Fostering data literacy by engaging in data cleaning

Jakim Eckert[1], Sarah Schönbrodt[2] and Martin Frank[1]
[1]Karlsruhe Institute of Technology, Germany; [2]Paris Lodron University Salzburg, Austria
Jakim.Eckert@kit.edu

The increasing societal relevance of data-driven technologies highlights the importance of fostering data literacy in education. One important part is data cleaning, which plays a crucial role in data-driven technologies and offers authentic opportunities to foster data literacy through critical engagement with real-world data. Despite its mathematical richness, data cleaning – particularly outlier detection – remains underrepresented in school curricula and educational research. This paper presents a design-based research project focusing on the mathematical foundations of outlier detection methods. Using the four-level approach by Hußmann and Prediger (2016), we specify and structure the mathematical topic of boxplots for outlier detection. We explore how these concepts can be meaningfully embedded in intended learning trajectories to promote students' understanding of variability, robustness, and the impact of assumptions. The material is based on real datasets and aims to support critical reflection on data-driven decision-making.

INTRODUCTION

The growing influence of data science and data-driven algorithms on our society and daily life, amplifies the relevance of data literacy. This trend is reflected in the increasing number of publications in this field, covering both statistics and computer science education, addressing different target groups from schools onwards (Daniel, 2017; Fleischer et al., 2022; Markulin et al., 2022; Msweli et al., 2023; Schönbrodt et al., 2022; Schüller, 2022; Witte et al., 2025). One important factor contributing to the relevance of data literacy is the development and omnipresence of AI technologies. These technologies, which are mostly based on machine learning methods, require large datasets and extensive training. A critical aspect of working with large datasets is the preprocessing and cleaning of messy data, including missing values, duplicates, outliers, and errors (Chu, 2019). Anaconda (2022) states that 38% of a data scientist's work involves preparing and cleaning data. According to Lohr (2014), the "data janitor work" of collecting and preparing data accounts for up to 80% of a data scientist's work.

Schüller (2022) emphasizes that data can be understood as abstractions of the real world, and that information and knowledge emerge through processes of cleaning and preparation. Data cleaning offers a promising starting point for promoting data literacy within mathematics education. Despite the growing relevance of this topic, data cleaning is rarely addressed in school curricula, and there is a noticeable lack of educational research in this area. To address this gap, we are conducting a design-based research project to explore the role of data cleaning in fostering data literacy competencies among high school students and to develop appropriate teaching and learning materials. Particular emphasis is placed on *outlier detection*, as it represents a mathematically rich and conceptually meaningful aspect of data cleaning.

THEORETICAL BACKGROUND

Definition – Outlier and Outlier Detection

As the focus in this paper is on outlier detection, we first provide a brief explanation of what constitutes an outlier. In addition, a short contextualization and a summary of potential outcomes from implementing an outlier detection algorithm will be provided. Hawkins (1980) defines an outlier as an observation deviating so much from the other observations that it suggests it was generated by a different mechanism. Aggarwal (2017) also refers to outliers as abnormalities, discordants, deviants or anomalies.

Outliers are not just statistical curiosities – they often carry meaningful information. Siebert et al. (2022) and Aggarwal (2017) contextualize that outliers can indicate real-world phenomena such as cyber-security attacks (intrusion detection), underlying diseases (healthcare), credit-card fraud (finance), or quality and production errors (manufacturing). Outliers can also provide insights about weather patterns and climate change (Aggarwal, 2017).

There are many mathematical methods for outlier detection, which are selected according to context and data set. Outlier detection can involve a test of the significance for an observation leading

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*. © ISI/IASE.

to the decision of rejecting or retaining an observation (Grubbs, 1950). In general, outlier detection is framed as a classification problem – either binary (outlier vs. normal data) or continuous (outlier scoring) (Aggarwal, 2017). A scoring can be interpreted as an increase in score from normal data to weak and strong outliers, which may be considered noise or anomalies (Aggarwal, 2017). In the context of this paper, the learning material developed within our design-research project focuses on a single numerical dataset and its corresponding context. It introduces various definitions and methods of outlier detection to offer a multifaceted view of the topic. The richness of the material becomes evident through discussions and transfer tasks that assess whether students apply their understanding in new contexts.

Mathematical overview – outlier detection

Outlier detection techniques can be categorized in various ways to structure the different concepts and procedures. Simple categorizations distinguish between statistical-based, distance-based and model-based outlier detection techniques (Chu, 2019). To provide a broader overview of the conceptual landscape, Table 1 offers a concise summary of common approaches. Distance-, density-, and clustering-based methods are types of proximity-based models (Aggarwal, 2017).

Table 1. Outlier detection methods and models (Aggarwal, 2017; Chu, 2019; Dean & Dixon, 1951; Dixon, 1950; Grubbs, 1950; Grubbs, 1969; Wang et al., 2019).

Category	Description	Examples
Statistical-based Methods	Provide outlier scores or probabilities of data points fitting to the chosen distribution.	Grubbs test, Dixon-type test, Boxplots, Trimmed Mean
Extreme-Value Analysis	Determines outliers based on deviations from normal patterns.	3σ rule, Boxplots, Depth-Based methods
Distance-Based Methods	Use the distance of a data point to the k-nearest neighbor to define the proximity between data points.	K-Nearest Neighbor, Instance-Specific Mahalanobis method
Density-Based Methods	Define local density by the number of other points within a specified local region.	Local Outlier Factor (LOF), Histogram-Based Techniques
Clustering-Based Methods	Quantify outliers by non-membership to any cluster, distance from other clusters, or size of the closest cluster.	Cluster Analysis, Extremely-Randomized Clustering Forest
Information-Theoretic Models	Implicitly generate a small summary of the data, and the deviations from this summary are flagged as outliers.	Frequent Pattern Mining, Histograms, Principal Component Analysis
Linear Models	Model data along lower-dimensional subspaces using linear correlations. The distance to a fitted model quantifies the outlier scores.	Linear Regression, Support Vector Machines
Learning-Based Methods	Learn different models through the application of these learning methods to detect outliers.	Deep Learning, Supervised Learning

Within our design-based research project, we will focus on *statistical-based methods*, *extreme-value analysis*, and *proximity-based models* (distance-based, density-based and clustering-based) as examples for outlier detection techniques. Those provide a broad perspective on outlier detection.

Educational research on data cleaning

There are some examples investigating data cleaning in mathematics education as part of data science projects. Fleischer et al. (2022) examine an extended data science project about machine learning with educationally designed Jupyter Notebooks. Markulin et al. (2022) present a statistics course for Business Administration. Both projects include data cleaning but address it as a minor component in their projects without specifically focusing on teaching the underlying mathematical methods. The approach of "Data Moves" from Erickson et al. (2019) encompasses analogous concepts, such as filtering, merging, and summarizing. However, it adopts a more comprehensive perspective that

does not prioritize mathematical data cleaning techniques. Wilkerson et al. (2021) examined the variability in data preparation processes for public datasets, which manifest as data moves. The paper focused on students performing data moves in CODAP for data preparation. The study is a re-analysis of data moves, which are about sorting, filtering, grouping, and calculating (Wilkerson et al., 2021). Therefore, the emphasis is not on the immediate detection of duplicates, missing values, or outliers. Fergusson et al. (2024) present a study about data cleaning and how high school teachers use data practices as they engage with messy data.

Design research project on outlier detection

Given the importance of data cleaning for fostering data literacy and the limited focus on mathematical aspects in existing studies, our research project addresses this gap explicitly. We aim to identify appropriate data cleaning methods and concepts, with a particular emphasis on outlier detection, to teach high school students. We investigate whether an intervention emphasizing mathematical aspects of data cleaning influences students' data literacy competencies according to Schüller's (2022) data literacy framework. Previous research has mostly investigated data cleaning and outlier detection in mathematics or statistics education as just one of many steps in the learning environment, or with a focus on high school teachers. Our design-based research project encompasses the development of teaching-learning materials on data cleaning, highlighting the mathematical foundations of outlier detection methods. The aim of this paper is to specify and structure the mathematical content of outlier detection as part of data cleaning.

METHOD

In the following, we highlight key concepts and methods of outlier detection. Building on the theoretical background and these concepts, we apply the four-level approach of Hußmann and Prediger (2016) to a selected method – boxplots. Specifically, we specify and structure the mathematical content of boxplots for outlier detection according to the four-level approach and focus on the first three levels: at the formal level, we identify relevant mathematical objects and methods; at the semantic level, we discuss key ideas, concepts and their interconnections; and at the concrete level, we present a generic approach for integrating outlier detection into teaching-learning materials, including a new visualization for introducing boxplots.

DIDACTICAL ANALYSIS

Each outlier detection method is based on distinct key ideas and assumptions about the underlying data distribution. Looking at the underlying ideas and concepts, we should focus on assumptions being made and key ideas. *Statistical-based methods* typically presuppose a specific distribution model. For example, tests such as Grubbs or Dixon rely on the assumption of a normal distribution to detect extreme values as potential outliers. Understanding the concept of a distribution model and recognizing that both global and local extreme values can indicate outlier therefore is crucial. These assumptions influence the distinction between normal and abnormal data behavior, emphasizing the role of human judgment in shaping detection outcomes and, consequently, data-driven decisions.

Distance-based methods, as articulated by Wang et al. (2019), rely on the underlying principle that normal data points are proximate to their neighbors, whereas outliers are characterized by a greater distance to their nearest points. The effectiveness depends on defining appropriate distance thresholds.

Cluster-based methods, discussed by Aggarwal (2017), Chu (2019), and Wang et al. (2019), require decisions about the number of clusters to be formed. These methods identify outliers based on their relationship to cluster structures, such as the distance to other clusters, the size of the nearest cluster, or the absence of cluster membership (Aggarwal, 2017).

Density-based methods, as delineated by Aggarwal (2017), segment the data space and determine local density based on the number of points within a designated region. The key idea is that outliers are more probable in regions of low density (Wang et al., 2019).

This brief overview of different detection methods shows how the assumptions and key ideas underlying each method influence which data points are detected as an outlier. In the following section, we will focus specifically on the use of boxplots as a statistical-based method for outlier detection, which constitutes a central topic in our design-research project and the associated teaching-learning

materials. We will follow the 4-level approach by Hußmann and Prediger (2016), addressing and adapting some of their guiding questions to further specify and structure the topic of boxplots in the context of outlier detection.

Boxplots: Introduction

The presentation of boxplots frequently used in school contexts focuses on the distribution of data but typically omits the detection of outliers. Even when outliers are visualized within boxplots, the thresholds applied to classify them are rarely made transparent. To address this gap in context of a teaching-learning arrangement, we have adapted the classical boxplot representation to emphasize the process of outlier classification.

A short review of the literature on boxplots reveals the focus on misconceptions, challenges in understanding and interpreting boxplots, and recommendations for exploratory data analysis in mathematics education. Studies have also explored the implementation of boxplots as a reversal task and their role in promoting key concepts related to variability (Abt et al., 2024; Bakker et al., 2005; Biehler & Steinbring, 1991; Lache et al., 2022; Lem et al., 2013; Ossadnik, 2022). Previous approaches have primarily focused on measures of location and dispersion, with comparatively less attention given to the role of outliers and their identification. In particular, the methods for detecting outliers are rarely discussed or integrated into didactic frameworks. Even when boxplots with outliers are used, explicit strategies for fostering students' understanding of outlier detection remain underdeveloped

Therefore, we will examine boxplots as a descriptive tool for visualizing datasets, with a particular emphasis on the detection of outliers. The aim of this paper is to provide an overview of how to meaningfully detect outliers using boxplots. After analyzing the mathematical content on the formal and semantic level, we propose a teaching approach designed to enhance students' understanding of boxplots in the context of outlier detection and to introduce these concepts in a pedagogically meaningful way.

Boxplots: Formal level

On the formal level we will address the guiding question "which concepts and theorems have to be acquired in the context of boxplots?" Understanding boxplots for outlier detection requires the acquisition of several concepts and theorems. First, students must grasp the descriptive elements of a boxplot, such as the minimum, maximum, and median. A dataset is typically described as a sample, which is a finite subset of a population, and the sample size, denoted as n, indicates the number of elements it contains.

Tukey (1977) introduced boxplots as graphical representation of the five-number summary (extreme values, lower quartile, upper quartile, and median). These values divide the dataset into four equally sized sections. Eichler and Vogel (2013) further formalize the definition of the median and the quartiles using quantiles. For a dataset $x_1, x_2, \ldots, x_n$ of metric data and $0 < p < 1$, the p-quantile $x_p \in R$ is defined as

$$x_p = x_{n \cdot p+1} \qquad for\ n \cdot p \notin Z$$

$$x_p = \frac{1}{2}(x_{n \cdot p} + x_{n \cdot p+1}) \qquad for\ n \cdot p \in Z.$$

Then they describe the quartiles and the median as:

- 0 - quantile is as (0. Quartile or) minimum of the sample (x_{min}),
- 0.25 - quantile is as 1. Quartile of the sample ($x_{0.25}$),
- 0.5 - quantile is as (2. Quartile or) median of the sample ($x_{0.5}$),
- 0.75 - quantile is as 3. Quartile of the sample ($x_{0.75}$),
- 1 - quantile is as (4. Quartile or) maximum of the sample (x_{max}).

The difference between the first and third quartile is known as the interquartile range (IQR), $IQR = x_{0.75} - x_{0.25}$, and serves as a robust measure of dispersion (Eichler & Vogel, 2013). The range, calculated as difference between the maximum and the minimum, is another measure of dispersion. Both dispersion measures are reflected in the five-number summary and the resulting boxplot.

In the context of outliers and boxplots, the IQR is crucial. Outliers are identified using fences defined by 1.5 times the IQR below the first quartile and above the third quartile. Any data point lying outside these bounds is classified as an outlier and is typically displayed individually in the boxplot.

How can the concepts, theorems, justifications, and procedures be structured in logical trajectories? The initial step is intuitively identifying the median and quartiles through ordering the sample (Eichler & Vogel, 2013). Familiarity with calculating the range provides a foundation for understanding dispersion, which naturally leads to the introduction of the IQR. Once all components of the five-number summary are introduced, students are capable of visually summarizing the distribution using boxplots. Outlier detection then becomes a straightforward extension as students apply the fences $((x_{0.25} - 1.5 \cdot IQR, x_{0.75} + 1.5 \cdot IQR))$ to detect outliers.

In summary, the formal understanding of boxplots for outlier detection involves descriptive statistics, quantile-based reasoning, and the use of robust dispersion measures. This process culminates in the graphical identification of outliers within a dataset. This logical progression facilitates the meaningful acquisition of the theoretical and practical aspects of boxplots in statistical education.

Boxplots: Semantic Level

At the semantic level, we highlight key ideas underlying the concepts, theorems, and procedures related to the use of boxplots for outlier detection. Boxplots aggregate information about a sample and provide a visual summary of distributional properties. This idea is crucial for boxplots and at the same time it introduces a limitation: individual data points become obscured (Bakker et al., 2005). To mitigate this, Bakker et al. (2005) recommend combining boxplots with dot plots. The simultaneous visualization of global patterns and individual cases can also prevent from the misinterpretation of the area in the box (Abt et al., 2024). This common misconception assumes that the size of the area is directly proportional to the number of data points. However, it is important to note that the size of the box actually reflects the density of the data inside this area (Lem et al., 2013).

Furthermore, understanding the concepts of spread and variability is crucial for engaging with boxplots (Abt et al., 2024; Bakker et al., 2005). The range can serve as a rudimentary measure of spread, and provides a useful start for introducing the interquartile range (IQR) as a more robust alternative. The IQR is interpreted not only as the spread of the middle 50% of the data, but as a measure of the overall spread (Bakker et al., 2005). Abt et al. (2024) describe the boxplot as an overlay of the range.

The IQR is a measure of spread that is particularly robust against outliers (Bakker et al., 2005). In contrast, the range and standard deviation, which are also employed in educational settings, are vulnerable to outliers. The notion of robustness becomes crucial for the construction of meaning through the use of the boxplot and the IQR in the context of outlier detection. In instances where outliers have a significant impact on the detection process, their presence may remain undetected. Consequently, the following question is posed: How does the robustness relate to other learning contents as the 3σ rule? This allows students to develop a better picture of robustness and the boxplot for outlier detection.

A comparison between the IQR and 3σ rule highlights their respective assumptions and implications for outlier detection. While the 3σ rule relies on the assumption of a normal distribution and uses the distance of 3 standard deviations as a cut-off point, the boxplot makes no assumption. Interestingly, the value of $1.5 \cdot IQR$ corresponds approximately to 2.7 standard deviations from the mean in case of a normal distributed sample (Aggarwal, 2017). In the intended learning trajectories both concepts should be compared with emphasis on outlier detection.

In terms of general outlier detection, classic boxplots are considered statistical methods that do not make direct distributional assumptions. This makes them more accessible, as they are mainly assigned to extreme value analysis, where the concept of global extreme values as potential outliers is crucial for the univariate case.

Boxplots: Concrete level

This section, along with the following ones, provides insights into our learning trajectory and demonstrates how key ideas and underlying structures can be realized. A closer examination of boxplots for outlier detection reveals the need for a slightly different representation. This representation emphasizes the 1.5-fold IQR as a criterion for outlier detection while retaining the essential information of the boxplot. A fundamental distinction from conventional representations lies in the extension of the

whiskers and the explicit visualization of individual data points. In this enhanced form, the whiskers are extended not only to the outermost points within the specified range, but also beyond it—up to the so-called "fences" that mark the threshold beyond which data points are considered outliers. Simultaneously, all data points are displayed individually, as illustrated in the boxplot in Figure 1. This approach will help students differentiate between noise, weak, and strong outliers. To strengthen this idea, different datasets can be sequenced and visualized in boxplots to illustrate the idea of the 1.5-fold IQR as a boundary. Therefore, it is crucial to direct students' attention to the whiskers as a boundary where the binary classification of an outlier occurs. For discussions, we suggest prompting students to consider data points near the fences and to discuss which data points might require a deeper analysis, thereby guiding the development of the concept of a whisker as a classification boundary.

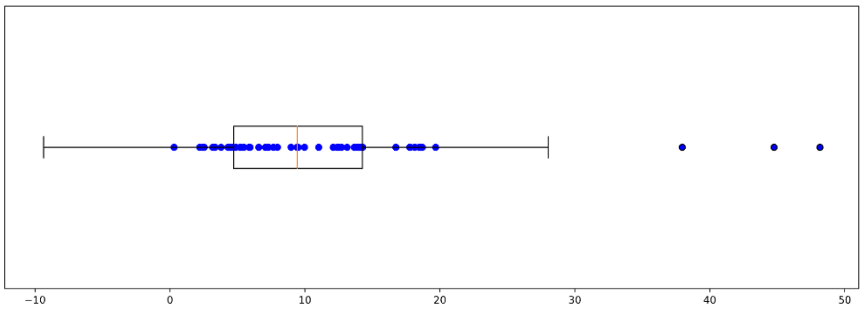

Figure 1. Enhanced boxplot for outlier detection.

Standard deviations calculated using the 3σ rule (particularly for normal distributions) and IQR depicted using boxplots can facilitate the visualization of data (see Figure 2). The detection of outliers is facilitated by the implementation of scatter measures. Human agency can be identified as a contributing factor within the system. Biehler and Steinbring (1991) describe the specification of the more general concept of dispersion in the form of quartile intervals (25% and 75% values) as pragmatically justified and, at the same time, arbitrary. The arbitrariness of the situation presents a valuable opportunity for students to engage in critical reflection and discourse regarding the assumptions that were initially made and the approaches that have evolved over time.

The standard deviation and the IQR are both used to detect outliers, defined as data points with a very low probability at this part of the distribution. On the one hand, this knowledge can be used to consider distribution assumptions that are sometimes made. At the same time, to move more strongly in the direction of statistical-based methods and deal with methods such as Grubbs or Dixon's test. This interplay between visual and statistical methods deepens students' conceptual understanding of outliers and lays the foundation for critical engagement with the assumptions and limitations of different detection methods.

OUTLOOK

In summary, outlier detection represents a mathematically rich topic that offers links to school curricula and to learning objectives from the field of data literacy. A comprehensive learning trajectory for outlier detection can be initiated with intuitive extreme value analyses and progressing toward statistical methods such as the 3σ rule, boxplots, and the Mahalanobis distance. Central questions could be: Which values are good candidates for potential outliers in order to check them with a detection method? What influence do the distribution and assumptions made have on the selection of outliers (also beyond the distribution assumption)? What influence does the method have on detecting outliers? At the same time, students should understand the influence of data cleaning and outlier detection on data-driven predictions and that carelessly removing outliers can have significant consequences.

The paper presents a first approach for integrating outlier detection into learning trajectories, which offers valuable opportunities to foster data literacy and to support students in becoming responsible, data-aware citizens. Further empirical research is needed, particularly on how students

interpret and use boxplots in the context of identifying outliers. Such studies have the potential to inform refinements to the learning trajectory and support the development of effective instructional strategies.

Following this structured approach not only opens up mathematical content that is rarely addressed in traditional classrooms but also highlights the role of data cleaning and outlier detection in machine learning contexts. The next step in this project is the empirical testing and analysis of the prototype materials, with a focus on how well the intended learning trajectory supports conceptual understanding and promotes data literacy competencies.

REFERENCES

Abt, M., Leuders, T., Loibl, K. & Reinhold, F. (2024). Ein verstehensorientierter Zugang zu Boxplots: Mit digitalen Explorationen zur Variabilität in Daten [A comprehension-oriented approach to box plots: Using digital explorations to variability in data]. *mathematik lehren, (243),* 41–45.

Aggarwal, C. C. (2017). *Outlier Analysis*. Springer. https://doi.org/10.1007/978-3-319-47578-3

Anaconda (2022). *State of Data Science 2022: Paving the Way for Innovation*. Anaconda Inc. https://www.anaconda.com/state-of- data-science-report-2022

Bakker, A., Biehler, R. & Konold, C. (2005). Should young students learn about box plots? Curricular development in statistics education. In G. Burrill & M. Camden (Eds.), Curricular development in statistics education. *International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 28 June–3 July 2004* (pp. 163–173). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. & Steinbring, H. (1991). Entdeckende Statistik, Stengel-und-Blätter, Boxplots: Konzepte, Begründungen und Erfahrungen eines Unterrichtsversuches [Exploratory statistics, stem-and-leaf, boxplots: Concepts, justifications and experiences of a teaching experiment]. *Der Mathematikunterricht*, *37*(6), 5–32.

Chu, X. (2019). Data Cleaning. *Encyclopedia of Big Data Technologies*. Springer. https://doi.org/10.1007/978-3-319-77525-8

Daniel, B. K. (2017). Big data and data science: A critical review of issues for educational research. *British Journal Of Educational Technology*, *50*(1), 101–113. https://doi.org/10.1111/bjet.12595

Dean, R. B. & Dixon, W. J. (1951). Simplified statistics for small numbers of observations. *Analytical Chemistry*, *23*(4), 636–638. https://doi.org/10.1021/ac60052a025

Dixon, W. J. (1950). Analysis of Extreme Values. *The Annals Of Mathematical Statistics, 21(4),* 488–506. https://doi.org/10.1214/aoms/1177729747

Eichler, A. & Vogel, M. (2013). *Leitidee Daten und Zufall: Von konkreten Beispielen zur Didaktik der Stochastik* [Key idea data and chance: From concrete examples to the didactics of stochastics]. Springer. https://doi.org/10.1007/978-3-658-00118-6

Erickson, T., Wilkerson, M., Finzer, W. & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12(1)*. https://doi.org/10.5070/t5121038001

Fergusson, A., Pfannkuch, M. & Budgett, S. (2025). Data cleaning doesn't happen in a vacuum: An initial exploration of high school statistics teachers' data practices with messy data. In J. Kaplan. & K. Luebke (Eds.). *Connecting data and people for inclusive statistics and data science education. Proceedings of the Roundtable conference of the International Association for Statistics Education(IASE), July 2024, Auckland, New Zealand. ISI/IASE.* https://doi.org/10.52041/iase24.301

Fleischer, Y., Biehler, R. & Schulte, C. (2022). Teaching and Learning Data-Driven Machine Learning with Educationally Designed Jupyter Notebooks. *Statistics Education Journal*, *21(2),* Article 7. https://doi.org/10.52041/serj.v21i2

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics, 21(1),* 27–58. https://www.jstor.org/stable/2236553

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21. https://doi.org/10.2307/1266761

Hawkins, D. M. (1980). *Identification of outliers*. Springer, https://doi.org/10.1007/978-94-015-3994-4

Hußmann, S. & Prediger, S. (2016). Specifying and structuring mathematical topics. *Journal für Mathematik-Didaktik*, *37*(S1), 33–67. https://doi.org/10.1007/s13138-016-0102-8

Lache, J., da Costa Silva, N. & Rolka, K. (2023). Individuelles Feedback und vielfältige Repräsentationen: Einsatz digitaler Mathematikaufgaben in der Schule [Individual feedback and diverse representations: Using digital mathematics tasks in school]. In: *Digitaler Mathematikunterricht in Forschung und Praxis. Tagungsband zur Vernetzungstagung 2022 in Siegen* (pp. 113–123). https://d-nb.info/128784488X/34

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

Lem, S., Onghena, P., Verschaffel, L. & Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction, 26,* 22–35. http://dx.doi.org/10.1016/j.learninstruc.2013.01.001

Lohr, S. (2014). *For Big-Data Scientists, "Janitor Work" is key hurdle to insights*. The New York Times. *https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html*

Markulin, K., Bosch, M., Florensa, I. & Montañola, C. (2022). The evolution of a study and research path in Statistics. *epiDEMES, 1*. https://doi.org/10.46298/epidemes-7584

Msweli, N. T., Mawela, T. & Twinomurinzi, H. (2023). Data Science Education – A scoping review. *Journal Of Information Technology Education Research, 22,* 263–294. https://doi.org/10.28945/5173

Ossadnik, H. (2022). Boxplots–einfach zu erstellen, schwer zu interpretieren: Interpretation mit Simulationen üben [Boxplots – easy to create, difficult to interpret: Practicing interpretation with simulations]. *digital unterrichten: Mathematik, 2022(1),* 6–7.

Schönbrodt, S., Wohak, K. & Frank, M. (2022): Digital Tools to Enable Collaborative Mathematical Modeling Online. *Modelling in Science Education and Learning*, *15*(1), 151–174, https://doi.org/10.4995/msel.2022.16269

Schüller, K. (2022). Data and AI literacy for everyone. *Statistical Journal of the IAOS*, *38*. 1–14. https://doi.org/10.3233/SJI-220941

Siebert, J., Schroth, C. & Groß, J. (2022). *Time Traveling with Data Science: Outlier Detection (Part 3)*. Frauenhofer IESE. *https://www.iese.fraunhofer.de/blog/outlier-detection/*.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Wang, H., Bah, M. J. & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964–108000. https://doi.org/10.1109/access.2019.2932769

Wilkerson, M. H., Lanouette, K. & Shareff, R. L. (2021). Exploring variability during data preparation: a way to connect data, chance, and context when working with complex public datasets. *Mathematical Thinking and Learning, 24(4),* 312–330. https://doi.org/10.1080/10986065.2021.1922838

Witte, V., Schwering, A., & Frischemeier, D. (2025). Strengthening Data Literacy in K-12 Education: A Scoping Review. *Education Sciences*, *15(1),* 25. https://doi.org/10.3390/educsci15010025