INVESTIGATION CYCLE FOR ANALYSING IMAGE-BASED DATA: PERSPECTIVES FROM THREE CONTEXTS

Sibel Kazak[1], Jill Fielding[2], and Lucía Zapata-Cardona[3]
[1]Pamukkale Üniversitesi, Turkey
[2]University of New England, Armidale, New South Wales, Australia
[3]Universidad de Antioquia, Colombia
skazak@pau.edu.tr

A traditional data investigation cycle includes problem posing, planning and collecting data, analysing data, and making conclusions. This research studies the data investigation cycle for analysing image-based data. In three independent research projects, students at different educational levels and from different countries were provided photographic data of families and their environments around the world from the Dollar Street project. Data collected included classroom video-recordings (Australia), individual student interviews (Colombia), and pre-service mathematics teachers' interviews (Turkey). Analysis focused on the sequence of actions that helped students when attempting to pose and answer questions based on the data set. Findings suggested a similar, iterative sequence of actions across all cohorts: context and data set familiarisation, variable identification/generation, problem posing and planning, data organisation and analysis, and drawing conclusions.

Everyday citizens encounter a variety of data from different sources that are openly available to make sense of the world, such as image-based data. This kind of data is inherently multivariate, contains or offers potentially non-numeric data, and is a source of rich information. Research suggests that the practice of statistics, as carried out by professional statisticians, can be introduced at school level by engaging students in solving problems with data (Watson et al., 2018). In our work, we seek to understand how students might engage in statistical investigations with image-based data.

The importance of conducting data investigations in K–12 statistics education has been emphasised for promoting students' statistical literacy (Franklin et al., 2007), and there are a number of different frameworks that describe student approaches to data investigations (Watson et al., 2018). A widely known approach is the PPDAC cycle, named as an acronym for the cycle's phases (Wild & Pfannkuch, 1999). To engage young students who have yet to learn about formal inference methods with the PPDAC cycle, a statistical data investigation is typically initiated by *Problem Posing,* which involves asking questions that can be answered through a data collection process. Context plays an important role in formulating investigative questions (Allmond & Makar, 2010). *Planning* for and *Collecting Data* often requires gathering a sample of data from a population with an understanding of sampling. In the planning phase, students need to decide how to obtain or generate data, what data is required, and how to measure the data. After data are collected and organised, *Data Analysis* involves cleaning data, constructing representations, and using numerical summaries to explore and compare distributions with an understanding of variation and central tendency. Finally, *Drawing Conclusions* entails making statistical interpretations and informal inferences with an articulation of uncertainty.

Most school-level statistical investigations focus on analysing quantitative data and data structured in a rectangular arrangement. However, modern data (large, messy, unstructured) encountered in daily life do not necessarily fit this traditional format. Several recent studies have focused on statistical investigations with less-traditional, large, and complex data from secondary sources (Gould et al., 2017; Wilkerson & Laina, 2018; Wilkerson et al., 2021). In these data investigations, the given multivariate data sets were typically repurposed to explore new questions, and thus there was more emphasis on interrogating data and data preparation. Gould et al. (2017) pointed out the importance of the back-and-forth movement between asking questions and considering data phases when investigating repurposed data. A further aspect of using available data collected by others with a specific purpose is limited access to the original context of data construction (Wilkerson et al., 2021). However, Wilkerson and Laina (2018) found that students working with repurposed, publicly available data used contextual knowledge to make connections between the patterns observed in the data and their personal experiences.

While using photographs as data in statistical investigations is rare and relatively new, *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) acknowledges the use of different data

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 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

types, including pictures, as part of developing students' statistical literacy. We explore the following question: *What sequence of actions supports students' statistical investigation with image-based data?*

METHOD

In this paper, we report on three research projects that originated independently. Preliminary sharing of these projects took place at a meeting of the Statistical Reasoning, Thinking, and Literacy (SRTL) Forum, at which time the authors became aware of the other studies. At that time, a discussion was held regarding possible implications for traditional investigative data cycles in light of changing data types. The authors elected to compare their separate studies to better understand any commonalities/differences in investigative approaches.

Participants

The three studies involve students at different educational levels and from different countries (Australia, Colombia, and Turkey). The Australian cohort consisted of a class of Year 4 students (9–10 years old), and their teacher, from a suburban mid-socio-economic government school. The class teacher had extensive experience working with statistical inquiry and investigation using the PPDAC cycle but had no training beyond her generalist degree. Students worked in small groups and as a whole class. The Colombian cohort consisted of two Year 9 students (16–17 years old) from different public schools. They had been exposed to statistics in the school curriculum that primarily involved basic exploratory data analysis of a small data set. The students were individually interviewed based on given tasks. The Turkish cohort consisted of two pre-service mathematics teachers interviewed together. They had taken a statistics course and teaching methods course in which they had experience with the PPDAC cycle and using the Common Online Data Analysis Platform (CODAP) to analyse data. The teachers worked with the researcher in an online environment.

Data Set

Image based data sets used for these projects were subsets of photographic data of families and their environments from around the world made available by the Dollar Street project (https://www.gapminder.org/dollar-street). Each image contained the name of the country, monthly income in U.S. dollars, and photographic information from a topic nominated by the Dollar Street project team (e.g., of a pet, a bedroom). The Australian class worked with 111 pre-selected printed images from the "most loved item" category. The Colombian students worked with 10 pre-selected printed images from the "toothbrush" subset. The Turkish pre-service teachers were provided access to the full 43,685 images in the online data set but worked with images based on their own filtering by topic and continent depending on the question explored, i.e., 35 images from "child rooms" category in the Americas.

Data Sources and Analysis

Our analysis focused on the sequence of actions that helped students when attempting to investigate the image-based data available to them. Multiple sources of data were used for analysis: video-recordings and transcripts of each lesson/interview session, student work samples, and field notes that identified critical points of interest. Video analysis was carried out following a process adapted from Powell et al. (2003). The videos were logged to link the video with student work samples and research notes. Each researcher used research logs to identify a sequence of events that took place with their respective research participants to address the specific research question, *what sequence of actions supports students' statistical investigation for image-based data?* These sequences were compared and contrasted across the three contexts, with discussion and clarification facilitating between-researcher understanding of each context. Discrete events were able to be identified according to purpose/activity and these were aligned with the more traditional sequencing of statistical investigations (e.g., Watson et al., 2018) for descripting/naming purposes. A snapshot of the sequences is provided below.

FINDINGS

Australian Context

The young age of the Australian cohort necessitated teacher guidance. The teacher guided initial exploration of one image, three images, and then the entire utilised set (111 images of families' most loved items) through asking questions about what students noticed and wondered about the images. This

was necessary to support students' contextual knowledge of the data set. Student conjectures about the images and subsequent discussions facilitated small groups of students to sort and group the images in multiple ways, which in turn enabled the identification of possible variables and values (e.g., income could be sorted into "low: 2-digit monthly income" to "very high: 5-digit monthly income").

After generating potential variables, groups were asked by the teacher to pose questions they thought they could answer by analysing some or all of the data images. These questions varied in simplicity from questions with simple yes/no responses (e.g., "*Are people's most loved items devices"?*) through those exploring associations (e.g., "*How is the monthly wage related to the country?"*), and one that sought to go beyond associations to make conjectures (e.g., "*If their monthly wage went up, would it affect the most loved item?"*). After deciding upon a question to address, each group sorted the printed images in ways that they thought would assist them in answering their question. This process was not straightforward for most groups because they continued to struggle with the categorization of some images (e.g., whether a goat was a pet or a food source). This led to richer discussions and exploration for a number of images. Three of the six groups chose to re-sort their data when they realised their initial sort was not going to enable them to answer their question efficiently. For instance, one group addressed the question, "*Do the most loved items match the country?"* They initially sorted images by 'type' of item. However, they had not clearly articulated what they meant by 'type' of item and struggled to sort the images meaningfully. They reconsidered their approach and sorted instead by country name, then looked across country groups to identify patterns in the data.

After the most loved item images were organised, students continued unprompted to consider ways in which they could represent their data to facilitate sharing their findings. All groups chose to represent their data groupings photographically. Three categories of data representations were noted (Figure 1): (a) dichotomous groupings (technology versus not technology); (b) groupings of like images (phones, dogs, books); and (c) clusters of images by variable and then by within variable patterns (by country and then by income within the country). Finally, students communicated their findings to the class, with reference to their data representation.

(a) Dichotomous grouping

(b) Like image grouping

(c) Sorting by variable, within variable

Figure 1. Student-generated data representations

Colombian Context

Students studied information from a single image to develop a sense of the multiplicity of attributes they could consider in the subsequent organisational process. After they were familiar with the information and recognized the variation and nature of the attributes, a set of 10 images was displayed for them to decide on a structure. Developing different criteria for organising the 10-image data set functioned as posing questions to the data in the investigative cycle. Participants posed questions such as, "*What does the data set look like according to the family income*?"; "*How could the data be organised according to the number of toothbrushes in the image*?"; "*How different is the quality of toothbrushes from families around the world*?"; "*What could the data set tell us about the level of technology inherent in the toothbrush*?"; and "*Is there any relationship between the quality of the toothbrush and the family income*?".

Students had to make decisions when applying different criteria to organise the data. Sometimes the selected criteria could not be applied in a straightforward manner, and they had to make decisions in the implementation. For example, applying the criterion related to the level of technology inherent in a toothbrush was challenging. There was an image from the country of Myanmar with a finger and an image from Cote d'Ivoire with a twig, both representing toothbrushes from these respective countries (Figure 2). One student had to decide which image would go first in his organisational strategy.

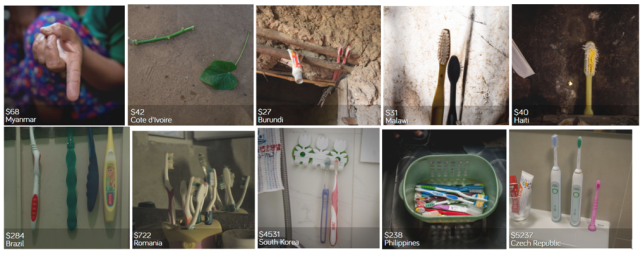

Figure 2. Data structure from a student

Students used information from the statistical investigation with image-based data to draw conclusions, propose generalisations, offer interpretations based on the context, and express limitations. One student expressed: "*it seems that the countries with higher income have more sophisticated technology in their toothbrushes.*" In these actions, students followed the different stages of the investigation cycle but not in a specific order. They went back and forth asking questions, analysing data, and making conclusions.

Turkish Context

The free exploration of image-based data helped the pre-service teachers get acquainted with the context and data involving more than 150 variables from the Dollar Street website. Working with these complex data required a considerable amount of exploration time to make observations and wonderings about the images that led to various statistical questions. The pair started by looking over the images of family snapshots by continents and selected countries that helped them see variation in attributes. Text provided for families was also considered.

The pair first focused on specific attributes in relation to income, e.g., number of children, technology devices, musical instruments, etc. After comparing these attributes between the families with similar and different incomes by country and continent, they generated various investigative questions, focusing on comparison or association between variables in relation to family income: Q1. "*How does the education status affect living spaces*?"; Q2. "*Do families have technological devices in their homes based on income level*?"; Q3. "*How do the things they want to buy compare by family income*?"; and Q4. "*How does the income level affect child room and space for a child*?".

To answer these questions, the pair needed to consider data based on images and text about families. Due to the complexity of the data available on the website, they tended to filter by topics and continents to have a manageable and reasonable sample size (e.g., 15 and 35 images for Q3 and Q4, respectively) as well as a range of variables available from the images. They also noticed that the variables they explored, except income, were all categorical. For example, when categorising their observations related to the images of 'child rooms' from 35 families in the Americas to answer Q4, they decided to focus on whether there is a bed, bookcase, enough personal space, and toy in the room. This seemed unusual to them because they mostly dealt with analysing numerical data in statistics.

(a)

Yeni Veri Seti

Durumlar (35 cases)

in-dex	Income level	Bed	Bookcase	Toy
28	1950	yes	yes	yes
29	2041	yes	yes	yes
30	2044	yes	yes	yes
31	3407	yes	yes	yes
32	3450	yes	yes	yes
33	3748	yes	yes	no
34	3767	yes	yes	yes
35	4650	yes	yes	yes

(b)

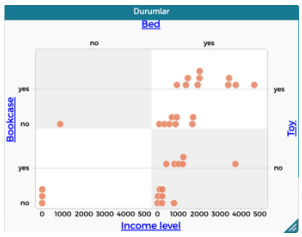

Figure 3. Organising image-based data in a table (a) and representing data (b) to answer Q4

For data analysis, the pair recorded the data (e.g., income, country, whether there is a bed, bookcase, enough personal space, and toy in the room) as text ordered by income in a shared Google document. In order to analyse the data in CODAP, they needed to organise their recorded data into a case-data table that required a column for each variable and their values in rows (Figure 3(a)). Then they

represented the data with dot plots showing all four variables (income, bookcase, bed, toy) at once in CODAP (Figure 3(b)) and drew conclusions, such as "*The families with low income generally have only a bed and no bookcase and a few with toys.*" When they were asked whether these conclusions were generalizable, they pointed out the small sample size and the sample selection as well as the observational nature of their data investigation as the limitation.

DISCUSSION AND CONCLUSION

Our findings suggested that students working with image-based data followed a common investigative sequence (Figure 4) with varying degrees of teacher/researcher prompts (that decreased with the increasing age of participants) during the tasks across the three contexts. All students commenced with *Data Familiarisation*—an initial immersion in the data set (or a portion thereof). Younger students were supported in this task through teacher-guided questions eliciting what they 'noticed' and what they 'wondered' about the images. The high school students were encouraged to ask questions of the data, and the pre-service teachers engaged in lengthy, self-directed exploration of the data set. These explorations supported familiarisation with both context and the parameters of the data set in use. The necessity for context familiarity is well documented. In investigations where students are required to collect data, context familiarisation occurs to support posing questions and to identify data to collect in planning and in later making sense of variation (Watson et al., 2018). By contrast, when analysing existing data sets, context familiarity is necessary to make sense of the variables and the range of values, and to identify questions that can be posed of the data set (Wilkerson & Laina, 2018). When addressing image-based data, identification of potential variables is more complex. Context familiarity is necessary to *generate* variables, many of which are not immediately evident due to the potential for interpretation, focus, bias, and background knowledge brought to analysing an image. In this study, we noted that data/context exploration took a different role in supporting the *Identification and/or Generation of Variables/Values* of interest. Although the younger students were supported in the process of identifying variables through repeated sorting of cards and justifying the sort (similar to Wilkerson et al., 2021), the older students and pre-service teachers identified potential variables of interest more readily and with less scaffolding.

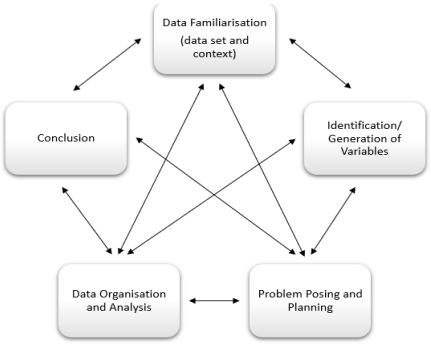

Figure 4. Statistical investigation cycle for image-based data

In the *Problem Posing and Planning* phase, students were able to consider relationships of potential interest between the variables identified. Across all student cohorts there was a range of question 'types' posed: Simple dichotomous ("Are people's most loved items devices?" [Australia]); association ("Is there a relationship between the quality of the toothbrush and the family income?" [Colombia]); and predictive ("How does the income level affect child room and space for a child?" [Turkey]). Having identified the problem(s) of interest, students planned how to sort or organise data to answer the question. In school data investigation cycles (e.g., Allmond & Makar, 2010; Fielding-Wells, 2018), students pose questions of interest very early in an investigation and use the question to plan for collecting the data. Conversely, in investigating image-based data, the examination of the data set supported students' planning for analysis, including data cleaning and organisation. The students then moved into *Data Organisation and Analysis,* reflecting on whether their approach supported answering their question, evaluating their interpretation/identification of contextualised variables/values, and adjusting methods accordingly, similar to Gould et al.'s (2017) description of back-and-forth movement

between questions and data components. Finally, students interpreted their data displays to draw conclusions to solve their problems within the given context in the *Conclusion* phase and presented their findings while drawing on their data. In this respect, the latter phase was not markedly different to concluding phases of existing data (e.g., Watson et al., 2018; Wild & Pfannkuch, 1999) except the limitations of data for generalisations.

In conclusion, student approaches to investigating image-based data appear substantially different from traditional data investigation cycles, where students typically commence with an investigative question, progress through collection and analysis of data, to drawing a conclusion. The problem posing phase of the PPDAC cycle was no longer the starting point when investigating the image-based data available for the students. Asking investigative questions was initiated after the students became familiar with the given data set (collected by others) and identified possible variables. By contrast, student approaches in the studies reported here demonstrated more commonalities with those focussed on investigations of data from secondary sources, in which large and complex data sets were given to students to analyse. Nonetheless, differences were evident; in particular, in the need to identify or generate potential variables from the imagery rather than these being evident in the data set. As such, we proposed a variant statistical investigation cycle aligned to supporting student investigation of image-based data. This investigation cycle can also be used with other available unstructured data sets involving sound, text, or video in which potential variables are identified based on contextual knowledge and situational interests.

REFERENCES

Allmond, S., & Makar, K. (2010). Developing primary students' ability to pose questions in statistical investigations. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics (ICOTS 8)*. ISI/IASE. http://icots.info/8/cd/pdfs/invited/ICOTS8_8A1_ALLMOND.pdf

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 II (GAISE II). A framework for statistics and data science education.* American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Fielding-Wells, J. (2018). Scaffolding statistical inquiries for young children. In A. Leavy, M. Mavrotheris-Meletiou, & E. Paparistodemou (Eds.), *Statistics in early childhood and primary education: Supporting early statistical and probabilistic thinking* (pp. 109–127). Springer. https://doi.org/10.1007/978-981-13-1044-7_7

Franklin, C., Kader, G., Mewborn, D. S., 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. https://www.amstat.org/docs/default-source/amstat-documents/gaiseprek-12_full.pdf

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

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior, 22*(4), 405–435. https://doi.org/10.1016/j.jmathb.2003.09.002

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

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.* Advance online publication. https://doi.org/10.1080/10986065.2021.1922838