Data collection as a catalyst for data literacy: A concept for building Smart City gadgets at school

Verena Witte, Angela Schwering and Daniel Frischemeier
University of Muenster, Germany
verena.witte@uni-muenster.de

In a data-driven world, fostering data literacy among students is essential - yet the sub-skill of data collection remains underrepresented in educational practice. This study explores how personal data collection using environmental sensors within a 'Smart City' project influences students' ability to critically evaluate corrupted datasets. In a quasi-experimental design, 79 high school students participated in an intervention, during which they developed sensor-based Smart City projects and collected their own data. A pre- and posttest measured their ability to identify outliers in a corrupted dataset. The results indicate an improvement in the experimental group compared to a control group that worked with externally sourced data. Findings suggest that engaging with personally collected data enhances students' understanding of data quality and supports the development of critical data reasoning skills. The study underscores the educational value of authentic data practices and calls for further research into their long-term curricular integration.

INTRODUCTION

Due to digitalization and technological advancement, data has become an integral part of our lives. As the foundation of artificial intelligence, data significantly influences our everyday lives and controls processes in ways we may not even realize (Kumar et al., 2023; Olari, 2023). An example of this can be found in Smart Cities, where real-time data is used to control technologies such as traffic lights and street lamps (Kaluarachchi, 2022). Environmental data, such as air quality measurements, are also relevant here to monitor public health (Allbach et al., 2014). In order to engage with such data in a reflective and competent way, comprehensive data literacy is essential (Engel et al., 2022). It should empower citizens to engage competently and critically with data, enabling them to recognize both its potential and risks and to make data-based decisions (Ridgway et al., 2022). To ensure all citizens acquire this skill and can participate actively in society, data literacy should be taught as early as possible at school level (Ben-Zvi et al., 2018; Leavy et al., 2018).

There are initial concepts and approaches for this; however, these often focus on the process of data analysis and neglect the planning and execution of data collection (Witte et al., 2024). At the same time, it has been emphasized that authentic and real-world data sets can enhance students' deep understanding of data (de Luca & Lari, 2011; Ucar & Trundle, 2011). Such data sets can, for instance, result from personally conducted data collection, which can have a significant impact on how a data set is interpreted (Wolff et al., 2019). The latter is also influenced by participants' available contextual knowledge, which is often used to support existing arguments (Fielding et al., 2025).

This paper addresses the discrepancy between the acknowledged importance of personal data collection and the lack of attention this sub-skill receives in existing educational concepts. On one hand, it investigates the research question: To what extent does personal data collection by learners influence their critical understanding of a corrupted data set? On the other hand, to foster the integration of the sub-skill 'data collection' in K-12 education, this paper presents an instructional concept focused on this sub-skill. Inspired by Wolff et al. (2019), the concept incorporates the topic of 'Smart City' and aims to guide 14- to 17-year-olds in designing and building a data-driven Smart City gadget. Thus, the sub-skill of data collection is explored both from a scientific perspective and in terms of practical implications for the classroom. The paper first takes a closer look at working with real-world data and offers recommendations for K-12 education. This is followed by a pedagogical intervention study that investigates the research question using an experimental pre-post design. This provides insights for future concepts aimed at promoting data literacy in K-12 education.

THEORETICAL BACKGROUND

Data literacy refers to the ability to critically collect, organize, evaluate, and apply data (Ridsdale et al., 2015). A look at existing competence models reveals that various sub-skills are required

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*.

throughout the process - from formulating a research question to communicating new findings - that contribute to competent data engagement (Bargagliotti et al., 2020; IDSSP Curriculum Team, 2019; Wolff et al., 2016). One of the most recent and comprehensive models is presented by Lee et al. (2022), which includes the following sub-skills: Frame Problem, Consider & Gather Data, Process Data, Explore & Visualize Data, Consider Models, Communicate & Propose Action. Similar to other existing models (Wild & Pfannkuch, 1999; Wolff et al., 2016), it emphasizes addressing real problems and incorporating relevant, authentic data sets. Regarding data collection, Lee et al. (2022) highlights the importance of understanding and questioning appropriate data collection methods, as well as assessing research conditions and data validity. This point is becoming increasingly important based on the assumption that data is now taking on unprecedented forms and is consequently being presented in a messy and unstructured way (Fielding et al., 2025). The following focuses on these aspects and considers the potential and challenges of working with real-world data sets.

Real World Data

Real Data – Real Learning – Real Data Literacy. With these three points, Erwin (2015) underscores the importance of working with real-world data, especially its impact on learners' intrinsic motivation. Moreover, students working with authentic data typically achieve a deeper understanding of both the topic and scientific concepts than those who do not have access to such data (de Luca & Lari, 2011; Ucar & Trundle, 2011). Open data, such as from citizen science or participatory sensing projects, can be a source of real-world data sets, but they also pose challenges for data analysis (Gould et al., 2017; Ridgway, 2016). One reason is their raw form and possibly hidden information regarding data collection procedures, which learners must examine in terms of data validity (Rubin, 2021). This non-trivial approach fosters critical and transdisciplinary thinking (Atenas et al., 2021). Wolff et al. (2019) suggest that these skills can be strengthened when learners conduct their own data collection. This assumed relationship between data familiarity and the ability to critically assess it will now be examined more closely.

Strengthening Data Literacy in K-12 Education

Various approaches exist to integrate the use of authentic and real-world data sets into both formal and informal education. Schreiter et al. (2024) recommend a project-based approach to introduce learners to working with data and empower them to solve specific problems through real-world applications. At the same time, content and methodology should be transdisciplinary (Friedrich et al., 2024; Schüller et al., 2021). Ridsdale et al. (2015) also emphasize clear learning goals, application-oriented learning through hands-on experimentation, modular learning with small successes, and the connection of theory and practice. Our scoping review on strengthening data literacy in K-12 education has shown that topics such as sustainability, climate change, or mobility transitions are particularly suitable (Witte et al., 2024). Wolff et al. (2019) also address this topic, focusing on Smart Cities. In this context, data collected through technology directly relates to the participants' environment and illustrates the process of data-driven decision-making. This approach is taken up in the following and serves as the basis for an instructional concept within which the research question introduced earlier is explored.

RESEARCH METHOD

As part of a pedagogical intervention study, the use of personal collected, real-world data sets was investigated more closely. An exemplary project day was developed, following the principles of project-based and transdisciplinary learning in the context of ‘Smart City’. Using a pre- and posttest in both an experimental and a control group, the relationship between data familiarity and the ability to critically analyze data could be examined in detail.

Participants

The project day on ‘Smart City’ was conducted in four classes at two different schools. Both schools were located in a city designated as a Smart City and featured various technologies that (often unconsciously) influenced students’ lives. The participating students were enrolled in a high school geography course and were between 15 and 17 years old. At this age, students are generally capable of working independently on projects and possess the mathematical skills to perform basic data analyses,

such as calculating averages (Ministry for Schools and Education of North Rhine-Westphalia, 2022). A total of $n = 79$ students participated in the project day: 49.4% male, 48.1% female, and 2.5% non-binary. Two of the four classes were randomly assigned to the experimental group ($n = 48$), while the other two were assigned to the control group ($n = 31$). The control group ensured internal validity of the intervention and minimized potential confounding factors.

Procedure

The goal of the project day was for learners to develop their own Smart City gadget that interacts with users based on the data collected (see figure 1). To provide the theoretical foundation, learners were first introduced to the topic of 'Smart City'. Their associations with the term were gathered and used to derive a definition. Using photos from their own city, a local context was created, and various technologies were identified as Smart City gadgets - such as adaptive lighting for bike paths or monitoring of lake water quality using sensors. Together with the students, data collection practices within the school were discussed, including CO2 traffic lights that provide information about air quality in classrooms and prompt ventilation when CO2 levels are too high. At this point, the pretest was conducted, where students analyzed data to assist the school janitor in deciding whether to install a new ventilation system.

Following this, the control group analyzed additional CO2 data from within the school building, obtained from an open environmental data platform, and transferred this data from graphs into tables. These data formed the basis for another round of data analysis during the posttest. Meanwhile, the experimental group personally collected data using CO2 sensors within the school building and used this data for their posttest analysis. After the posttest, students were tasked with developing their own Smart City gadget. They were introduced to the use of the senseBox, which combines a programmable microcontroller with environmental sensors - making it suitable for building such gadgets and offering insights into the digital world's 'black box' (Biehler et al., 2018; Podworny et al., 2022).

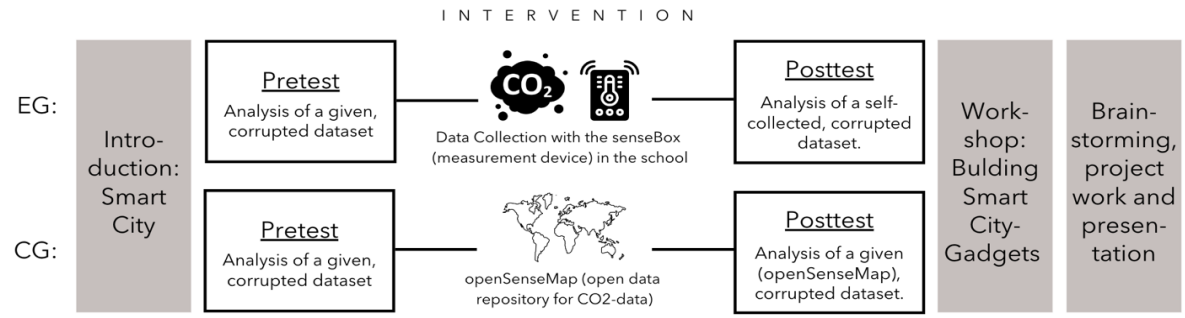


Figure 1. Procedure of the project day and the pedagogical intervention research.

Using supporting materials and project plans, the students developed their own ideas for smart technologies within their school or city and considered what direct consequences the collected data could have for life in the city. The personal data collection using technology that took place in this context can offer a practical and reflective approach to the field of 'Big Data' (Biehler et al., 2023). Throughout this process, various teachers provided advisory support and assisted with idea generation or project implementation if needed. At the end of the day, students presented their projects to one another and voted on the most creative project and the one with the best technical implementation (see figure 2).

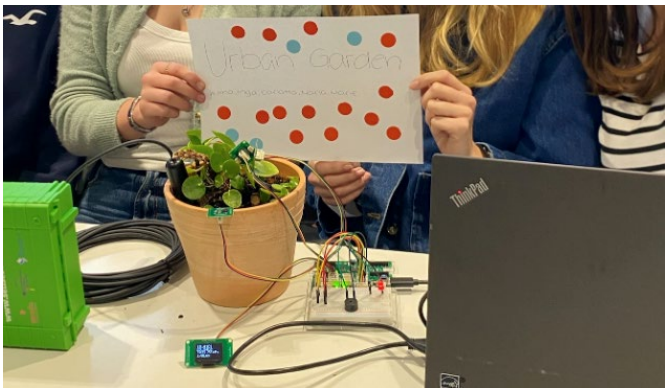
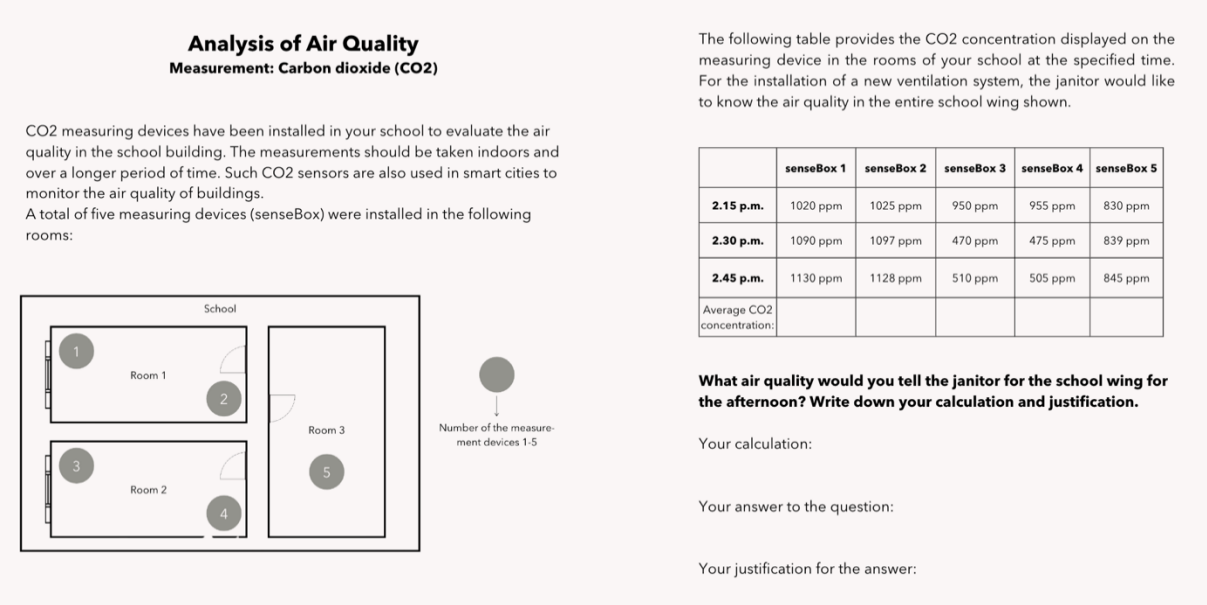


Figure 2. A smart city project by learners on the data-based cultivation of plants, called 'Urban Garden'.

Study Instrument

The pre- and posttests of the study focus on the analysis of a corrupted dataset. In both cases, students must use data from a CO2 sensor to assess whether the installation of a new ventilation system in the school is necessary. The test specifies that data must be collected over a longer period and in enclosed rooms. However, the provided data is corrupted (e.g., due to open windows during data collection), so the challenge for students is to identify and eliminate some data points as outliers in order to calculate the average air quality (see figure 3). The structure and difficulty of the pre- and posttests are identical. The difference lies in the fact that while both groups receive a dataset in the pretest, the experimental group enters their own self-collected (corrupted) data into the table in the posttest, whereas the control group works with equally corrupted data from the open environmental data platform.



Analysis of Air Quality

Measurement: Carbon dioxide (CO2)

CO2 measuring devices have been installed in your school to evaluate the air quality in the school building. The measurements should be taken indoors and over a longer period of time. Such CO2 sensors are also used in smart cities to monitor the air quality of buildings.
A total of five measuring devices (senseBox) were installed in the following rooms:

The following table provides the CO2 concentration displayed on the measuring device in the rooms of your school at the specified time. For the installation of a new ventilation system, the janitor would like to know the air quality in the entire school wing shown.

	senseBox 1	senseBox 2	senseBox 3	senseBox 4	senseBox 5
2.15 p.m.	1020 ppm	1025 ppm	950 ppm	955 ppm	830 ppm
2.30 p.m.	1090 ppm	1097 ppm	470 ppm	475 ppm	839 ppm
2.45 p.m.	1130 ppm	1128 ppm	510 ppm	505 ppm	845 ppm
Average CO2 concentration:					

What air quality would you tell the janitor for the school wing for the afternoon? Write down your calculation and justification.

Your calculation:

Your answer to the question:

Your justification for the answer:

Figure 3. Pretest for analyzing the CO2 concentration in the afternoon. The values of stations 3 and 4 have been falsified by ventilation and must be sorted out.

Statistical Approach

To evaluate the results of the pre- and posttests, students' responses were coded and rated on a scale of zero, one, or two points. One point was awarded if corrupted data sets were excluded from the calculation and the correct average air quality was determined. An additional point could be earned if the result was critically reflected upon or a justification was given for excluding certain data points. This was the case when terms such as "outlier", "unusual", "ventilated", "window", or "polluted" were used. If neither criterion was met, the response received zero points.

A frequency analysis was initially conducted to gather information about the sample. This was followed by descriptive statistics and a bootstrapping procedure with 1,000 iterations to ensure robust estimates. Since the data were not normally distributed, the Mann–Whitney *U* test was used to calculate mean ranks and effect size. All analyses focused on the difference between pre- and posttest results while accounting for baseline conditions. A multiple test correction was applied by adjusting the confidence interval to 97.5% and the significance level to $p < .025$. A key focus was the relationship between the intervention and the change from pre- to posttest, particularly in comparison between the experimental and control groups. All analyses were conducted using *IBM SPSS Statistics*, version 29.0.2.0 for macOS.

RESULTS

The pre- and posttest results reflect the initial status of students' ability to critically evaluate a dataset. Among the total of $n = 79$ participants, 94.5% scored zero points in the pretest, 3.9% scored one point, and 1.3% scored two points. The initial conditions between the experimental and control groups were nearly identical. These pretest results highlight the need for change, leading to a closer look at the effect of the intervention: While the control group averaged $M = 0.1$ points in both pre- and posttests, the experimental group improved from $M = 0.04$ in the pretest to $M = 0.31$ in the posttest. This corresponds to an increase of $M = 0.27$ with a 97.5% confidence interval of [0.10; 0.49] (see figure 4). This difference is supported by the Mann–Whitney *U* test: the control group had a mean rank of $M_{Rank} = 35.08$, while the experimental group had a mean rank of $M_{Rank} = 42.26$, with $U = 587.5$, $Z = 2.250$. The rank-biserial correlation of $r = 0.255$ indicates a small to medium effect, as does Cliff's delta of $\delta = 0.149$.

A closer look at the qualitative responses reveals four categories: (i) Students who did not question the data critically and simply calculated the average provided a purely mathematical justification, e.g., "*On average, the CO_2 concentration in the afternoon is 857.93 ppm. This value results from the average of all 5 rooms*" (score: 0 points). (ii) Students have excluded the outliers from their calculation but have not provided a substantive reason for doing so, e.g. "*The average of stations 1, 2 and 5 is 1000,4 ppm*" (score: 1 point). (iii) The learners calculated the average of all five stations but critically questioned and reflected on the result, for example: "*The result is 857.93 ppm. However, it is*

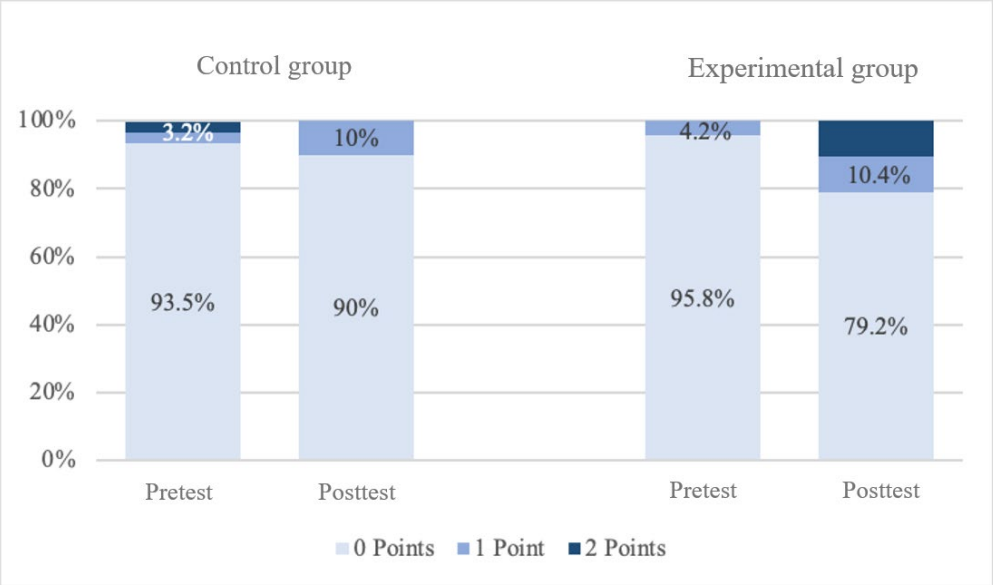


Figure 4. Percentage of students who achieved zero, one or two points in the pretest and posttest.

actually important to consider the average for each room individually, as each room has different conditions. For instance, it could be that the rooms with stations 3 and 4 were ventilated, or that there was a different number of people present, etc." (score: 1 point). (iv) Other participants noticed the unusually low values, excluded them from their calculation and critically reflected on the result and its

generalizability: "*I would cite the average value of the stations 1, 2 and 5 (1000.4 ppm). Since the first two stations drop so suddenly, it suggests, for example, that the windows were opened. These values would distort the result and measure the fresh air instead - not the actual condition"* (score: 2 points). The last case therefore applies to 10.4% of the students in the experimental group, as they provided an optimal answer and thus scored two points. The remaining 10.4% either presented a correct calculation without justification or, alternatively, an incorrect calculation (including outliers) along with a critical evaluation of the result.

DISCUSSION & FURTHER STUDY

The presented results highlight the poor initial ability of students to engage critically with data. Only 5.2% of the participants not only performed a mathematical operation but also questioned and critically examined the provided data. Authentic datasets inherently pose greater challenges in analysis due to their non-trivial structure, compared to pre-cleaned data (Gould et al., 2017). In evaluating such datasets, learners must identify and eliminate potential confounding factors based on the data itself and supplementary information (Rubin, 2021). This ability was only evident in the experimental group following the intervention involving personal data collection. As a result, over 20% of the students in this group scored one or two points in the posttest, indicating that they identified outliers or critically reflected on the calculated result. The average improvement between pre- and posttest scores differed by $M = 0.27$ points between the experimental and control groups. Combined with the calculated effect sizes, the results suggest that data collection as part of the intervention led to greater improvement in critical data analysis than using existing datasets from the openSenseMap platform. This finding aligns with the assumptions by Wolff et al. (2019), which indicate a correlation between familiarity with a dataset and the ability to engage with it critically.

The effect sizes are small to moderate, indicating that while a difference exists, it is not particularly large. One possible explanation is the students' limited familiarity with the environmental phenomenon of CO2, despite the introductory session, leading them to interpret the values more as abstract numbers than meaningful information. Repeating the study with more familiar, everyday data - such as weather-related metrics (e.g., temperature, UV radiation) - would be a worthwhile next step. Furthermore, students' familiarity with data could be increased through longer-term engagement, such as a week-long project, rather than a single-day intervention. In light of these findings, the presented 'Smart City' concept appears well-suited to introduce learners to the complex world of data, to demonstrate the real-life implications of sensor-collected data, and to foster critical reflection during data analysis. It is important to note that this was an experimental setting using a task specifically developed for this purpose. Future research should include the development of a comprehensive test instrument that ensures validity, reliability, and fairness at all levels. Such a test could also be implemented within a research design that moves beyond the slightly variable conditions of pedagogical intervention studies, thereby minimizing potential confounding variables.

CONCLUSION

To strengthen data literacy, especially in terms of the often overlooked area of data collection, a concept on the topic of 'Smart City' was developed using the senseBox. This concept formed the basis for pedagogical intervention research that explored the connection between familiarity with data - for example through collecting data personally - and the ability to critically evaluate a dataset during analysis. The findings support this connection and reinforce existing assumptions in the field. At the same time, this study is only a starting point. There is potential for improvement, particularly regarding sample size, test design, and the choice of dataset. The project day, during which learners created their own Smart City gadgets using microcontrollers and sensors, helped to strengthen their understanding of data collection and emphasized its importance. It can serve as a basis for a broader educational framework that includes all aspects of data literacy. Further research is needed to evaluate and develop such concepts.

REFERENCES

Allbach, B., Henninger, S., & Deitche, E. (2014). An urban sensing system as backbone of smart cities. REAL CORP 2014. Proceedings of 19th International Conference on Urban Planning, Regional Development and Information Society, pp. 55–64.

Atenas, J., Havemann, L., & Priego, E. (2021). Open data as open educational resources: Towards transversal skills and global citizenship. *Open Praxis*, *7*(4), 377–389. https://doi.org/10.5944/openpraxis.7.4.233

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II)*. American Statistical Association and National Council of Teachers of Mathematics.

Ben-Zvi, D., Makar, K., & Garfield, J. (2018). *International handbook of research in statistics education*. Springer.

Biehler, R., Frischemeier, D., Gould, R., & Pfannkuch, M. (2023). Impacts of digitalization on content and goals of statistics education. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 1–37). Springer International Publishing. https://doi.org/10.1007/978-3-030-95060-6_20-1

Biehler, R., Frischemeier, D., Podworny, S., Wassong, T., Budde, L., Heinemann, B., & Schulte, C. (2018). *Data science and big data in upper secondary schools: A module to build up first components of statistical thinking in a data science curriculum*. https://doi.org/10.5445/KSP/1000087327/28

de Luca, V., & Lari, N. (2011). The GRIDc project: Developing students' thinking skills in a data-rich environment. *Journal of Technology Education*, *23*(1), 5–18. https://doi.org/10.21061/jte.v23i1.a.2

Engel, J., Nicholson, J., & Louie, J. (2022). Preparing for a data-rich world: Civic statistics across the curriculum. In J. Ridgway (Ed.), *Statistics for Empowerment and Social Engagement* (pp. 445–475). Springer International Publishing. https://doi.org/10.1007/978-3-031-20748-8_18

Erwin, R. W. (2015). Real-World learning through problem-solving with datasets. *American Secondary Education*, *43*(2), 18–26.

Fielding, J., Makar, K., & Ben-Zvi, D. (2025). Developing students' reasoning with data and data-ing. *ZDM Mathematics Education, 57,* 1–18. https://doi.org/10.1007/s11858-025-01671-6

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. *International Journal of STEM Education*, *11*(1), 58. https://doi.org/10.1186/s40594-024-00517-z

Gould, R., Bargagliotti, A., & Johnson, T. (2017). An analysis of secondary teachers' reasoning with participatory sensing data. *Statistics Education Research Journal*, *16*(2), 305–334. https://doi.org/10.52041/serj.v16i2.194

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science*. http://idssp.org/files/IDSSP_Frameworks_1.0.pdf

Kaluarachchi, Y. (2022). Implementing data-driven smart city applications for future cities. *Smart Cities*, *5*(2), 455–474. https://doi.org/10.3390/smartcities5020025

Kumar, S., Sharma, R., Singh, V., Tiwari, S., Singh, S. K., & Datta, S. (2023). Potential impact of data-centric AI on society. *IEEE Technology and Society Magazine*, *42*(3), 98–107. https://doi.org/10.1109/MTS.2023.3306532

Leavy, A., Meletiou-Mavrotheris, M., & Paparistodemou, E. (Eds.) (2018). *Statistics in early childhood and primary education: Supporting early statistical and probabilistic thinking*. Springer Singapore.

Lee, H., Mojica, G., Thrasher, E., & 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

Ministry for Schools and Education of North Rhine-Westphalia. (2022). *Kernlehrplan für die Sekundarstufe 1, Gesamtschule/ Sekundarschule in Nordrhein-Westfalen. Mathematik.*

Olari, V. (2023). Data Literacy as a fundamental component of artificial intelligence education in schools. *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, 1–2. https://doi.org/10.1145/3631802.3631839

Podworny, S., Hüsing, S., & Schulte, C. (2022). A place for a data science project in school: Between statistics and epistemic programming. *Statistics Education Research Journal*, *21*(2), 1–15. https://doi.org/10.52041/serj.v21i2.46

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

Ridgway, J., Campos, P., & Biehler, R. (2022). Data science, statistics, and civic statistics: Education for a Fast Changing World. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement* (pp. 563–580). Springer International Publishing.

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]. Dalhouse University.

Rubin, A. (2021). What to consider when we consider data. *Teaching Statistics*, *43*, 23–33. https://doi.org/10.1111/test.12275

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

Schüller, K., Koch, H., & Rampelt, F. (2021). *Data literacy charter*. Stifterverband.

Ucar, S., & Trundle, K. C. (2011). Conducting guided inquiry in science classes using authentic, archived, web-based data. *Computers & Education*, *57*(2), 1571–1582. https://doi.org/10.1016/j.compedu.2011.02.007

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Witte, V., Schwering, A., & Frischemeier, D. (2024). Strengthening data literacy in K-12 education: A scoping review. *Education Sciences*, *15*(1), 25. https://doi.org/10.3390/educsci15010025

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, *12*(3), 9–26.

Wolff, A., Wermelinger, M., & Petre, M. (2019). Exploring design principles for data literacy activities to support children's inquiries from complex data. *International Journal of Human-Computer Studies*, *129*, 41–54. https://doi.org/10.1016/j.ijhcs.2019.03.006