



# Data-ing process with image-based data: variable identification and generation

Sibel Kazak<sup>1</sup>

Accepted: 8 January 2025 / Published online: 21 January 2025  
© The Author(s) 2025

## Abstract

Using photographs as data, which involves making observations from images and organizing them into variables to answer statistically investigative questions, is recommended for K–12 level statistics education. Research is needed to support pre-service mathematics teachers' experiences with exploring such image-based data. With this task-based interview study, the goal was to shed light on (1) the pre-service mathematics teachers' data-ing actions during identifying and generating variables in relation to data familiarization, question posing, and data organization components and (2) how pre-service mathematics teachers identified and generated variables in the process of data-ing. Data from video recordings, transcripts of the interview sessions for each pair, and their work with photos on the shared online document, that is, groupings and questions posed, were analyzed using a progressive focusing approach. The results showed that the data-ing actions during identifying and generating variables with data familiarization, question posing, and data organization included observing, interpreting, conjecturing, inferring, comparing, grouping, ordering, questioning/question posing, relating variables, categorizing variables, and measuring variables. The pairs used various combinations of multiple actions while data-ing. There were two types of variable identification: (1) observational variables based on visual judgment or metadata and (2) inferential variables based on personal interpretation. The latter type presented a tension between the variable defined and how to measure it objectively, as well as challenges in writing clearly defined variables when posing questions.

**Keywords** Data investigation · Image-based data · Variable identification · Variable generation · Pre-service mathematics teachers

## 1 Introduction

In statistics, data can be defined as “a collection of numbers or other pieces of information to which meaning has been attached” (Utts, 1996, p. 16) and which need to be interpreted to become useful. In relation to this definition, data-ing can be viewed as engaging with data to describe and organize quantities or qualities, either by working with predefined variables or through making observations, to explore statistically investigative questions before analyzing data. How we experience this process yet might differ with respect to the data types we encounter in school or daily life.

On one hand, traditional data are generally structured forms of data that are typically numerical or categorical and organized in a tabular format with rows and columns, where each column represents a specific variable, and each row corresponds to an observation that can be expressed as a number or category. As traditional data are collected with predefined variables, data-ing involves identifying and measuring variables during data collection, and organizing and structuring data in preparation for analysis based on statistical questions posed.

On the other hand, contemporary non-traditional data tend to be messy, unstructured, and repurposed, and come in various forms, such as image/text/video/sound-based data from social media posts and other digital platforms. *The Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education* (Bargagliotti et al., 2020) recognize the need for students to understand the concept of data, including these new forms as they make better sense of the

---

✉ Sibel Kazak  
skazak@metu.edu.tr

<sup>1</sup> Department of Mathematics and Science Education,  
Faculty of Education, Middle East Technical University,  
06800 Ankara, Türkiye

world around them. For example, the Dollar Street website ([www.gapminder.org/dollarstreet](http://www.gapminder.org/dollarstreet)) provides visual documentation of the living conditions of families worldwide, such as photographs/videos of household items, living environments, daily activities, and so on. Each image is also associated with a specific monthly income in US dollars, location (country), and topic, such as gardens, toys, plates of food, sitting and watching TV, shampoo, and so forth.

The GAISE II report illustrates how to engage students in using photographs as data from Dollar Street through an investigative process, which involves formulating a statistically investigative question, collecting/considering data, analyzing data, and interpreting results (Bargagliotti et al., 2020). In this example, data-ing requires extensive questioning and exploration for making observations and organizing those observations into variables to explore statistically investigative questions before analyzing data. Within the Dollar Street data set, the variables are captured by photographs available under each topic. So, students can make various observations about these images, define variables, and pose investigative questions that can lead to data analysis. For instance, when given the photos of families from Dollar Street to explore “What does the typical family photograph look like for the Dollar Street families?” (Bargagliotti et al., 2020, p.64), students are expected to notice that the family members stay close together in all the photos, young children are taken in the arms of a family member, and the sizes of the families vary. Here, Dollar Street photos can be considered cases that provide specific context for making observations to identify variables and generate image-based data about those variables to explore a statistically investigative question. Hence, data-ing with images can be complex.

According to the GAISE II report, there is a need to develop school students’ understanding of and engagement with image-based data and other non-traditional data types. However, pre-service teachers often have no experience with such data types in their schooling. It seems then that pre-service teachers need to be equipped with knowledge and skills to guide their students through data-ing with different data types, and teacher educators need to understand better how to support them in developing these skills. Hence, this study aims to delineate pre-service mathematics teachers’ understanding of photographs as data and their data-ing process, focusing on variable defining, that is, identification and generation of variables when investigating a set of selected images from the Dollar Street website.

## 2 Theoretical background

This section provides an overview of data sources and relevant data investigation cycles in the literature, a review of previous studies on conceptualization of variability and

variables, and a perspective on image-based data and data-ing process.

### 2.1 Data sources and relevant data investigation cycles

The data sets used in statistics education have changed over the years (Rubin, 2021), especially with the availability of new data sources (Lee & Wilkerson, 2018) and the implementation of data science in K–12 education around the world (e.g., Introduction to Data Science (IDS) curriculum ([www.introdatascience.org/](http://www.introdatascience.org/)), International Data Science in Schools Project (IDSSP) curriculum framework (<http://www.idssp.org/>), and the ProDaBi project learning materials (<https://www.prodabi.de/en/>)). Within this context, the approaches to data investigations at the school level tend to show slight variations in using primary and secondary data.

The PPDAC (Problem, Plan, Data, Analysis, Conclusion) cycle (Wild & Pfannkuch, 1999) provides a general framework often used in teaching statistical investigation involving data collection in K–12 education. Within the PPDAC cycle, a data investigation process typically starts with defining the problem and posing questions that can be answered through collecting data. The investigation involves collecting and recording data (primary data) after planning for what variables (attributes) to measure, how to measure them, and data collection procedures. In the data analysis phase, appropriate data representations and statistical summaries are used to explore and compare distributions. Finally, the conclusion phase entails interpreting data and making inferences about the population. Arnold (2013) adapted the PPDAC cycle for investigations when the data collected by others are given to the students (secondary data). In this modified version, a data investigation begins with interrogating the data and the original data collection plan: how the data were collected, what variables were recorded, what was measured, whom the data were collected from, and so on. Subsequent stages are the problem, analysis, and conclusion phases of the PPDAC cycle.

Similarly, the statistical problem-solving process described in the GAISE II report (Bargagliotti et al., 2020) has four components that are interlinked: (1) formulating statistically investigative questions, (2) collecting/considering the data, (3) analyzing the data, and (4) interpreting the results. The “collecting/considering data” component emphasizes interrogating the available (primary or secondary) data. When provided with secondary data, students are expected to ask questions about “how the variables differ by type, the possible outcomes of each of the variables, and how the data were collected” (ibid., p.14) and then pose statistical questions about the data set.

From a data science education perspective on solving problems using complex, large data sets, the data cycle

presented in the IDS curriculum at the secondary school level entails a four-stage process of a statistical investigation: pose questions, consider data, analyze data, and interpret data. The “collect data” stage in the traditional data investigation process is replaced with the “consider data” in this data cycle (Gould et al., 2017). More specifically, this phase involves observing and recording data through participatory sensing<sup>1</sup> or other means or interrogating previously collected data by asking questions such as: What was the purpose of this study? Who/what is the data about? What variables were measured, and how were they measured?

A comprehensive framework for data investigations has also been proposed in data science education at the K–12 level (Lee et al., 2022). This data investigation process involves experiences with large, complex data often collected by others for other purposes. It includes six inter-related phases: frame a problem, consider and gather data, process data, explore and visualize data, consider models, and communicate and propose actions. Like the other cycles mentioned above, this data investigation starts with posing an investigative question within a context. The “Consider and gather data” phase involves considering types of data, primary or secondary, required to answer the question and identifying which data are relevant and useful for solving the problem. Understanding the variables and how they are measured is also needed. Since data in the real world do not come in tidy table formats (Lee et al., 2024), the “Processing data” phase involves obtaining data in a usable form, creating new variables from existing ones, sorting/grouping/filtering data, and so on.

These different approaches to data investigation at the school level have several commonalities relevant to data-ing (the process before analyzing data), such as posing a question that can be answered with data and interrogating the data/data collection. Depending on the form of data obtained (primary or secondary data), students need to either plan what variables to measure and how to measure them or ask questions about what variables were recorded, how they were measured, the types of the variables, and how data were collected. Moreover, back-and-forth movements between question posing, collecting/considering data, and exploring data are noted in these data investigation cycles. On the other hand, the more recent ones related to data science education pay particular attention to complex, non-traditional data forms and managing and processing these data, including creating new variables.

<sup>1</sup> According to the IDS Curriculum, Participatory Sensing is a data collection method in which individuals or groups use mobile devices or other sensors to explore various aspects of their lives systematically (<https://www.introdatascience.org/>).

## 2.2 Conceptualization of variability

The aforementioned data investigation cycles support the process of students’ learning and adopting the practice of statistics at the school level, which includes (1) asking, understanding, and refining statistical questions, (2) focusing on samples and sampling for data collection, (3) using visual representations of data and summarizing data for data analysis, and (4) making informal inferences (Watson et al., 2018). Recognizing, understanding, and quantifying variability in data, that is, “the [varying] characteristic of the entity that is observable” (Reading & Shaughnessy, 2004, p.202), is salient in these practices. According to the GAISE II report (Bargagliotti et al., 2020), formulating a statistical question requires anticipating variability; planning data collection involves an acknowledgment of variability in data; data analysis includes accounting of variability with the use of distributions; and making interpretations entails considering the variability in the data. So, understanding how to attend to variability in all components of the data investigation process is essential in teaching how to work with data (Bargagliotti & Eubanks-Turner, 2024).

Previous studies appeared limited to how pre-service teachers use the notion of variability only when analyzing numerical data presented in graphs and making informal inferences. For example, research showed pre-service teachers’ tendency to attend to variation as a characteristic of distributional shape rather than as a measure when comparing numerical data distributions (Makar & Confrey, 2005). In another study, pre-service teachers’ use of graphical representations highlighting distributional features of the numerical data appeared to support their comparison of distributions in light of the variability (Leavy, 2006). Since real-world data tend to be multivariate, including many categorical variables (Higgins et al., 2023), and come in different forms, including images (Lee & Wilkerson, 2018), we need research addressing such non-traditional, mainly categorical data and providing opportunities to consider variability also prior to analyzing data, that is when posing questions and considering variables during the data-ing process with images.

## 2.3 Focus on variables

There is a growing interest in engagement with non-traditional data generated from a variety of secondary data sources, such as participatory sensing data (Gould et al., 2017), public data sets (Wilkerson & Laina, 2018; Wilkerson et al., 2021), and complex multivariate civic data (Engel, 2017). These types of investigations generally involve repurposing the presented multivariate data sets to explore new questions and lead to more emphasis on considering data and data preparation in the statistical investigation process.

What seems common in these studies is examining the fundamental details about the variables available in the data set by asking questions to interrogate the data: What variables are measured? How are the variables defined? How are they measured? What values can a variable assume? How do variables co-vary? These interrogative questions help to make sense of the variables in the data set and understand their context before formulating statistically investigative questions and analyzing the given data.

In a study conducted by Gould et al. (2017), secondary teachers were provided with a big and complex data set (17 variables and 2600 observations) collected on mobile devices by others. Researchers focused on teachers' actions and approaches to asking statistical questions as they explored these repurposed data within the IDS data cycle. The findings showed that statistical questions that could be answered with the available data were likely to come from examining variables in the "consider data" phase and their analyses through visualizing distributions in the "analyze data" phase to refine their question. These back-and-forth movements indicated the importance of clearly stated variables in the statistical questions.

When using complex available data to address specific statistical questions, understanding and working with variables are also necessary to create new groupings or measures, known as data moves (Erickson et al., 2019), before data analysis. Wilkerson et al. (2021) focused on this phase, called data preparation, which is "the evaluation and manipulation of a dataset in preparation for analysis" (p.315) as a component of analyzing large, public data sets. Their research suggested that students' certain data moves, such as grouping data by a variable or converting existing variables (measures) into new ones in preparing public data sets for data analysis, helped them consider the variability in the data. For example, "[s]orting emphasized natural variation, filtering highlighted interrelationships between measures, and calculating made existing patterns of variation more interpretable and revealed new patterns of variation to explore." (Wilkerson et al., 2021, p. 16).

More recent studies have focused on statistical investigations with other types of non-traditional data, such as data produced from text (Horton et al., 2023) and image (Buehring & Grando, 2023; Fergusson & Pfannkuch, 2024; Kazak, 2022; Kazak et al., 2022). In Horton et al.'s (2023) study, data production was viewed as feature creation from text. The study addressed identifying features in the need for sorting email/web page headlines into clickbait and news. The work of the pairs of undergraduate students in a series of tasks revealed that the features were identified in three categories related to function, content, and form of clickbait. The design of the tasks also elicited different types of features, such as human-perceivable features and computer-detectable features, necessary in understanding text as data

and text classification models. For instance, some features of clickbait identified by students, such as "making people curious," can only be perceived by humans, whereas the features that the computer can detect tend to be related to the structural elements of the text and some design rules and more objective (Horton et al., 2023).

Fergusson and Pfannkuch (2024) reported on a study of secondary teachers developing a decision rule to classify the given sets of grayscale photos as light or dark or as high or low contrast by dichotomizing a numeric variable based on the distributions of greyscale values provided as dot plots and box plots. In the initial task of linking the photos to the sample distributions of grayscale values, teachers tended to use visual proportional and distributional reasoning to estimate the proportions of different shades of gray for each image with qualitative descriptions like "more white" and "pure black" rather than numeric values. However, in the classification model of high/low contrast grayscale photos, the task required them to use an aggregate measure of a quantitative variable, such as mean, median, and standard deviation, to classify image data as one of the two categorizations. As in Horton et al. (2023) study, teachers' approach to classifying photos involved connecting features based on human judgment of a high contrast photo to features of the distributions like a skewed or bimodal shape, and to computer detectable features, such as the difference between the mean and median for a random sample of grayscale values.

Other studies by Buehring and Grando (2023) and Kazak et al. (2022) examined how students investigate the multivariate data set of Dollar Street images (<https://www.gapminder.org/dollar-street>). In Buehring and Grando (2023), Brazilian children (7–8-year-old) explored the images on the Dollar Street website to draw conclusions about how people live around the world. Children's explorations of a particular topic available in Dollar Street, such as people's pets, tend to involve looking at specific and general cases, identifying similarities and differences, comparing and classifying images, and expressing perceived patterns and unusual cases in the multivariate data set. Classifying the images of pets into distinct groups appeared to generate discussions among children based on perceived attributes. For instance, while one classification involved groups of birds and chickens, another categorized birds based on whether they lived in a cage or not.

Moreover, Kazak et al. (2022) described the sequence of actions supporting students' statistical investigation for Dollar Street image data in three studies conducted at different educational levels (primary, secondary, and university) from different countries (Australia, Colombia, and Türkiye). The findings showed that identifying possible variables and values was supported during sorting and grouping the images in the data familiarization process, posing questions that can be answered with the set of available images, and organizing

images to analyze the data to answer those questions. Similar to the results of Buehring and Grando (2023), this study also revealed challenges in categorizing some images, such as whether considering a goat as a pet or a food source and defining the level of technology inherent in two different toothbrush images displaying a finger and a twig. Another study (Kazak, 2022) showed that when working with the photos of different butterfly types observed in Türkiye, a group of pre-service mathematics teachers generated variables after an extensive familiarization with context and data set and while posing questions to be answered with a set of images. Moreover, the variable generation appeared to be subjective and depended on observers' perception of attributes related to the images of butterflies within their contextual knowledge.

#### 2.4 Image-based data and data-ing

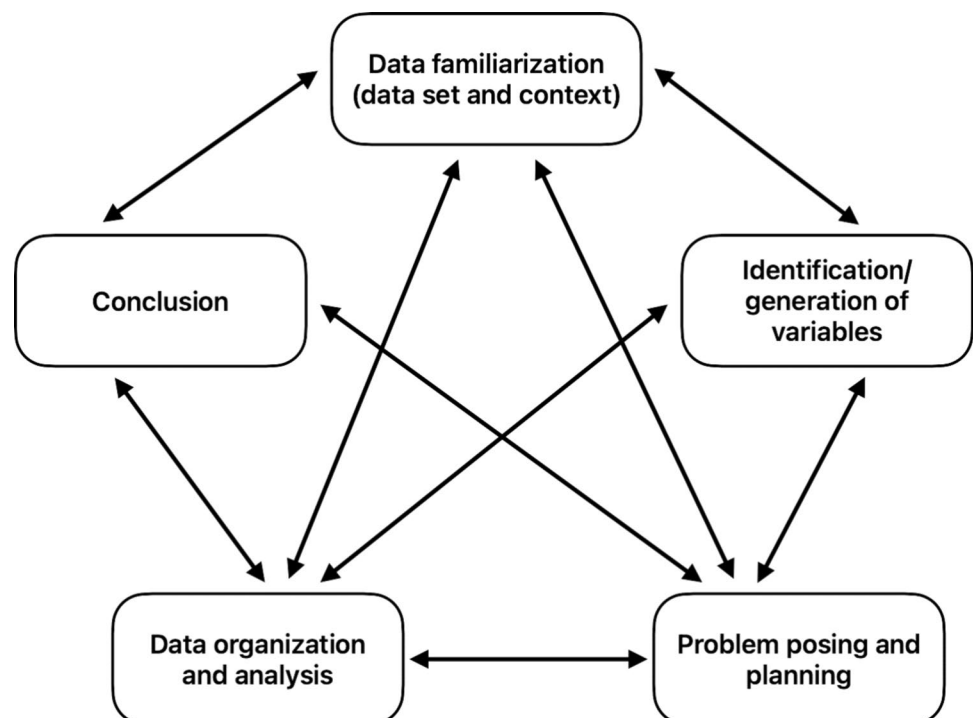
In the current study, data become available by using and interpreting photographs related to a specific topic from the Dollar Street website. This image-based data type can be considered between primary and secondary data. Since others collect these photographs for specific purposes, they come from a secondary data source. However, turning these cases into data that can be analyzed to answer specific questions posed entails somewhat subjective decision-making, as seen in previous studies (Buehring & Grando, 2023; Kazak, 2022; Kazak et al., 2022) because such data arise from the knowledge, purpose, and experience of the individuals engaging with them. So, this decision-making process

involving identifying and generating variables and their values seems similar to obtaining primary data through planning what variables to measure and how to measure them, but with the possibility of subjectivity.

Kazak et al. (2022) proposed a data investigation cycle for exploring image-based data (Fig. 1). The proposed cycle was based on the common investigative sequences of actions students followed with varied levels of prompts from the researcher/teacher when working with photos from Dollar Street. This iterative sequence of interlinked actions includes context/data set familiarization, variable identification/generation, problem posing and planning, data organization and analysis, and drawing conclusions.

In this iterative cycle, data familiarization is essential to the statistical investigation before and after problem posing and planning and during data organization and analysis. In connection with these three components, forming categories based on observations about photographs and criteria for identifying categorical values for them, that is, the variable identification and generation stage, is recognized as part of the statistical investigation process for image-based data. Moreover, other studies using non-image-based secondary data (e.g., Gould et al., 2017; Wilkerson et al., 2021) indicate the close connection between defining variables, formulating statistical questions, and data preparation before analysis. As a result, in designing the present study I anticipated that focusing on the identification and generation of variables component, in relation to data familiarization, question posing, and data organization, should provide insights into our understanding of the data-ing process with image-based

**Fig. 1** Statistical investigation cycle for image-based data (Kazak et al., 2022)



data. That process would involve making observations from photos to describe a quantity/quality and organizing those observations into variables to explore statistical questions with the recognition of variability in the data before analyzing data.

### 3 Methodology

In order to shed light on the data-ing process with image-based data and variable identification and generation in that process, this qualitative study was guided by the following research questions: (1) What is the nature of the data-ing process when pairs of pre-service mathematics teachers work with Dollar Street images? (2) How do pairs of pre-service mathematics teachers come to identify and generate variables in data-ing? To investigate these questions, task-based interviews were conducted with pairs of volunteering pre-service mathematics teachers (ages 21–23, females) who were comfortable collaborating. This set-up allowed the researcher to observe and listen to the explanations participants shared with each other, similar to students collaboratively solving a mathematical problem (Schoenfeld, 1985). One of those pairs (pair 1) was finishing their third year, and the other two (pair 2, pair 3) were at the end of their fourth year. While all participants had taken a two-semester Introduction to Probability and Statistics course, pairs 2 and 3 had additional coursework, including a Research Methods in Education course and a Methods of Teaching Mathematics course with topics on teaching statistics. In the latter two courses, participants engaged in formulating research questions or statistical questions and the statistical investigation cycle with quantitative data. None of the participants had experience working with image-based data or Dollar Street.

#### 3.1 Design of the task-based interview

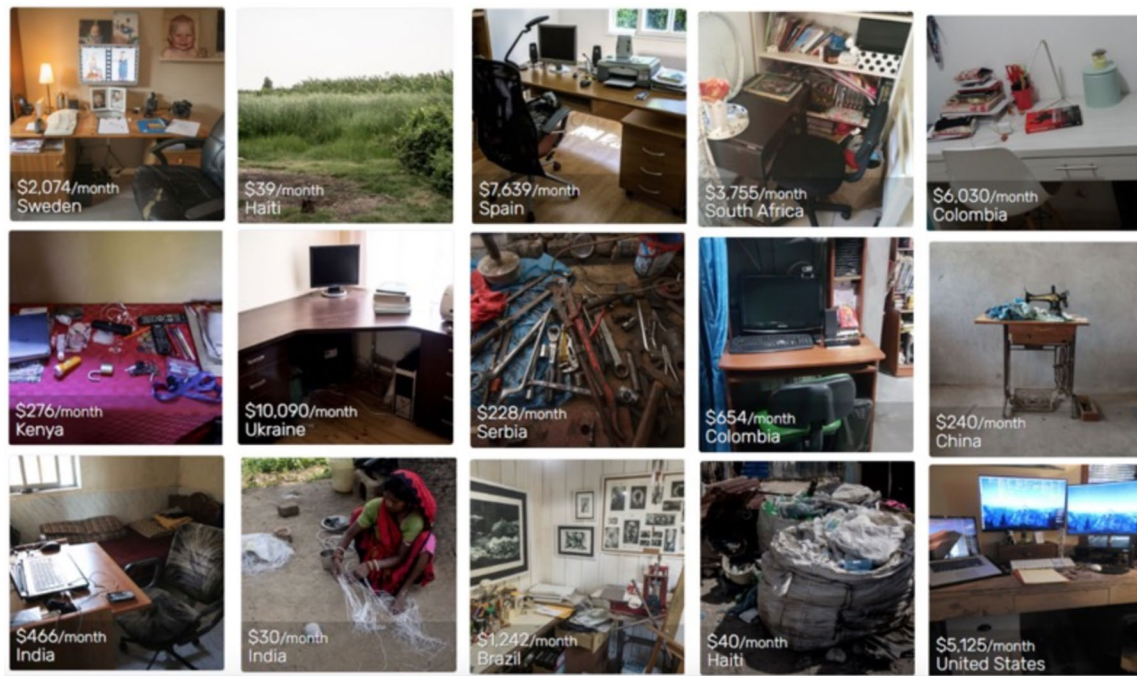
This study examines data-ing with a new data form in the context of Dollar Street images. Dollar Street provides a rich array of photos of living rooms, beds, roofs, pets, earrings, and all sorts of items, and videos of brushing teeth, children playing outside, throwing food trash away, and so on by various topics from 460 homes in 66 countries from Africa, America, Asia, and Europe (44,916 photos and 8296 videos in total). Descriptive metadata, such as income and country, are also attached to these photos. These data are collected manually by teams of photographers who visit volunteer families across different cultures and income levels. The aim is to better understand the living conditions of people from various income groups worldwide through photos as data. The website

displays the world as a street ordered by monthly income per family member (the lowest incomes to the left and the highest incomes to the right). The income information (\$ per month) is assigned to each family not based on salary but according to the consumption accessible to each adult in the household and adjusted for purchasing power parity (Lindgren, n.d.). Due to voluntary participation and limited access to certain areas, the families and regions selected might not be representative of different parts of the world.

Since familiarization with data set and context has been shown as an essential component of data-ing with image-based data in previous studies (Kazak, 2022; Kazak et al., 2022), each pair first participated in an approximately 20-min session meant to introduce the Dollar Street website. They were provided with information about how and from whom the data were collected and the topics related to the photographs of families and their living spaces described above. Next, they did a free exploration of the images on a specific topic and were asked to share their observations from these pictures (what they noticed and what they wondered). Then, an interview with each pair (about 90 min) was conducted through a recorded Zoom meeting with screen sharing.

The task-based interview aimed to study how the pre-service teachers recognized and generated variables in connection with data familiarization, question posing, and data organization in data-ing. Therefore, starting with data familiarization, the pairs were initially asked to examine selected photographs of the Work Areas of the participating families with different income levels (ranging from \$30/month to \$10,090/month) from different countries and continents. The individual photos, which can be moved around and resized if needed (this feature supported the participants' data familiarization), were organized unordered and in arrays using Google Jamboard pages, as seen in Fig. 2. In this shared document, 90 images over six pages were available to the participants.

During the interview, each pair was asked the following questions: (1) "What do you notice after examining the photos of work areas given on this page? What do you wonder about?" (2) "Can you pose an investigative question that can be answered with these photos? Can you think of another question?" (3) "How would you arrange the photos to answer this question? Can you explain why you did that?". After working with the photos on the first page of the Jamboard document and answering the questions above, the pairs could move to the following pages and reconsider their responses/revise their investigative questions if needed. They also used the sticky note tool in Jamboard to write down their investigative questions as they moved through the pages.



**Fig. 2** Page 1 of the Jamboard with a sample of photographs of Work Areas in Dollar Street (<https://www.gapminder.org/dollar-street?topic=work-areas&media=image>)

### 3.2 Data collection and analysis

The interviews were designed to capture the pre-service teachers' data-ing process with Dollar Street photographs, specifically, how they recognized and generated variables when prompted for data familiarization, question posing, and data organization, as these could support students' investigations of image-based data (Kazak et al., 2022). The data for analysis consisted of Zoom video recordings, transcripts of the interview sessions for each pair, and their work with photos (i.e., groupings and questions posed) on the Jamboard pages. Using progressive focusing (Parlett & Hamilton, 1972), qualitative analysis was carried out to understand participants' data-ing strategies when exploring photographs of the workspaces in Dollar Street. The progressive focusing approach was introduced as a method where "researchers systematically reduce the breadth of their enquiry to give more concentrated attention to the emerging issues" (Parlett & Hamilton, 1972, p.18). With this approach, the researcher carries out multiple stages of analysis and adapts the focus to salient issues as new insights progressively emerge from data.

In the first stage of the analysis, the author/researcher transcribed the video recordings of the interview sessions and the screenshots of the critical moments, such as grouping and ordering the photos. In the second stage, the author/researcher immersed herself in the data to gain insights about the pairs' data-ing process with the following two foci

on variable identification and generation: (1) what variables the pair defined during the whole interview session and (2) how the variables were identified and generated. By continually going over the transcripts for each pair and watching the video recordings when needed, the author/researcher identified the words and phrases that each pair used for a variable or attribute and its values in the transcript and then labeled the chunks of these occurrences in relation to the data familiarization, question posing, and data organization. In the third stage, the coding on the transcripts was done as insights emerged from the following three foci on the data-ing process: (1) what actions, related to data-ing, each pair took when identifying and generating variables during data familiarization, question posing, and data organization, (2) how these actions occurred during data-ing, and (3) what characteristics of these variables were identified and generated during data-ing. In the fourth stage, the themes emerged when the author/researcher tried to describe and interpret these foci in relation to the definition of data-ing used in this study.

To illustrate, after Pair 1 was asked, "What question can you answer with these photos?" their attempt to group them into two categories, 'home environment' and 'work environment' (Fig. 3), was initially labeled as "grouping" related to data organization. Then, variables were identified and coded as "income," "home vs. work environment," and "eye-pleasing vs. basic needs" in the excerpt shown in Fig. 3. The codes related to data-ing actions (observing,

**Fig. 3** Excerpt of Pair 1's grouping of photographs of Work Areas on the second page of the Jamboard



- Those with higher **income** have environments that are more **appealing to the eye**.
- In the lowest ones [**income**], the environments are mostly aimed at **meeting basic needs**. Like food, agriculture.

interpreting, inferring, and relating variables) were assigned to these responses. After similar coding in all transcripts, the themes “observational variables based on visual judgment/metadata” and “inferential variables based on personal interpretation” emerged from the codes “income” and “home vs. work environment/eye-pleasing vs. basic needs,” respectively.

## 4 Results

The findings are presented in two subsections corresponding to the research questions addressed in the study: (1) What is the nature of the data-ing process when pairs of pre-service mathematics teachers work with Dollar Street images? (2) How do pairs of pre-service mathematics teachers come to define variables in data-ing? Throughout the section, P1.1 and P1.2, P2.1 and P2.2, and P3.1 and P3.2 refer to the participants in Pairs 1, 2, and 3, respectively.

### 4.1 Pre-service teachers' data-ing process

This section gives an overview of the findings related to the data-ing process of the pairs. As mentioned earlier, data-ing is conceptualized as the process of making observations to describe a quantity or quality and organizing those observations into variables to explore statistical questions before analyzing data. Accordingly, several common actions emerged across the three pairs' data-ing process with the given Dollar Street images of work areas during identifying and generating variables related to data familiarization, question posing, and data organization. These actions included *observing*, *interpreting*, *conjecturing*, *inferring*, *comparing*, *grouping*, *ordering*, *questioning/question posing*, and *relating variables*. While *grouping* and *question posing* were prompted through the interview questions (e.g., “Can you pose an investigative question that can be answered with these photos? Can you think of another question?” and “How would you arrange the photos to answer this question?”),

others appeared to occur spontaneously during free explorations of images. In addition, *categorizing variables* and *measuring variables* were observed as part of Pairs 2 and 3's data-ing. Of these 11 actions, five (*observing*, *interpreting*, *conjecturing*, *inferring*, *comparing*) occurred during data familiarization; one (*questioning/question posing*) happened during problem posing; two (*grouping*, *ordering*) took place in data organization; and three (*relating variables*, *categorizing variables*, *measuring variables*) were relevant to variable identification and generation phase of the statistical investigation cycle for analyzing image-based data (Kazak et al., 2022) shown in Fig. 1. While each action was connected to identifying and generating variables, several actions tended to occur in various combinations during the pairs' data-ing process. Some examples of these data-ing actions are presented for each pair next.

#### 4.1.1 Pair 1

During the initial data familiarization phase (until they were prompted to pose a statistical question that could be answered with the photos), usually *observing* and *interpreting* the photographs of work areas on page 1 of the Jamboard were seen. For instance, after P1.1 said, “There is an area [Haiti \$39/month] like a garden. Could be agricultural area. There are areas related to repairing [Serbia \$228/month]. There is a repair sector. There is tailor [China \$240/month].”, P1.2 added, “[photos] may also indicate occupations.” So, they observed what was in the picture and interpreted what they saw, such as “tailor” as an occupation mentioned based on the sewing machine in one photo. These responses revealed the critical role of *interpreting* in connection with *observing* in using photographs as data. As part of *question posing*, their first statistical question was: “What is the relationship between the occupation of the different families shown in the photographs and their income?” The question was posed in a way that related two variables (*relating variables*), and next, the pair organized data by *ordering* the photos according to income. In this

example, *ordering* the images seemed to help them see if it made sense to relate the two variables in their question. After further data familiarization on page 2 of the Jamboard, the pair started *grouping* the photos into clusters based on *observing* and *interpreting* the photographs: “On one hand, desk jobs with technology and more modern places. On the other hand, there are places that require more agriculture and hand skill jobs” (P1.2). During the *grouping*, they also paid attention to *ordering* images by income. These groups of photos ordered by income made any relational patterns in the image-based data more apparent.

#### 4.1.2 Pair 2

In the initial data familiarization phase, P2.2 began to say, “I think what is asked here is for people to take photos of the places they work. In high-income families, the work area generally includes a desk or a technological device, computer, or book on the desk. While the work in families with higher incomes is more related to IT and technology, in those with lower incomes, it requires muscular strength. I do not know if this is true. However, it seems like the jobs done in families with high incomes can be done with an increase in education level, while in lower-income families, they are like physical jobs, where education is not very necessary.” As seen in this quote, both *observing* items and *interpreting* them in relation to the jobs requiring more education vs. not much education seemed to be combined with *relating variables* when considering income simultaneously. The data-ing process again appeared to involve a combination of several actions, as the pair described several qualities in the photos in relation to income level through observation/meta-data and interpretation. In the problem-posing phase, P2.1 started with *conjecturing* (“As their income increases, they work in jobs that require more education.”) and suggested: “We can obtain data by categorizing the physical conditions of the work area with respect to income level” (*relating variables*). Once the pair conceived relating the two variables as a way to get data from the photos, they posed their question: “What is the relationship between individuals' monthly incomes and the physical conditions of their work areas? Physical conditions are evaluated on a scale.” (*question posing*). In this example, *categorizing variables* and *measuring variables* using a scale suggest an inclination for measuring an observation to obtain data by the somewhat objectively defined measuring instrument.

#### 4.1.3 Pair 3

After several iterations of viewing specific photos by enlarging them, *observing* items, and *interpreting* them

during the initial data familiarization, the pair engaged in *inferring* and *grouping to categorize* the work areas as ‘home office’ and ‘workplace’:

##### Excerpt 1

P3.1: Maybe we can infer something. Some of them are home offices. Some workplaces are outside the home. I think this is the house of someone in Colombia (\$654/month).

P3.2: That [in Colombia] might not even be a home. For example, South Africa is a house, but Colombia, with a 654 income, is not a house.

P3.1: Hmm. I think that one is a house [India \$466/month]

P3.2: It is a house.

P3.1: This [India \$466/month] works from home, for example.

P3.2: I think you call it home because of those things (*a bed or couch*) in the back, right?

P3.2: Uh-huh.

All these interpretations of observed spaces in the photos that led them to categorize the work areas appear to be part of using photos as data that can be analyzed. Towards the end of their initial data familiarization on page 2 of the Jamboard, *comparing* also seemed to be used to consider variability in data-ing. For instance, P3.1 stated, “We said they were both India [India \$466/month and India \$30/month], and we could make a comparison. We can either compare different occupations from the same country or compare different countries that we think are the same occupation. For example, this can be a software engineer in India (\$466) and this one in the US (\$5125), too. Moreover, their work areas are very different, and their incomes are very different, but of course, we cannot be sure exactly whether they are [software engineers].” They had a plan for comparing different occupations from the same country or different countries with the same occupations (through controlling one variable and seeing the variation in the other), but determining one’s occupation solely based on a computer in the picture was not sufficient, indicating high uncertainty. When the pair was prompted to pose a statistical question (*question posing*), their first reaction was related to *measuring variables*, and they started to talk about variables in terms of their measurability:

##### Excerpt 2

P3.1: I am thinking about how measurement will be done when the data is a photograph. Let us say I write a question about a word called comfort. Maybe comfort is something that can be measured with a photograph. However, I am thinking about what we can measure with a photograph so that I can pose a question.

P3.2: Maybe not the comfort but the quality of the materials in the work area. When we say quality, can we fully capture what it means? For example, when we look at the work area in Colombia, India, or Haiti, is it the quality of the material? How exactly can we specify it?

This exchange about what observations can be measured when working with photographs indicates the challenges of using photographs as data in data-ing.

## 4.2 Pre-service teachers' variable identification and generation

This section focuses on how the descriptions of the qualities observed in the photos in relation to data familiarization, question posing, and data organization led to variable identification and generation during data-ing. The data-ing actions across the pairs presented above tended to lead or be linked to the variable identification/generation process based on the image-based data. The variables identified by the three pairs were classified into two groups: (1) observational variables based on visual judgment/metadata and (2) inferential variables based on personal interpretation of image-based data. The first group, including 'income' and 'having computer' variables, appeared to be either metadata or easily measurable visually from the photos and be the same across the pairs. While income was already shown on each photo in the data set, the computer or laptop was apparent in the pictures. In the latter group, however, defining the variables was more challenging. The variables identified by the pairs became divergent due to the nature of variable generation, requiring personal interpretation or inference.

In the second group of variables, only the occupation variable was common in all three pairs but with slightly different categories. For instance, Pair 3 grouped the photos on page 4 of the Jamboard by 'computer-related jobs,' 'office jobs,' and 'agriculture/land-related jobs' when they were prompted by the researcher about how they could organize the photos to answer the question they just posed ("Do the occupational groups in the countries differ according to the countries they are in?"). On the other hand, Pair 1 initially identified 'desk jobs with technology' and 'agriculture and hand skill jobs' categories for the occupation variable as a result of grouping the photos on page 2 of the Jamboard based on what they could infer about the jobs. However, when the pair noted the image of Cambodia (\$60/month) as not fitting in either group, they re-grouped the photos into categories: 'home' and 'workplace.' However, they still had difficulty distinguishing the photos as home or workplace outside the home. Although, at first, they thought that the photos with a computer could be on-site offices, after closely examining the

pictures and discussing the items/furniture in the workspace, they moved some images with a computer into the 'home' category.

Pair 2 focused on the physical conditions of the workspace, including features like sanitary, hygiene, safety, and tidiness in the photos. When they moved to page 6 of the Jamboard, observing one particular photo (Sweden, \$3057/month) prompted them to discuss a new variable identification.

### Excerpt 3

P2.1: Sweden [\$3057/month] is very good. My dream work environment!

Researcher: Why did you say that?

P2.1: (*speaking with excitement*) Organized, flowers. The presence of things that appeal to the eye other than work actually encourages [you] to work. We did not pay any attention to these. We have kept thinking about computers. Most of them did not have these, however. P2.2: But as we said, people's priorities change. For example, the person's priority here (Sweden) might be that the work area should be peaceful. However, the person in Indonesia (\$153/month) would not need such an environment. He might say, "I just need to earn money and make a living."

P2.1: Even the presence of a window can lead us to an interpretation [about the work area], like having a bright environment. While a person is valued in one place, he can be completely worthless in another. Based on work areas, this can also be reflected on the countries.

Later, this unique observation and the pair's inference from it led them to pose the following question relating two variables, 'country location' and 'amenities in the work area': "What is the relationship between the locations of countries and the amenities available in their work areas?". However, they did not notice the lack of clarity in the second variable, that is, how the amenities are defined and measured. When the researcher asked them to clarify that variable, they wrote a note after the question: "Work areas are evaluated based on the tidiness of the items and the setting, the presence of materials that encourage work, etc." This explanation gave some indication of how the variable was defined, but was insufficient for measuring it. P2.2 emphasized the subjectivity in the 'presence of materials that encourage work' but mentioned some 'de facto features,' such as 'being bright, having flowers, large table, spacious setting,' when identifying the variable.

When Pair 3 posed a question about the image-based data about work areas on the last page of the Jamboard, they began by identifying the variables first.

### Excerpt 4

P3.1: You have said work setting. Let us go back to that. Let us use the word comfort, though comfort...

P3.2: But the better working conditions at work are also very subjective. For example, working in an outdoor or garden is perhaps more comfortable for me than being stuck in a room.

P3.1: Yes, and we cannot fully capture the working conditions with a photo. I mean, can we know how many hours they work?

P3.2: Then let us say the same occupational groups. As the income of the same occupational groups increases, do they tend to work in better conditions, or can it be associated with their work?

P3.1: What does better condition mean?

P3.2: Let us write that question first. Now, the same occupational groups. We can describe it as, for example, better quality materials. Or we can say more functional tools like computers are getting better. We can categorize it as such. Because when we do different occupations, it is not very useful. However, when I classify the same occupations, comfort means something.

In this exchange related to data-ing, Pair 3 acknowledged the subjectivity in defining variables that can be measurable when posing their statistical question. They finally agreed to focus on the same occupational group for which ‘comfort’ can mean (more or less) the same thing. They wrote the following question: “Is there a relationship between the monthly income of the same occupational groups and the working conditions (quality of materials in the environment, tidiness of the setting, etc.)?”

## 5 Discussion and conclusion

This paper aimed to explain pre-service mathematics teachers’ understanding of photographs as data, that is, image-based data, and the nature of their data-ing process with a focus on variable defining when investigating 90 images of work areas around the world in Dollar Street. The findings revealed new insights into (1) the pre-service teachers’ data-ing actions during identifying and generating variables in relation to data familiarization, question posing, and data organization and (2) how the pre-service teachers identified and generated variables in data-ing.

### 5.1 Data-ing with image-based data

Data-ing in general refers to engaging with data to describe and organize quantities or qualities, either by working with predefined variables or through making observations, to explore statistically investigative questions before analyzing

data. Compared to what we know about these in traditional data investigations used in schools mentioned in Sect. 2.1, more needs to be known about data-ing with image-based data. This study showed various common actions emerging in the data-ing process across the three pairs, such as *observing*, *interpreting*, *conjecturing*, *inferring*, *comparing*, *grouping*, *ordering*, *questioning/question posing*, and *relating variables*. However, *categorizing variables* and *measuring variables* were only seen in Pairs 2 and 3. This could be related to the senior pre-service teachers’ additional coursework, such as Research Methods in Education. Most of the data-ing actions found in this study, that is, *observing*, *interpreting*, *comparing*, *grouping*, *ordering*, and *question posing*, were consistent with the prior research results reported by Buehring and Grando (2023) and Kazak et al. (2022).

These actions provided further insights about data-ing with Dollar Street images in relation to the phases of data familiarization, problem posing, variable identification and generation, and data organization described for analyzing image-based data in Kazak et al. (2022). Similar to the connections between these phases recognized by Kazak et al., the different combinations of various data-ing actions in pairs’ work, such as *observing-interpreting-relating variables*, *question posing-relating variables*, and *inferring-grouping-categorizing variables*, show an iterative sequence of interlinked moves between phases of data familiarization, problem posing, variable identification and generation, and data organization.

Recognition of these combinations of data-ing actions identified in this study can also reframe how to look at the data-ing process when working with Dollar Street images, as illustrated in the GAISE II report (Bargagliotti et al., 2020). More specifically, in Erickson et al.’s (2019) terms, these individual and combined data-ing actions can be considered as “data moves” in data-ing through back-and-forth transitions among data familiarization, problem posing, variable identification and generation, and data organization before data analysis. Similar to the finding of Wilkerson et al. (2021) that certain data moves in preparing public data sets for data analysis enabled students’ reasoning about variability in the data, one of the data-ing moves seen in this study, that is *comparing* photos while controlling one variable and seeing the change on the image-based data related to the other variable (Pair 3, Sect. 4.1), appears to be a way of acknowledging and considering the variability as expected in the data investigation (Bargagliotti et al., 2020). However, this notion of variability with image-based data inherently differed from the pre-service teachers’ observations about the visual aspects of distributions on a numeric axis to reason about variability in comparing quantitative data distributions (Leavy, 2006; Makar & Confrey, 2005). In using the photographs as data, grouping photos according to one variable, such as the same occupations, and comparing them

with another aspect of the image-based data, such as the working conditions, seemed helpful in data-ing. This could also be considered part of the “processing data” phase in the framework based on data science practices (Lee et al., 2022).

## 5.2 Variable identification and generation in data-ing with image-based data

With regard to variable defining, the two types of identifying variables, that is, observational variables based on visual judgement/metadata and inferential variables based on personal interpretation of image-based data, emerged from the analysis. These appeared to share some similarities with the different types of features, such as computer-detectable and human-perceivable features, in the classification tasks involving text (Horton et al., 2023) and photographs (Fergusson & Pfannkuch, 2024) as data. In particular, the variables identified based on the pairs’ interpretation or inference from the photos of work areas, such as occupation, cleanness, tidiness, comfort, and quality of items, tended to be human-driven decisions that can lead to uncertainty (Fergusson & Pfannkuch, 2024). Even observational variables related to Dollar Street images, such as ‘having a computer’, involve a human decision. Still, the uncertainty tends to be less when identifying an object in the photo. One way to overcome the uncertainty in defining inferential variables is to reduce the variability in the data, as Pair 3 decided to group the same occupations to define ‘comfort’ as a variable (Excerpt 4, Sect. 4.2).

Distinguishing observational and inferential variables can help address the tension between the variable defined and how to measure it objectively due to the subjectivity in identifying variables based on human judgment. This tension was apparent as Pairs 2 and 3 struggled and made various attempts to define the variables in the questions they posed more clearly and to make them measurable. The findings suggest that deciding the quality values for an item or a setting can be more challenging than generating an observational variable with ‘yes/no’ values, such as ‘having a bookcase in child’s room’ as seen in pre-service teachers’ investigation of Dollar Street images in Kazak et al. (2022). As Gould et al. (2017) pointed out, for considering variables in other non-traditional data from a secondary source, clearly stated variables in the statistical questions posed by the pairs seemed to be related to this challenge in identifying and generating variables during data-ing with image-based data.

## 5.3 Concluding remarks

Working with image-based data is an emerging area with the recent developments of data science education in K–12 education settings. However, research in this area is still rare.

This study contributes to the literature by providing insights into how pre-service mathematics teachers engage with a set of photos as if they were data and their data-ing actions, focusing on variable defining when investigating image-based data. More specifically, variable identification and generation with image-based data tend to depend on multiple data-ing actions, such as *observing, interpreting, conjecturing, inferring, comparing, grouping, ordering, questioning/question posing, relating variables, categorizing variables, and measuring variables*. Using different combinations of various actions during back-and-forth moves between several phases of investigating image-based data (data familiarization, problem posing, and data organization) underpins variable defining. Moreover, the study reveals the complex process of developing and negotiating variables based on personal interpretation of image-based data and the related challenges due to the subjectivity of such experiences.

Hence, providing opportunities for pre-service teachers to engage with image-based data is essential to enhance their understanding of the data-ing process perceived beyond traditional data investigations during their teacher preparation programs. Such experiences with image-based data can also broaden pre-service teachers’ understanding of key statistical concepts, such as data, variables, and variability, while making them aware of the affordances of using image-based data in their classrooms. In turn, these can support the practices of pre-service teachers related to designing statistical data investigations with Dollar Street or similar image-based data for students at different educational levels and scaffolding their data-ing process. Moreover, the study findings can help teacher educators appreciate the challenges of developing pre-service teachers’ statistical knowledge and skills required to extend practice aligned with real, non-traditional data forms as recommended in the GAISE II report (Bargagliotti et al., 2020) for teaching statistics in Pre-K–12 settings.

## 5.4 Limitations and recommendations for future research

Related to the exploratory nature of this research, the small number of participants who volunteered for the interview are not representative of the general population of pre-service mathematics teachers. Also, the participants’ cultural backgrounds could not be identified since relevant information was not collected in compliance with the ethical approval for the study. Another limitation of the task design is that the photographs on different pages of the Jamboard could not be moved between the pages when the participants wanted to group the photos by the same country or occupation. Providing only the images of Work Areas from Dollar Street also posed a restriction about occupation details, which could

be accessed in the family description text available on the Dollar Street website.

Future research is needed to address how learners' data moves in data-ing with images can be supported through digital and non-digital tools at different educational levels. Moreover, an extension of this study would be to compare the data-ing process with pictures and videos in various contexts available on Dollar Street and with other groups of pre-service teachers from different cultural backgrounds.

**Funding** Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

## Declarations

**Conflict of interest** There is no conflict of interest.

**Ethical approval** The study was approved by the Human Subject Ethics Committee at Middle East Technical University (0276-ODTUI-AEK-2023).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Arnold, P. (2013). *Statistical investigative questions. An enquiry into posing and answering investigative questions from existing data*. [Unpublished doctoral thesis, The University of Auckland].
- Bargagliotti, A., & Eubanks-Turner, C. (2024). Teacher preparation in statistics: Focusing on variability through attending to precision. *Journal of Statistics and Data Science Education*. <https://doi.org/10.1080/26939169.2024.2350935>
- 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.
- Buehring, R. S., & Grando, R. C. (2023). Reading and writing the world with children: Statistical thinking and multivariate data. *Statistics Education Research Journal*, 22(2), 6. <https://doi.org/10.52041/serj.v22i2.446>
- 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. H., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education*, 12(1). <https://escholarship.org/uc/item/0mg8m7g6>
- Fergusson, A., & Pfannkuch, M. (2024). Using grayscale photos to introduce high school statistics teachers to reasoning with digital image data. *Journal of Statistics and Data Science Education*. <https://doi.org/10.1080/26939169.2024.2351570>
- 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>
- Higgins, T., Mokros, J., Rubin, A., & Sagrans, J. (2023). Students' approaches to exploring relationships between categorical variables. *Teaching Statistics*, 45, S52–S66. <https://doi.org/10.1111/test.12331>
- 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>
- Kazak, S., Fielding, J., & Zapata-Cardona, L. (2022). Investigation cycle for analysing image-based data: perspectives from three contexts. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11)*. International Association for Statistical Education. [https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11\\_253\\_KAZAK.pdf?1669865544](https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_253_KAZAK.pdf?1669865544). Accessed 1 Aug 2024.
- Kazak, S. (2022). *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-03751833v2
- Leavy, A. (2006). Using data comparison to support a focus on distribution: Examining preservice teachers' understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal*, 5(2), 89–114.
- Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects*. Commissioned paper for the National Academies of sciences, engineering, and medicine, board on science education, committee on science investigations and engineering Design for grades 6–12. Washington, D.C.
- Lee, H. S., Mojica, G. F., & Wilkerson, M. H. (2024). *Data investigations as a framework for stem integration*. Paper presented at the 15th International Congress on Mathematical Education (ICME-15), Sydney, 7–14 July, 2024.
- Lee, H. S., Mojica, G. F., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, 21(2), 3. <https://doi.org/10.52041/serj.v21i2.41>
- Lindgren, M. (n.d.). *Detailed income calculations for Dollar Street*. <https://drive.google.com/file/d/0B0HB08a-a9MbZFJZMTFEUkx0RWc/view?resourcekey=0-XFuTX8d2FyhmrHMP35LDA>. Accessed 1 Aug 2024.
- Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1), 27–54.
- Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovative programs, Occasional Paper no 9*. University of Edinburgh. Edinburgh: Centre for Research in the Educational Sciences. Retrieved from ERIC database (ED167634).
- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). Kluwer. <https://doi.org/10.1007/1-4020-2278-6>
- Rubin, A. (2021). What to consider when we consider data. *Teaching Statistics*, 43, S23–S33. <https://doi.org/10.1111/test.12275>
- Schoenfeld, A. (1985). *Mathematical Problem Solving*. Academic Press. <https://doi.org/10.1016/C2013-0-05012-8>
- Utts, J. M. (1996). *Seeing through statistics*. Duxbury Press.

- Watson, J., Fitzallen, N., Fielding-Wells, J., & Madden, S. (2018). The practice of statistics. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 105–137). Springer International Publishing. [https://doi.org/10.1007/978-3-319-66195-7\\_4](https://doi.org/10.1007/978-3-319-66195-7_4)
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical inquiry. *International Statistical Review*, 67(3), 223–248. <https://doi.org/10.1111/j.1751-5823.1999.tb00442.x>
- Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM Mathematics Education*, 50(7), 1223–1235. <https://doi.org/10.1007/s11858-018-0974-9>
- 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>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.