WHAT QUESTIONS DO NOVICES POSE ABOUT CATEGORICAL DATA?

Malia S Puloka, Stephanie Budgett & Maxine Pfannkuch
The University of Auckland, New Zealand
m.puloka@auckland.ac.nz

Posing questions about data is an important aspect of statistical thinking and enquiry. While there has been research on students posing questions about data, there seems to be no research on students posing questions about categorical data involving two variables. In an exploratory study conducted with 15 students in a mixed-ability class in a low socio-economic school, a pre-test on posing questions of categorical data was given. In this paper, the questions posed by students are described. The findings revealed a range of responses including that some novices consider the purpose of the survey and a few consider questions involving quantification. Identifying novices' thinking processes through questions they naturally pose could help teachers to design learning strategies.

INTRODUCTION

Citizens of the 21st century are bombarded every day with data-derived information. Data on important societal topics such as migration, social inequality, health and safety, and education are increasingly available to the general public. However, data literacy requires interrogation of data-based evidence that is presented. Consider the following scenario involving categorical data: a media outlet publishes data ranked from highest to lowest (11,748 for dairy farming to 1,130 for logging) on the *frequencies* of claims made in the period July 2013-July 2014 for accident cover for a variety of industries in New Zealand. The media headline stated that the data highlighted New Zealand's riskiest industries (Stuff.co.nz, 2015). While the numbers used in the media article may be correct, the headline is misleading. When reading about such a comparison between categories, we should interrogate the data and the context in which it is embedded. For example, how many people do these industries employ? Should comparisons be based on proportions rather than frequencies? The importance of asking questions of data-based information is paramount. Before novices can query data-based information in the media, they need to learn how to interpret and pose questions about categorical data presented in datasets, two-ways tables and bar graphs. The focus of this paper is therefore: when tasked with posing questions about categorical data, how do novices naturally respond?

LITERATURE

The statistical enquiry cycle (problem, plan, data, analysis, conclusion) is a fundamental part of the New Zealand school statistics curriculum (Ministry of Education, 2007). Questioning permeates the statistical enquiry cycle. Arnold (2013) stated that there were two types of questions, *question posing* and *question asking*. *Question posing* occurs at the problem and plan stages of an investigation and results in formally structured questions for investigative and survey/data collection purposes. *Question asking* involves *analysis questions* and *interrogative questions*. "Analysis questions are those asked about statistics, graphs and tables" (Arnold & Franklin, 2021, p. 2). For graphical displays Friel et al. (2001) classified analysis questions as, read the data, read between the data, and read beyond the data. Reading the data is a question that asks for a number to be extracted from a graph, reading between the data requires several pieces of information to be combined, whereas reading beyond the data involves an inference. Shaughnessy (2007) added a further category, read behind the data, to capture the idea that questions can ask the reasons or causes behind the patterns occurring in data.

Another issue to consider regarding students asking questions based on graphs is their ability to interpret them. Because graphical displays involve an abstraction of data, Friel et al. (2001) explained that students need to learn how to decode graphs by comprehending the context and the meaning of features such as axes and variables before interpreting them. For novices, their natural inclination is to focus on individual cases (Konold et al., 2015) and to want to see each case plotted.

Interrogative questions occur throughout the enquiry cycle and "act as a check … to ensure the consideration of all available information before proceeding to take action" (Arnold & Franklin, 2021, p. 2). For example, when students are given an existing dataset, to understand and interpret it, they need to interrogate the background to the data such as where the data come from and what the variables mean. We named this type of interrogative question as *reading beneath the data*, as the questions relate to the

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

foundations on which the data source sits. Thus, we added a fifth category, *reading beneath the data*, to Friel et al.'s (2001) classification scheme. Once students have interrogated the dataset, to investigate and explore the data, the next step is to pose an investigative question, but again they need to interrogate, for example, whether the question posed can be answered with the available data.

Despite the need for students to learn how to effectively use question *posing* and *asking* during the enquiry cycle, there is very limited research on student *question asking* for graphs and two-way tables, as researchers (e.g., Böcherer-Linder et al., 2016; Watson & Callingham, 2014) typically give students questions to answer and analyse their responses. Furthermore, Budgett and Puloka (2019), in a pilot study on undergraduate students, and Puloka and Pfannkuch (2018), in a study on 17-18-year-old school students, found that students had difficulty articulating analysis questions for categorical data in datasets, two-way tables and their corresponding bar graphs. With regard to students being given datasets, Arnold and Franklin (2021) stated that students, with appropriate teaching interventions, could learn how to pose investigative questions. However, research has focused on summary, comparison and relationship *investigative questions*, not on association-type investigative questions and *question asking* about two categorical variables.

METHODOLOGY

Participants and data collection

The data that forms the basis of this paper comes from an exploratory study conducted in a mixed-ability Year 9 class (13-14 year-olds) in a low socio-economic school in Auckland, New Zealand. The study was conducted over a period of three weeks and involved a pre-test, a teaching intervention, and a post-test. The entire Year 9 class participated in the study activities and the students had no prior knowledge or experience of working with datasets or posing statistical questions about data. Of this Year 9 class, 15 students (three females and 12 males) of Pasifika and Māori ethnicities gave consent for their data to be collected and completed the pre-test. The data presented in this paper is based on the written responses to three pre-test tasks. The following research question was established to guide this study:

What types of questions do Year 9 students naturally ask when given three different representations of categorical data?

Tasks

The pre-test took about 50 minutes to complete. The first three tasks (Tasks A, B and C) of the pre-test involved asking students to pose questions about the categorical data presented in a dataset, a two-way table and bar graphs respectively (see Figure 1). One task was given at a time to complete. Once a student has finished, their script was collected before they received the next task.

Prior to students sitting the pre-test, they were briefed on what the pre-test tasks were about. The students were told that they would be presented with tasks involving data that they would use to complete the tasks. When asked if they had seen statistical data before, they responded by referring to internet data and mentioning words such as megabytes. The researcher, the first author, then explained what statistical data was and the meaning of some of the statistical terms used in the tasks such as the term survey. The data used in the pre-test were taken from Census@School (https://new.censusatschool.org.nz/), a biennial online survey for Year 3 to Year 13 school students in New Zealand. A random sample was drawn from the Census@School database (Figure 1a), with selected variables used in the tasks. The selected data was used without alteration in Tasks A and C. However, for the purposes of understanding students' reasoning, the numbers in the two-way table for Task B (Figure 1b) were modified.

Because of the students' lack of familiarity with datasets, the researcher used the rows of the dataset snippet (Figure 1a) to describe several survey respondents' characteristics, explaining the variables, and how they were recorded. Task A asked the students to pose six questions about the data in the dataset (Figure 1a); Task B asked the students to pose three questions about the data in the two-way table (Figure 1b); and Task C asked the students to pose three questions about the data in the bar graphs (Figure 1c). A prompt, "I wonder…", was stated in each task to help the students form their question should they need, so they could either write a question or an 'I wonder' statement. In total, each student had the opportunity to pose up to 12 questions across the three tasks.

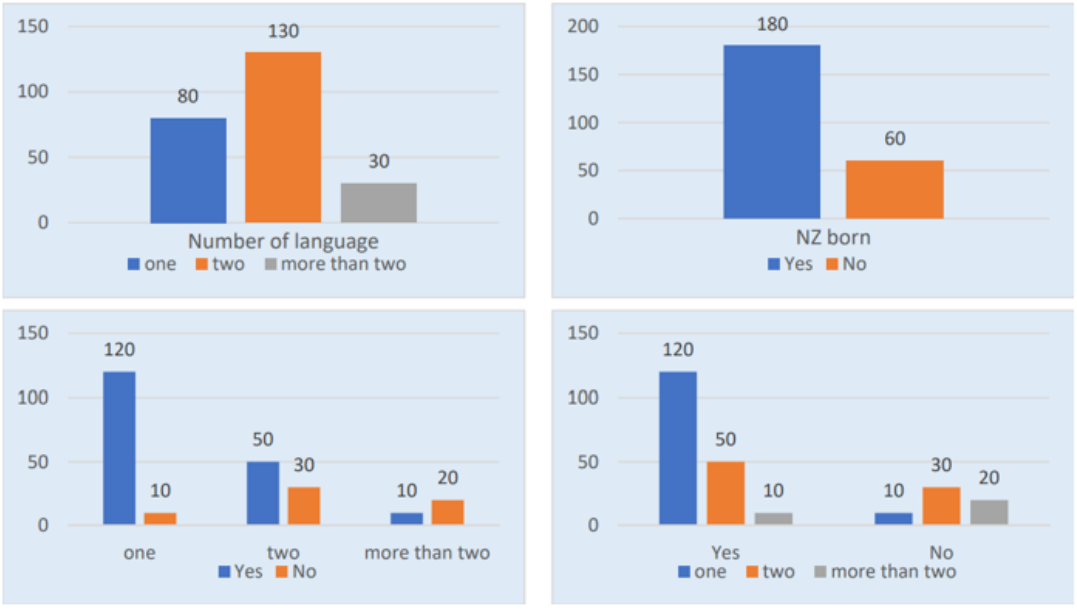
Table 1. Pre-test tasks: Three representations of categorical data

	gender	NZ born	Ethnicty	languages	handed	travel	lunch	sport	celltype	celluse	pressure	region
125	girl	Yes	NZ Euro	one	ambi	motor	home	Netball	iPhone	Social media	some or a lot	Canterbury
126	boy	Yes	Pasifika	two	right	motor	other	Basketball	Other	Other	none or very little	Auckland
127	girl	No	Asian	two	right	bus	home	Other	iPhone	Social media	none or very little	Auckland
128	boy	Yes	NZ Euro	one	left	walk	home	Other	iPhone	Other	some or a lot	Canterbury
129	boy	Yes	Maori	one	right	bus	home	Basketball	iPhone	Social media	some or a lot	Canterbury
130	girl	Yes	Pasifika	more than two	right	walk	home	Other	Other	Sending txt, SMS or other instant messages	some or a lot	Otago
131	girl	Yes	NZ Euro	one	right	motor	home	Netball	iPhone	Sending txt, SMS or other instant messages	some or a lot	Waikato
132	girl	Yes	NZ Euro	two	right	bus	home	Other	iPhone	Listening to music	some or a lot	Canterbury
133	boy	Yes	NZ Euro	one	right	bus	home	Other	Samsung	Other	some or a lot	Canterbury
134	girl	Yes	Asian	two	right	motor	tuckshop	Other	iPhone	Social media	none or very little	Auckland
135	boy	Yes	NZ Euro	one	right	motor	none	Other	Samsung	Social media	some or a lot	Auckland
136	boy	Yes	NZ Euro	one	right	walk	home	Other	Samsung	Other	some or a lot	Otago

(a) Task A: A snippet of the dataset

Pressure	Girls	Boys	Total
None or very little	15	60	75
Some or a lot	15	20	35
Total	30	80	110

(b) Task B: A two-way table of Gender vs Homework Pressure



(c) Task C: Bar graphs of Number of languages spoken vs NZ born

RESULTS

An analysis of the students' written responses to all three tasks revealed four main categories of questions were posed. The definition of each category is discussed below. Table 1 shows the number of questions posed by each student under each category in each task. Within each task, a total of 32 questions were posed for Task A, 25 questions for B and 22 questions for C.

Table 2. Number of questions posed by each student under each category

	Category 1 Task terminology			**Category 2** Survey background			**Category 3** Reasons behind data			**Category 4** Quantifying			
Student	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**TOTAL**
Asa	1	1		1		1		1					**5**
Cam					1								**1**
Cyd				2	1					5	2	3	**13**
Dax				1		3		1					**5**
Dru			1	4	1	2	1	2					**11**
Jay											1		**1**
Joe				3		2		2					**7**
Kiu					2	2							**4**
Mou				5		3		2					**10**
Ned										1			**1**
Nia											3	1	**4**
Pio				1	2	2							**5**
Qee				6		1		1					**8**
Soa								1					**1**
Vin				1	1				1				**3**
TOTAL	**1**	**1**	**1**	**24**	**8**	**16**	**1**	**10**	**1**	**6**	**6**	**4**	**79**
		3			**48**			**12**			**16**		

Across the three tasks, 15 students posed a total of 79 questions. The number of questions posed by each student ranged from 1 to 13 with an average of about 5 questions posed per student, out of a possible 12 questions that they could have posed. Notably, Cyd posed a total of 13 questions, seven in Task A and three each for Tasks B and C. However, it appears that most of these students were not aware of how to ask questions of the data given.

Category 1 Questions: Understanding task terminology

Although only three written questions in Category 1 were posed by two students, students had many questions prior to sitting the pre-test about the tasks and terminology used including the meaning of the word "data." During the pre-test some students continued to ask the meaning of words such as proportion. Therefore, it was decided to include *Understanding the task and the terminology used* as a separate category. For example, for Task A (Figure 1a), Asa posed the question, "What is data?" Understanding what it meant to pose questions about data was also problematic, with Asa asking "Why is it called homework?" for Task B (Figure 1b) and Dru querying, "Do you have to answer the question?" for Task C (Figure 1c).

Category 2 Questions: Background to the survey

A total of 48 questions were posed by 11 students in relation to the background to the survey and individuals surveyed. The preponderance of questions posed in this category compared to the other categories shows that these students were naturally interested in finding out more about each student or the students surveyed and understanding what the variables meant and why the survey was conducted. Sixteen questions focussed on (a) the purpose of the survey, (b) ethical issues, and (c) practical considerations. That is, the questions were delving *beneath the data* collected. Respective written examples are:

Qee (Task A): I would like to know what this [the survey] is for? (a) I wonder if these people felt uncomfortable sharing this data. (b)

Vin (Task A): How long does it [the survey] take? (c)

Notably, Qee was the only student concerned about ethical issues and posed four questions about this aspect. Thirty-two questions were in relation to the individuals surveyed or individual cases through students wanting to know more information about them:

Joe (Task A): I wonder if 127 and 128 [row numbers of students] play the same sports (as both 127 and 128 recorded Other for Sport).
Mou (Task A): I wonder what social media they use.
Dru (Task B): What homework do they receive?
Mou (Task C): How many students were Islanders?
These are questions that cannot be answered with the available data that was given for the task.

Category 3 Questions: Reasoning behind the data

In Category 3 eight students seemed to be reading, decoding or interpreting the raw data, which led them to reason *behind the data* through wondering or making recommendations about the situation. For example, in Task B, Soa appeared to interpret from the data that girls were feeling less pressure than boys to do homework as she stated, "I wonder why girls don't get as pressured than boys." Vin seemed to have decoded and interpreted the data in Task C as showing that not many students speak more than one language as the question he posed recommended that: "there should be more kids that speak more than one language." Dru for Task A wrote, "I wonder if our school can or have the same data as this table", which seems to suggest he was wondering whether their school would produce similar data. The questions posed are similar to the responses expected when *reading behind the data*.

Category 4 Questions: Quantifying

Four of the 15 students posed quantifying questions with Cyd posing 10 of the 16 questions in this category. We considered the questions posed in Category 4 as statistical analysis questions because they captured questions that invite responders to *read the data* or *read between the data* when interpreting graph displays. Furthermore, the questions involved asking about frequencies rather than proportions. All questions used the words "how many" and involved (a) finding a total, (b) extracting a number from one variable, or extracting two numbers from two variables and finding the (c) sum, (d) joint sum, or (e) difference. Respective examples are:
Cyd (Task A): How many students in total were surveyed? (a)
Cyd (Task C): How many students can only speak two languages? (b)
Cyd (Task A): How many languages does Pasifika and Asian students speak? (c)
Ned (Task A): I wonder how many boys that can make their own lunch at home. (d)
Nia (Task C): How many more students were born in NZ than born elsewhere? (e)
Jay (Task B): How many more boys were there than girls? (e)
Although these students would have experienced answering questions about bar graphs they did not seem to be able to create questions apart from Cyd and perhaps Nia.

DISCUSSION

The focus of this small exploratory study was on learning about the types of questions novice students posed about categorical data involving three different representations. The findings suggested there were four distinct categories, namely, (1) understanding task terminology, (2) background to the survey, (3) reasoning behind the data, and (4) quantifying.

Contextual knowledge is an integral part of statistical thinking and needs to be drawn upon throughout the statistical enquiry cycle. The findings in Category 2 and 3 indicated that these students had many questions about contextual aspects. That is, they were attempting to understand issues related to the conduct of the survey, what individuals surveyed meant when they gave a certain response and why the data suggested a particular finding. About 40% of the total number of questions posed were related to an individual case perspective, which is not surprising in cognisance of Konold et al.'s (2015) research that found students initially focus on individual cases when learning to plot data. The propensity of almost 50% of these students to ask about the purpose of the survey and other issues suggests there is a good foundation on which to build, as querying the source of the data is the first question that should always be asked (Arnold & Franklin, 2021).

Category 1 questions, which also occurred before and during the pre-test, suggest that learning strategies need to focus on the language of statistics. In Category 4, only about 20% of the students attempted to pose a question involving quantification, indicating that asking analysis questions was not part of their repertoire. Furthermore, the average of five questions per student out of a possible 12 and the quality of students' responses suggest that posing questions about data, interrogating data or

knowing what questions can be answered with the available data needs to be addressed in learning. Overall, the findings indicated that *question posing* and *question asking* has been a neglected area of teaching and research.

The implications for teaching suggest the next steps in these students' learning should focus on a number of strategies. First, the teaching could start with a context with which students are familiar, such as collecting data on themselves, which could enhance their understanding of variables, what was measured and how and the purpose of the study and how it was conducted. That is, enabling them to *read beneath the data*. Second, these students need to be transitioned from a focus on individual cases to a focus on the aggregate and to be exposed to a range of questions that can be posed, for example, simple, joint, conditional and comparison questions (Budgett & Puloka, 2019; Puloka & Pfannkuch, 2018). Because two categorical variables are involved, students also need to consider both frequency and proportion questions, which may not be easy as proportional reasoning can be problematic. Thirdly, considering the students' current knowledge, teaching should not include inference, reading beyond the data, rather the focus should be on *reading the data* and *reading between the data* (Friel et al., 2001).

REFERENCES

Arnold, P. (2013). *Statistical investigative questions: an enquiry into posing and answering investigative questions from existing data.* (Doctoral thesis). Retrieved from https://researchspace.auckland.ac.nz/handle/2292/21305

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? *Journal of Statistics and Data Science Education*, *29*(1), 122 – 130. https://doi.org/10.1080/26939169.2021.1877582

Böcherer-Linder, K., Eichler, A., & Vogel, M. (2016). The impact of visualization on understanding conditional probabilities. *Proceedings of the 13th International Congress on Mathematical Education*, Hamburg (pp. 1–4). International Statistical Institute. https://iase-web.org/documents/papers/icme13/ICME13_S1_Boechererlinder.pdf?1472724247

Budgett, S., & Puloka, M. (2019). Making sense of categorical data – Question confusion. In S. Budgett (Ed.), *Decision Making Based on Data Proceedings of the Satellite conference of the International Association for Statistical Education (IASE), August 2019, Kuala Lumpur, Malaysia.* http://iase-web.org/Conference_Proceedings.php

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124-158. https://doi.org/10.2307/749671

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, *88*(3), 305-325. https://doi.org/10.1007/s10649-013-9529-8

Ministry of Education. (2007). *The New Zealand curriculum*. Learning Media.

Puloka, M. S., & Pfannkuch, M. (2018). Year 13 students' reasoning from an eikosogram: An exploratory study. In M. A. Sorto, A. White, & L. Guyot (Ed.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics, Kyoto, Japan.* International Statistical Institute. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_6B3.pdf?1531364279

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning*, (Vol. 2, pp. 957–1009). Information Age Publishers.

Stuff.co.nz (2015). ACC statistics show New Zealand's riskiest industries. Retrieved from: https://www.stuff.co.nz/business/74034875/acc-statistics-show-new-zealands-riskiest-industries

Watson, J. M., & Callingham, R. (2014). Two-way tables: issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning*, *16*(4), 254–284. https://doi.org/10.1080/10986065.2014.953019