Organising and Representing Grouped Data

Steven Nisbet
Griffith University
<s.nisbet@griffith.edu.au>

Two classes of Year 8 students were asked to organize and represent sets of numerical data with large variation of scores, to determine the extent to which the students had learned to produce grouped data and represent them in displays such as a histogram (as stated in the current Queensland syllabus for Year 7). Analysis of the students' responses in terms of a statistical-thinking framework revealed wide variation in students' ability to re-organise the data and represent them in organised graphs. Only 21% of students were able to complete the task successfully. Interviews with other students indicated that a little prompting in terms of grouping the data assisted many to produce grouped data with convenient interval sizes. The results of the study have implications for the teaching and learning of the organisation and representation of grouped data.

This paper reports on a study of the ability of Year 8 students to organise data as grouped data and represent the grouped data appropriately (in a histogram or other similar graph). These skills belong to the Chance & Data strand of contemporary Australian curricula e.g., the new draft syllabus in Queensland (Queensland Studies Authority, 2003) and to the Statistics topic in previous syllabus documents.

The teaching of statistics in the primary school has been a relatively late starter in Queensland (Nisbet, 2002c) and Australia generally, compared with topics such as number and measurement. Topics in statistics and graphing appeared first in Australian syllabuses in the mid-sixties (ACER & NSW Education Department, 1965), with revisions and extensions made in subsequent syllabus documents. The skills of organising grouped data and representing the data as histograms have been included explicitly in the Queensland syllabus for Year 7 (Department of Education, Qld., 1987) for the last 27 years (Department of Education, Queensland, 1975 & 1987), and are incorporated (less explicitly) in the Chance & Data strand of the National Statement (Australian Education Council, 1991). Similarly, many countries, including the USA, have only recently begun to teach data analysis at pre-tertiary levels and hence research on the reasoning of students in these areas at the school level has not been a priority (Konold & Higgins, 2000).

Nevertheless, research activity in data handling has been growing in recent years, and a number of key studies have helped shape the nature of this investigation. Jones et al. (2000) developed a statistical-thinking framework after extensive research, development and validation with elementary school pupils. The framework consists of four constructs (Describing Data Displays, Organising and Reducing data, Representing Data, & Analysing and Interpreting data), which are described at four developmental levels (Idiosyncratic, Transitional, Quantitative & Analytical). The skill of organising data into groups falls within the construct *Organising and Reducing Data*, and the drawing of a histogram or similar graph falls within the construct *Representing Data*. This framework was used to assist in the analysis of the students' responses to the tasks in the present study.

A number of research studies relate specifically to organising and representing data, some of which have found that the re-organisation of data does not come easily for many students. (Please note that the words *organisation* and *re-organisation* can be used

interchangeably in this context. It is merely semantic as to whether a dataset in its raw form is described as organised or unorganised. It could be seen to be organised in terms of the order in which the original scores were collected, in which case the data are subsequently re-organised into groups, for instance.)

First, it has been established that students from primary to tertiary level find numerical data more difficult to organise than categorical data (Nisbet, 2001; Nisbet, Jones, Thornton, Langrall & Mooney, in press). Results of other studies confirm that organisation of numerical data is a real obstacle for middle school students (Bright & Friel, 1998) and elementary school students (Jones et al., 2001), and demonstrate that many students at these levels have difficulty in accepting the meaning of the term *organising* as *sorting*. Although some bright students in primary school are able to organise numerical data into classes (Konold & Higgins, 2000; Nisbet, 2001; Nisbet, Jones, Thornton, Langrall & Mooney, in press), most attend to the characteristics of individuals rather than the group, and "see the trees rather than the forest". Further, a teaching experiment by Jones et al. (2001) showed that children's thinking in organising and reducing data was problematic. Children were reluctant to use paper and pencil to reorganise data, but technology proved helpful in stimulating their strategies.

In relation to representing data, it has been shown (Nisbet, 2002a, 2002b) that (a) size of dataset is a factor in students' representations of data in that students are more likely to represent large numerical data sets in an organised way compared to small data sets, (b) mathematically-able students are more likely to organise numerical data validly than their less able counterparts, and (c) teachers' prompts to organise data are effective in assisting students produce valid organised representations.

In line with the results from the research studies quoted above, the Queensland Syllabus Sourcebook (Department of Education, Qld., 1990) comments that (a) students will need help with drawing a histogram, and (b) not all students will reach this level of proficiency in Year 7. The present study was conducted at the start of the year with two classes of Year 8 students; the goals were to determine the proportion of students who had reached this level of proficiency by the end of Year 7, and to investigate the issues that the students face when confronted by the task of organising data as grouped data and representing the data in a histogram or similar graph. The analysis of students' responses to the tasks and the probing of the students' thinking have the capacity to highlight issues to be addressed in the planning, teaching and learning of the organisation and representation of data.

The research questions which were specifically addressed in this study were:

1. To what extent can Year 8 students organise data into grouped data and then draw valid histograms, as suggested by the Year 7 syllabus?
2. What levels and types of organisation do Year 8 students demonstrate when asked to organise a large data set (50 scores)?
3. What levels and types of representation do Year 8 students demonstrate when asked to represent a large data set (50 scores)?
4. To what extent does the level and type of data organisation influence the level and type of data representation?

5. What factors influence students' thinking in their organisation of data into groups and the representation of such grouped data?

Method

The participants for this study were 43 students from two Year 8 classes early in their first year at a government secondary school in a Brisbane suburb. As this was a study about what students had learned about handling data during Year 7 (the last year of primary school in Queensland), the students were questioned informally about where they had completed Year 7. Approximately 50% of the students had attended the local government primary school, and the other 50% had attended seven other primary schools—some in adjacent suburbs, the others further away. It was reasonable therefore to assume that the results from the study would be indicative of the situation in a number of schools in the district, not just one. All 43 students participated in the first phase of the study (a written task involving the organisation and representation of a dataset), and following an initial analysis of their work, 12 of these students were selected for the second phase (an interview about how they organised and represented the data).

In the first phase of the study, the participants were each presented with a dataset of 50 scores in a class situation, and were asked to organise and represent the data. To minimise collusion between students, two data sets were prepared—one with the heights (in cm) of 50 students, the other with estimates by 50 people of the number of countries in the world—and the two datasets were handed out to alternate students. The datasets were presented on A4 paper as an array of 5 columns with 10 scores in each column. There was ample space on the paper for the students to organise the data. Each student was also given an A4 sheet of graph paper (1cm x 1cm grid) on which to draw the graph. The students were allowed 30 minutes to respond to the task. Most students needed all this time to complete the task, and some did not finish.

Following the researcher's initial analysis of the students' responses, 14 students were selected for the second phase—individual interviews a few days later—but 2 of those students were unavailable for interview on the day. Hence the interview data relates to 12 students. The interview consisted of questions designed to gain an insight into the students' thinking about how they organised and represented the data and why they did it that way, and to present another data organisation and representation task. The second task related to another set of 50 scores (estimates of how many cars were in a car park) and included more specific directions about organising the data ("organise the data into groups"), and a series of questions to see how much prompting was required to assist the students to produce grouped data and draw a valid histogram. The data was formatted as 5 rows with 10 scores in each row, with less space between scores, to minimise the tendency that some students exhibited in the first task to group the data into 5 groups of 10 by making each column a group.

Initially, the students' graphs and written responses to the tasks were analysed according to the framework (Jones et al., 2000). Levels of organisation and representation were assigned to the methods of organisation and the graphs drawn (respectively). Further classification of responses was undertaken because some levels covered a number of different types of responses, which were significant to note. For instance, for organising

data, Level 1 responses included *no organisation* as one type and *irrelevant grouping* as another. Similarly, Level 3 responses included *ordering* as one type and *grouping* as another. For representing data, Level 3 responses varied according to the type of re-organisation demonstrated—ordered data or grouped data. (There were no Level 2 responses for organisation.)

Results

Research Question 1: To what extent can Year 8 students re-organise data into grouped data and then draw valid histograms, as suggested by the Year 7 syllabus?

Only 21 % of the students were able to re-organise the data into grouped data with convenient interval sizes, and only 14% were able to draw valid histograms. Not all those who produced grouped data drew histograms.

Research Question 2: What levels and types of organisation do Year 8 students demonstrate when asked to organise a large data set (50 scores)?

As indicated above, 21 % of the students were able to re-organise the data into groups with convenient interval sizes, 33% organised the data by putting them in order from smallest to largest or vice versa, 16% made irrelevant groups (5 columns of 10 scores), and 30% did not re-organise the data at all.

Research Question 3: What levels and types of representation do Year 8 students demonstrate when asked to represent a large data set (50 scores)?

As indicated above, 14% were able to draw histograms of the grouped data, 23% drew valid graphs with some type of re-organisation, 37% drew graphs with no re-organisation, and 26% did not draw any display at all.

Research Question 4: To what extent does the level and type of data organisation influence the level and type of data representation?

Level and type of data organisation has only limited influence on the level and type of data representation. Nine students were able to organise the data into groups with convenient interval sizes (Level 3), but only 5 of those students then drew valid histograms (Level 3). One drew a graph of frequencies against individual scores (also Level 3, but not as suitable as a histogram), and the other 3 drew bar graphs with no re-organisation at all (Level 2).

Research Question 5: What factors influence students' thinking in their organisation of data into groups and the representation of such grouped data?

Results from the interviews suggest that asking many students to "organise the data into groups" did prompt them into producing grouped data and then draw valid histograms of the grouped data. Other helpful prompts were "Could the groups be based on size of the scores?", "What size could the groups be: 10s, 50s, 100s?".

Results in Detail

The students' responses were initially analysed in terms of the level of organisation and representation according to the Framework (Jones et al,, 2000). However, for data organisation, it was noted that some levels were exemplified by a number of types. For

instance, Level 1 included 3 types: (a) no organisation at all, (b) mere numbering of scores, and (c) irrelevant grouping. Level 3 included 2 types: (a) put into rank order, and (b) grouped into convenient interval sizes. Similarly, for data representation, Level 3 included 2 types: (a) ordered bar/line graph, and (b) histogram or polygon.

Students' heights dataset. Eight students out of 23 were allocated to Level 1 on Data Organisation. Four students revealed no organisation of the data, and they all drew valid displays but demonstrated no re-organisation of the data. One student merely numbered the scores in ascending order, and he also drew a graph which was valid but had no re-organisation of the data. Three students made irrelevant groupings of the data (5 columns) and 2 of them drew valid bar graphs of the group averages.

Altogether 15 students were allocated to Level 3 on Data Organisation. Twelve students ordered the data from smallest to largest or vice versa, of whom only one drew a histogram with 10cm interval sizes, 6 drew bar graphs with the bars ordered in size, and 5 did not draw any display.

Three of the Level 3 students grouped the data using a convenient 10 cm interval size, 2 of whom drew valid histograms with 10 cm interval size, and the other had the bars for each interval out of order.

"Number of countries in the world" dataset. Altogether 12 students out of 20 were allocated to Level 1 on Data Organisation. Eight students revealed no organisation of the data, six of whom drew valid displays but demonstrated no re-organisation of the data, and 2 drew invalid graphs. Four students made irrelevant groupings of the data (5 columns); one of that group drew a valid bar graph of the group averages, one drew a drew valid display but demonstrated no re-organisation of the data, and 2 did not draw displays at all (merely tabulated the column means). Eight students were allocated to Level 3 on Data Organisation. Two students ordered the data from smallest to largest or vice versa, of whom only one drew valid bar graph with no re-organisation, and one did not draw any display. Six of the Level 3 students grouped the data using convenient interval sizes (25, 50, or 100). Two of that group drew valid histograms; one reverted to a line graph of frequencies of individual scores, and three drew valid bar graphs with no organisation.

A comparison of the results from the two datasets revealed similar patterns of responses, even though the number-of-countries dataset had a greater variation of scores (35 to 400) than the students' heights dataset (132cm to 180cm). Table 1 shows the combined results of organisation levels and types versus representation levels and types.

The results show that level of organisation does not necessarily dictate the level of representation. Although 9 students altogether organised the data into groups with convenient interval sizes, only 5 drew valid histograms, and 4 did not draw graphs based on that organisation, but reverted to bars for each individual. In Konold & Higgins' (2000) words, they were still focused on the trees rather than the forest. It was interesting to note that one of the six students who drew histograms had not organised the data into grouped data, but ordered them from smallest to largest. His graph demonstrated a higher level of thinking for representation than for initial organisation. Nevertheless, levels of organisation and representation were associated, to the extent that none of the students at Level 1 for "organisation" were able to draw a histogram or similar Level 3 display. Apparently, Level 3 in "organisation" is a necessary but not sufficient condition for producing a histogram at

Level 3 in “representation”. The extra condition required is the ability to recognize the need to produce grouped data in order to represent the dataset as a whole.

The interviews with the 12 selected students confirmed the methods of organisation and representation revealed in the students’ written work. They also revealed that for those who demonstrated no organisation of the data, this was a difficult exercise for which they had no experience, and which consumed a lot of the allotted time. The interviews also confirmed that the format of the data presented to the students (5 columns of 10 scores) lead many to assume that there were 5 groups of data to analyse separately.

Table 1.
Cross-Tabulation of Organisation and Representation for Initial Datasets Combined, Showing Numbers of Students in Each Cell

Organisation level and type	Representation level and type			
	Level 1: No display or invalid display	Level 2: Valid display but no re-organisation	Level 3: Valid graph of ordered data or group averages	Level 3: Histogram or polygon
Level 1: No organisation	2	11	0	0
Level 1: Irrelevant groups	3	1	3	0
Level 2: Inconsistent	0	0	0	0
Level 3: Ordered	6	1	6	1
Level 3: Grouped	0	3	1	5

After enquiring about the students' performance in the initial task, the researcher asked each of the students interviewed to examine another dataset of 50 scores (people's estimates of the number of cars in a shopping-centre car park) and organise the data into groups. (These were the students who were not able to produce grouped data in the first task.) The researcher had a series of questions (prompts) ready for each student, and noted during the interview how many prompts he had to provide before the student came up with grouped data with intervals of a convenient size (e.g. 1-50, 51-100, 101-150, etc.). Table 2 shows a summary of the results from this part of the interview, with the pseudonyms of the students, their responses to Task 1, and the number of prompts required in Task 2.

Table 2.
Cross-Tabulation of Organisation for Task 1 (Initial Written Exercise) With Number of Prompts Required to Produce Grouped Data in Task 2 (Follow-Up Interview), Showing Pseudonyms of Interviewed Students for Each Cell

Organisation level and type in Task 1	Number of prompts required in Task 2		
	1 prompt	2 prompts	3 prompts
Level 1: No organisation	Julio, Lorraine		Felicity
Level 1: Irrelevant grouping		Tim, Pedro, Barry, Jack	Sally

Level 2: Inconsistent		
Level 3: Ordered	Tan, Lucy	Clarissa
Level 3: Grouped	Winston	

The results from the interviews showed that most students did not require much prompting to think of how to group the data in the second task. Most thought of how to produce grouped data with convenient interval sizes after only 1 or 2 prompts. The phrase "organise into groups" helped them realize what was meant by "organise the data". Inspection of Table 2 shows that there is little relationship between the students' level and type of organisation in Task 1 and the number of prompts required to produce grouped data and a valid histogram in Task 2. Those 5 students who needed only 1 prompt for Task 2 were spread across Levels 1 and 3 of organisation in Task 1. However, those requiring 3 prompts in Task 2 were only from Level 1 of organisation Task 1.

Discussion

Although many students found the organisation and representation of grouped data difficult initially, it seems as though these skills could be taught effectively to Year 7 students, as required by the current syllabus. It is possible that the reason many students were not successful at Task 1 was that they had not been taught the skills in Year 7. Hence, it is feasible that most students would respond positively to well-planned instruction on these techniques.

It appears that in teaching these topics there is a need to establish a reason for grouping data. Some students do not understand that statistics allows us to describe and represent the data overall, not just individual scores. As Konold & Higgins (2000) describe, these students don't see the forest for the trees. Organisation and representation of data helps fulfill the main purpose of descriptive statistics i.e. to give an overall picture of the dataset.

The study found that the formatting of the data influences the method of organization students employ. For example, 5 columns of 10 scores in Task 1 lead some students to assume that there were 5 groups, and they proceeded to calculate the means for the 5 groups. In other words, the formatting caused them to embark on an irrelevant method of organization. This was not a problem in Task 2 where the data was arranged into 10 columns of 5 scores and the columns were less obvious as separate groups.

The grouped data produced by students in this study varied in two characteristics—interval size and interval limits. Interval sizes varied enormously in this study and included 2s, 10s, 20s, 50s and 100s. Some students set the actual interval limits by the lowest score in the dataset e.g. 35-135, 135-235, over 235, etc. Other students set the interval limits using round numbers e.g. 0-100, 100-200, 200-300 (sic).

As hinted at in the previous sentence, the interval limits often showed a duplication of the minimum and maximum points of the ranges e.g. the score of 50 lies in both ranges 0-50 and 50-100. This did not seem to be a problem to the students until either they were faced with having to assign a score of, say, 50 to an interval, or it was pointed out to them. Most students were not sure how to deal with this situation, but were satisfied with the interviewer's suggestion of having intervals of 1-50, 51-100, 101-150, etc.

With recognition of the caveats noted above, it appears that even though some students may find organizing and representing grouped data difficult initially, it is feasible that a well-planned unit of lessons would assist most Year 7 students to perform such tasks successfully.

References

Australian Council for Educational Research & NSW Department of Education (1965). *Background in Mathematics: A guidebook to elementary mathematics for teachers in primary schools.* Sydney: Author.

Australian Education Council (1991). *National statement on mathematics for Australian schools*. Melbourne.: Curriculum Corporation.

Bright, G. & Friel, S. (1998). Graphical representations: Helping students represent data. In S. Lajoie (Ed.), *Reflections on statistics: Learning, teaching and assessment in Grades K–12*. Mahwah, NJ: Erlbaum.

Department of Education, Queensland (1990). *Sourcebook for Year 7, Years 1 to 10 Syllabus in Mathematics.* Brisbane: Author.

Department of Education, Queensland (1987). *Years 1 to 10 Syllabus in Mathematics.* Brisbane: Author.

Department of Education, Queensland (1975). *Program in Mathematics.* Brisbane: Qld Government Printer.

Jones, G., Langrall, C., Thornton, C., Mooney, E., Wares, A., Perry, B., Putt, I., & Nisbet, S. (2001). Using statistical thinking to inform instruction. *Journal of Mathematical Behavior, 20*, 109–144.

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.

Konold, C. & Higgins, T. (in press). Working with data: Highlights of related research. In S. Russell, D. Schifter & V. Bastable (Eds,), *Working with data. A module of the K–6 professional development curriculum, Developing mathematical ideas.*

Nisbet, S. (2002a). Year seven students' representation of numerical data: the influence of sample size. In B. Barton, K. Irwin, M. Pfannkuch, & M. Thomas (Eds.), *Mathematics education in the South Pacific, Proceedings of the 25th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 520–527). Auckland: MERGA.

Nisbet, S. (2002b). Representing numerical data: The influence of sample size. In A. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th Annual Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 417–424). Norwich, UK: University of East Anglia.

Nisbet, S. (2002c, January). *The development of students' thinking in Chance & Data.* Paper presented to Biennial Conference of Australian Association of Mathematics Teachers, Brisbane.

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: MERGA.

Nisbet, S., Jones, G., Thornton, C., Langrall, C., & Mooney, E. (in press). Children's Representation and Organisation of Data. Manuscript accepted for *Mathematics Education Research Journal 15* (1).

Queensland Studies Authority (2003). *Years 1 to 10 Mathematics Syllabus, March 2003 Draft.* Author.