Representing association: Children manipulating data sets

Helen L. Chick
University of Melbourne
h.chick@unimelb.edu.au

Introduction

The ability to deal with multi-variate data and to understand associations between variables is an important component of statistical literacy. The teaching of these topics is often delayed until late high school, yet there is evidence that younger students can comprehend aspects of association, including interpreting data where an association exists. Representing associations, however, has been less extensively studied for these students. There may be an assumption that students lack techniques for such representations; furthermore, young students may not appreciate that claims about data ought to be supported by some form of evidence, thus necessitating representation (cf., Chick, 2000). This study considers these issues and examines how the ability to represent statistical associations develops.

Background

Transnumeration

One of the important purposes of statistics is to provide evidence for stories that are contained within data sets. All the calculations or representations that are produced—from simple frequency tables to correlation coefficients to *t*-tests and more advanced techniques—are intended to present the data in such a way that the reader is convinced about the message within. The process of taking a data set and producing a more informative representation of it is vitally important. Wild and Pfannkuch (1999) introduced the term *transnumeration* for this process, with Pfannkuch and Rubick (2002) identifying three particular instances in statistical thinking. For the current report two are pertinent: transforming raw data into other representations (e.g., sorted data, graphs, means) in order to find meaning in the data, and communicating this meaning to others. Chick (2003) suggested that there is an "art" to this process, and that success is dependent upon an ability to identify the message, an appreciation of the need to present the data in a way that provides evidence for that message, and then the ability to select and produce an appropriate representation. This becomes critical when the messages in the data are complex, such as with associations, and depends on the statistical skills available in one's repertoire.

The Challenge of Representation

Watson and her colleagues have investigated the approaches used by school-age students to represent data, using the "Data Cards protocol" first reported in Watson, Collis, Callingham and Moritz (1995). The Data Cards protocol provides students with a data set involving seven variables (both categorical and numerical) about 16 students and asks students to explore and represent the data. In one study Chick and Watson (2001) found, perhaps not surprisingly, that representation is generally more difficult for students than finding meaning in the data. The Grade 5/6 students in their study were asked to produce posters and graphs to represent aspects of the supplied data set. Although the students could often identify relationships within data, or interpret others' representations, their own representations were not always at the same high level. Pfannkuch, Rubick and Yoon (2002) similarly found that Grade 7 and 8 students using the Data Cards protocol were able to identify quite complex "messages" in the data, but had only a limited repertoire of strategies available for organising data to represent these messages clearly.

The Place of Association

The idea of association is an important one. Moritz (2000, p. 440) points out that whereas formal numerical statistics—such as *t*-tests and correlation coefficients—can be used to convey association, graphical approaches offer a useful and accessible visual alternative. He has

conducted a number of studies examining the representation and interpretation of association. In one of these (Moritz, 2000) he gave Grade 4-6 students the description of an association, such as "people grow taller as they get older", and asked the students to draw a graph to show this. Note, however, that no actual data were supplied. Students' representations varied from idiosyncratic pictures that failed to show the association to reasonably standard graphs successfully illustrating the indicated relationship. In another study, Moritz (2003) examined interpretation of an already existing representation. In both studies, one of the interesting outcomes was evidence that young students can appreciate aspects of association, despite many curriculum documents placing it late in the curriculum. It seems timely to examine further the development of students' understanding of this important aspect of statistical thinking.

The present study

With these issues in mind, the purpose of this study is to start investigating the techniques used by students of different ages and statistical backgrounds to represent associations within data. In contrast to the studies of Moritz (2000, 2003), participants are supplied with an actual data set, based on the Data Cards protocol of Watson et al. (1995), but with new data incorporating a smaller set of variables. The focus is on representation and the transnumerative processes undertaken while producing a representation, rather than on the interpretation of the data.

The Participants

The participants in this first phase of the study were approximately 100 Year 7 girls (ages 11-13) at a private school in a major Australian city. The study was conducted shortly after the start of the school year, so students' experience of data handling would have been in elementary school in earlier years. Curriculum guidelines suggest that they should have completed work on producing bar and line graphs, tables, and time series data, with some work on grouping and ordering data, and computing simple statistics up to and including the mean.

The Task

The data set given to the participants was presented as a table. The names of 16 children were listed, together with each child's favourite activity, number of hours of exercise per week, and the weekly number of fast food meals consumed. Two pairs of variables exhibited associations: the two numerical variables (hours of exercise and number of fast food meals), and the categorical variable (favourite activity) with hours of exercise (a numerical variable). The students were informed of possible associations among the variables, and then asked to produce a representation that would convince someone else of the association. The following prompts were given to students to encourage them to notice and represent these associations.

1. A group of people looked at this data set and said that they thought that people who ate lots of fast food didn't seem to do much exercise. Can you draw a graph or something similar to show this?
2. They also said that they thought that people who had more active favourite activities did more exercise during the week. Can you use the data to draw a graph or something similar to demonstrate this so that you could convince your friends?

Participants had about 40 minutes to work on the task and were supplied with graph paper marked with a 1cm grid. It is acknowledged that supplying graph paper may have prompted graphical responses but participants were reminded that it was not necessary to draw a graph if they felt some alternative representation showed the requested relationship.

Data Analysis

Students' responses to each of the two prompts were sorted according to the extent to which the representations portrayed the indicated relationships. This study reports only a preliminary analysis of this data, through an examination of the qualitatively different approaches taken to representing data sets. In reporting the results, the word "student" will refer to the participants in the study, and the words "child" and "children" to the individuals in the task's data set.

Results and Discussion

For each of the two sets of graphs the degree of sophistication of the graphs—and the level of success in portraying the desired association—was dependent on the extent of transnumeration that took place. Some students simply duplicated the data in graphical form, whereas others undertook sorting, grouping, other rearranging, or calculation of means to produce more convincing representations. Some examples are presented to illustrate the changes in effectiveness of the transnumeration undertaken to produce the different representations.

Dealing with Two Numerical Variables

As suggested above and illustrated in Figure 1, a number of students essentially duplicated the data associated with the two variables, without any reordering. The only transnumeration—or change in representation of the data—that took place was to "transliterate" the tabular data as a graph. To "see" the association between exercise and fast food consumption it is necessary for the viewer to do a pairwise comparison between the top graph and the bottom to see that when one value is high the other is low. This comparison needs to be conducted across the whole data set in order to gain evidence for the association.

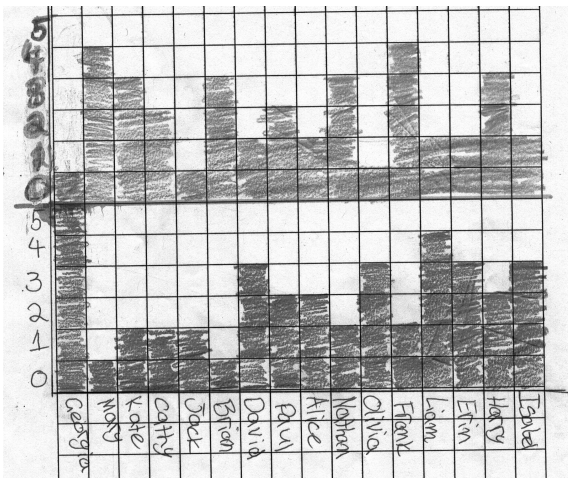

Figure 1. Duplicating the data: Two variables on two separate graphs, aligned vertically. (A key on the graph indicated that darker shading is exercise, lighter

Those students who produced graphs like that in Figure 2 undertook a second stage of transnumeration. Again there has been no sorting of the data, but this time the values have been incorporated on a single graph, by placing each child's exercise and fast food values side by side. The reader must still undertake a visual comparison between the two, again looking for one being high and the other low, but this is much easier because of the juxtaposition of variables.

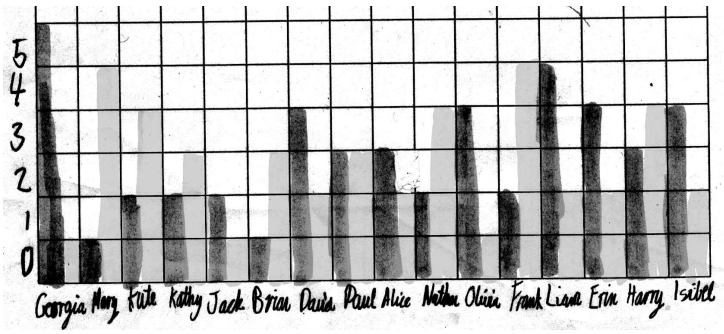

Figure 2. Duplicating the data but with clearer comparisons between the values of exercise (dark columns) and fast food (light columns).

A few students sorted the data on one variable—so that, for example, fast food consumption values were increasing—and then listed the corresponding values of the other variable. Provided the reader of the resulting table understands what has been done, a scan of the second column—in this case reporting hours of exercise—can be conducted to determine the associated trend in that variable. For this example the fact that the trend is generally decreasing supports the claim that those who eat more fast food do less exercise.

Finally, a small number of students were able to construct a conventional scatter graph, as illustrated in Figure 3. Although this is a standard approach for presenting data associated with two numerical variables it is not taught in Australia until later in high school; nevertheless some Year 7 students have either been taught the approach or have seen examples of it and been able to apply the principles to the production of their own graphs. Some of the students labelled the points with the name of the child; the student in Figure 3 has not, suggesting recognition that the names are not actually relevant to the question of showing association and that data associated with the name variable can be "lost" to the representation. Many students want to retain all details of the data for as long as possible, and seem reluctant to compress or omit data, even when doing so would make the message in the data clearer.

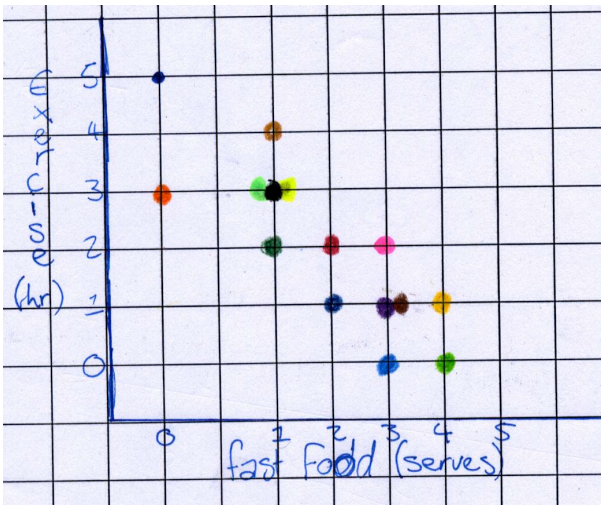

Figure 3. A conventional scatter graph illustrating the association between hours of exercise and fast food.

Dealing with Two Variables, One Categorical and One Numerical

The different levels of transnumeration undertaken by the students for the second part of the task parallel those of the first. Since one of the variables was categorical, however, some unconventional strategies had to be implemented to represent this aspect of the data. Some students used an approach similar to that in Figure 1, with the numerical variable (hours of exercise) represented as a bar graph, and the favourite activity represented on a separate graph directly above, with the activities listed up the y-axis of the second graph. For some students the second graph had bars going up to the height of the activity; other students coloured in the single square across from the activity. In both cases, the students' transnumeration is insufficient to see the message in the data, as the viewer has to undertake additional visual cross-comparison in order to determine the relationship between exercise and favourite activity.

Other students, paralleling the examples illustrated by Figure 2, produced graphs similar to Figure 4. Again the viewer must make the comparisons across the data, but this is easier to examine with the activity name and the hours of exercise in close proximity. The lack of grouping or ordering of the data obscures the relationship between the two variables, thus making this representation ineffective for showing the desired relationship.

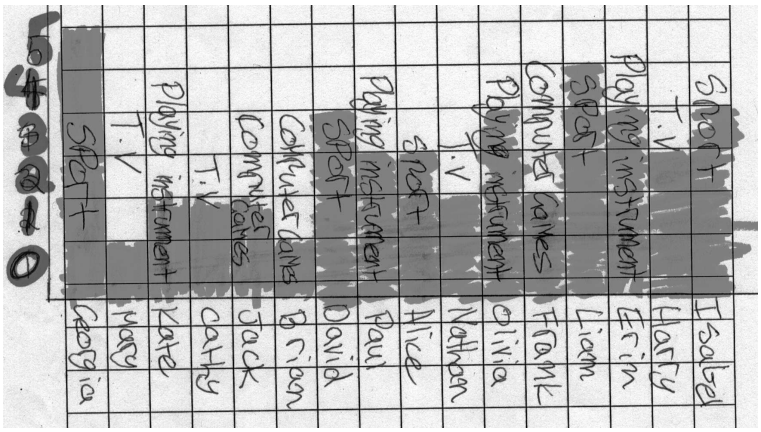

Figure 4. Duplicating the data, with exercise shown as a column graph and corresponding favourite activities written on the columns.

More successful approaches for showing the relationship between exercise and favourite activity were achieved by students who grouped the data by activity category. Having accomplished this first stage of transnumeration, the next problem for these students was to determine what to do with the grouped data. Some students calculated the total hours of exercise for all the students in each of the activity categories, and produced a bar

graphs of these totals. Such a representation obviously fails to take into account the number of students contributing to the total. One of the students represented this frequency data in a pie graph, requiring additional (but unnecessary) transnumeration to work out the number of hours of exercise in each of the categories as a proportion of the overall total.

Other students produced representations that allowed a more appropriate comparison across the categories. The student who produced Figure 5 constructed separate bar graphs for all the students in each of the favourite activity categories. This makes possible a "by eye" comparison across the groups, to give a visual determination that those who play sport do more hours of exercise than those liking computer games or watching TV. Similarly, the quasi-scatter graph in Figure 6 resulted in the sportspeople ending up in a group in the top right hand corner, with higher hours of exercise than the other groups. The student supported the graphical representation with some explanatory text:

> Those whose favourite activity was sport had more hours of exercise. The 5 whose favourite activity was sport all had at least 2 hours of exercise. Those whose favourite activity was playing a musical instrument seemed to be the next active, while those who liked either watching TV or playing computer games were the least active.

Comparisons with the other groups are also possible, with the comparison made a little easier than in Figure 5, thanks to the fact that the quasi-scatter graph has rows of data associated with each activity that can then be compared vertically.

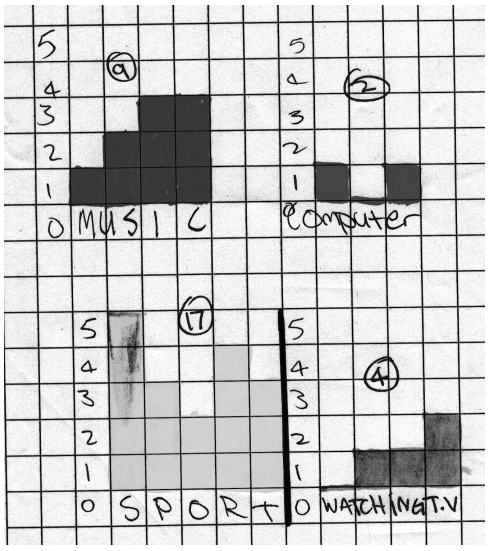

Figure 5. Data grouped by favourite activity, with sets of bar graphs showing the number of hours of exercise for each student in the activity category.

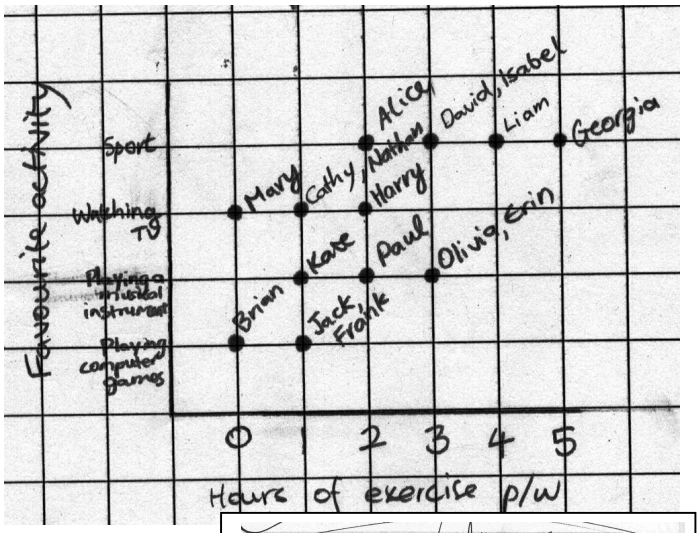

Figure 6. Scatter
favourite act

One student used an approach similar to that in Figure 5, but with only two of the favourite activity categories, sport and playing a musical instrument. This allowed the student to show a contrast between two of the groups, but was insufficient for fully supporting the claimed association.

Finally, a few students appreciated one of the main purposes for calculating means: namely, to allow

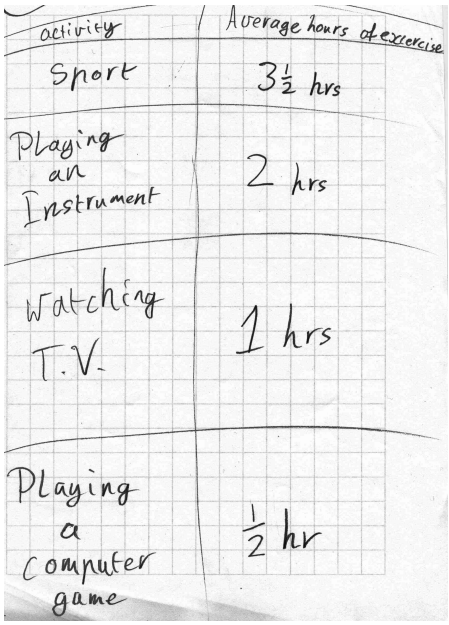

Figure 7. A student's table of average hours of exercise for each of the favourite activities.

comparisons of numerical variables among groups of unequal size. These students, as illustrated in Figure 7, calculated the average number of hours of exercise undertaken by the children in each of the favourite activity categories, thus allowing a numerical rather than graphical/visual comparison. Means quantify the differences visible in graphs like Figures 5 and 6, and, in this case, allow students to rank the activities according to the average number of hours of exercise, as seen in Figure 7. This level of response requires both grouping of data and then compression of data through computation. On the basis of the data obtained in this research, both of these techniques seem to be uncommon among students of this age. It should be noted that according to the curriculum documents the students in the study should have had some experience with calculating means; it is intriguing that so few students could apply the idea to this situation (cf. similar findings by Watson & Moritz, 1999).

It should be noted that some students responded to both parts of the task with a single representation. Many of these responses duplicated the data with no sorting, grouping, or compression, and were essentially a combination of the graph shown in Figure 1 and its parallel for the second task. This produced a set of three aligned bar graphs (or position graphs, showing a coloured square opposite the category), with the viewer having to do the bulk of the work in identifying the existence of the claimed associations.

Conclusions

The results reported here comprise only a preliminary and qualitative analysis of the data set. Nevertheless, they highlight that students can be quite creative in their attempts to deal with data, but also reveal some of the issues that may need consideration in the classroom. The results thus lend support to the suggestion of Pfannkuch, Rubick, and Yoon (2002) that students should be encouraged to create their own representations before being introduced to conventional ones, while also indicating some areas that need to be addressed explicitly. These include:

- The importance of a representation as evidence of some claim about data;
- The importance of grouping and sorting data as strategies for both understanding and representing data;
- The importance of compressing data in order to convey a message more effectively, with recognition that the resulting "loss" of data is acceptable or even advantageous;
- The importance of the mean for comparing groups, noting that often classroom activities only involve calculating the mean for an entire data set, rather than for several data sets or for subgroups of the data (cf., Watson & Moritz, 1999); and
- The power of conventional representations, and what types of data sets they suit.

This research suggests that students in the first year of high school do have the capacity to represent association, even if awkwardly or unconventionally. Use of the mean—presumed taught to these students—was rare, so it will be interesting to see the frequency of other taught strategies—such as scatter graphs—for other year levels. Extensions of this study will examine the representations and transnumerative strategies used by students of other ages, as the range of tools in students' statistical tool kits changes.

References

Chick, H. L. (2000). Young adults making sense of data. In J. Bana & A. Chapman (Eds.), *Mathematics Education Beyond 2000* (Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia, Fremantle, pp. 157-164). Sydney: MERGA.

Chick, H. L. (2003). Transnumeration and the art of data representation. In L. Bragg, C. Campbell, G. Herbert, & J. Mousley (Eds.), *Mathematics Education Research: Innovation, Networking, Opportunity.* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, Geelong, pp. 207-214). Sydney: MERGA.

Chick, H. L., & Watson, J. M. (2001). Data representation and interpretation by primary school students working in groups. *Mathematics Education Research Journal*, *13,* 91-111.

Moritz, J. (2000). Graphical representations of statistical associations by upper primary students. In J. Bana & A. Chapman (Eds.), *Mathematics Education Beyond 2000* (Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia, Fremantle, pp. 440-447). Sydney: MERGA.

Moritz, J. (2003). Interpreting a scattergraph displaying counterintuitive covariation. In L. Bragg, C. Campbell, G. Herbert, & J. Mousley (Eds.), *Mathematics Education Research: Innovation, Networking, Opportunity.* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, Geelong, pp. 523-530). Sydney: MERGA.

Pfannkuch, M., & Rubick, A. (2002). An exploration of students' statistical thinking with given data. *Statistics Education Research Journal, 1*(2), 4-21.

Pfannkuch, M., Rubick, A., & Yoon, C. (2002). Statistical thinking and transnumeration. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics Education in the South Pacific* (Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia, Auckland, pp. 567-574). Sydney: MERGA.

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation, 1,* 247-275.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.