# ENHANCING DATA LITERACY THROUGH (VIRTUAL) REALITY: A PEDAGOGICAL INTERVENTION RESEARCH ON THE SUB-SKILL OF DATA COLLECTION

VERENA WITTE
*University of Münster*
*verena.witte@uni-muenster.de*

ANGELA SCHWERING
*University of Münster*
*schwering@uni-muenster.de*

DANIEL FRISCHEMEIER
*University of Münster*
*dfrische@uni-muenster.de*

**ABSTRACT**

*Proficient handling of data is a skill gaining importance with the increasing amount and availability of data. Therefore, promoting data literacy should begin in everyday school life. The foundation for this is formed by competence models for data literacy, which include the sub-skill of data collection that has so far been inadequately considered in K–12 education. This study investigates the relevance of data collection through pedagogical intervention research with learners aged 14 to 17 years. The evaluation highlights the benefits of personal data collection—data collection in one's own environment and in virtual reality—as part of a holistic approach to teaching data literacy. This work extends existing approaches regarding the importance of data collected by students for their own use through sensors for a reflective and critical understanding of data.*

***Keywords:*** *Data literacy; Citizen science; Data collection; Virtual reality; K–12 education*

## 1. INTRODUCTION

Living in today's society requires skills that respond to the developments of digitalization and globalization, such as the 21st-century skills (Vuorikari et al., 2022). These skills include not only certain basic competencies and character traits but also the "4Cs" of communication, collaboration, creativity, and critical thinking (Fadel et al., 2015). Critical thinking, especially, is of great importance in order to make informed decisions in times of fake news and social media. Data form the basis for many decisions as well as for innovation and progress (Engel et al., 2022). Therefore, a reflective and critical approach to data is essential for participating responsibly in societal life (Ridgway, 2016). This approach is described as data literacy and complements the forward-looking competencies of the 21st-century skills (Van Audenhove et al., 2024). At present, the availability of datasets is growing rapidly —keyword big data—yet the necessary knowledge to deal with the data is not being developed comprehensively (Engel et al., 2022; German Informatics Society, 2024). To address this knowledge gap, the teaching of data literacy should begin in early childhood education (Matthews, 2016). The foundation for this is provided by international competence models that help shape the educational landscape and take into account current issues of digitalization, such as the European Commission's (2024) initiative, *European Skills, Competences and Occupations* (*ESCO*), and *DigComp 2.2: The Digital Competence Framework for Citizens* (Vuorikari et al., 2022). *The Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II* (GAISE II; Bargagliotti et al., 2020) also emphasized the importance of statistical thinking and data literacy in a data-driven world. GAISE II highlighted that the process of statistical problem-solving is of significant importance and encompasses

the following four components: (a) formulate statistical investigative questions, (b) collect/consider the data, (c) analyze the data, and (d) interpret the results. These four components should be integrated into school education to impart competent handling of data and thereby enable responsible participation in society (Arnold & Franklin, 2021; Gould, 2021). A prerequisite for comprehensive and transdisciplinary teaching of data literacy is the establishment of respective school or country curricula. In the four countries of Ireland, the Netherlands, Austria, and Lithuania, the aspects mentioned above that are relevant for competent data handling are considered in their K–12 education guidelines (German Informatics Society, 2024). Curricula for a comprehensive strengthening of data literacy can only be found to some extent in Germany and Finland, whereas in the other European Union (EU) countries, data literacy is not mentioned at the political level (German Informatics Society, 2024). Moreover, the school curriculum for statistics does not adequately prepare students for dealing with real data because the focus is on mastering mathematical techniques rather than understanding and interpreting data (Ridgway & Nicholson, 2010).

Discrepancies can be identified between the curricula of different countries mentioned, as well as between the curricula and the overarching guidelines for promoting data literacy. Research also reveals different priorities when it comes to strengthening data literacy in education. Within research reported in the last 10 years, a large percentage of the studies focused on the teaching of data literacy in K–12 education limited their designed and evaluated project units to the area of data analysis (Witte et al., 2024). In contrast, significantly fewer articles covered data collection by learners—a focus that, in addition to the area of data evaluation, would make it possible to address the entire field of data literacy and the statistical problem-solving process (e.g., De Oliveira Souza et al., 2020; Frischemeier, 2020; Lee et al., 2022). Personal data collection—data collected by students for their own use—represents a way to produce real datasets. Working with real and authentic datasets offers potential for students to gain a deeper understanding of the data (De Luca & Lari, 2011; Ucar & Trundle, 2011) than working with contrived data. In addition, there presumably is a connection between learners' familiarity with a dataset, such as through data collection, and their ability to analyze it critically.

The imbalance in the consideration of different sub-skills of data literacy, particularly between data collection and data analysis, should be critically examined. After all, active participation in society requires competent handling of data, including all sub-skills, which can be accomplished through interdisciplinary and comprehensive teaching of data literacy (Engel et al., 2022). Therefore, the question arises as to what significance personal data collection can have in terms of a comprehensive teaching of data literacy. Additionally, the question is whether alternative approaches to regular data collection, such as data collection in virtual reality, can open up new potential for students' learning. Although the increased effort and the difficult integration of personal data collection into the school routine may be possible reasons for the infrequent consideration of this sub-skill, new technologies offer more flexibility and time savings (Shute et al., 2017). With the help of Immersive Virtual Reality (IVR), activities from the real world can be transferred to the virtual world (Maresky et al., 2019). However, the insights into the effectiveness of using IVR are predominantly subject-specific at this point and cannot be generalized, thus opening up another area of investigation (Liu et al., 2017).

This paper demonstrates the extent to which the aspect of personal data collection by learners can impact other areas of skills within areas of data literacy, particularly in terms of a deeper understanding of data in the context of data analysis. Thus, the relevance and influence of the sub-skill of data collection on students' acquisition of comprehensive data literacy are highlighted. For this purpose, pedagogical intervention research is used to evaluate (a) data collection by learners in reality, and (b) data collection in IVR, in order to enable data collection in the classroom and simultaneously at virtual locations that would otherwise be inaccessible. This study creates a research-based foundation for subsequent modifications in the teaching of data literacy in both in-school and extracurricular settings. To explore this topic in greater detail, this paper first focuses on the sub-skill of data collection within the construct of data literacy, using examples such as citizen science projects. Subsequently, IVR is presented as an alternative to traditional classroom scenarios. The presentation of the pre-post study design, as well as the presentation and discussion of the results, ultimately provide implications for sustainable teaching of data literacy in everyday school life. This article focuses on the generation of real data using sensor-equipped measuring devices but does not address learners' design of methodology, nor their selection and placement of sensors. These latter aspects are considered within the domain of consider and gather data. The consider and gather data domain encompasses a variety of

sub-skills, with personal data collection within the context of citizen science projects or similar initiatives being one of them (Lee et al., 2022).

## 2. THEORETICAL BACKGROUND

### 2.1. DATA COLLECTION AS PART OF COMPETENCE FRAMEWORKS

Data literacy involves the ability to engage with datasets and data visualizations in a reflective and critical manner, thereby creating the foundation for informed and participatory involvement in the digital society (Ridsdale et al., 2015; Schüller, 2022). The evaluation of existing data is particularly necessary for an individual to identify misinformation and refute fake news (Carmi et al., 2020). However, to be able to generate information from data oneself, one must carry out an analysis and interpretation of available datasets (Engel, 2017). Data literacy not only encompasses dealing with existing datasets, but also includes the personal production of data, for example, in the context of personal data collection and acquisition (Lee et al., 2022). Rubin (2021) emphasizes the importance of critical reflection throughout the whole data collection and interrogation process. The ability to handle data is, therefore, very comprehensive and characterized by interdisciplinarity, meaning the connection of different disciplines (Engel, 2017). Statistical literacy builds the foundation for critically engaging with data and functions as a key competence of data literacy (Ben-Zvi & Garfield, 2004; Schield, 1999). It combines two main components: knowledge elements and dispositional elements, which together encompass both mathematical skills and context-based critical thinking (Gal, 2002). Over time, researchers have shifted focus from the concept of statistical literacy toward data literacy, with a growing focus on data within the process of inquiry (Burrill & Pfannkuch, 2024; Friedrich et al., 2024; Schreiter et al., 2024). According to Gould (2017), statistical literacy and data literacy remain closely connected. Moreover, they share a common overarching objective: fostering a competent and critical approach to data and information, allowing individuals to navigate and act reflectively in a data-driven world (De Veaux et al., 2022; Gould, 2021; Guler et al., 2016). A detailed look at various definitions and competence models shows that different sub-skills are required to meet the demands of competent data handling (Bargagliotti et al., 2020; Gould et al., 2016; International Data Science in Schools Project (IDSSP) Curriculum Team, 2019). For example, Lee et al. (2022) broke down working with data into six interconnected areas in their *Data Investigation Process* framework: frame problem, consider and gather data, process data, explore and visualize data, consider models, and communicate and propose action. Among other things, they emphasized the consideration of real-world problems and questions, the inclusion of relevant data, and the identification of potential biases and ethical issues. Regarding data collection, they specifically addressed understanding and questioning the appropriateness of a data collection method, as well as evaluating the research conditions and the validity of the data (see Figure 1).

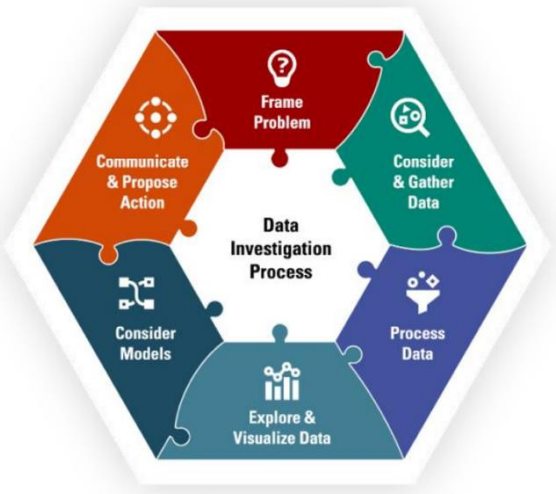



*Figure 1. Data Investigation Process framework (Lee et al., 2022)*

In contrast, Wolff et al. (2016) focused on five sub-skills in a linear sequence in their competence framework for data science education: problem, plan, data, analysis, and conclusions. Their framework was based on the investigative cycle of Wild and Pfannkuch (1999), which served as a foundation for teaching statistical thinking and problem-solving with data. In their model, Wolff et al. (2016) viewed solving a real-world problem and establishing an ethical framework as prerequisites for using self-collected or acquired data to answer a question. Here, too, the evaluation of the validity of the data used was mentioned. The *Digital Competence Framework for Citizens* (DigComp 2.2; Vuorikari et al., 2022) also highlighted the role of data literacy in a digitalized society and approached it from two perspectives: (a) using data and (b) understanding data. The latter perspective includes the sub-skills of observing, analyzing, evaluating, and reflecting, whereas the former perspective includes sub-skills of interpreting, navigating, collecting, processing, and presenting. However, these two perspectives are not equally weighted. In their review of the DigComp 2.2 report, Van Audenhove et al. (2024) noted that the area of understanding data, particularly in terms of security and privacy, was mentioned significantly more often and was thus given more importance than the area of using data in the mathematical and scientific sense. The sub-skill of data collection was mentioned only once in the report (Van Audenhove et al., 2024).

With regard to other competence models in the area of competent data handling, the unequal consideration of different sub-skills can also be observed. For example, Büscher's (2022) competence model, *Selective and Imaginative Reading of Statistical Information,* focused on the sub-skills of data organization, data analysis, data visualization, data interpretation, and data evaluation, thereby emphasizing the reading of existing data rather than data production. The latter also applies to the *Theoretical Framework of Statistical Literacy* (Kurnia et al., 2023), which included the individual steps between the sub-skills of data analysis and evaluation. The *Statistical Literacy Process* (Koga, 2024) likewise did not include the sub-skill of data collection. In addition to the competence models discussed earlier by Lee et al. (2022), Wolff et al. (2016), and Wild and Pfannkuch (1999), the sub-skills of define question, plan study design, and collect/acquire data were also emphasized in the statistical problem-solving process from GAISE II (Bargagliotti et al., 2020) and the *IDSSP Framework* (IDSSP Curriculum Team, 2019).

Although the competence models mentioned vary in some aspects, they all address the same goal of using data to find well-founded solutions to real-world problems for which critical and statistical thinking are fundamental prerequisites (Gould, 2021). However, as mentioned at the outset and supported by academic research, the sub-skill of data collection is often neglected in theoretical models and receives little attention in the context of practical implementation in school. Because the basic skills of data literacy are expected to be taught in K–12 education (Robertson & Tisdall, 2020), this imbalance must be questioned in light of the goal of teaching a critical approach to data. Therefore, the following section takes a closer look at the relevance of data collection. Although some competency models include the selection of methodology and measuring instruments within this sub-skill, the term 'data collection' is hereafter understood as the act of gathering data by learners themselves, without necessarily incorporating the preceding or subsequent processes.

## 2.2 DATA SCIENCE EDUCATION WITH REAL WORLD DATASETS

The foundation for data literacy lies in working with data. The sources of these data are diverse, ranging from personally collected data through sensors, surveys, or observations, to existing data tables in school textbooks for mathematics instruction. Studies suggest that the origin and thematic focus of data can influence learners' motivation and work habits. For example, Ucar and Trundle (2011) highlighted that using archived data not collected by students was logistically practical and enabled effective teaching in the context of data-based decision-making in schools. However, working with authentic datasets provides learners with a deeper understanding of the dataset compared to the traditional approach, where archived, printed data is used (Ucar & Trundle, 2011).

Additionally, learners working with real datasets tend to learn more thoroughly and have a better grasp of scientific concepts, such as data-based decision-making, than those who do not have access to authentic data (De Luca & Lari, 2011). Cross-disciplinary reinforcement of computational skills and efficient preparation for everyday life and the workplace are two further aspects associated with project-based work using authentic datasets (Erwin, 2015). Moreover, the aspect of increased intrinsic

motivation among learners through work with real data should not be overlooked. Erwin (2015) highlighted three pillars: real data, real learning, and real data literacy. In this regard, open data—freely available datasets—allows access to real and authentic data for working with learners (Gould, 2017; Ridgway, 2016). Atenas et al. (2021) saw potential in working with open data, particularly in terms of inquiry-based learning and the strengthening of transdisciplinary skills, such as critical thinking, collaboration, and the use of digital media (Atenas et al., 2021). The origin of open datasets is variable and can be traced back to, for example, citizen science or participatory sensing projects, which focus on involving citizens in data collection (Gould, 2017). However, personal collection of data results in the datasets being in uncleaned form, which can pose challenges during non-trivial analysis (see, for example, Gould et al., 2017 and their study with teachers). Rubin (2021) also emphasized that already collected and provided datasets may contain hidden information that learners need to identify before analysis. The questions of who, when, how, where, and why can be used to check the validity of the data (Rubin, 2021). For this reason, Wolff et al. (2019) highlighted that personal data collection can lead to a more critical perspective on datasets. However, he emphasized that controlled studies were needed to understand the relationship between familiarity with data and the ability to question it critically (Wolff et al., 2019). Snee (1993) additionally emphasized that personal data collection enabled a connection between the learning process and reality. With regard to our data-driven world, the experience of personal data collection is therefore beneficial for answering context-related questions more reflectively based on the acquired experience (Teixeira et al., 2022). Additionally, mastering the sub-skill of data collection is essential for actively participating in public decision-making processes, for example, as part of a citizen science project (Engel et al., 2022; Teixeira et al., 2022).

Active participation in the data collection process can be supported by technology-assisted measuring instruments, which can offer an advantage in terms of collecting large amounts of data while also providing insight into the black box of the digital world (Ben-Zvi et al., 2018; Biehler et al., 2023). Additionally, the collection of personal data using technologies can provide a practical and reflective approach to the field of big data (Biehler et al., 2023). One possible implementation of personal data collection is the use of a senseBox, as demonstrated by Podworny et al. (2022) and Biehler et al. (2018) in their evaluated teaching units to promote data literacy. The senseBox is a do-it-yourself kit that allows for the assembly and programming of a modular weather station. It is based on a microcontroller to which various sensors, such as temperature, ultraviolet (UV) radiation, and humidity, can be connected. These data can then be collected and displayed on a screen or stored on a secure digital card (Pesch et al., 2022). Additionally, there is the option to upload the collected data to the openSenseMap, an open environmental data platform where weather stations can be registered to store and visualize collected data (Pfeil et al., 2015). As a technology, the senseBox thus offers the possibility to collect and store a large number of authentic datasets, making it a suitable tool for personal data collection with sensors by students (Witte et al., 2023).

## 2.3 EFFECTS OF VIRTUAL REALITY IN EDUCATIONAL SETTINGS

Limited time and increased effort may explain why the data collection sub-skill is not adequately emphasized in education. With the help of the latest technologies, such as virtual reality, challenges related to data collection can be addressed, and teaching can shift to otherwise inaccessible environments (Maresky et al., 2019). In particular, IVR—where learners feel part of the computer-generated 3D world through a head-mounted display and interact with it—offers potential for designing forward-thinking education (McGrath et al., 2018; Slater & Sanchez-Vives, 2016). As a result, space and time no longer need to be barriers to learning success, for example, during geographical excursions to hard-to-reach places (Shute et al., 2017). The limits of what is possible in the classroom are thus overcome, and scenarios that are difficult to replicate in reality, such as historical events or molecular processes, can be replicated using IVR (Fransson et al., 2020). Through the possibility of interaction and virtual participation, IVR can promote a deep understanding of certain processes in imaginary worlds (Slater & Sanchez-Vives, 2016). Especially in direct comparison between the learning experience in reality and in IVR, a shift from students assuming an observer role in traditional teaching-learning scenarios to their direct involvement in IVR can lead to better learning outcomes (Wu et al., 2020). These outcomes manifest in the form of additional knowledge and enhanced skill development among K–12 learners using a head-mounted display (Wu et al., 2020). Furthermore, IVR can counteract

loss of motivation and attention, contributing to increased interest and intrinsic motivation towards the subject matter (Melinda & Widjaja, 2022; Serrano-Ausejo & Mårell-Olsson, 2024).

However, when learners are confronted with novel technologies, the resulting motivation and fascination can also increase their cognitive load and, consequently, lower their learning outcomes—an effect known as the novelty effect (Miguel-Alonso et al., 2024). In addition, schools that implement IVR must commit resources, particularly for purchasing expensive hardware and for training the teachers who will use it (Stranger-Johannessen & Fjørtoft, 2021). The results of recent studies show that the use of IVR has a positive effect on learning outcomes in teaching and learning scenarios compared to less interactive methods (Hamilton et al., 2021; Papanastasiou et al., 2019; Villena-Taranilla et al., 2022; Wu et al., 2020). However, further research on the use of IVR in K–12 education is needed because current results do not allow for generalizations due to the possible influence of various factors in IVR studies (Liu et al., 2017). One of the most significant parameters in this context is subject matter, which has often been classified as extracurricular to date (Matovu et al., 2023). Therefore, an assessment of the effectiveness of IVR should be subject-specific. In the context of school education, it is particularly important to consider topics from the curriculum that have been previously neglected (Matovu et al., 2023). The current study builds on this point. At the present time, there is no known evidence that learners have used measuring devices within an IVR to collect data on occurrences in the virtual world. Therefore, a separate environment has been developed using the Unity Real-Time Development Platform (https://unity.com/), which accurately reflects the scenario of reality (see Section 4.2) (Unity Technologies, 2025). We used a ready-to-use virtual environment of a school from the Unity Asset Store and added virtual models of the senseBox to the Unity world. The movement and navigation on virtual school grounds to the senseBoxes was implemented in the GeoGami software, developed at the Institute for Geoinformatics (2024) of the University of Muenster (see www.geogami.org).

## 3. RESEARCH QUESTION

The previous discussions have underscored the importance of the sub-skill of data collection and the work with authentic datasets for a comprehensive promotion of data literacy. It was emphasized that more research is needed in the teaching of data literacy, particularly focusing on the relationship between familiarity with data and the ability to critically evaluate it (Song & Zhu, 2016; Wolff et al., 2019). Furthermore, the potential of IVR for enhancing deep understanding was noted. Based on these insights, the following research questions arose, focusing on the influence of personal data collection—in both real and virtual worlds—on the reflective and critical handling of data. Specifically, the focus is on the critical evaluation of a dataset after data collection has been conducted in different learning settings. The research questions are as follows.

1. To what extent does personal data collection by learners in the real world influence their critical evaluation of the dataset?
2. To what extent can the insights gained from data collection in real-world settings also be observed during data collection in IVR?

As previously mentioned, working with authentic and realistic datasets can positively impact learners' motivation. Because high motivation, in turn, has a positive effect on performance (Steinmayr & Spinath, 2009), it was hypothetically assumed that personal data collection by learners would positively influence the critical evaluation of a corrupted dataset in relation to question one. At the same time, learning settings in IVR can lead to a deeper understanding and higher motivation among learners compared to learning settings in less interactive scenarios. With regard to research question two, it was initially assumed that data collection through IVR would lead to a deeper understanding of data than if no personal data collection were to take place. However, during data collection in IVR, only the senses of sight and hearing can be engaged, whereas in reality, all five senses can be utilized by learners (Biocca, 1995). Therefore, the hypothesis related to the second research question was extended to suggest that the positive effects on data understanding would be higher in real-world data collection than in IVR-based data collection. The hypotheses set the stage for a study that examined these variables in depth, contributing valuable insights into how different modes of data collection might enhance critical data literacy in educational contexts.

# 4. RESEARCH METHOD

## 4.1. STUDY DESIGN

To address the research questions, a pedagogical intervention study employing a pretest and posttest design was conducted. The study followed a field-based, quasi-experimental research design involving two experimental groups and one control group. Experimental Group 1 (EG 1) conducted data collection on a school campus in the real world; Experimental Group 2 (EG 2) conducted data collection on a school campus in IVR; and the control group did not conduct any personal data collection. This study design allowed for testing the effectiveness of a pedagogical intervention, in this case, personal data collection under two different conditions. The aim was to pursue and evaluate a targeted change (McBride, 2016).

***Participants*** The study's sample consisted of students from three different high schools in Germany. The three schools have a good to very good school social index, indicating a homogeneous student body and low levels of child and youth poverty, a small proportion of young people who do not speak the national language, and a low number of students with special educational needs (Ministry for Schools and Education of the State of North Rhine-Westphalia, 2025). To prevent any cross-group influence between the two experimental groups and the control group, each group was assigned to a separate school located at a significant distance from the other participating schools. Potential differences in performance between the groups were identified using a pretest to provide a uniform baseline. Additionally, due to the standardized curriculum within the federal state, it was assumed that the 14- to 17-year-old students had a similar level of knowledge. Regarding the study's thematic focus, prior knowledge in European climate criteria and the temperate climate zone was expected (Ministry for Schools and Education of the State of North Rhine-Westphalia, 2019). The selected age group helped to ensure that students possessed basic mathematical competencies in statistical thinking and working, such as calculating relative frequencies, arithmetic means, and measures of spread (Ministry for Schools and Education of North Rhine-Westphalia, 2022), which were partly relevant for answering the items on the pretest and posttest.

A total of $n = 128$ students participated in the study, comprising 40.6% females, 54.7% males, 0.8% non-binary, and 3.9% who did not specify their gender. EG 1 consisted of $n = 43$ students, EG 2 of $n = 31$ students, and the control group of $n = 54$ students. The control group contributed to the internal validity of the intervention study and minimized the influence of confounding factors.

***Procedure*** The study was integrated into the everyday school curriculum within the subject of social studies, focusing on the topic of weather combined with the internationally required methodological competency of personal information gathering, including data collection (Bargagliotti et al., 2020; European Commission, 2024). Although both the control group and EG 1 participated in the study as a whole class, the implementation in EG 2 had to be carried out with one-on-one supervision due to the use of VR headsets. Nevertheless, the procedure was identical across groups. Initially, the topic of weather data was introduced by having students check the weather app on their smartphones and discuss where this data comes from and who collects it. The discussion led to the conclusion that the German Weather Service (DWD) collected and provided this data. Based on this, an image of a DWD weather station was shown to students, and the class discussed which weather data was collected under which regulations and conditions, and with which sensors. The focus was on environmental phenomena, such as temperature, UV radiation, precipitation, and air quality in the form of particulate matter. To be able to collect and publish data similarly to the DWD, the senseBox was introduced as a tool for personally collecting weather data (see Section 2.2). As an example of data collection using the senseBox, the fine dust sensor was presented, and a table of measured data on air fine dust pollution (particulate matter) at two different times from three different measuring stations was shown. The students then calculated the average for each of the three measuring stations. They decided which measurements appeared valid and which were skewed and needed to be discarded, for instance, due to the sensor's proximity to a sandbox. This short lesson provided an opportunity, on one hand, to ensure that students had sufficient content

knowledge regarding standards for collecting environmental data. On the other hand, the methodological skills for analyzing (corrupted) environmental data were taught.

Subsequently, the students' individual baseline conditions were assessed through an anonymous pretest in pen-and-paper format for all three groups (see Figure 2). This pretest provided initial insights into the critical evaluation of datasets during data analysis by the students. For example, data series on temperature measurements were provided for evaluation, which were corrupted due to the different positioning of the measurement devices in the sun and in the shade. These data series were excerpts from a comprehensive dataset. This data choice ensured a focus on the topic relevant to the study and helped to avoid overwhelming students with working on big data without much prior experience.

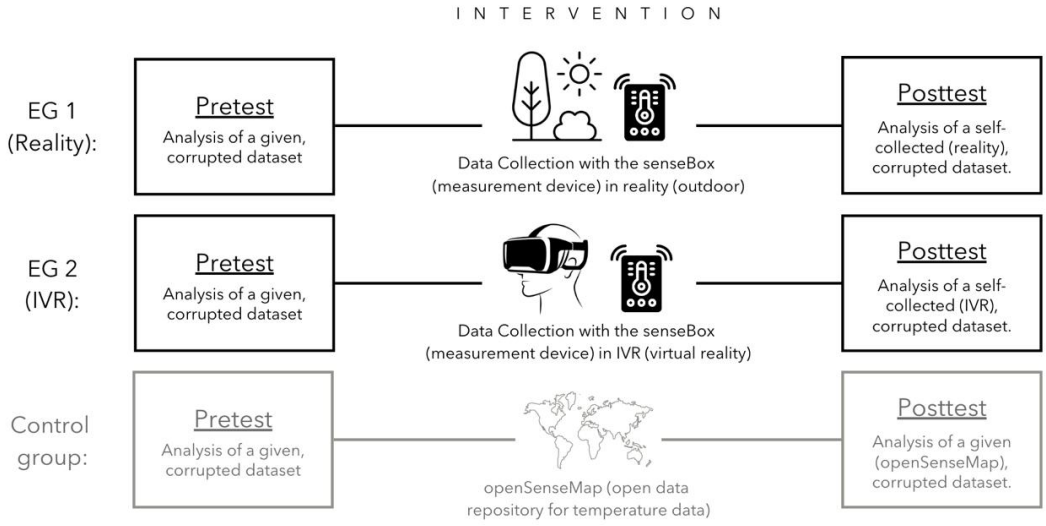
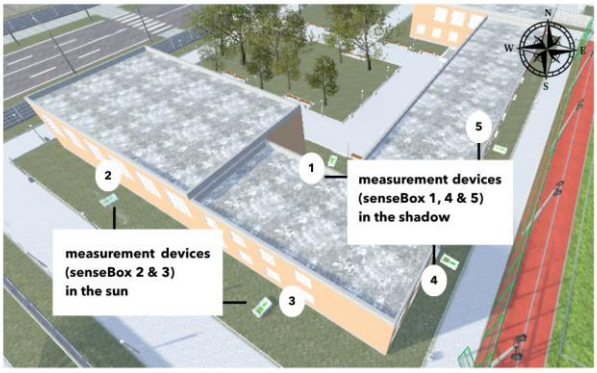



*Figure 2. Study design of the pedagogical intervention research*

For the experimental groups, the intervention followed one week after the pretest was conducted: personal data collection on the schoolyard (EG 1) or on a computer-generated schoolyard in an IVR environment using a head-mounted display (EG 2). In both scenarios, the teacher had previously positioned five senseBox measurement devices equipped with temperature sensors at five different locations in the schoolyard. The different locations of the sensors were deliberately chosen so that by the afternoon, two of the devices would be exposed to direct sunlight, introducing the confounding factor of sun exposure that led to corrupted temperature data. The remaining three sensors were positioned in the shade, thereby adhering to the guidelines for accurate temperature data collection and providing valid measurements.

During the intervention, the students read the temperature values from the displays of the senseBox measurement devices and recorded them in a pre-prepared table. In the IVR scenario, students navigated to the five locations, collecting data from virtual senseBox measurement devices. Like in the real-world condition, two sensors were placed in the direct sun, and three sensors were placed in the shade. Students verbally reported the readings and then recorded them in the table after removing the head-mounted display. The confounding factor of sunlight was thus actively experienced. In reality, the senseBox readings varied depending on whether they were exposed to sunlight, which could be both observed and felt. In the IVR, the position of the sun for the afternoon scenario was predetermined, and accordingly, the readings on the displays of the senseBoxes placed in the sun and shade were fixed. Although warmth on the skin could not be felt, the position of the sun and the effects of sunlight were still visually observable (see Figure 3).




Basis for data analysis & evaluation:
Measured temperature of students in the afternoon

| | Box 1 | Box 2 | Box 3 | Box 4 | Box 5 |
|---|---|---|---|---|---|
| 2.00 p.m. | 27.8 °C | 65.7 °C | 65.8 °C | 28.0 °C | 27.7 °C |
| 2.30 p.m. | 28.2 °C | 64.9 °C | 65.2 °C | 27.9 °C | 28.0 °C |
| 3.00 p.m. | 27.7 °C | 64.7 °C | 64.9 °C | 28.1 °C | 27.6 °C |

*Figure 3. Experimental set-up (Screenshot of IVR) and temperature values recorded in the afternoon*

To ensure that the amount of time spent on the study-related topic was identical across all three groups, the control group continued working with the provided datasets in the following week instead of participating in the intervention. The open data repository for environmental data, openSenseMap, was utilized for datasets (see Section 2.2). Here, five sensors were pre-registered at five different locations on the school grounds. Instead of personally collecting data and actively experiencing the confounding factor of sunlight, the control group used the temperature data provided on the openSenseMap. The provided temperature data was read from a line graph together with the class and entered into the prepared table under instruction so that no errors would occur. Additionally, the positioning of the five senseBoxes on the school ground, the timestamp of the collected data, and the cardinal direction could be identified on the openSenseMap. The data provided there had the same structure as the personally collected temperature data to avoid deviations in the study design. The data for two sensors resembled the measurements taken in the sun, and the data for the remaining three sensors resembled the data measured in the shade (see Figure 3; measurement devices 2 and 3 deliver much higher temperature values than measurement devices 1, 4, and 5 in the shade). In addition, by starting the data collection in reality (EG 1), it was possible to create almost identical conditions for the other two scenarios so that the datasets for the posttest were almost identical and would have no influence on the results.

The potential effect of the intervention was then evaluated by administering the posttest. Similar to the pretest, students once again analyzed the temperature data they had personally collected or read from the graphs on the openSenseMap. By calculating the average temperature and answering the corresponding questions, it was possible to assess whether the students had identified the confounding factor of sunlight and had conducted a reflective analysis of their data.

***Study Instrument*** The study instrument was divided into two versions following an identical structure: pretest and posttest. First, students provided personal information, such as gender and age, as well as an anonymous code to match the two tests. To introduce the tasks related to data analysis, a map of the school grounds, including cardinal directions and the positions of the five measurement devices, was presented. The map was embedded in a fictional scenario: classmates had set up these sensors and collected temperature data. The subsequent table contained the data collected by all five sensors at three different times. Students then answered the following question from the pretest, providing a calculation and reasoning, "What average temperature would the German Weather Service report for the morning?" (see Figure 4). The posttest contained the same question but referred to a period in the afternoon.

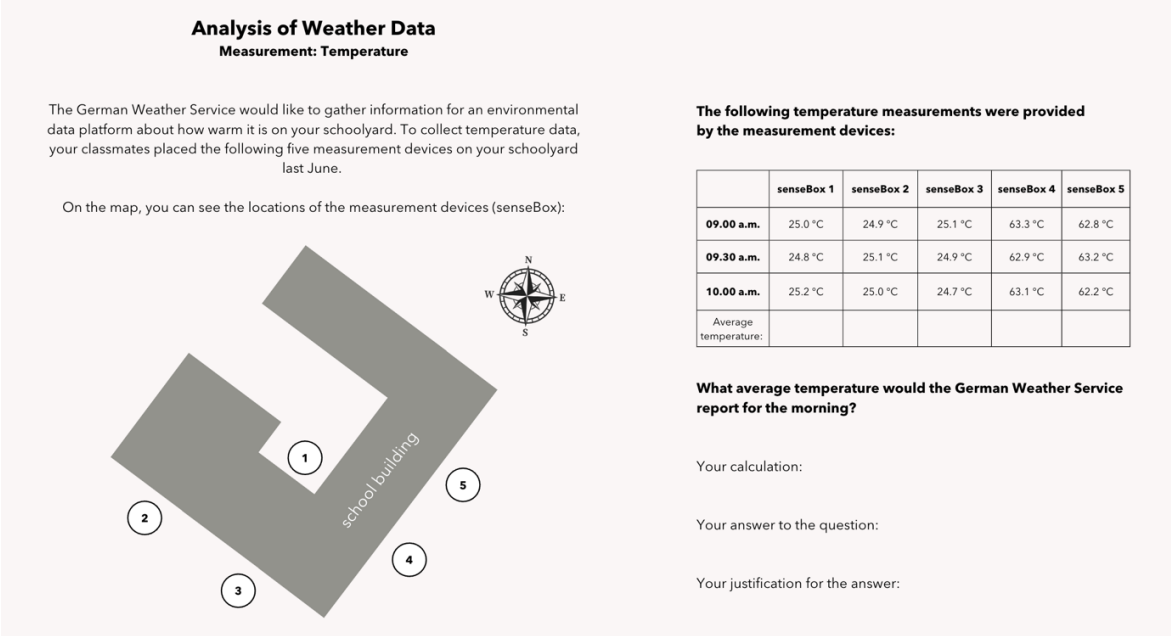



*Figure 4. Task from the pretest (translated from German). The map on the left explains the set-up of the measurement devices. The table shows the temperature measured for each device.*

The challenge in answering the pre- and posttest questions was identifying sunlight as a confounding factor, and consequently, cleaning the dataset by filtering out distorted temperature values (e.g., 63 °C in the morning). These tasks allowed for assumptions to be made about the learners' critical and reflective understanding of corrupted datasets. The table from the posttest, which referred to the afternoon temperature values, was initially incomplete and had to be completed by the students. They could use either the data they collected themselves (e.g., EG 1 and EG 2) or the data from openSenseMap (control group). After learners transferred their data into the table, the posttests became similar to the pretest. The procedure described was discussed in the plenary session in order to avoid errors that could bias the results. Additionally, an information sheet with an explanation and an example on calculating an average was provided with the test so that a lack of mathematical skills would not disadvantage the students in completing the test tasks.

To provide evidence that the task design could yield data sufficient for answering the research questions, a pilot study of the pre- and posttest was conducted with $n = 21$ students under the same conditions as described in the study. Due to their age, school type, and background, they shared similar social characteristics as the students in the main study. The pilot study served two purposes. First, the pilot study verified the stability of the measurement results. Test-retest reliability was calculated using Pearson's correlation coefficient. The results of the two test points were significantly correlated, $r = 0.94$, $p < .001$, indicating a high stability of the measurement results over time. Second, the pilot study allowed for an evaluation and validation of students' understanding of the task instructions. It is important to emphasize that the pre- and posttests represent an exemplary task designed to answer the stated research questions for the target group under consideration, aiming to identify initial trends and findings. The experiment was reviewed by an ethics committee of the respective university, and the learners, as well as their legal guardians, gave their informed consent to participate.

### 4.2. STATISTICAL APPROACH

To evaluate the pretests and posttests, a scoring system ranging from zero to two points was applied. The following coding was used for the distribution of points. If the corrupted data series were identified and excluded from the calculation of the result, one point was awarded. Another point could be earned if a justification for excluding the data series was provided. To be considered correct, this justification had to include at least one of the following words: sun, shade, unrealistic, unusually high, or atypical. Therefore, two points could be achieved if both aforementioned criteria were met. Zero points were awarded if neither criterion was met. Scoring on a scale from zero to two points allowed for an assessment of the extent of the learners' reflective understanding of the data (see Table 1).

*Table 1. Categorization and rating of answers for the pretest and posttest*

| Category | Definition | Example |
|---|---|---|
| 1) No critical consideration of the corrupted dataset | The calculation to provide the average temperature for the morning is incorrect, meaning the corrupted data were not identified/filtered out. Additionally, there is no justification, or reasoning about the result is from a mathematical perspective.<br>Score: 0 points | Result: 42.8 degrees Celsius<br>*"I added the averages of all stations and then divided by the number of stations."* |
| 2) Partial critical consideration of the corrupted dataset | 2a) The calculation to provide the average temperature for the morning is correct, meaning only valid data were included, and falsified data were identified/ filtered out. However, there is no justification, or reasoning about the result is from a mathematical perspective.<br>Score: 1 point | Result: 27.9 degrees Celsius<br>*"27.9 degrees Celsius is the result of calculating the average."* |
| | 2b) The calculation to provide the average temperature for the morning is incorrect, meaning the corrupted data were not identified/filtered out. However, there is justification for the result from a contextual or thematic perspective.<br>Score: 1 point | Result: 42.8 degrees Celsius<br>*"The values from stations two and three are unusually high (perhaps they were in the sun). Therefore, I filtered out this data."* |
| 3) Critical consideration of the corrupted dataset | The calculation to provide the average temperature for the morning is correct, meaning only valid data were included, and falsified data were identified/filtered out. Additionally, there is justification for the result from a contextual or thematic perspective.<br>Score: 2 points | Result: 27.9 degrees Celsius<br>*"I filtered out two datasets (S2 and S3). They are far too high for our region, and temperature is usually measured in the shade."* |

For the analysis of these data, a frequency analysis was first conducted to obtain information about the sample. Descriptive statistics and bootstrapping with 1000 repetitions were then performed to provide robust estimates of the means and confidence intervals. Because the data were not normally distributed, the Mann-Whitney *U* test served as the basis for calculating rank means and *p*-values. All analyses focused on the difference between the pre- and posttest scores, taking the initial conditions into account. Consequently, the difference between pre- and posttest scores calculated for each student was considered for the Mann-Whitney *U* test. Additionally, a multiple test correction was applied by adjusting the confidence interval to 97.5% and *p* to be less than .025. A key interest was the relationship between the intervention and the difference between the pre- and posttest, particularly in comparing differences between (a) EG 1 (Reality) and the control group, (b) EG 2 (IVR) and the control group,

and (c) EG 1 (Reality) and EG 2 (IVR), which is presented in the following section. The analyses were conducted using IBM SPSS Statistics Software (version 29.0.2.0 for macOS).

## 5. RESULTS

The evaluation of results from the pre- and posttests provides insight into both the effect of personal data collection in different learning settings on data analysis and the students' baseline understanding of critical reflection on datasets during the data analysis process. The latter can be assessed after analyzing the pretest results. Out of a total of $n = 128$ pretests, two were invalid. Among the remaining $n = 126$ pretests, 86.5% of the responses received zero points, 5.6% received one point, and 7.9% received two points. Small differences were observed between the initial conditions of the two experimental groups and the control group. EG 2 (IVR) had the strongest initial conditions, with seven out of $n = 31$ students (22.6%) scoring two points on the pretest. The control group followed with three students (6%), each scoring two points (see Figure 5). The low scores in the pretest underscore the need for changes in the teaching of data literacy and called for a closer examination of the intervention's effect. An overview of the results is found in Figure 5, which will be discussed in more detail below.

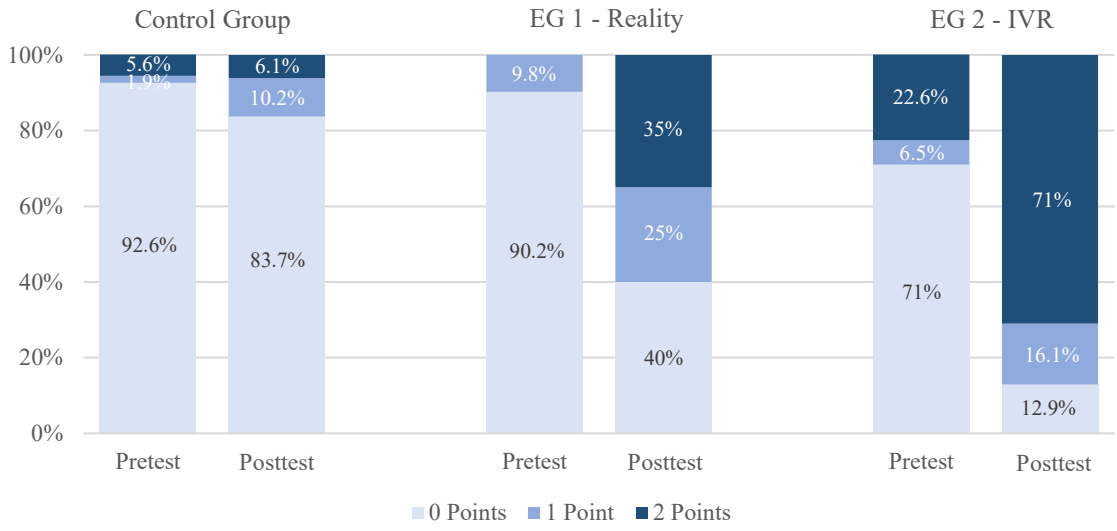



*Figure 5. Percentage of students who achieved zero, one, or two points for the pretest and posttest*

### 5.1. DATA COLLECTION IN REALITY VERSUS NO DATA COLLECTION

A comparison between the pretests and posttests for EG 1 (Reality) and the control group demonstrated how personal data collection on the school grounds can lead to a more critical and reflective understanding of data. With $n = 6$ invalid posttests, the sample size available for analysis was $n = 89$ students. Of these, $n = 49$ belonged to the control group, and $n = 40$ belonged to EG 1 (Reality). EG 1 (Reality) achieved a mean score of $M = 0.10$ for the pretest and improved to $M = 0.95$ for the posttest. This corresponded to an increase in means of 0.85, meaning that participants scored on average 0.85 points higher in the posttest compared to the pretest. The 97.5 % confidence interval of [0.55, 1.16] supports the conclusion that the average increase is between 0.55 and 1.16 points with high confidence. In contrast, the control group showed a smaller increase, moving from $M = 0.14$ for the pretest to $M = 0.22$ for the posttest. Thus, the increase in means amounted to 0.08 with a 97.5% confidence interval of [-0.02, 0.22], indicating that no meaningful improvement occurred between the pretest and the posttest. This difference in development between the pretest and posttest was further supported by the Mann-Whitney $U$ test. Although the control group had a mean rank of $M_{Rank} = 34.74$, EG 1 showed a mean rank of $M_{Rank} = 57.56$ with $U = 477.5$, $Z = 5.060$, $p < .001$. The mean difference of MD = 0.77 and the corresponding $p$-value indicated an improvement of students in EG 1 compared to the control group,

highlighting the effectiveness of the intervention. Moreover, the mean ranks in combination with the effect size confirmed the stronger average improvement of EG 1 compared to EG 2.

Figure 6 shows the development of posttest scores compared to the pretest scores for each group. A decline in performance occurred in only 2% of cases in the control group, whereas the majority (89.8%) showed no change. In contrast, in EG 1 (Reality), 30.0% of students gained one point and 27.5% gained two points.

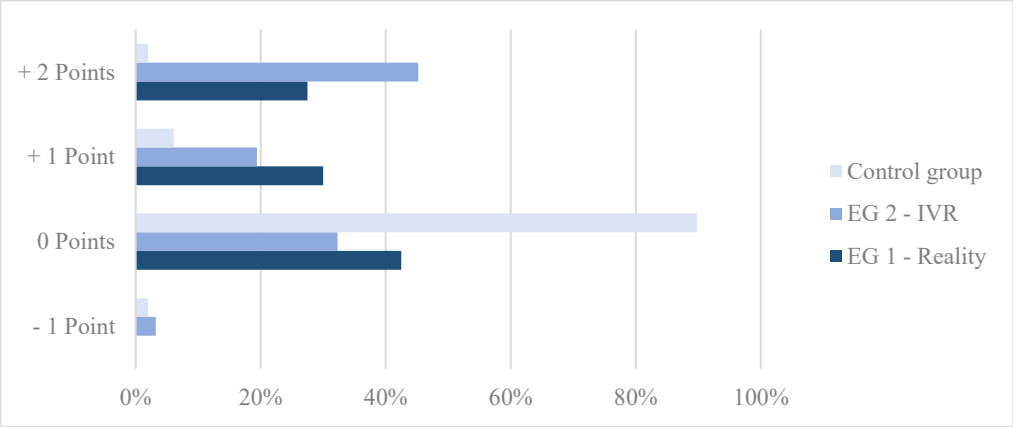



*Figure 6. Score changes from pretest to posttest in EG 1, EG 2, and the control group*

A qualitative analysis of the students' responses shows that students in the control group predominantly justified their answers from a mathematical perspective: "I added all the stations and divided by five." They thus described their calculation process for determining the average without embedding it in a broader thematic context (see the first category in Table 1). In the posttest, 84% of the control group and 40% of the EG 1 scored zero points. When students identified the sun as a confounding factor, their explanations were thematic: "The values from the stations whose temperatures were affected by the sun are not considered." Alternatively, they pointed out that the temperatures for their region were unrealistically high, "The temperatures are unusually high and don't make sense. Therefore, I discarded the two data series with unrealistic values." This type of content-based reasoning, combined with correct calculations, was found in responses from 6% of the control group and 35% of EG 1 from the posttest (see the third category in Table 1). Ten percent of the students in the control group and 25% in EG 1 provided a thematic justification for the posttest without adjusting their calculations accordingly, or they gave a correct calculation without sufficient reasoning (see the second category in Table 1).

## 5.2. DATA COLLECTION IN IVR VERSUS NO DATA COLLECTION

In the analysis of the posttest results, considering EG 2 (IVR) and the control group, a total of $n = 80$ students were included after accounting for $n = 5$ missing or invalid tests. Among these, $n = 49$ students were from the control group and $n = 31$ from EG 2 (IVR). In the control group, an increase of one or two points compared to the pretest was observed in four tests (8.2%). Although the mean difference between the pre- and posttest was 0.08 points, students from EG 2 (IVR) achieved an increase of 1.06 points with a 97.5% confidence interval of [0.67, 1.46], indicating a reliable improvement in performance. An increase of one or two points was achieved by 64.5% of the students in EG 2 (see Figure 6). The difference between pre- and posttest was further analyzed using the Mann-Whitney $U$ test. A comparison of the two groups showed a lower mean rank for the control group ($M_{Rank} = 31.74$) and a higher mean rank for EG 2 (IVR) ($M_{Rank} = 54.34$) with $U = 1188.5$, $Z = 5.121$, $p < .001$. The higher average increase in points for EG 2 is evident from the difference between the mean ranks of both groups. In combination with the effect size and the p-value, the results indicate an effect of the

intervention in EG 2. A qualitative look at the students' justifications and calculations once again highlights two contrasting perspectives. Figure 7 shows the calculation and mathematical justification for including all datasets. This approach was awarded zero points and was observed in 13% of the posttests from EG 2. In contrast, Figure 8 shows a calculation using only valid datasets, accompanied by a justification addressing the influence of sun and shade positions on the data. This approach received two points and was achieved by 71% of EG 2 for the posttest. The same number of students scored zero points in the pretest, resulting in an increase in the number of students from EG 2 who provided a content-based contextualization of the topic after the intervention in the IVR.

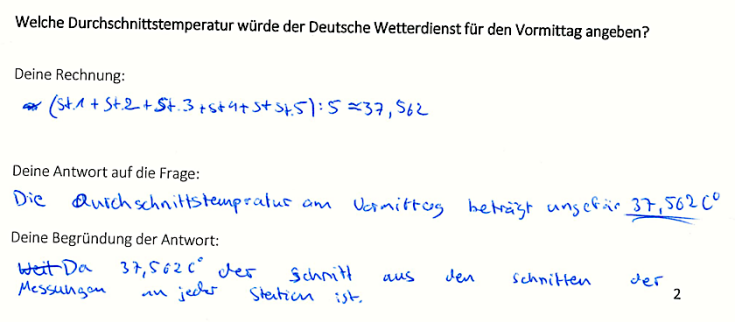



*Figure 7. Zero points in pretest*
*(mathematical explanation, "37.562 degrees Celsius is the result, as it is the average of the average of the measurements from each station.")*

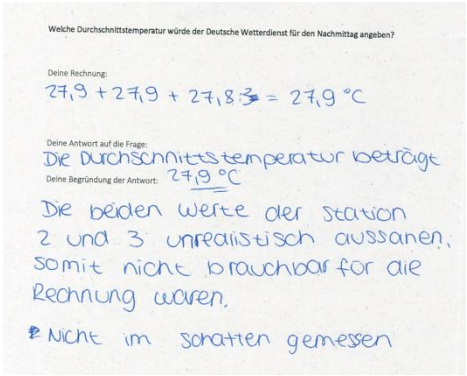



*Figure 8. Two points in posttest*
*(thematic explanation, "Since the values from stations 2 and 3 appeared unrealistic (not measured in the shade), they were not suitable for the calculation.")*

### 5.3. DATA COLLECTION IN REALITY VERSUS DATA COLLECTION IN IVR

The comparison of posttest results between EG 1 (Reality) and EG 2 (IVR) involved a total of $n = 71$ students. In EG 1 (Reality), $n = 23$ out of $n = 40$ students (57.5%) improved by one or two points on the posttest. In EG 2 (IVR), $n = 20$ out of $n = 31$ students (64.4%) showed similar improvements. Notably, EG 2 (IVR) exhibited a higher rate of two-point improvement (EG 1: 27.5%; EG 2: 45.2%), whereas EG 1 (Reality) predominantly saw a one-point improvement (EG 1: 30%; EG 2: 19.4%) (see Figure 6). Thus, the intervention contributed to a better performance on the posttest and potentially encouraged a more critical view of data, whether in IVR or in real-life situations. With regard to the differences in development between pre- and posttests in both experimental groups, no meaningful effects could be observed. The Mann-Whitney $U$ test showed similar mean ranks, with $M_{Rank} = 33.85$ (EG 1) and $M_{Rank} = 38.77$ (EG 2) with $U = 706$, $Z = 1.060$, $p = .289$. The mean ranks indicate similar average scores for both groups and, together with the low effect size, also suggest that there is no meaningful difference in the development of EG 1 and EG 2 based on the intervention. Table 2 presents

a summary of the average development of the scores achieved by each group between the pre- and posttest.

*Table 2. Score differences between pretest and posttest: Mean, standard error, and confidence interval for EG 1, EG 2, and control group*

| | Mean | SE | 97.5% CI |
|---|---|---|---|
| Control Group | .08 | .06 | [-0.02, 0.22] |
| EG 1 (Reality) | .85 | .14 | [0.55, 1.16] |
| EG 2 (IVR) | 1.06 | .17 | [0.67, 1.46] |

## 6. DISCUSSION AND CONCLUSION

### 6.1. SUMMARY OF KEY FINDINGS

The results of the pretest revealed that a large portion of learners who did not experience personal data collection were unable to critically interpret the provided datasets or recognize outliers in corrupted datasets. With only 13.6% of all students achieving one or two points for the pretest, significant knowledge gaps in reflective data understanding during data analysis were evident. However, personal data collection by students proved to be an effective measure for raising learners' awareness of a more critical view of temperature data. The posttest results for EG 1 (Reality) showed that personal data collection improved the ability to critically interpret datasets and recognize confounding factors. After the intervention, more than half of the students in EG 1 improved their scores by one or two points, gaining a more critical and reflective perspective on the dataset affected by confounding factors. In contrast, only 8% of students in the control group improved their posttest scores. The mean difference, as well as the calculated *p*-value for the group comparison, indicated an improvement in the students in EG 1 compared to the control group, highlighting the effectiveness of the intervention. This result aligns with research on the use of open and authentic datasets, which can pose difficulties in analysis due to their unprocessed form, such as the presence of outliers in the dataset (Gould et al., 2017). The challenge of working with corrupted datasets is supported by insights gained from the pretests. Additionally, Rubin (2021) emphasized that learners needed to check the origin and validity of all datasets and identify potential confounding factors before analysis. The latter was easier for students in EG 1 compared to the control group, allowing them to critically assess, clean, and attribute outliers to identified confounding factors. The results for EG 1 confirm the hypothesis of Wolff et al. (2019) that there is a relationship between familiarity with data, e.g., through knowledge of the conditions of data collection or through personal data collection, and the ability to critically question it. Thus, the hypothesis posed in Section 3 regarding the first research question—that personal data collection by learners can lead to a deeper understanding of data and a more critical view of datasets—was validated.

The activities of EG 2 (IVR) allowed us to assess whether data collection in IVR can produce similar effects to data collection in reality. The results suggest that two-thirds of EG 2 students showed an improved from the pretest to the posttest. Notably, nearly half of the students achieved an increase of two points. In connection with the presented results, it suggests that personal data collection by learners within the IVR facilitates a more reflective and comprehensive understanding of the dataset during the data analysis process compared to conditions where learners do not engage in personal data collection. Because IVR allows for realistic simulation of learning scenarios, this assumption was supported by the results of the intervention in the real environment (EG 1). McGrath et al. (2018) highlighted the potential of a computer-generated and interactive 3D world, which was also recognized in this pedagogical intervention research. Additionally, the results can be explained by the fact that IVR can contribute to increased interest and intrinsic motivation towards the learning subject (Melinda & Widjaja, 2022; Serrano-Ausejo & Mårell-Olsson, 2024), which can, in turn, have a positive impact on performance on the posttest (Steinmayr & Spinath, 2009). The increased motivation and excitement of

using a novel technology for the first time may also explain the comparative results of the two experimental groups. The difference in posttest results is slight. In terms of score increase, EG 2 (IVR) had a small advantage of $M = 0.21$ points over EG 1 (Reality). The $p$-value and the confidence interval also indicated that the difference between the two experimental groups may not necessarily be attributable to the intervention. Similar results from both experimental groups can be explained by the fact that interaction with the learning subject occurred in both cases. IVR has an advantage over reality when direct involvement in IVR contrasts with a passive observer role in reality (Wu et al., 2020). Because a passive observer role of the students is not the case and interaction with the learning subject occurs in both scenarios, the similar results from the posttest can be accounted for. The hypothesis mentioned in Section 3, that the number of senses engaged can affect the impact of the intervention, cannot be confirmed. With regard to the second research question, however, the hypothesis can be verified in such a way that the effects of the intervention in the real world can be transferred to the virtual world. Consequently, personal data collection in IVR may provide an alternative means of fostering a reflective understanding of data.

In terms of promoting data literacy in schools, it can be concluded that learners do not uniformly apply a critical perspective to corrupted datasets. With the increasing amount of open data, citizen science projects, and fake news, the challenge of taking a reflective view of data must be addressed. Pedagogical intervention research has shown that personal data collection by learners can be a means to foster critical and reflective handling of non-trivial or corrupted datasets. Whether the data collection took place in the learners' real environment or in IVR with a head-mounted display was not relevant, as it allowed for the transfer of these results into school practice with two viable options. These findings underscore the importance of the sub-skill of data collection within the overall construct of data literacy and provide a recommendation for greater consideration of this sub-skill in the future design of teaching materials in K–12 education.

### 6.2. LIMITATIONS

The study has contributed valuable information for the teaching of data literacy in education. However, there are also several limitations to note. Some limitations pertain to the study design. Pedagogical intervention research is characterized by being embedded within the students' school environment, which can lead to differing learning settings between the two experimental groups (Kraft, 2020). Due to the randomization at the classroom level within the school context, clustering effects within the data cannot be ruled out. In future studies, students within classes should therefore be randomly assigned to groups, or the approach applied here should be taken into account in the data analysis. For instance, the implementation of the intervention in the real world (EG 1) was dependent on the weather conditions of the data collection day. Selected days in June allowed for implementation on days with high sunlight and warm temperatures, making the confounding factor noticeable for students. Nevertheless, the extent of sunlight was uncontrollable and varied slightly with each implementation.

In contrast, this factor could be controlled during the intervention in IVR (EG 2) and in the control group. The use of the head-mounted display in EG 2, however, resulted in the study being conducted in a 1:1 setting rather than in a classroom environment, which could lead to more thorough test handling and consequently better test results. Also, the learners who had already completed the pretest with one point or more did not have as much potential for improvement as students whose pretest scores were zero points. The use of a single task with a three-level evaluation limits the significance of the results but can serve as an initial indication for further investigations (see section 6.3). Additionally, solving mathematical tasks sometimes lacked real-world considerations, and task solutions might have been superficial (Wisenöcker et al., 2024). Although the tasks were not of a mathematical origin, calculations were necessary for solving them. It is possible that these mathematical operations became the focus for learners, thereby leading to less attention being paid to contextualizing the data and considering additional datasets.

### 6.3. FURTHER STUDY

The identified limitations provide a basis for further studies. One avenue is to examine and compare the motivational effects experienced by both experimental groups to determine if this factor influenced the results. Ensuring external validity and replicating the findings would benefit from including a broader randomized target group. Moreover, the test task developed in this process could be further refined into a comprehensive testing instrument. For further investigations, test development following the *Standards for Educational and Psychological Testing* is recommended (American Educational Research Association et al., 2014) to ensure that the validity of interpretations about test results and fairness are assessed and considered at all levels. Additional tasks should be included in order to gain a more differentiated view of student performance beyond the evaluation of this single task. A test like this would facilitate an investigation into the impact of data collection on other sub-skills of data literacy, with particular emphasis on the skills learners require to independently conduct data collection, including the design of the methodology and the selection of appropriate instruments.

Individual interviews with the learners could provide further insights into knowledge gains that were not apparent in the task used here. Additionally, there are thematic connections to the results that warrant exploration. For instance, future research could investigate other data categories to separate the findings from the specific context of temperature data. The latter could involve examining environmental data (e.g., precipitation), traffic data, or personal data, as well as exploring the transferability of skills from analyzing one type of dataset to another. Moreover, the insights gained about the importance of personal data collection offer potential for enhancing data literacy among students. Evaluated teaching units designed to strengthen the sub-skill of data collection could encourage teachers to incorporate the sub-skill into school curricula more thoroughly. That represents just one of many possible options for transferring research findings on data literacy into practical applications in education, thereby supporting the effective teaching of data literacy.

## ACKNOWLEDGEMENTS

We thank the three schools that cooperated with us for the study. In particular, we are grateful to the students who participated in the study. We also thank the reviewers and editors for their constructive and valuable feedback, which has helped to improve the quality of this article. The data collected in the study is available online: https://osf.io/a73gq/?view_only=db41f2a411044558ae01b1830b1579df

## REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education*, *29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

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; National Council of Teachers of Mathematics.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–15). Springer. https://doi.org/10.1007/1-4020-2278-6_1

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473–502). Springer. https://doi.org/10.1007/978-3-319-66195-7_16

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. 547–583). Springer. 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. *Archives of Data Science, Series A, 5*(1). https://doi.org/10.5445/KSP/1000087327/28

Biocca, F., & Levy, M. R. (Eds.). (1995). *Communication in the age of virtual reality*. Lawrence Erlbaum Associates.

Burrill, G., & Pfannkuch, M. (2024). Emerging trends in statistics education. *ZDM Mathematics Education*, *56*(1), 19–29. https://doi.org/10.1007/s11858-023-01501-7

Büscher, C. (2022). Design principles for developing statistical literacy in middle schools. *Statistics Education Research Journal*, *21*(1), Article 8. https://doi.org/10.52041/serj.v21i1.80

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1481

De Luca, V., & Lari, N. (2011). The $GRID_C$ 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

De Oliveira Souza, L., Espasandin Lopes, C., & Fitzallen, N. (2020). Creative insubordination in statistics teaching: Possibilities to go beyond statistical literacy. *Statistics Education Research Journal*, *19*(1), 73–91. https://doi.org/10.52041/serj.v19i1.120

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213

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. https://doi.org/10.1007/978-3-031-20748-8_18

Erwin, R. W. (2015). Data literacy: Real-world learning through problem-solving with datasets. *American Secondary Education*, *43*(2), 18–26. http://www.jstor.org/stable/43694208

European Commission. (2024). *ESCO*. https://esco.ec.europa.eu/en

Fadel, C., Bialik, M., Trilling, B., & Schleicher, A. (2015). *Four-dimensional education: The competencies learners need to succeed*. Center for Curriculum Redesign.

Fransson, G., Holmberg, J., & Westelius, C. (2020). The challenges of using head mounted virtual reality in K–12 schools from a teacher perspective. *Education and Information Technologies*, *25*(4), 3383–3404. https://doi.org/10.1007/s10639-020-10119-1

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), Article 58. https://doi.org/10.1186/s40594-024-00517-z

Frischemeier, D. (2020). Building statisticians at an early age—Statistical projects exploring meaningful data in primary school. *Statistics Education Research Journal*, *19*(1), 39–56. https://doi.org/10.52041/serj.v19i1.118

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

German Informatics Society. (2024, March 25). *TrainDL Interactive Policy Monitor*. https://traindl-policymonitor.ocg.at/

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics*, *43*(S1), 11–22. https://doi.org/10.1111/test.12267

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

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & 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)*. ISI/IASE. https://iase-web.org/documents/papers/rt2016/Gould.pdf

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, *8*(1), 1–32. https://doi.org/10.1007/s40692-020-00169-2

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

Institute of Geoinformatics. (2024). *GeoGami App* (Version 5.1.1) [Mobile app]. https://geogami.ifgi.de/app_en.html

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kurnia, A. B., Lowrie, T., & Patahuddin, S. M. (2023). The development of high school students' statistical literacy across grade level. *Mathematics Education Research Journal*, *36*(S1), 7–35. https://doi.org/10.1007/s13394-023-00449-x

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

Liu, D., Bhagat, K. K., Gao, Y., Chang, T.-W., & Huang, R. (2017). The potentials and trends of virtual reality in education. In D. Liu, C. Dede, R. Huang, & J. Richards (Eds.), *Virtual, augmented, and mixed realities in education* (pp. 105–130). Springer. https://doi.org/10.1007/978-981-10-5490-7_7

Maresky, H. S., Oikonomou, A., Ali, I., Ditkofsky, N., Pakkal, M., & Ballyk, B. (2019). Virtual reality and cardiac anatomy: Exploring immersive three-dimensional cardiac imaging, a pilot study in undergraduate medical anatomy education. *Clinical Anatomy*, *32*(2), 238–243. https://doi.org/10.1002/ca.23292

Matovu, H., Ungu, D. A. K., Won, M., Tsai, C.-C., Treagust, D. F., Mocerino, M., & Tasker, R. (2023). Immersive virtual reality for science learning: Design, implementation, and evaluation. *Studies in Science Education*, *59*(2), 205–244. https://doi.org/10.1080/03057267.2022.2082680

Matthews, P. (2016). Data literacy conceptions, community capabilities. *The Journal of Community Informatics*, *12*(3), 47–56. https://doi.org/10.15353/joci.v12i3.3277

McBride, N. (2016). *Intervention research*. Springer. https://doi.org/10.1007/978-981-10-1011-8

McGrath, J. L., Taekman, J. M., Dev, P., Danforth, D. R., Mohan, D., Kman, N., Crichlow, A., & Bond, W. F. (2018). Using virtual reality simulation Environments to assess competence for emergency medicine learners. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *25*(2), 186–195. https://doi.org/10.1111/acem.13308

Melinda, V., & Widjaja, A. E. (2022). Virtual reality applications in education. *International Transactions in Educational Technology*, *1*(1), 68–72. https://doi.org/10.33050/itee.v1i1.194

Miguel-Alonso, I., Checa, D., Guillen-Sanz, H., & Bustillo, A. (2024). Evaluation of the novelty effect in immersive virtual reality learning experiences. *Virtual Reality*, *28*(1), Article 27. https://doi.org/10.1007/s10055-023-00926-5

Ministry for Schools and Education of the State of North Rhine-Westphalia. (2022). *Kernlehrplan für die Sekundarstufe 1, Gesamtschule/Sekundarschule in Nordrhein-Westfalen. Mathematik.* [Core curriculum for secondary level 1, comprehensive school/secondary school in North Rhine-Westphalia. Mathematics].

Ministry for Schools and Education of the State of North Rhine-Westphalia. (2019). *Kernlehrplan für die Sekundarstufe I, Gymnasium in Nordrhein-Westfalen. Erdkunde.* [Core curriculum for secondary level 1, comprehensive school/secondary school in North Rhine-Westphalia. Geography].

Ministry for Schools and Education of the State of North Rhine-Westphalia. (2025). *School social index. The school-specific social index.* https://www.schulministerium.nrw/schulsozialindex.

Papanastasiou, G., Drigas, A., Skianis, C., Lytras, M., & Papanastasiou, E. (2019). Virtual and augmented reality effects on K–12, higher and tertiary education students' twenty-first century skills. *Virtual Reality*, *23*(4), 425–436. https://doi.org/10.1007/s10055-018-0363-2

Pesch, M., Bartoschek, T., & Schwering, A. (2022). *Presenting showcases for "senseBox and openSenseMap" as a learning suite for computer-, data- and scientific literacy*. ISLS Annual meeting, Hiroshima, Japan.

Pfeil, M., Bartoschek, T., & Wirwahn, J. A. (2015). Opensensemap—A citizen science platform for publishing and exploring sensor data as open data. *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*, *15*(39), 122–139. https://doi.org/10.7275/R56971SW

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), Article 6. 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., & Nicholson, J. (2010). Pupils reasoning with information and misinformation. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics.* ISI/IASE. https://iase-web.org/documents/papers/icots8/ICOTS8_9A3_RIDGWAY.pdf

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]. Dalhousie University.

Robertson, J., & Tisdall, E. K. M. (2020). The importance of consulting children and young people about data literacy. *Journal of Media Literacy Education*, *12*(3), 58–74. https://doi.org/10.23860/JMLE-2020-12-3-6

Rubin, A. (2021). What to consider when we consider data. *Teaching Statistics*, *43*(S1), S23–S33. https://doi.org/10.1111/test.12275

Schield, M. (1999). Statistical literacy: Thinking critically about statistics. *Of Significance. 1*(1), 15–21.

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. (2022). Data and AI literacy for everyone. *Statistical Journal of the IAOS*, *38*(2), 477–490. https://doi.org/10.3233/SJI-220941

Serrano-Ausejo, E., & Mårell-Olsson, E. (2024). Opportunities and challenges of using immersive technologies to support students' spatial ability and 21st-century skills in K–12 education. *Education and Information Technologies*, *29*(5), 5571–5597. https://doi.org/10.1007/s10639-023-11981-5

Shute, V., Rahimi, S., & Emihovich, B. (2017). Assessment for learning in immersive environments. In D. Liu, C. Dede, R. Huang, & J. Richards (Eds.), *Virtual, augmented, and mixed realities in education* (pp. 71–87). Springer. https://doi.org/10.1007/978-981-10-5490-7_5

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00074

Snee, R. D. (1993). What's missing in statistical education? *The American Statistician*, *47*(2), 149–154. https://doi.org/10.1080/00031305.1993.10475964

Song, I.-Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*(1), 80–90. https://doi.org/10.1016/j.lindif.2008.05.004

Stranger-Johannessen, E., & Fjørtoft, S. O. (2021). Implementing virtual reality in K–12 classrooms: Lessons learned from early adopters. In V. L. Uskov, R. J. Howlett, & L. C. Jain (Eds.), *Smart education and e-Learning 2021* (Vol. 240, pp. 139–148). Springer. https://doi.org/10.1007/978-981-16-2834-4_12

Teixeira, S., Campos, P., & Trostianitser, A. (2022). Data sets: Examples and access for civic statistics. In J. Ridgway (Ed.), *Statistics for empowerment and social engagement* (pp. 127–151). Springer. https://doi.org/10.1007/978-3-031-20748-8_6

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

Unity Technologies. (2025). *Unity Engine* [Computer software] (Version 6.0). https://unity.com/products?c=unity+engine

Van Audenhove, L., Vermeire, L., Van Den Broeck, W., & Demeulenaere, A. (2024). Data literacy in the new EU DigComp 2.2 framework how DigComp defines competences on artificial intelligence, internet of things and data. *Information and Learning Sciences*, *125*(5/6), 406–436. https://doi.org/10.1108/ILS-06-2023-0072

Villena-Taranilla, R., Tirado-Olivares, S., Cózar-Gutiérrez, R., & González-Calero, J. A. (2022). Effects of virtual reality on learning outcomes in K-6 education: A meta-analysis. *Educational Research Review*, *35*, Article 100434. https://doi.org/10.1016/j.edurev.2022.100434

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2—The digital competence framework for citizens: With new examples of knowledge, skills and attitudes* (EUR, Issue JRC128415). Publications Office of the European Union. https://doi.org/10.2760/115376

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

Wisenöcker, A. S., Binder, S., Holzer, M., Valentic, A., Wally, C., & Große, C. S. (2024). Mathematical problems in and out of school: The impact of considering mathematical operations and reality on real-life solutions. *European Journal of Psychology of Education*, *39*(2), 767–783. https://doi.org/10.1007/s10212-023-00718-0

Witte, V., Schwering, A., Bartoschek, T., & Pesch, M. (2023). Zukunftsweisender MINT-Unterricht mit dem senseBox-Ökosystem. Die Plattform für partizipative Data Science mit Physical Computing. [Future-oriented STEM education with the senseBox ecosystem. The platform for participatory data science with physical computing]. *MNU-Journal*, *76*(4), 296–301.

Witte, V., Schwering, A., & Frischemeier, D. (2024). Strengthening data literacy in K-12 education: A scoping review. *Education Sciences*, *15*(1). 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. https://doi.org/10.15353/joci.v12i3.3275

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

Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology*, *51*(6), 1991–2005. https://doi.org/10.1111/bjet.13023

VERENA WITTE
University of Münster, Germany
Institute of Geoinformatics
Heisenbergstr. 2
48149 Münster