Pre-service mathematics teachers' experiences of acquiring and organizing image-based data

Sibel Kazak

▶ **To cite this version:**

Sibel Kazak. Pre-service mathematics teachers' experiences of acquiring and organizing image-based data. Twelfth Congress of the European Society for Research in Mathematics Education (CERME12), Feb 2022, Bozen-Bolzano, Italy. hal-03751833v1

HAL Id: hal-03751833

https://hal.science/hal-03751833v1

Submitted on 15 Aug 2022 (v1), last revised 19 Nov 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pre-service mathematics teachers' experiences of acquiring and organizing image-based data

Sibel Kazak

Pamukkale University, Turkey; skazak@pau.edu.tr

There is a growing interest in engagement with non-traditional data generated from a variety of sources, such as sensory devices, smartphones/watches, and social media tools. This paper reports on experiences of a group of pre-service mathematics teachers in acquiring and organizing image-based data to analyze categorical data as part of a course assignment. To identify their actions to formulate statistical questions and to transform observations from butterflies' photos to cases when answering these questions, the video recordings of retrospective interview and artifacts generated during the data exploration were analyzed. The findings suggest that considering data/data handling became an important component of the statistical investigation before formulating statistical questions. While the use of hierarchical tables in structuring data by hand appeared to be intuitive, it could be challenging to use such a structure already built in a digital data organizing tool.

Keywords: Statistical inquiry, data acquiring, data organizing, image-based data, categorical data.

Introduction

Understanding and reasoning with data become increasingly essential part of our everyday lives as we need to handle information related to global issues on health, environment and so on, such as the COVID19 pandemic and wild fires due to extreme weather conditions. The nature of data also evolves with the advanced digital technologies. Large quantities and different forms of data are generated from a variety of sources, such as sensory devices, smartphones/watches, social media tools etc., every day. To analyze and make predictions from these data require new skills. Therefore, an important goal of statistics education is to develop necessary data skills starting from early school years.

Statistics is typically taught as part of school mathematics curriculum and there has been a shift from simply computing numerical and graphical representations of data to an inquiry-based approach (Watson, Jones, & Pratt, 2013). In this approach to the teaching of statistics, the focus is on the statistical problem-solving process (Franklin et al., 2007; Bargagliottiet al., 2020). According to *the Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*, the purpose of this process is "to collect and analyze data to answer statistical investigative questions" (Bargagliotti et al., 2020, p. 13). As seen in Figure 1, it has four components that are interlinked: (1) Formulate statistical investigative questions, (2) Collect/consider the data, (3) Analyze the data, and (4) Interpret the results. In the Turkish National Mathematics Curriculum in middle school (Milli Eğitim Bakanlığı, 2018), "Data Handling" strand include these components across different grade levels. In the 5th grade, the learning outcomes involve formulating research questions, collecting data (restricted to one variable and discrete data), displaying data with frequency table and bar graph, and interpreting the results from these representations. In the 6th grade, this process involves comparing two data sets (limited to discrete data) with the use of range and arithmetic mean. In the 7th and 8th grades, the emphasis is on analyzing

data that includes representing and interpreting data using appropriate graphs (line graph, pie chart, and bar graph) as well as calculating and interpreting the measures of central tendency (mean, median and mode).

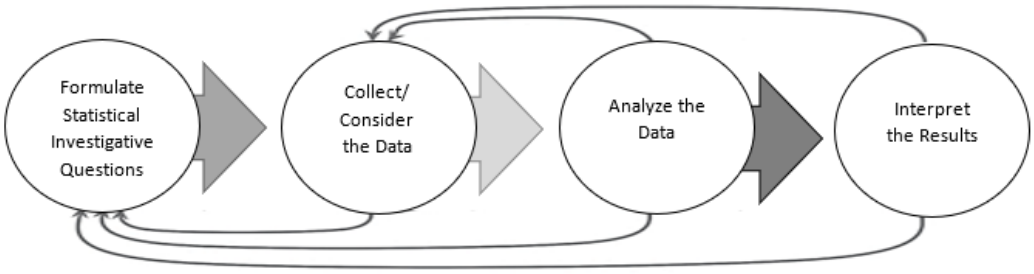

Figure 1: Statistical problem-solving process (reproduced from Bargagliotti et al., 2020)

Theoretical background

Traditional statistical education focuses on data from random samples and making inferences from a sample to an unknown population. In today's digital age, however, data come in various forms (big, messy, unstructured data, repurposed data, image/text/sound-based data etc.) and do not necessarily comply with the structure required for statistical inference in the traditional sense (Gould, Bargagliotti, & Johnson, 2017). Along with the emerging field of data science, statistics education tends to focus more on the engagement with such non-traditional data sets.

Image-based data, i.e., the use of photographs as data, becomes increasingly common in daily encounters. The GAISE II report (Bargagliotti et al., 2020) acknowledges the use of this kind of non-traditional data as part of the statistical problem-solving process and helping students to make sense of these non-traditional data. As an example, Bargagliotti et al. use data from Dollar Street website as part of the Gapminder Foundation project (www.gapminder.org/dollar-street). The Dollar Street website displays the world as a street ordered by monthly income per person in the family from different countries and continents, and currently uses 43685 photos of 422 families in 66 different countries. Using these image-based data, students are expected to investigate "How are people's concepts of family and living spaces similar or different across the world?" (Bargagliotti et al., 2020, p. 63). The guidelines in the GAISE II report offers some instructional ideas with regard to the use of these image-based data through each four components of the statistical problem-solving process mentioned above. Bargagliotti et al. (2020) suggest that this kind of non-traditional, multivariate data sets requires a great amount of data exploration time to make observations and wonderings to analyze data. Engagement with other types of non-traditional (large and complex) data from secondary data sources, such as participatory sensing data (Gould et al., 2017) and public data sets (Wilkerson, Lanouette, & Shareff, 2021), also appears to require more emphasis on considering data and data preparation phases in statistical investigation process. As seen in these studies, such existing multivariate data sets are constructed and made publicly available by others with a particular purpose and investigators often repurpose them to explore new questions. To do so, data handling and structuring becomes an important part of statistical investigations.

The Collect/Consider the Data component of the statistical problem-solving process involves recording/acquiring, measuring and organizing the data to answer statistical questions with the

acknowledgment of variability in data (Bargagliotti et al., 2020). Konold, Finzer, and Kreetong (2017) call attention to this important process of transforming observations into data and focus on table format as a typical means of data recording and structuring either by hand or using software. Two common table formats in statistics are case-data table and summary table as called by Konold et al. (2017). While case-data tables are mainly used for collecting and storing raw data, summary tables usually include only some of the information gathered and are organized in a way to compare groups or detect trends in the data. The data analysis software, such as TinkerPlots (Konold & Miller, 2011) and CODAP (https://codap.concord.org/), uses attribute-based structure for recording data where each column includes a variable and each row holds one observation, called a 'case', as described by Konold et al. (2017). However, entering the information collected to organize the data in such a format can be challenging for the learners. Konold et al. point out the need for thinking of data based on attributes which requires considering properties of the observations along with their values case by case. Their study revealed that students and adults tended to create two types of case tables, flat and hierarchical, when they were asked to construct data sheets to record and organize the data by hand from the pictures showing traffic along two road segments (including information about the vehicle type, speed, direction etc.) in a given time and date. Compared to the flat tables (attributes as columns and cases as rows) as seen in most data analysis software, the hierarchical tables include the cases at more than one level created by nested structure. This type of hierarchical structure of observations is available in CODAP and allows exploration of multilevel data sets.

Given the scarcity of research on how students/adults come to record and organize data, especially with non-traditional image-based data, as part of a statistical problem-solving process, the aim of this paper is to present a case of three pre-service mathematics teachers who chose to use photos of butterflies as data for analyzing categorical variables. More specifically, the following question is investigated in this exploratory study: What actions does the group of pre-service teachers take to transform observations of butterflies' photos to cases to answer statistical questions related to categorical data?

Method

As part of a 3rd year course, titled 'Teaching of Statistics and Probability', in the mathematics education program at a public university in Turkey, the pre-service teachers were given a group assignment to design a statistical investigation activity that is aligned with the learning outcomes in the middle school mathematics curriculum mentioned earlier. More specifically, small groups of pre-service teachers were asked to plan a class activity at a particular grade level including all components of the statistical problem-solving process shown in Figure 1 by acquiring real data in a context interesting to middle school students and using CODAP software. The task assigned to the groups varied by (1) the nature of statistical investigation question (beyond students themselves and classmates (van de Walle, Karp, & Bay-Williams, 2019), experimentation, and change over time), (2) variable types (categorical and discrete/continuous-numerical), and (3) number of groups (single and two or three groups for comparison). Based on these given conditions, they identified the grade level according to the curricular objectives. Two of these groups were expected to focus on questions beyond students themselves and classmates, analyzing categorical data only, and comparing two or three groups using existing data sets in publicly available internet sources. One of these groups,

including Arya, Yelda and Hale (pseudonyms), chose image-based data using Kelebek-Türk (Butterflies-Turk) group's website (https://www.kelebek-turk.com/) while the other group took a more traditional approach and designed their data investigation activity involving categorical data collected through survey questions on readings books.

Due to the unique nature of transforming image-based data into a statistical analysis of categorical data with group comparisons, a retrospective interview was conducted with these three pre-service teachers to gain insights into their approaches when acquiring and organizing information from photographs of butterflies. Other artifacts, such as the documents created during the data exploration and organization, were also collected for analysis to answer the research question. Recordings of group interview and the student artifacts were analyzed using the perspectives on collect data/consider data suggested by Bargagliotti et al. (2020) and Konold et al. (2017). First, the critical moments during formulating and answering statistical questions were identified in the data. Then, these processes were summarized to give an overview of pre-service teachers' actions related to acquiring and organizing information from the image-based data.

Findings

The group of pre-service teachers, Arya, Yelda and Hale, began with considering possible data appropriate for analyzing categorical data with comparing groups, which was given in the course assignment. Due to the difficulty in finding at least three categorical variables in data sets from publicly available data sources, such as Turkish Statistical Institution website, they stated that some of the data explored during the course, such as the ladybug data (Bargagliotti et al., 2020, p. 32), inspired them to consider similar data about other animals. Yelda came up with the idea of searching butterfly data on the internet and found the group of Kelebek-Türk that provides photographs and observational data provided by their members from different age and occupation groups. After reading the "About us" information on the website, the pre-service teachers got their initial information about how the data were collected. Then, they decided to use some of these photographs to construct their own observations involving categorical data.

According to the Kelebek-Türk website, there are 416 species of butterflies under nine different families in Turkey and their members have photographed 393 of them. As seen in Figure 2, there are nine butterfly families with different number of species, such as the Family Hesperiidae has 43 species, observed in Turkey and Figure 3 shows a sample of photos of different butterfly species belonging to the Family Hesperiidae.

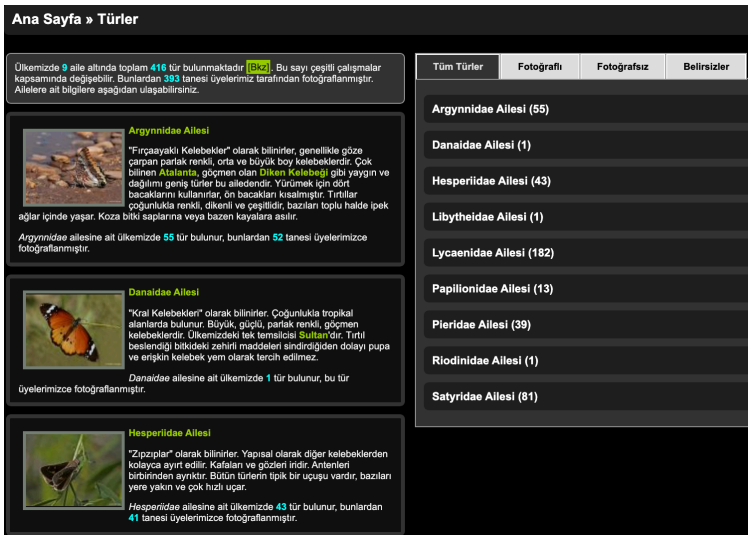

Figure 2: Screenshot of the website displaying the information about all nine different butterfly families

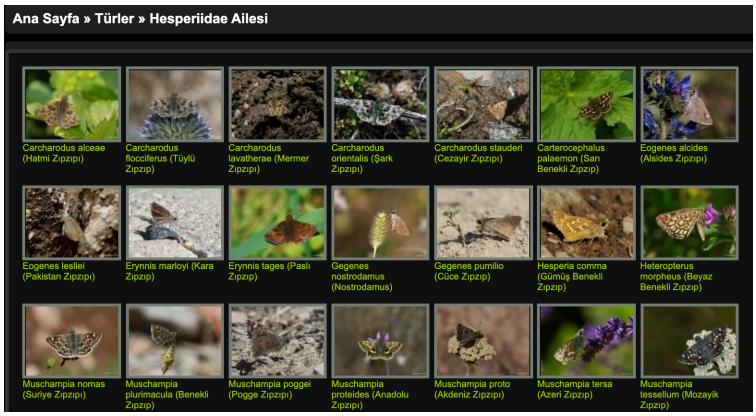

Figure 3: Screenshot of the website displaying the photographs of the butterfly species belonging to the Family Hesperiidae

Following the exploration of the information given in these pages, the pre-service teachers decided to focus on two families with a similar number of butterfly types (the Family Hesperiidae with 43 types and the Family Pieridae with 39 species) for comparison purposes given in the assignment. To collect information about these two families, they then began to click on each photo as seen in Figure 3. As an example, Figure 4 shows various information about Muschampia poggei from the Family Hesperiidae, including numerical (frequencies of photographs by cities), geographic (distributions of butterflies across Turkey), photographic (captured images of butterflies with/without identified gender), tabular (months they were seen) and graphical (distribution of altitudes seen by months) representations. From this information, they selected to focus on the image of butterfly with unidentified gender (since not all types had gender information) to generate attributes, such as wing color and wing appearance, which were considered as distinguishing characters of butterflies, and the seasons the butterflies seen, which involved grouping of months data given on the website.

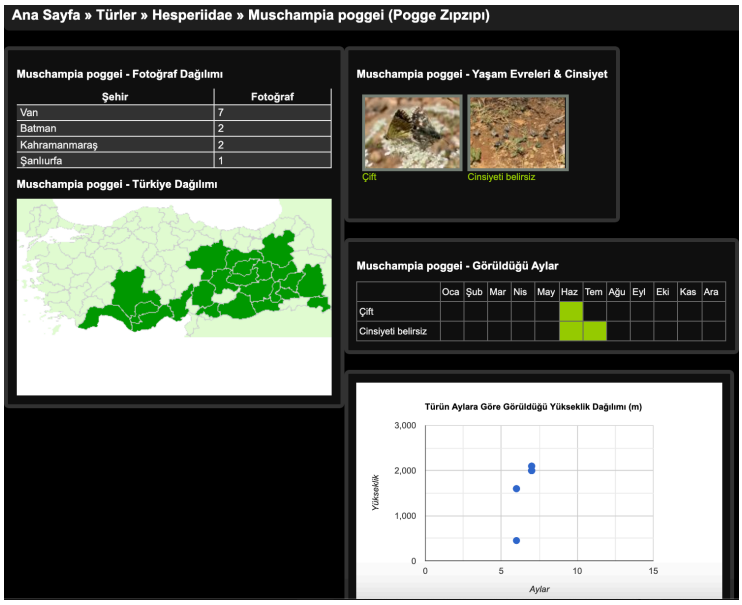

Figure 4: Screenshot of the website displaying various information about Muschampia poggei from the Family Hesperiidae

After this extensive exploration on the website, the pre-service teachers began to make an ordered list for each species of butterflies in both families, including the name of the butterfly, its photo, and the season it is seen, on a Word document. As part of the course assignment, they needed to use CODAP to analyze the data to answer questions considered in the Formulate statistical investigative questions phase of the statistical problem-solving process in Figure 1. Based on the conditions of the assigned task (beyond the classroom questions, categorical variable, and two or three groups) three statistical questions were formulated during this exploration: 1) What colors are the wings of the Family Hesperiidae and the Family Pieridae? 2) How do the appearances of the wings of the Family Hesperiidae and the Family Pieridae look like? 3) What seasons do the Family Hesperiidae and the Family Pieridae tend to be seen? Then, they constructed a table, as seen in Figure 5, displaying the family name, species name, wing color, wing appearance, and season seen in columns respectively and the case values in rows.

Kelebek Aileleri	Kelebek türleri	Kanat Rengi	Kanat görünümü	Yaşadıkları Mevsimler
Hesperidae (Zıpzıplar ailesi)	Pyrgus cinarae (Güzel Zıpzıp)	Siyah	Desenli	Yaz
	Pyrgus serratulae(Zeytuni zıpzıp)	Kahve	Desenli	İlkbahar-Yaz
	Eogenes alcides (Alsides Zıpzıpı)	Beyaz	Düz	Yaz
	Pyrgus sidae (Sarıbandlı Zıpzıp)	Siyah	Desenli	İlkbahar-yaz
	Pyrgus jupei (Kafkasya Zıpzıpı)	Kahve	Desenli	Yaz

Figure 5: Part of the table created for organizing the data

Organizing the data as shown in Figure 5 made entering them into CODAP table relatively easier due to its case-based structure. One difficulty the pre-service teachers encountered in this process was having more than one value for season variable for some cases, such as Pyrgus serratulae seen in both

spring and summer. Therefore, they chose to create two different data sets, one including 70 cases of butterflies in two families with family names, wing color and wing appearance variables (Figure 6, table on the left) and the other including 120 cases with season variable for two families (Figure 6, table on the right). Due to the repeated cases for each season a butterfly seen, the number of cases was increased in the second data set. The pre-service teachers decided to separate multiple seasons since they wanted to answer their third question: What seasons do the Family Hesperiidae and the Family Pieridae tend to be seen? In either of the data sets, they also did not include species name which was displayed in the table in Figure 5 as they considered that having the species names would not allow to see two categories (the families) which were compared in the data analysis.

Yeni Veri Seti

Durumlar (70 cases)

index	Kelebek Aileleri	Kelebeklerin Kanat Renkleri	Kelebeklerin Kanat Görünümü
1	Hesperidae (Zıpzıplar ailesi)	Siyah	Desenli
2	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
3	Hesperidae (Zıpzıplar ailesi)	Beyaz	Düz
4	Hesperidae (Zıpzıplar ailesi)	Siyah	Desenli
5	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
6	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
7	Hesperidae (Zıpzıplar ailesi)	Boz	Desenli
8	Hesperidae (Zıpzıplar ailesi)	Turuncu	Desenli
9	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
10	Hesperidae (Zıpzıplar ailesi)	Boz	Desenli
11	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
12	Hesperidae (Zıpzıplar ailesi)	Boz	Desenli
13	Hesperidae (Zıpzıplar ailesi)	Siyah	Düz
14	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
15	Hesperidae (Zıpzıplar ailesi)	Beyaz	Desenli
16	Hesperidae (Zıpzıplar ailesi)	Boz	Desenli
17	Hesperidae (Zıpzıplar ailesi)	Boz	Düz
18	Hesperidae (Zıpzıplar ailesi)	Siyah	Desenli
19	Hesperidae (Zıpzıplar ailesi)	Kahve	Desenli
20	Hesperidae (Zıpzıplar ailesi)	Boz	Düz

Yeni Veri Seti

Durumlar (120 cases)

index	Kelebek Aileleri	Kelebeklerin Uçtuğu Mevsimler
1	Hesperidae (Zıpzıplar ailesi)	İlkbahar
2	Hesperidae (Zıpzıplar ailesi)	İlkbahar
3	Hesperidae (Zıpzıplar ailesi)	İlkbahar
4	Hesperidae (Zıpzıplar ailesi)	İlkbahar
5	Hesperidae (Zıpzıplar ailesi)	İlkbahar
6	Hesperidae (Zıpzıplar ailesi)	İlkbahar
7	Hesperidae (Zıpzıplar ailesi)	İlkbahar
8	Hesperidae (Zıpzıplar ailesi)	İlkbahar
9	Hesperidae (Zıpzıplar ailesi)	İlkbahar
10	Hesperidae (Zıpzıplar ailesi)	İlkbahar
11	Hesperidae (Zıpzıplar ailesi)	İlkbahar
12	Hesperidae (Zıpzıplar ailesi)	İlkbahar
13	Hesperidae (Zıpzıplar ailesi)	İlkbahar
14	Hesperidae (Zıpzıplar ailesi)	İlkbahar
15	Hesperidae (Zıpzıplar ailesi)	İlkbahar
16	Hesperidae (Zıpzıplar ailesi)	İlkbahar
17	Hesperidae (Zıpzıplar ailesi)	İlkbahar
18	Hesperidae (Zıpzıplar ailesi)	Yaz
19	Hesperidae (Zıpzıplar ailesi)	Yaz
20	Hesperidae (Zıpzıplar ailesi)	Yaz

Figure 6: The data set tables created in CODAP

Concluding remarks

This paper presented insights from a group of pre-service teachers' experience of acquiring and organizing image-based data to analyze categorical data by comparing two groups as part of a course assignment. As pointed out by Bargagliotti et al. (2020), the pre-service teachers spent a great deal of time to explore the non-traditional data source including photographs and other types of data collected by others for specific purposes as mentioned in the Kelebek-Türk website. In accord with the findings of Gould et al. (2017) and Wilkerson et al. (2021), considering data and data handling became an important component of the statistical investigation before and after formulating statistical questions. They needed to consider attributes/variables and criteria for identifying categoric values for them, such as assigning values, plain or pattern, for wing appearance based on the dots or lines seen in the photos of butterflies. The pre-service teachers did not seem to have any issue with preparing the data for the analysis in CODAP as they were able to consider the data as attribute-based. This could be due to their prior experiences with using CODAP for data analysis during the Statistics course taken in the previous semester. However, transforming the initial table created by hand (Figure 5) into a single data set table in CODAP was challenging for them. Since their initial table included subsets of data within the season variable, the flat table did not work. Instead, they

needed to create hierarchical data sets, which is possible in CODAP, but this feature was not used in the course before. It was also evident in the work of Konold et al. (2017) that the use of hierarchical tables in structuring data by hand was more intuitive, but it could be challenging to use such a structure that is already built in CODAP environment.

References

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* American Statistical Association.

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

Konold, C., & Miller, C. D. (2011). *TinkerPlots 2.0: Dynamic data exploration*. Key Curriculum.

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

Milli Eğitim Bakanlığı (2018). *Matematik Öğretim Programı (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar)[Mathematics Curriculum (Primary and Middle Schools 1, 2, 3, 4, 5, 6, 7 and 8 grades)]*. MEB Yayınları.

Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2019). *Elementary and middle school mathematics: teaching developmentally (10th edition)*. Pearson Education UK.

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*, 1–19. https://doi.org/10.1080/10986065.2021.1922838