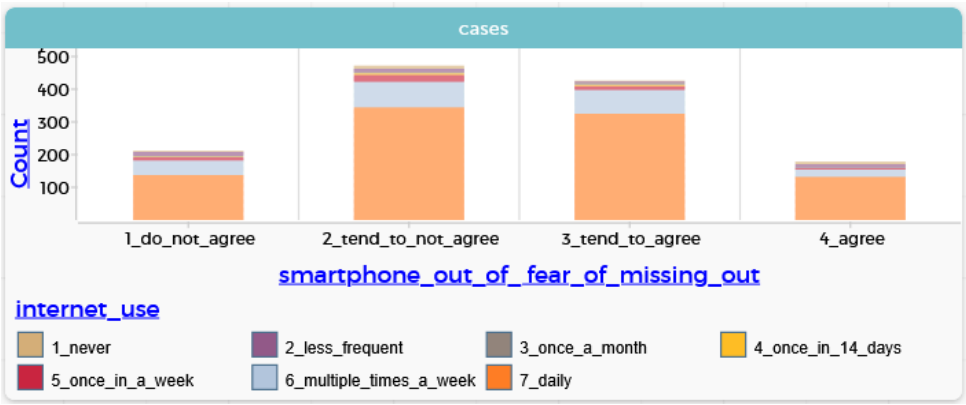

Figure 2. CODAP graph of group 1.

Table 4. Results for category 3: "Model/Analyze the data".

Group	1	2	3	4	5	6	7	8
Number of graphs	1	5	13	5	1	1	2	1
Description for each graph	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes

As data science practices, we interpreted that the groups used *big data technology* while using CODAP, *exploratory data analysis*, *data visualization*, and *communication skills* in the process of modeling/analyzing the data.

Category 4: Communicate Results and Engage in an Action Plan

Most groups directly linked their findings to their original questions or hypotheses, though the descriptions remained relatively brief. Only a few groups addressed the "action plan" explicitly. Where it was mentioned, the recommendations remained general and broad (see group 1).

Examples: "Social media activities differ between male and female users, but leisure activities in the real world are balanced. Female users use social media platforms (Instagram, etc.) more. Male users use streaming platforms (Twitch, Twitter, etc.) more." as interpretation with regard to the posed question and presumed hypotheses (group 3); "No clear link between daily internet use and FOMO.

Other variables need to be investigated. There seem to be other reasons why some young people are afraid of missing out" as an interpretation with regard to the posed question in section *Model/Analyze the data* and "Outlook: Investigation of the underlying reasons for fear of missing out is necessary, as this is common among many young people." as part of a possible action plan (group 1).

Table 5. Results for category 4: "Communicate results and engage in an action plan".

Group	1	2	3	4	5	6	7	8
Interpretation related to question	Partly*	Yes	Yes	Yes	No	Yes	Yes	Yes
Action plan addressed	Yes	Yes	No	Yes	No	Yes	Yes	No

*Group 1 answered their question posed in section *Analyze the data*, but not their initial questions formulated in section *Formulate the problem*.

As data science practices, we interpreted that the groups used *communication skills* to communicate the results and engage in an action plan.

SUMMARY AND DISCUSSION

In Category 1 (*Get and explore the data*), all groups described the dataset, identifying between 2 and 7 aspects (variables, values), but only three groups used graphs to support their exploration.

In Category 2 (*Formulate the problem*), every group developed at least one statistical question or hypothesis, mostly unrelated to the module topic *gaming*, showing consistent engagement in question formulation. Preservice teachers' formulation of the questions suggests that posing questions worked quite well in our study, in contrast to previous research indicating that formulating statistical questions is often a challenge for learners (Leavy & Frischemeier, 2022). We interpret that this may be due to the initial, quite extensive exploration of the data before formulating a problem. The challenge here was to relate the question to the topic of gaming, but all formulated questions fit the dataset. Surprisingly, direct reference to the topic of "gaming" was rare. While all groups set a clear focus in their questions, most diverged from the central theme, e.g., exploring media use more broadly rather than specifically gaming. We interpret that the complexity of a large, multivariate dataset posed a challenge, as preservice teachers can become overwhelmed by numerous variables.

In Category 3 (*Model/Analyze the data*), all groups created graphs (1–13 graphs per group), with most of them also providing explanations, indicating a strong focus on data visualization and interpretation. A deep analysis of the participants' statistical competencies remains open at this stage, but we have two points of discussion here. First, all participants were able to create meaningful visualizations from the data that helped to answer their questions. Second, looking at some visualization such as Fig. 2, a more successful interpretation might have been possible when displaying relative instead of absolute frequencies. We assume that participants in this case were challenged by comparing groups or the use of the digital tool CODAP. A support for this is that most descriptions in this category were brief and focused only on selected aspects of the visualizations (e.g., notable peaks or differences), suggesting again limited analytical depth. This aligns with the findings of Frischemeier et al. (2021), in which students in grade 10 (age 16-17) also worked with this dataset and experienced challenges interpreting the often highly complex graphs, although participating in several teaching lessons on the statistical background and the use of digital tool beforehand.

In Category 4 (*Communicate results and engage in an action plan*), most groups linked findings to their questions, but action plans were often missing or too general, showing weaker engagement with applying findings to recommendations.

CONCLUSION

The results show that students followed the DataSETUP module effectively and engaged with a wide range of data science practices while working through the four data science processes. While not every module is intended to cover all data science practices outlined in the framework, many of these practices naturally emerged through working on this module. Although the module was framed around the topic of gaming, this topic was rarely taken up directly by the students. Instead, they used the rich dataset as a springboard to formulate their own questions and hypotheses based on personal interests.

In terms of the framework's goal—to introduce students to data science—one key insight is that a rich dataset, closely connected to students' everyday experiences (in this case through variables related

to leisure activities and social media use), can spark meaningful engagement. Several data science practices were frequently applied.

The simplicity of the DataSETUP framework appears to have been a strength in this context: despite having little prior experience with data, the students were able to explore the dataset productively within a 90-minute session—often making personal connections to their own lives— and engage with the different data science processes and practices, which strengthens the general setup of the DataSETUP framework. Future investigations will examine how other DataSETUP modules are used and which data science practices can be further supported through their implementation.

ACKNOWLEDGEMENT

The Promoting Data Science Education for Teacher Education at the University level (DataSETUP) project (2023-1-DE01-KA220-HED-00160333) is co-funded by the European Union's Erasmus+ Cooperation Partnership Programme. All views expressed are those of the authors and do not necessarily reflect those of the European Commission

REFERENCES

Bargagliotti, A., Arnold, P., & Franklin, C. (2021). GAISE II: Bringing data into classrooms. *Mathematics Teacher: Learning and Teaching PK-12*, *114*(6), 424–435. https://doi.org/10.5951/MTLT.2020.0343.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A curriculum framework for K-12 statistics education*. Alexandria, VA: American Statistical Association.

Friedrich, A., Schreiter, S., Vogel, M., Becker-Genschow, S., Brünken, R., Kuhn, J., Lehmann, J., & Malone, S. (2024). What shapes statistical and data literacy research in K-12 STEM education? A systematic review of metrics and instructional strategies. *IJ STEM Ed*, *11*(58). https://doi.org/10.1186/s40594-024-00517-z.

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). https://doi.org/10.1111/test.12283.

Fry, K., & Makar, K. (2021). How could we teach data science in primary school? *Teaching Statistics*, *43*(S1). https://doi.org/10.1111/test.12259.

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., ... & Zanontian, L. (2016). Teaching data science to secondary students: The mobilize introduction to data science curriculum. In J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE)*, July 2016, Berlin, Germany.

Horton, N. J., & Hardin, J. S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, *29*(sup1), 1–3. https://doi.org/10.1080/10691898.2020.1870416.

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science*. The international data science in schools project. http://www.idssp.org/files/IDSSP_Frameworks_1.0.pdf.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing data science: A framework and case study. *Harvard Data Science Review*, *2*(1). https://assets.pubpub.org/l534v3hn/28d7ddde-a9ab-47a6-9521-fc7105b716db.pdf.

Lee, H. S., Mojica, G. F., Thrasher, E. P., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, *21*(2), Article 3. https://doi.org/10.52041/serj.v21i2.41.

Leavy, A., & Frischemeier, D. (2022). Developing the statistical problem posing and problem refining skills of prospective teachers. *Statistics Education Research Journal*, *21*(1), 1–27. https://doi.org/10.52041/serj.v21i1.226

Mayring, P. (2015). Qualitative content analysis: theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365–380). Springer. https://doi.org/10.1007/978-94-017-9181-6_13.

Oliver, J. C., & McNeil, T. (2021). Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context. *PeerJ Computer Science*, *7*, e441. https://doi.org/10.7717/peerj-cs.441.

Ow-Yeong, Y. K., Yeter, I. H., & Ali, F. (2023). Learning data science in elementary school mathematics: a comparative curriculum analysis. *International Journal of STEM Education*, *10*(8). https://doi.org/10.1186/s40594-023-00397-9.

Podworny, S., Fleischer, Y., Stroop, D., & Biehler, R. (2022). An example of reach, real and multivariate survey data for use in school. In J. Jodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)* (pp. 940–947). Free University of Bozen-Bolzano and ERME. https://hal.science/CERME12/hal-03751842v1.

Ridgway, J. (2022). *Statistics for empowerment and social engagement. Teaching civic statistics to develop informed citizens*. Springer. https://doi.org/10.1007/978-3-031-20748-8.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., & Wuetherick, B. (2015). *Strategies and best practices for data literacy education: Knowledge synthesis report*. http://hdl.handle.net/10222/64578.

Schreiter, S., Friedrich, A., Fuhr, H., Malone, S., Brünken, R., Kuhn, J., & Vogel, M. (2024). Teaching for statistical and data literacy in K-12 STEM education: a systematic review on teacher variables, teacher education, and impacts on classroom practice. *ZDM–Mathematics Education*, *56*(1), 31–45. https://doi.org/10.1007/s11858-023-01531-1.

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.