DATA CLEANING DOESN'T HAPPEN IN A VACUUM: AN INITIAL EXPLORATION OF HIGH SCHOOL STATISTICS TEACHERS' DATA PRACTICES WITH MESSY DATA

Anna Fergusson, Maxine Pfannkuch and Stephanie Budgett
Waipapa Taumata Rau | University of Auckland, New Zealand|
a.fergusson@auckland.ac.nz

Cleaning data is an important facet of statistical practice. The research literature on examining data practices of learners when dealing with messy data that needs cleaning, however, is scarce. As part of a larger study, six Grade 12 high school statistics teachers engaged with a height estimation task, for which the data were drawn from a publicly available website containing 39,195 rows of text entries in a variety of measurement systems. The teachers' observed data practices were characterised as inspecting, ideating, sorting, sampling, converting, visualising, creating, and describing. The implications of the findings with regard to statistical enquiry pathways are discussed.

INTRODUCTION

The data used for teaching statistics needs diversifying to better reflect the pervasiveness of digital technologies in our modern world (e.g., Engel, 2017). Large volumes of data are now available for teaching but often these data are messy, unstructured, or inaccessible without the use of automated approaches. In contrast, data sets used traditionally in statistics courses are often smaller and pre-processed to make them immediately usable with standard statistical software tools (e.g., Hardin, 2018). The advent of data science education has led to calls to teach "thinking with data" (e.g., Gould, 2021; Hardin et al., 2015; Nolan & Temple Lang, 2010). Consequently, education researchers are re-thinking inquiry approaches for modern data (e.g., Fry & Makar, 2021; Perez & Lionberger, 2023) and proposing new frameworks that explicate data science investigative or thinking practices (e.g., Fergusson, 2022; Lee et al., 2022). Despite an increased focus on data science education and data literacy at the high school level, research-based literature that examines the data practices of learners as they engage with messy data is scarce. In this paper, we present our initial exploration of high school statistics teachers' data practices as they engage with messy data.

ENGAGING LEARNERS WITH MESSY DATA

The use of real or authentic data for teaching statistics often involves datasets that are already cleaned before students use them (Engel, 2017; Hardin, 2018). Messy data, on the other hand, can be poorly structured and contain missing or incorrect values, often caused by issues in the data collection process (Chai, 2020; Kjelvik & Schultheis, 2019). Engaging with managing, and pre-processing of data for analysis is an important aspect of the *Data* stage of the PPDAC (Problem-Plan-Data-Analysis-Conclusion) statistical enquiry cycle (Wild & Pfannkuch, 1999) and can help make the data feel more real to learners (Cummiskey et al., 2012). Using messy data for statistical enquiry introduces students to the uncertainties and complexities of creating data (D'Ignazio, 2017; Gafney & Ben-Zvi, 2023) and to authentic data practices for analysing complex data (Dvir & Ben-Zvi, 2022; Kjelvik & Schultheis, 2019; Lee et al., 2022; Rosenberg et al., 2020). Messy data provide interesting opportunities for data exploration (Hammett & Dorsey, 2020; Kjelvik & Schultheis, 2019) to support student-driven investigations of sources of variation (Gould et al., 2014). Engaging learners with cleaning data may encourage more creativity in the statistical enquiry (e.g., Yue, 2012) and can highlight the impact of human decision making on the data set produced (Cummiskey et al., 2012).

Considering that data scientists can spend up to 80% of their time cleaning data (Lohr 2014), it is important that learners develop effective statistical and computational practices to process messy data (Finzer & Reischman, 2018; Lee et al., 2022; Perez & Lionberger, 2023; Wickham, 2014). These data practices include organising the data in a way that both humans and computer programs can read, structuring the data with cases as rows and variables as columns, using statistical techniques to identify outliers, and recording the changes made to the raw data (Broman & Woo, 2018; Cummiskey et al., 2012; Holcomb & Spalsbury, 2005; Wilson et al., 2017). Although data cleaning is acknowledged as important (e.g., Erickson et al., 2019), there is limited research that focuses on teaching statistical enquiry involving processing messy data (e.g., Cummiskey et al., 2012; Holcomb & Spalsbury, 2005;

In: Kaplan, J. & Luebke, K. (Ed.) (2024). *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.*

Konold et al., 2017; Musoyka et al., 2017). Advice from science educators is to start with a small set of messy data and then branch out to bigger data sets, introducing more and more complex data sets over time (e.g., Hammett & Dorsey, 2020; Kjelvik & Schultheis, 2019).

The scarcity of research for teaching data cleaning is perhaps not surprising considering a recent survey of US-based undergraduate statistics instructors found that data cleaning practices were not taught by 54% of respondents, with an additional 37% indicating they were only taught in a minor way (Legacy et al., 2022). Science education researchers also report limited teacher engagement with large messy data (e.g., Rosenberg et al., 2022) and database education researchers state it is rare to find research reports about teaching data cleaning (Yue, 2012). Data cleaning does not happen in a vacuum, independent from other aspects of a statistical enquiry pathway. Learners may only identify issues with their data or find the need to change the data in some way once they start exploring it (Chai, 2020; Lee et al., 2022) and may need to use a wider range and combination of data practices or moves such as creating hierarchy, creating new variables, or randomly sampling rows of data to make it more manageable to process the data using statistical software tools (cf. Erickson et al., 2019).

Research question

Aotearoa New Zealand has one national mathematics and statistics curriculum taught at nearly all high schools (Ministry of Education, 2007), with the PPDAC statistical enquiry cycle (Wild & Pfannkuch, 1999) being a core component of the statistics curriculum. Cleaning data is mentioned specifically in the curriculum as an achievement objective. Senior high school level assessment materials, however, indicate that although students are required to use real data to carry out statistical investigations, they are not required to engage with messy data, nor demonstrate data practices related to cleaning data. Consequently, high school statistics teachers are unlikely to have experience with designing and implementing assessment tasks that involve students engaging with messy data, although they may have experience cleaning data when creating data sets for use in assessment tasks. Given the need for new research to support the implementation of data science at the high school level, the purpose of the research was to learn more about the messy data practices used by high school statistics teachers, to inform future curriculum and professional development projects within Aotearoa New Zealand. The research question is: *What data practices are observed when high school statistics teachers engage with messy data as part of a statistical enquiry?*

RESEARCH APPROACH

The larger study within which this research paper sits (Fergusson, 2022) used a design-based research approach (e.g., Bakker & van Eerde, 2015) to explicate task design principles for statistical modelling from a data science perspective. The participants were four female and two male Grade 12 statistics teachers. The teachers had, on average, 10.5 years high school teaching experience (mean = 10.5, min = 7, max = 14). All the teachers had previous experience with spreadsheet tools and data visualisation tools. The teachers were participants in the larger study that involved four full-day professional development workshops.

Teaching experiment

For the study, high school teachers were positioned as the learners. The teaching experiment took place during the afternoon session of the third day of the professional development workshops. The overall theme for the workshop was *Humans vs Computers*, and in the morning session teachers completed a task that investigated whether humans' estimates of heights could be influenced by external prompts (see Fergusson & Pfannkuch, 2022). During the morning session task, the teachers were asked to provide three estimates for the height of an unknown man shown in a photo: one that was too high, one that was too low, and their final estimate for his height. Each teacher, and two members of the research team, completed their estimates on data cards, three of which are shown in Figure 1. At the beginning of the teaching experiment, the eight height estimate data cards were given back to the teachers. The researcher, the first author, then showed the teachers the website from which the height estimate activity had been obtained (estimation180.com), and the true height of the unknown man was revealed to be 1.93 metres. At the time of the teaching experiment, the website provided a *Google form* with similar questions as the data cards shown in Figure 1, and the data collected from this form was publicly available as a *Google sheet* containing nearly 40,000 rows (see Appendix).

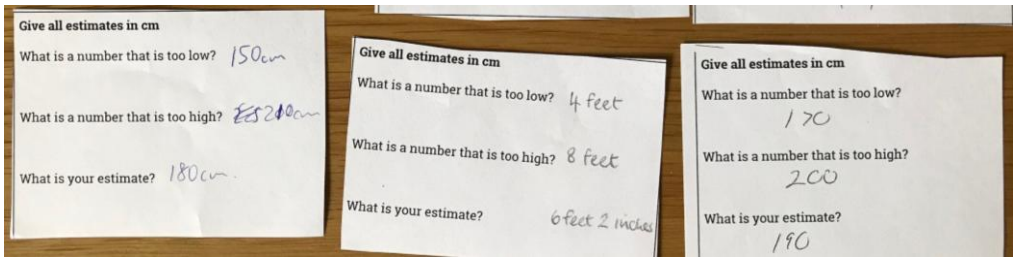

Figure 1. Three of the eight data cards completed by the teachers and two members of the research team in response to being asked to estimate the height of an unknown man shown in a photo

Each row in the *Google sheet* represented one response to the *Google form*. The data within the *Google sheet* provided by the website were unable to be used "as is," however, because the "too low", "too high" and final height estimates were essentially text data. Just like the data cards shown in Figure 1, the height estimates made by respondents were messy as they were written using a variety of imperial, metric, and other measurement systems, often included symbols, units, and words, and in some cases missing values. In addition to the columns related to the different height estimates, the *Google sheet* contained columns providing the timestamp of when the response was received, the name of the respondent, and the respondent's reasoning for their final height estimate (see Appendix).

A *Google document* was provided to the teachers, which described the two data sources: the data cards representing their "too low", "too high" and final height estimates, and all the responses to the task on the *estimation180.com* website. The document also contained a link to a copy of the publicly available *Google form* data as an editable *Google sheet*. The teachers were instructed to investigate the relationship between the "too low", "too high" and final estimates, and to keep notes as they investigated. As the teachers were already familiar with the data context from the morning task and had expressed curiosity about how humans make estimates, we expected them be creative and investigate questions involving all three height estimates, such as: *Are the final estimates centred between the too low and too high height estimates?* Due to the teachers' familiarity with the PPDAC statistical enquiry cycle and Grade 12 statistics assessment tasks, we conjectured that their engagement with the messy data would take the following statistical enquiry path: developing a specific investigative question using the eight data cards shown in Figure 1, taking a random sample of rows from the provided data after realising how messy it was, manually cleaning these data for any relevant variables, carrying out appropriate inferential analysis (e.g., constructing a confidence interval), and making a conclusion that answered their investigative question.

Data collection and analysis

Teachers worked in pairs, with each pair sharing one laptop computer to complete the task. Their actions and conversations were recorded as they engaged with the task, using the screen recording tool *Screencastify*. The teacher pairings for the task were Amelia and Harry, Alice and Nathan, and Ingrid and Naomi (pseudonyms have been used). Due to technical issues, the last eight minutes of Ingrid and Naomi's engagement with the task were not recorded, and the teachers were asked to record on a piece of paper what they would have done if their laptop had not crashed.

We had planned to use the PPDAC statistical enquiry cycle as the theoretical framework to interpret and label teachers' verbalisations and actions as they engaged with the task, inspired by the analytical approaches described by Barker and Elrod (2023) and Gould et al. (2017). Because none of the teacher pairs followed our conjectured statistical enquiry path, however, we decided it would be more informative to identify the main focuses of the observed *data practices* and use these to create a visualisation that compared the different statistical enquiry paths taken. Our decision to adopt a more open and "theory-free" analysis approach was consistent with the way one of the teachers described their engagement with the task, stating it was, "a different D [data], a different approach to D [data]. It's the data science lens of the PPDAC cycle compared to what we currently do."

A retrospective task-oriented qualitative analysis (Bakker & van Eerde, 2015) was used to characterise the *data practices* observed as the teachers engaged with the messy data. First, the screen recordings of each pair of teachers were used to create a sequence of timestamped "blocks", consisting

of transcripts of what was said (verbalisations) and notes describing what was done computationally (actions). Each block was then reviewed, and initial labels assigned by considering the actions taken with the data (e.g., *typing over a value given in feet and inches and replacing it with a value given in centimetres*). Through a process of constant comparison (e.g., Bakker & van Eerde, 2015; Creswell, 2012), bigger timestamped blocks of data practices were created by merging adjacent blocks that contained complementary data actions, and these "super blocks" were labelled to characterise the main focus of the data practices they contained.

RESULTS

The data analysis process resulted in eight main focuses for the observed *data practices*:

- *Inspecting* provided data (discussing the number of rows, column headers, messiness of data)
- *Ideating* with provided/sample data or data cards (creatively generating ideas for analysing data)
- *Sorting* provided/sample data by height estimate
- *Sampling* rows from provided data
- *Converting* height estimates to a standard unit
- *Visualising* sample data (including discussing interesting features of data)
- *Creating* new variables with sample data (e.g., differences, ratios)
- *Describing* relationships between variables (using measures such as correlation coefficient)

The timestamped super blocks and their labels characterising the main focuses of the observed data practices within the super block were used to create a visualisation that compared the different statistical enquiry paths taken by pairs of teachers (Figure 2).

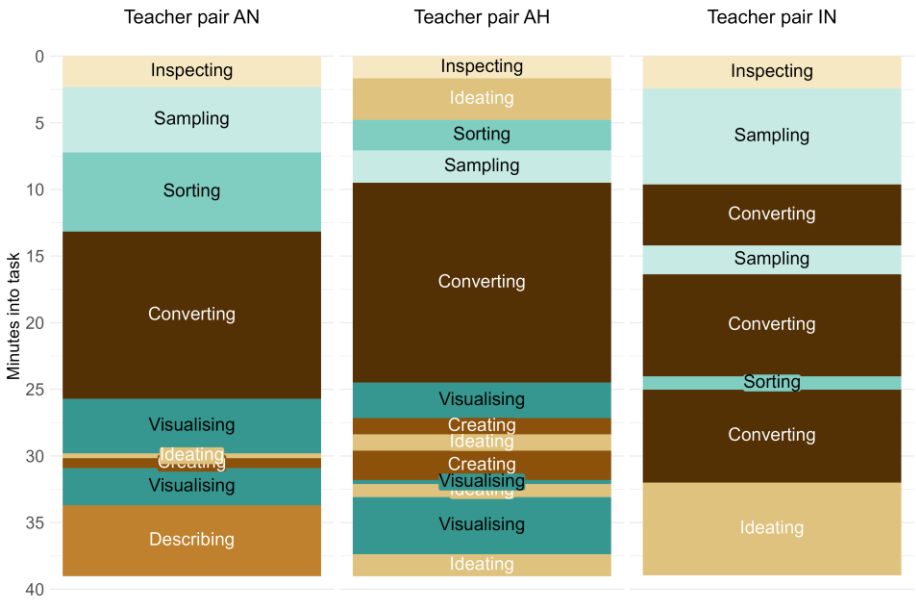

Figure 2. A comparison of the statistical enquiry paths followed by each pair of teachers, using the order and time spent on each of the main focuses for the observed data practices

Figure 2 supports our earlier statement that none of the teacher pairs' engagement with the messy data followed our conjectured statistical enquiry path. It appeared the novelty of being provided with a new and large data set motivated their attention to first *inspecting* its features rather than specifying an investigation question. Initial excitement about the data almost immediately turned to despair, however, when they noticed the messiness of the data, aptly captured by Naomi's reaction, "*Oh, we need to clean it? Oh my god, that's so annoying, I hate that*!" All teachers used random *sampling* as a strategy to reduce the amount of *converting* required. For the two teacher pairs (AN & AH) that *visualised* or *described* relationships between variables in their sample data, however, neither considered sampling variation when discussing features of the data distributions. All teachers tried to use the *sorting* function within *Google sheets* or *Microsoft Excel* to group height estimates from the same measurement system together, and discovered this approach was not that helpful. We observed that the teachers' data

practices focused on *sampling* and *converting* also utilised specific spreadsheet-based data practices, for example the use of the *randomise range* function in *Google sheets*, but due to paper length constraints these are not discussed. Figure 2 also illustrates the different statistical enquiry paths followed, the salient features of which will now be described.

Teacher pair AN (Alice and Nathan)

The statistical enquiry path followed by Alice and Nathan (Figure 2) had some aspects of similarity to the one we conjectured. Although the teachers did not explicitly state an *idea* for data analysis at the start of their enquiry, it appeared this may have been because the task instruction *to investigate the relationship between the "too low", "too high" and final estimates* was interpreted by the teachers as providing the specific investigative question. Specifically, the use of the word "relationship" indicated to them that they needed to investigate the relationship between two numeric variables using linear regression. Hence, after *sampling* rows of data from the provided messy data set, attempting to *sort* them, and *converting* the height estimates into a common unit (inches), Alice and Nathan *visualised* their sample data using scatterplots with fitted lines, for each combination of the three height variables. It was only when they did not find clear relationships between any of the height estimates, even after deleting many values they considered outliers using the data visualisation tool, that they discussed the *idea* of exploring the difference between the "too high" and "too low" height estimates. After *creating* a new "difference" variable, the teachers again focused on producing visualisations of scatterplots, plotting the "difference" variable against the final, the "too low", and then the "too high" height estimates. This time, they added the display of the correlation coefficient.

Alice and Nathan appeared surprised to discover that the correlation coefficient between the "difference" variable and the "too high" variable for their sample data was 0.99. The teachers were the only pair that clearly *described* the relationship between the variables. However, they did not consider sampling variation, nor the mathematical relationship between the "difference" and "too high" variables, stating, "… almost perfect for a prediction if we were to use this model." Although they didn't always discuss their reasons for removing values at the time, it appeared the teachers were motivated by finding stronger relationships, although they could have also been removing unreasonable values for heights. Our interpretation is consistent with the notes the teachers wrote at the end of their enquiry, which stated, "*there was an indication of a strong relationship once a couple of outliers (and they were outliers in the context of things i.e. 1000cm individual) were removed."*

Teacher pair AH (Amelia and Harry)

The statistical enquiry path followed by Amelia and Harry (Figure 2) also had some aspects of similarity to the one we conjectured. Although they did not brainstorm *ideas* for data analysis immediately at the start of their enquiry, after a couple of minutes *inspecting* the provided data Harry stated, "A purpose first might be nice." The two teachers then engaged in a discussion where they both shared what they were interested in exploring. For instance, Amelia stated that she wanted to know, "Is there a bigger difference between the max and the 'estimate' compared with the min and the 'estimate'. Are you better at picking something too low or something too high?", which were characterised as *ideating*. After the teachers attempted to *sort* the height estimates in the provided messy data set, they *sampled* some of the rows instead and *converted* the height estimates for their sample into a common unit (inches). Amelia and Harry then *visualised* and discussed the variation of each of the three height estimate variables separately, before remembering that they had wanted to analyse differences, which led them to *create* new variables based on differences. The nature of their statistical enquiry from this point onwards followed a similar pattern of moving between *ideating*, *visualising*, or *creating* and consequently they spent the longest time focused on these data practices compared to the other teachers. At the end of their enquiry, they brainstormed further *ideas* for analysis, which included *categorising the differences between the "true" height of the man and final height estimate by "Close" and "not Close", to account for those playing around and those who are bad at estimating.*

Teacher pair IN (Ingrid and Naomi)

Ingrid and Naomi's engagement with the messy data (Figure 2) was focused on developing automated data practices for *converting* the height estimates to a common unit, so that they could "clean the data" more efficiently than manual approaches. Ironically, they spent the longest of the three teacher

pairs processing the messy data and would have spent even longer if their computer had not crashed before the end of the task as at the time of the crash, they had only converted one of the height estimates into centimetres for 40 responses. When the teachers first *inspected* the provided data, however, they did not agree about using automated data practices, as demonstrated in the following exchange:

Ingrid:	We have to do conversions as well!
Naomi:	Yeah, it's ridiculous. So, we need to take a sample. So, that gives us a reason to sample, because we'll only clean the sample.
Ingrid:	OK, that's what you want to do, not clean the whole dataset?
Naomi:	No, we're not going to clean the whole dataset because this is huge.
Ingrid:	Nah, but if you wrote functions …
Naomi:	Yeah, that's true but it's going to be pretty complicated.

Naomi was aware that *Google sheets* functions could assist their data processing, and even suggested using "text to columns" later in the enquiry. However, it appeared in this exchange that she was not convinced that they would be able to develop an effective automated approach due to the large variety of ways the height estimates had been provided. When their laptop crashed, they turned their attention to the height estimate data cards and used these data to discuss creative ideas for data analysis (*ideating*), including: *Do people informally/intuitively centre their final estimate, like with a prediction interval? Are people trying to make "realistic" low and high boundaries?*

DISCUSSION

The purpose of our research was to conduct an initial exploration into the data practices used by high school statistics teachers as they engaged with messy data. After recognising the teachers did not follow the conjectured statistical enquiry path, a decision was made to re-orient our data analysis approach to be more open and to focus on *how* different data practices were used by the teachers within each of their statistical enquiry paths. Following teaching recommendations to start with a small messy data set and then go bigger (Hammett & Dorsey, 2020; Kjelvik & Schultheis, 2019), we intended for the teachers to use the height estimate data cards to generate ideas for data analysis first. We did not make this instruction clear for the task, however, and consequently none of the teachers generated ideas for analysing the data using the smaller set of messy data. Rather than following a statistical enquiry path focused on confirmation or problem solving, two of the teacher pairs (AN & AH) appeared to be operating in an exploratory or discovery mode (cf. Finzer & Reischman, 2018), adopting a creative perspective where they were open to be surprised (McKenney & Reeves, 2018). From a technical perspective, their statistical enquiries can also be characterised as exploratory, not confirmatory, as they used the same height estimates several times to create data visualisations and to generate more ideas for analysis (Wickham et al., 2023). Their approaches, while creative, did not always appear to be statistically sound, for example Alice and Nathan's data practice of removing outliers (cf. Holcomb & Spalsbury, 2005). The opportunity to be creative with their statistical enquiry, however, also resulted in productive iterations of *ideating, visualising* and *creating* data for Amelia and Nathan (cf. Yue, 2012), providing space for them to ask questions driven by curiosity rather than constrained by formality. In contrast, Ingrid and Naomi pursued an automated approach for cleaning the messy data, demonstrating more focused thinking about text data at the computer-extraction level (cf. Horton et al., 2023). Their disagreement about how to process the messy data could be viewed as them having different goals for their enquiry, that is statistical versus computational (cf. Thoma et al., 2018). We are also not sure that high school students would be as interested in spending so much time cleaning data computationally as Ingrid and Naomi, without having a personal and motivating "end goal" in sight. It was observed for all teachers that they articulated uncertainty about the quality of the data, but once in an "exploratory and discovery" mode, uncertainty due to sampling variation was not considered (cf. Gafney & Ben-Zvi, 2023). We don't consider this a weakness of the task, or for using messy data for statistical inquiry, but instead propose that the frameworks used for interpreting how learners engage with data may need re-thinking. As the use of statistical enquiry expands to include sources of data not traditionally used for teaching, more research is needed to better understand how teachers can engage their learners with messy data as part of statistical enquiry (cf. Kjelvik & Schultheis, 2019).

REFERENCES

Bakker, A., & van Eerde, D. (2014). An introduction to design-based research with an example from statistics education. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 429-466). Springer. https://doi.org/10.1007/978-94-017-9181-6_16

Barker, H., & Elrod, E. (2023). An analysis of K-8 pre-service teachers as data storytellers. In: E.M. Jones (Ed.), *Fostering Learning of Statistics and Data Science Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* International Association for Statistics Education.

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician, 72*(1), 2-10. https://doi.org/10.1080/00031305.2017.1375989

Chai, C. P. (2020). The importance of data cleaning: Three visualization examples. *Chance, 33*(1), 4-9. https://chance.amstat.org/2020/02/data-cleaning/

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Cummiskey, K., Kuiper, S., & Sturdivant, R. (2012). Using classroom data to teach students about data cleaning and testing assumptions. *Frontiers in Psychology, 3.* https://doi.org/10.3389/fpsyg.2012.00354

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal, 23*(1), 6-18. https://doi.org/10.1075/idj.23.1.03dig

Dvir, M., & Ben-Zvi, D. (2022). Students' actual purposes when engaging with a computerized simulation in the context of citizen science. *British Journal of Educational Technology, 53*(5), 1202-1220. https://doi.org/10.1111/bjet.13238

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

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. (2022). *Towards an integration of statistical and computational thinking: Development of a task design framework for introducing code-driven tools through statistical modelling*. PhD Thesis, University of Auckland. https://hdl.handle.net/2292/64664

Fergusson, A., & Pfannkuch, M. (2022). Introducing teachers who use GUI-driven tools for the randomization test to code-driven tools. *Mathematical Thinking and Learning, 24*(4), 336-356. https://doi.org/10.1080/10986065.2021.1922856

Finzer, W., & Reichsman, F. (2018). Exploring the essential elements of data science education. https://concord.org/newsletter/2018-fall/exploring-the-essential-elements-of-data-science-education/

Fry, K., & Makar, K. (2021). How could we teach data science in primary school? *Teaching Statistics, 43(*S1), S173-S181. https://doi.org/10.1111/test.12259

Gafny, R., & Ben-Zvi, D. (2023). Students' articulations of uncertainty about big data in an integrated modeling approach learning environment. *Teaching Statistics, 45,* S67-S79. https://doi.org/10.1111/test.12330

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics, 43,* S11–S22. 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., Sunbury, S., & Dussault, M. (2014). In praise of messy data. *The Science Teacher, 81*(8), 31. https://www.proquest.com/scholarly-journals/praise-messy-data/docview/1627727600/se-2

Hammett, A., & Dorsey, C. (2020). Messy data, real science. *The Science Teacher, 87*(8), 40-48. https://www.jstor.org/stable/27048170

Hardin, J. (2018). Dynamic data in the statistics classroom. *Technology Innovations in Statistics Education, 11*(1). https://doi.org/10.5070/T5111031079

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D. & Ward, M. (2015). Data science in statistics curricula: Preparing students to "think with data". *The American Statistician, 69*(4), 343-353.

https://doi.org/10.1080/00031305.2015.1077729
Holcomb, J., & Spalsbury, A. (2005). Teaching students to use summary statistics and graphics to clean and analyze data. *Journal of Statistics Education*, *13*(3). https://doi.org/10.1080/10691898.2005.11910567
Horton, N. J., Chao, J., Palmer, P., & Finzer, W. (2023). How learners produce data from text in classifying clickbait. *Teaching Statistics*, *45*, S93-S103. https://doi.org/10.1111/test.12339
Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE—Life Sciences Education, 18*(2), 1–8. https://www.lifescied.org/doi/10.1187/cbe.18-02-0023
Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal*, *16*(2), 191-212. https://doi.org/10.52041/serj.v16i2.190
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). https://doi.org/10.52041/serj.v21i2.41
Legacy, C., Zieffler, A., Fry, E. B., & Le, L. (2022). COMPUTES: Development of an instrument to measure introductory statistics instructors' emphasis on computational practices. *Statistics Education Research Journal, 21*(1). https://doi.org/10.52041/serj.v21i1.63
Lohr, S. (2014, August 18). For Big-Data Scientists, "Janitor Work" Is Key Hurdle to Insights. New York Times.
McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research*. Routledge. https://doi.org/10.4324/9781315105642
Ministry of Education. (2007). *The New Zealand Curriculum*. Learning Media.
Musyoka, J., Lunalo, J., Garlick, C., Ndung'u, S., Stern, D., Parsons, D., & Stern, R. (2017). Embedding Data Manipulation in Statistics Education. In: A. Molnar (Ed.), *Teaching Statistics in a Data Rich World Proceedings of the Satellite conference of the International Association for Statistical Education (IASE)*, International Association for Statistics Education.
Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician, 64*(2), 97-107. https://doi.org/10.1198/tast.2010.09132
Perez, L. & Lionberger, K. (2023). Opening the door to data science in STEM classrooms: How can we help all students navigate our data-rich world? https://ngs.wested.org/doortodatascience/
Rosenberg, J., Edwards, A., & Chen, B. (2020). Getting messy with data. *The Science Teacher, 87*(5), 30-35. https://www.jstor.org/stable/27048120
Rosenberg, J. M., Schultheis, E. H., Kjelvik, M. K., Reedy, A., & Sultana, O. (2022). Big data, big changes? The technologies and sources of data used in science classrooms. *British Journal of Educational Technology, 53(*5), 1179-1201. https://doi.org/10.1111/bjet.13245
Thoma, S., Deitrick, E., & Wilkerson, M. (2018). "It didn't really go very well": Epistemological framing and the complexity of interdisciplinary computing activities. In J. Kay & R. Luckin (Eds.), *Rethinking learning in digital age: Making the learning sciences count. Proceedings of the 13th International Conference of the Learning Sciences (ICLS)*, London, UK, (Vol. 2, pp. 1121–1124). International Society of the Learning Sciences.
Yue, K. -B. (2012). A realistic data cleansing and preparation project. *Journal of Information Systems Education, 23(*2), 205-216.
Wickham H. (2104). Tidy Data. *Journal of Statistical Software. 59*(1), 1–23. https://doi.org/10.18637/jss.v059.i10
Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science*. O'Reilly Media, Inc.
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
Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology, 13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

APPENDIX

Screenshot of survey from estimation180.com/day-1 (no longer available).

What's too LOW? *

What's too HIGH? *

Your estimate. *

Your reasoning. *

Do better than "I guessed."

Your name. *

Screenshots from shared Google sheet demonstrating responses from estimation180.com/day-1 survey (no longer available).

	A	B	C	D	E	F
1	Timestamp	LOWER limit of your range.	UPPER limit of your range.	Your strategic estimate.	Your reasoning.	Your name.
2	6/27/2014 13:36:55	4	7	5 ft 10in	The fence is probably 3 ft	Sam
3	6/29/2014 11:49:03	2"	10'	6"2"	I used the railing as a marker. Railings are about 3" tall	Scott
4	6/29/2014 20:09:32	5 feet	7 feet	6' 4"	He is taller than the fence and is very thin.	Derek
5	6/30/2014 5:55:48	3 feet	8 feet	6 feet	fence is probably 3 feet high	Bob
6	6/30/2014 7:51:21	72 in	90 in	80 in	compared to fence	k
7	6/30/2014 15:32:52	1 Ft	10 Ft	6 Ft	Fences are usually about 3 ft	I
8	7/2/2014 8:53:37	100 in	10 in	72 in	compared to the fence	Strohl
9	7/2/2014 11:10:35	Arms	Head	204	He is tall	Evan
10	7/2/2014 15:08:21	10	40	28	no gray hair	Bubba
11	7/2/2014 18:38:10	5 feet	7 feet	6' 2"	You look tall	Blythe
50380	6/17/2019 17:39:06	5 feet	8 feet	6 foot 1	i've seen other fences of that type that are around 4 feet tall, so I added about half the height of the fence to make my guess.	Laura
50381	6/17/2019 18:33:21	6'	10'	8	he looks tall	Jess
50382	6/18/2019 1:34:06	30 cm	10 m	199cm	I am 174cm and he quite a bit taller than me	Mrs Roberts
50383	6/18/2019 2:48:31	5 feet	7 feet	6 feet	Yard length and multiplied it	Harry
50384	6/18/2019 7:22:55	4ft	8ft	6ft	the fence looks about 4ft and he is taller than that.	Elle