REPRESENTING NUMERICAL DATA: THE INFLUENCE OF SAMPLE SIZE

Steven Nisbet, Griffith University, Australia

Abstract

Twenty secondary school students (in Grades 9 & 11) were given two datsets to represent graphically – one with 10 pieces of numerical data, and one with 30. Students were more likely to represent the large dataset in an organised form than the small dataset. The more mathematically able students found it easier to organise the data than their less able counterparts. Grade level had no effect. Possible explanations for the results are explored and the implications for teaching and the curriculum are discussed.

Introduction

"A picture is worth a thousand words"

This old saying might explain why many people find it worthwhile to use graphs to represent data. However, just as the capacity of a picture to convey the meaning of "a thousand words" depends on the technical ability of the artist, so too, the capacity of a graph to communicate messages depends on the ability of the drawer of the graph to represent the data appropriately. This paper concerns the ability of secondary students to represent numerical data in a graph, and explores factors which assist students to organise the data before drawing the graph. The term *numerical data* refers here to counts or measures such as the number of pencils a student has, whereas *categorical data* refers to categories such as eye colour.

The ability to draw an organised graph is one in a suite of skills expected of all students according to recent curriculum documents. The Australian Numeracy Benchmarks (Curriculum Corporation, 2000) include the ability of primary school students to organise, summarise, and display information in graphs. Similarly, the National Council of Teachers of Mathematics Standards (NCTM, 2000) has highlighted the need for students at all levels to organise and represent data.

Research into children's ability to draw graphs has included the development of a framework for statistical thinking by Jones, Thornton, Langrall, Perry, & Putt (2000) (Framework). The third construct in the Framework - *Representing Data* - is the main issue under consideration in this study, and incorporates constructing representations that exhibit different organisations of the data. As with the other constructs, four levels of thinking have been proposed for this construct. The levels are defined by statements describing students' data displays in terms of the validity of a display when asked to complete a partial graph, and the degree of reorganisation of data when asked to produce a display. The evidence obtained in this study relates to the latter – the degree of reorganisation of data shown in the display.

According to the Framework, at Level 1 the student produces an idiosyncratic display that does not represent the data set. At Level 2, the student produces a display that represents the data but does not attempt to reorganise the data. At Level 3, the

student not only produces a display that represents the data but also shows some attempt to reorganise the data. At Level 4, the student produces multiple valid displays, some of which reorganise the data.

Research in the area of students' representation of data is not extensive, however a small number of studies provide some background for this study. Lehrer and Schauble (2000) investigated the process of data organisation with elementary school children in grades 1, 2, 4 and 5. They examined how these children developed and justified models to categorise (by grade level) drawings made by children in the same grade levels as themselves. Their results suggest that, at higher grades, children use more sophisticated strategies for organising data.

Nisbet (1999) examined the representations of categorical data generated by teacher-education students. The majority (99%) drew representations of the data showing some reorganisation of the data. However, the data was categorical, not numerical. Nisbet, Jones, Langrall & Mooney (submitted) analysed children's representations of categorical and numerical data. The study revealed that numerical data was significantly harder for children to organise and represent than categorical data. Children beyond Grade 1 can make connections between organizing and representing data when the data are categorical but generally not when the data are numerical. Whereas 60% exhibited Level 3 thinking with categorical data by reorganising the data, only 20% exhibited Level 3 thinking with numerical data. Two of the three Level-3 thinkers produced a tally table while the third drew a pictograph.

Another study (Nisbet, 2001) found that teacher-education students had similar difficulties with organising numerical data. All could produce an organised graph from categorical data, but only 19% could produce an organised graph from numerical data. For the latter, the majority of students merely drew separate bars for each individual piece of data without organising the data into numerical categories.

Why do more students find it difficult to represent numerical data in an organised way, compared to categorical data? It could be that the way to organise categorical data is obvious, but less obvious for numerical data. Maybe the need for organisation is not perceived to be great when there are only 10 items in the dataset. Perhaps, if the dataset was made larger, then the students would be more likely to see the need to organise the data, and subsequently draw an organised graph based on numerical categories. This proposition was the motivation behind the current study.

This study was therefore designed to test the hypothesis that if students were presented with two data sets, one small (say 10 items) and the other significantly larger (say 30 items), then the students would be more likely to draw an organised graph of the larger dataset. It was decided to conduct this investigation with secondary students as the statistical thinking of this band of the age/grade spectrum had not been investigated by the researcher. Students in Grades 9 and 11 were given the task of drawing two graphs – one for a small data set, and a second for a larger set. The Framework was used to evaluate the graphs produced by the students' organisations and representations of the data, and an interview protocol was employed to ascertain the extent of prompting required before students realised that the larger set needed to be organised before it could be represented meaningfully.

Method

Participants

A sample of 20 students at a suburban secondary school was drawn from Grades 9 and 11 – eight students in Grade 9, six students from the Grade 11 Mathematics A classes, and six students from the Grade 11 Mathematics B classes. At the school, all Grade 9 students do a common mathematics course, but Grade 11 students can select either Mathematics A ("life-skills mathematics") or Mathematics B, a higher-level course which includes algebra and calculus. The latter however is recommended only for students who achieved highly in Grade 10 mathematics. Mathematics B students in Grade 11 are more mathematically able than their Mathematics A counterparts. The students were selected for this study by the Head of Mathematics such that a spread of achievement (high, medium & low) was included in all three groups.

Tasks

Participants were given two tasks. The first required them to draw a graph representing the information in the following scenario.

> *Ten students were asked about the number of novels they read during the term. These are their answers.*
> *NUMBER OF NOVELS: 5, 4, 1, 7, 5, 0, 3, 4, 5, 6*
> *Draw a graph which represents this data.*

After the students had drawn their graphs, they were asked the following questions:

> *(i) What sort of graph did you draw? and*
> *(ii) Why did you draw it that way?*

The second task required the students to draw a graph representing this information.

> *Thirty students were asked about the number of CDs they bought during the year. These are their answers.*
> *NUMBER OF CDs: 2, 4, 2, 7, 5, 0, 3, 4, 5, 1, 5, 4, 1, 7, 5, 0, 3, 4, 5, 6, 3, 4, 8, 6, 3, 2, 3, 4, 5, 6*
> *Draw a graph which represents this data.*

If a student had drawn an organised graph successfully, he/she was asked:

> *(i) What sort of graph did you draw? and*
> *(ii) Why did you draw it that way?*

If the student was experiencing difficulty in working out what to do, a protocol comprising a series of prompts was available, and the extent of prompting necessary for the student to embark on the correct course of action was noted.

> *1. How many students bought no CDs? How many bought 1? Does that help with your graph?*
> *2. Could you fill in a table of values like this? (Show blank table of values as in Table 1)*

3. Could you draw a graph with this table of values? (Show completed table of values.)

Table 1: Prompt No. 2 – Blank table of values (No. of CDs)

No. of CDs									
No. of people									

If those prompts were not sufficient, then the student would be shown a graph of the data set (Figure 1) and asked the following question.

4. Here's a graph drawn by someone else. What does it tell you?

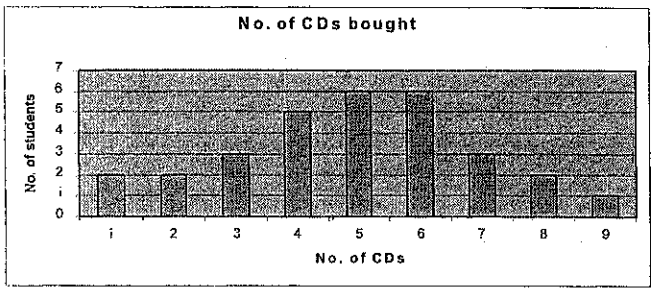

Figure 1: Graph shown in Prompt 4.

Each student was interviewed individually away from any class distractions, and all interviews were audio-taped. The researcher also kept brief notes of the interviews. Sheets of graph paper, rulers, pens and pencils were supplied by the researcher, and all graphs drawn by the students were collected by the researcher for analysis.

Results

Overview of results

1. The effect of size of data set: With the small data set (10 items) most students drew graphs showing no organisation of the data. However, increasing the size of the data set to 30, lead more students to organise the data and draw a representation based on number of CDs rather than individual measures.
2. The effect of mathematics ability: Most Grade 11 students in the higher ability group (Mathematics B) were able to organise and represent the data in an organised way without any prompting. However, only one student in lower ability group (Mathematics A) completed Task 2 without any prompting. There was no similar ability effect for the students in Grade 9.
3. The effect of Grade level: Grade level had no effect on performance at organising and representing numerical data.

Results in detail

Increasing the size of data set lead more students to organise the data and draw a representation based on categories rather than individual measures. In Task 1, which had only 10 items of data the majority of participants (95%) did not organise the data, but without much hesitation, drew a bar graph based on the individual pieces of data – one bar for each person in the dataset, showing how many novels read (Figure 2).

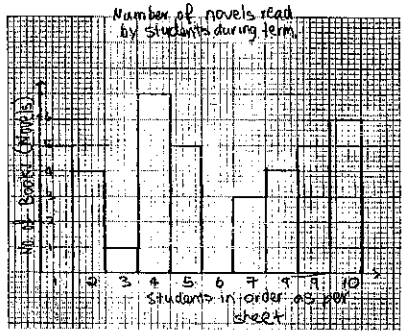

Figure 2: Typical bar graph drawn for data in Task 1

According to the Framework, one student produced a response at Level 1 (Idiosyncratic), 18 students produced responses at Level 2 (Transitional), and one student produced a response at Level 3 (Quantitative). The response patterns were fairly uniform across the three groups of students.

(See Table 2)

Table 2: Numbers of students and level of thinking in Task 1 by group.

Mathematics group	Level of thinking (representing data)			Total students
	1 (Idiosyncratic)	2 (Transitional)	3 (Quantitative)	
Grade 9	1	7	0	8
Grade 11A	0	6	0	6
Grade 11B	0	5	1	6
Total students	1	18	1	20

Only one student organised the data into categories, e.g. the number of novels read. The dependent variable in her graph was the number of people who read that many.

The pattern of responses was quite different for Task 2, which had 30 items of data (number of CDs bought by 30 people). In response to this task, 10 of the participants (50%) organised the data into categories without any prompting and drew graphs based on the how many CDs people bought. The other 10 participants required varying degrees of prompting before they realised how the data could be organised first and then represented graphically. Table 3 shows the number of students requiring prompts to represent the data in Task 2 in an organised fashion.

Table 3: Number of students requiring prompts for Task 2 by group.

Mathematics group	Number of prompts			Total students
	0	1	2	
Grade 9	4	0	4	8
Grade 11A	1	1	4	6
Grade 11B	5	1	0	6
Total students	10	2	8	20

There was no significant effect of mathematics group, nor of Grade level. However, there was a significant difference between the Grade 11 Mathematics A students and the Grade 11 Mathematics B students [$\chi^2(2, N = 20) = 6.67$, $p < .05$) indicating an effect of mathematics ability. Students in Mathematics B produced organised data

representations with fewer prompts than students in Mathematics A. Interestingly, there was no similar ability effect for students in Grade 9. Of the four Grade 9 students who produced an organised graph without any prompts, two were A students (high achievement) and two were D students (low achievement). Further, the other four students who required two prompts to assist in organising the data were spread across the achievement spectrum – their results were A, B, C, and E.

Discussion

The effect of dataset size seems to imply that when students see the need to organise data they are more inclined to draw an organised graph. The nature of statistics vis-a-vis data handling comes to the fore when students are faced with a large dataset. Unfortunately, not all students understand how to organise the data in such situations.

What factors come into play in determining how well students are able to organise the data? This study showed that there was an ability effect demonstrated, but it evident in Grade 11 students but not Grade 9 students. The ability effect is quite understandable, but why only Grade 11 students? Either the levels of achievement allocated by the school for the Grade 9 students are not valid measures of their true mathematical ability, or the ability effect is more to do with mathematical processes such as categorising than passing standard mathematics exams. The study concluded that there was no age effect – Grade 11 students did no better than Grade 9 students. So it appears that performance at drawing organised graphs is related to some extent to a general mathematics ability rather than age.

Another factor relevant to some students' difficulties in organising numerical data may be the mathematics curriculum in its various forms – the formal syllabus, the school work program, the intentions of the teacher, and students' experiences in the classroom. Hopefully secondary students would have had extensive experience in collecting, organising and representing data over seven years of primary mathematics. The current state-wide Mathematics syllabus, which has been in official use since 1987, incorporates statistics from Grade 3 onwards. Topics include collecting, organising, and representing data with various forms of graphs – picture graphs, bar graphs, line graphs, histograms and circle graphs. However, most of the examples shown in the support booklets involve categorical data, and very few involve numerical data. With such an emphasis on categorical data in the syllabus, it would be of no surprise if teachers had a similar emphasis, giving students more experience with situations involving categorical data compared to numerical data.

Another syllabus-related issue is the approach taken by teachers in the teaching of statistics. At one end of the spectrum of teaching approaches is a *mechanistic* approach (Ernest, 1989) which implies teaching rules and formulae (e.g., for finding mean) and using data out of a text book. At the other end of the spectrum is a *dynamic* approach (Russell & Friel, 1989) in which students investigate an issue of interest to them by collecting, analysing and representing primary data. It could be that during the students' school careers, most of their teachers have used the mechanistic approach in preference to the dynamic approach. In the mechanistic approach, teachers would have given the students secondary data that are already grouped. Hence the students wouldn't have had to think through the reorganisation step, thus missing a crucial stage of the data-reduction process. The author contends that if more teachers used a dynamic approach, and if an expanded range of techniques for data organisation was specified in the school curriculum and taken up by teachers, then students' would find organising and representing data easier.

A number of other issues did arise from the study. The first issue concerns the participants' choice of direction of axes in the process of constructing the graphs. Many hesitated over which variable was the independent variable and which was the dependent variable, especially in Task 2. In Task 1, most chose "students" as the independent variable and put it on the horizontal axis. Then they chose "the number of novels read" as the dependent variable and put it on the vertical axis, thus producing a series of vertical bars – one bar for each observation. (See Figure 1.) However, in Task 2, when faced with data that had to be grouped according to how many CDs were bought, many participants had difficulty in determining that "the number of CDs" was the independent variable and "the number of students buying that number of CDs" was the dependent variable. This apparent reversal of axes caused some difficulty for the number of students.

Perhaps this difficulty is associated with an inherent complication in the data-reduction process. In collecting the raw data, such as how many CDs people bought during the year, "the number of CDs" is the dependent variable – it *depends* on who you ask! However, when organising the data, "the number of CDs" becomes the dependent variable, and "the number of students" becomes the dependent variable: the number of students varies according to how many CDs they bought! The data have been transformed by the organisation process. Related to this complication was the choice between horizontal or vertical bars in drawing their bar graphs. For most, it was determined by the labels they gave the axes. If they wrote "number of CDs" on the horizontal axis, then the bars drawn were vertical. If, however, they wrote "number of CDs" on the vertical axis, then the bars drawn were horizontal.

Another observed difficulty was coping with zero – zero novels or zero CDs. Firstly, some students did not realise initially that zero was a legitimate piece of data, and that they had to allow for it in their organisation of the data. (In the Task 2 scenario, two people bought zero CDs.) Secondly, in representing the fact that two people bought zero CDs, the bar can't be located at the intersection of the two axes – the vertical axis has to be located away from zero on the x-axis.

The last significant issue noted during the interviews was how helpful the table of values was to most of those who had difficulty initially in working out *how* to organise the data. Seeing a blank table of values seemed to "turn on the light" for them. They immediately knew what to do, filled in the table, and drew the graph.

A number of implications for teaching arise out of the results of this study. Firstly, it is clear that the formal syllabus needs to distinguish between categorical and numerical data, and to place more emphasis on situations involving numerical data. Categorical variables have their place in the early Grades but beyond the

middle-primary Grades, they should take a lesser role. Teachers then need give students greater exposure to organising and representing numerical data. Further, teachers need to highlight the process of organising and summarising data as a pre-requisite to representing the data. Use of the term *summarising* as a key word could be useful for some students – it conveys one of the purposes of statistics! Secondly, the examples used by teachers should involve large datasets, with sample sizes of the order of 30 rather than 10, so that students are challenged to think about organising the data. Teachers should not waste too much time on drawing graphs based on individual measures; they should get onto organising the data e.g. rank ordering, grouping, and tabulating. During this procedure it would be beneficial for teachers to offer *scaffolding* to the students (Wood, Bruner & Ross, 1976) by talking through the processes and their reasoning behind the steps.

In conclusion, it is worth reiterating that a *dynamic* approach to teaching statistics would be the means of integrating all of these curriculum and teaching issues to improve students' skills in representing numerical data.

References

Curriculum Corporation (2000). *Australian Numeracy Benchmarks*. Melbourne, Vic.: Author.

Ernest, P. (1989). The impact of beliefs on the teaching of mathematics. In P. Ernest (Ed.), *Mathematics Teaching: The state of the art*. New York: Falmer.

Jones, G., Thornton, C., Langrall, C., Mooney, E., Perry, B., & Putt, I. (2000). A framework for characterizing students' statistical thinking. *Mathematical Thinking & Learning, 2*, 269-307.

Lehrer, R., & Schauble, L. (2000). Inventing data structures for representational purposes: Elementary grade students' classification models. *Mathematical Thinking and Learning, 2* (1 & 2), 51-74.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Nisbet, S. (2001). Representing categorical and numerical data. In J.Bobis, M.Mitchelmore, B.Perry (Eds.) *Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia*, Sydney, NSW, 1-4 July 2001. Sydney: MERGA.

Nisbet, S. (1999). Displaying statistical data. *Teaching Mathematics, 24*, (1).

Nisbet, S., Jones, G., Langrall, C., & Mooney, E. (submitted). Children's representation of data.

Russell, S., & Friel, S. (1989). Collecting and analyzing real data in the elementary school classroom. In P. Trafton & A. Schulte (Eds.) *New Directions ofr Elementary School Mathematics, 1989 Yearbook*. Reston, Va: National Council of Teachers of Mathematics.

Wood, D., Bruner, J., Ross, S. (1976). The role of tutoring in problem solving. *British Journal of Psychology, 66*, 181-191.