Tools for Transnumeration: Early Stages in the Art of Data Representation

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

This paper considers the skills needed for data representation. A framework of transnumerative techniques that facilitate data representation is proposed and applied to the responses of 73 year 7 students to two tasks involving association. Students' responses were classified according to their levels of success in representing the association, and the types of techniques used. It was found that while students had techniques for representing data, their choices of graph type were not always suitable, and they overlooked simple techniques such as ordering and grouping data that could have made their representations clearer. Implications for doing and teaching data analysis are discussed.

Konold and Pollatsek (2002) describe data analysis as searching for signals in noisy processes. It can also be thought of as finding and revealing messages among data. In a data-drenched world statistical literacy—particularly the capacity to identify such signals or messages—is essential. Statistics is now recognised as a more important component of the mathematics curriculum than ever before (e.g., Board of Studies, 2000; National Council of Teachers of Mathematics, 2002). For data presented in the form of tables and graphs the ability to read that data, and read between and beyond the data (Curcio, 1987) is vital, and these skills have been the focus of many studies. It is the process of representation rather than interpretation, however, which is the focus of the current study, particularly in order to convey messages that are within data. Evidence from at least one study (e.g., Chick & Watson, 2001) suggests that, for small data sets at least, students can interpret trends and facts about the data, but may have difficulty representing these clearly. It has also been suggested (e.g., Chick, 2000) that students may not appreciate that claims about data ought to be accompanied by evidence in the form of a convincing representation.

That data representation can be difficult is known. Tufte's seminal work (1983) gives numerous examples of poorly designed graphs produced by professionals. Turning to students' attempts at data representation, research with the open-ended Data Cards task of Watson and colleagues showed that bar graphs of frequency data (univariate data) are done well, but that making comparisons or representing relationships (multivariate data) is more difficult (e.g., Chick & Watson, 2001; Pfannkuch, Rubick, & Yoon, 2002; Watson & Callingham, 1997; Watson, Collis, Callingham, & Moritz, 1995). Nisbet's studies (e.g., 2001, 2003) suggest that categorical data are more readily represented than numerical, and that students do not often decide to group data. In considering older students' attempts to represent data, Chick (2003) found that whereas none of the representations lied about the data, some were better at depicting messages than others, and she suggested that there is an art to choosing the most appropriate representational technique.

The search for associations among variables is an important aspect of statistics, often delayed until late high school. Evidence suggests, however, that younger students can comprehend aspects of association and interpret data where an association exists; the representation of such associations, on the other hand, has been less extensively studied. Moritz (2000) asked primary schools students to graphically depict a verbally described association, finding that many could produce reasonable, if unconventional,

representations. He did not, however, supply students with actual data. Few of the primary school students using the Data Cards data set, which allows students choice of what aspects to consider, chose or knew how to represent relationships between variables, with most concentrating on univariate analysis (see, e.g., Chick & Watson, 2001). In contrast, the able Year 7 and 8 students in the study of Pfannkuch, Rubick, and Yoon (2002) produced some multivariate representations—not always conventional—that allowed comparisons between groups and the identification of associations.

What is it, then, about representation that makes it difficult? Given a data set, it is clear that *something* must be done to it to produce a representation. Even merely re-presenting the raw data in a different format, such as a table, requires an understanding of that format. Of course, better representations usually result from additional manipulation of the data that makes clearer the messages therein. Wild and Pfannkuch (1999) introduced the term *transnumeration* for the process of "changing representations to engender understanding" (p. 227). Their description included three aspects: (i) capturing measures from the real world, (ii) reorganising and calculating with data, and (iii) communicating data through some representation. The latter two are the focus here. Success with representation—particularly when messages in the data are complex, such as with association—is clearly dependent on knowing what types of representation are useful and having a range of techniques for transforming data into forms conducive to such representations.

The purpose of the current study is two-fold. First, it will examine what skills are important in the early stages of data representation at the upper primary and secondary school levels, by investigating the applicability of a framework of transnumerative techniques. Second, it will consider how effectively these techniques are used by young high school students in the specific case of representing association.

Framework

As a first attempt to identify the strategies that might be used to transnumerate data—particularly for straightforward data sets encountered at school—a framework of transnumerative techniques is proposed. This framework, shown in Table 1, lists a number of techniques that might be applied to data in order to find and display the message within. Each technique involves some "change in representation", by creating a new variable, organising the data differently, or representing them in a graphical way. The examples used to illustrate the techniques in Table 1 are based on the data set used by participants in this study. The initial variables are the number of hours of exercise, the number of fast food meals consumed per week, and favourite activity.

The techniques of graphing and tabulation, included in Table 1 as "graphing", change representations and thus are types of transnumeration. They are often final steps, to display the results of data analysis, but may also occur during data exploration. In most cases, however, other transnumerative steps precede graphing. These will involve the other techniques, and transform data into a form suitable for graphing. A frequency bar graph, for example, needs frequency data to determine the bar heights. For small data sets, some of this transnumeration may take place concurrently with the graph production, as when tallying. For larger data sets, however, these transformations are done in advance of graphing. Once data are in a suitable form, the graph (or table) then can be produced without further transnumeration, apart from the transnumerative mechanics of recording the data in graphical form. This is essentially what *Excel*'s Chart Wizard does: the user specifies the data and representation type—and sometimes has to transform the data

beforehand—and the program carries out the final transnumeration that produces the graph.

Table 1
Transnumerative Techniques

Technique	Key	Description	Example
Sorting	SRT	The data are sorted on some criterion. No new variables arise.	The data are sorted by hours of exercise, from lowest to highest.
Grouping	GRP	The data are grouped according to some criterion. This creates a new variable. This may involve the change variable type transnumeration beforehand.	A new variable "level of consumption" is created using the fast food data, with values "low" (0-1 fast food meals/week), "medium" (2-3 meals/week), and "high' (≥4 fast food meals/week).
Subset selection	SSS	A subset of the data is selected for further transnumeration.	Data associated with "low" and "high" levels of consumption are considered ("medium" is not).
Change variable type	CVT	A numerical variable is thought of in categorical terms or a categorical variable is thought of in numerical or ordinal terms.	Favourite activity (a categorical variable) can be given ordinal status, by ordering activities from most to least active.
Frequency calculation	FRQ	The frequencies of occurrence of values of a categorical variable are determined. Creates new variable.	The numbers of people in each of the "level of consumption" categories are determined.
Proportion calculation	PRP	Proportions (e.g., fractions) are determined in relation to a whole. This creates a new variable.	The percentage of people in each of the activity categories is determined.
Graphing	GRF	Some or all of the variables in the data (in their current form) are graphed or tabulated.	A scatter graph of hours of exercise v number of fast food meals consumed is constructed.
Central tendency calculation	CEN	A measure of central tendency (e.g., mean) is determined for a variable. May create new variable.	The average number of fast food meals consumed per week is determined.
Measure-of-spread calculation	MOS	Some measure of the spread of values associated with a numerical variable is determined. May create new variable.	The range of values associated with the number of hours of exercise is noted.
Other calculation	OTH	Generic term, recognising that other statistical calculations on the data are possible (e.g., sum, correlation coefficients, etc.).	A line of best fit is determined for the data from hours of exercise and number of fast food meals consumed.

Method

The task. Participants in the study were given a table showing a data set (similar to the Data Cards used by Watson and colleagues, but with fewer variables). The data set listed the names of 16 children, together with each child's favourite activity, number of hours of exercise per week, and the weekly number of fast food meals consumed. The set was constructed so that there were associations between the two numerical variables (hours of exercise and number of fast food meals), and between the categorical variable (favourite

activity) and hours of exercise (a numerical variable). Two Questions, shown below, informed the participants of the possible existence of these associations and asked them to produce a representation that would convince others of the relationship.

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?

The task was also discussed with the participants, before they had about 40 minutes to work on the task. Graph paper marked with a 1cm grid was supplied, but the participants were also reminded that it was not necessary to draw a graph if they felt some alternative representation showed the requested association.

Participants. The participants were 73 Year 7 girls (ages 11-13) at a private school in a major Australian city. The study took place early in the school year, so students' main prior experience of data handling would have been in primary school. Such experience should have included work with bar and line graphs, tables, and time series data; grouping and ordering data; and computing simple statistics, including the mean.

Data analysis. Students' responses to the two Questions were classified according to the extent to which the representations portrayed the indicated relationships. Five levels were identified. Level 0 responses did not deal with multivariate data; Level 1 responses did not make association apparent, often reproducing the original data in an alternative form; Level 2 responses made some effort to highlight values indicating association; Level 3 responses sorted or grouped data so that association was partially apparent; and Level 4 responses clearly depicted association. The transnumerative techniques used were also recorded. In reporting results the data have been compressed: a few representations exhibited minor variations from the category in which they have been placed, such as the use of an extra technique. In this report, "student" refers to the participants in the study, and "child" and "children" to the fictitious individuals in the data set.

Results

There were 133 responses to the Questions. Representations depicted one, two, or all three variables, and 85% were graphical. The Level 0 responses, which did not address the multivariate nature of the tasks, are reported first, followed by those responses that directly addressed each of the two Questions by representing bivariate data. Finally, those representations that incorporated all three variables, and so could be used to address both Questions simultaneously, are discussed.

Representations Not Dealing With Bivariate Data

There were 15 responses that did not clearly address either of the Questions, classified as Level 0. Six of these were from three students, who claimed to be responding to each of the two Questions. One of these students wrote narrative responses discussing the effect of, for example, fast food and exercise on weight, but without reference to the data and so no transnumerative techniques were used. The other two students, together with four others, graphed the values of the data for one of the variables (e.g., one student did a bar graph showing the consumption of each of the 16 children in the data set). The only transnumerative technique applied was graphing—a "literal translation" of the original

data—for a single variable. A further three graphs were frequency bar graphs of activity, requiring a frequency calculation before graphing. These only considered one of the variables and so association was not depicted. The final two representations each incorporated three graphs that were frequency bar graphs for each of the three variables. To produce these for fast food consumption and hours of exercise, the students had to treat the numerical values as categories (cf., Nisbet, 2001), prior to frequency calculation and graphing. The univariate nature of these representations prevents depiction of association.

Association Between Two Numerical Variables

Of the remaining 118 representations, 47 involved only the two numerical variables—fast food consumption and hours of exercise—and were clearly responses to Question 1. Table 2 indicates the types and frequency of occurrence of the more common responses.

Table 2
Representations Depicting Association between Two Numerical Variables

Level	Description of representation (FF = fast food consumption variable, Ex = hours of exercise variable)	Techniques used	Number (n=47)[1]
1	Two aligned graphs: FF for each child, Ex for each child	GRF	9
1	Single graph, two columns (FF and Ex) for each child	GRF	10
1	Other paired FF with Ex for each child (e.g., graph of Ex values with corresponding FF listed)	GRF	8
2	Paired FF and Ex, but for only a subset or full set with a subset highlighted (e.g., high FF values listed, with corresponding Ex graphed above)	GRP, SSS, GRF	4
3	Listed FF high to low, listed Ex low to high, and noted occurrence of people at top of both lists	SRT, GRF	1
4	Data grouped by categories in one variable, other variable listed/graphed (e.g., 5 graphs, one per FF category, showing Ex for children in that category)	CVT, GRP, GRF	3
4	Average Ex calculated for each FF category, and graphed	CVT, GRP, CEN, GRF	1
4	Scatter graph (or equivalent) of Ex v FF or FF v Ex	GRF (GRP)	6

[1] Not all representations are described, but all successful ones (Levels 3 and 4) are included.

Over half of these representations were literal re-depictions of the original raw data, with no other transnumeration apart from that required to place the data into the chosen graphical form. These representations, classified as Level 1 in Table 2, show the association no better than the original data. The reader can skim the representation to look for high occurrences of one variable with low occurrences of the other, but the comparison has to be made for each individual, just as would be done if examining the original data set. Four students highlighted some subset of the data to show high values of fast food consumption with low hours of exercise, for example, but did not contrast this with low fast food consumption. This made the association partially apparent, but not fully so (Level 2).

There were 11 successful or nearly successful representations. Most of the nearly successful representations undertook some sorting or grouping of the data (Level 3), so that the association is apparent provided the viewer can interpret the representation appropriately. Finally, the successful representations took advantage of some more

advanced techniques. Scatter graphs, produced by six students, depict association without requiring any more transnumeration than merely graphing the data. No data reorganisation is required; the secret is to know and use this graph type. Only one student appreciated the power of calculating averages for comparing groups (cf. Watson & Moritz, 1999), and used a graph of the average hours of exercise for each category of fast food consumption. This required the fast food consumption variable to be regarded as categorical, and the data to be grouped into these categories, before the mean was calculated and graphed.

Association between categorical and numerical variables

A total of 49 representations were clearly responses to Question 2, involving hours of exercise and favourite activity. Table 3 describes common responses.

Table 3
Representations Depicting Association between Categorical and Numerical Variables

Level	Description of representation (Act = favourite activity variable, Ex = hours of exercise variable)	Techniques used	Number (n=49)[1]
1	Paired data duplication (cf. Table 2, Level 1) (e.g., Ex graphed, Act listed; twin columns Ex and Act with Act given diff. heights for diff. categories)	GRF	19
2	Subset of data considered (e.g., Ex graphed for sport subgroup); not enough data for association	GRP, SSS, GRF	4
2	Totals of Ex graphed for each category of Ex	GRP, OTH, GRF	4
3	Scatter graph: Act categories indicated by position on axis but not ordered by degree of activity	GRF	7
4	Ex values graphed for each category of Act (or enough categories to convince about association)	GRP, GRF (SSS)	6
4	Average Ex computed for each Act category	GRP, CEN, GRF	4

[1] Not all representations are described, but all successful ones (Levels 3 and 4) are included.

As with the responses to Question 2 a large proportion of representations (about 40%) duplicated the data in an unsorted paired fashion, so that the association is only apparent by doing a value-by-value manual analysis of the representation (Level 1). Four responses presented representations of a subset of the data, such as the hours of exercise of those whose favourite activity is sport. Although allowing a focus on the most active group, they failed to show that the hours of exercise are higher than any of the other activity groups because these were not depicted. A further four students produced totals of the number of hours of exercise for each activity category, failing to take into account the different numbers of data values contributing to these (Level 2). The scatter graphs that were presented used positions on one of the axes to represent the activity categories, but because these were not ordered by degree of activity the graph was not successful at revealing the association clearly (Level 3). The Level 4 responses either involved average hours of exercise for each activity category, or presented graphs of hours for exercise for each activity group, with differences among the graphs revealing the association.

Representations Involving All Three Variables

There were 22 representations that incorporated all three variables, and for ten of the students this was their only representation. Eleven responses were Level 1 representations

that duplicated the values of the three variables, by producing a dual column graph (or aligned pairs of graphs) for the numerical variables, and writing or position-graphing the corresponding activity categories alongside, with no sorting or grouping of the data. Two students totalled the numerical variables for each activity category, with one then undertaking additional but unnecessary transnumeration by producing pie charts for these subtotals as proportions of the overall totals (Level 2). Three students listed all values of the data, but sorted by fast food consumption, so that trends in the other variables could be seen as a consequence, whereas another student grouped by activity category and listed the other variables (Level 3-4).

Discussion and Conclusion

Space constraints prevent a consideration of all the representations produced, however the framework contains all the strategies used by students. This suggests that for this type of statistical task at least—and for more straightforward representation tasks—the framework adequately identifies the transnumerative techniques useful for and likely to be used when representing data. Furthermore, the results indicate that many Year 7 students are able to produce representations of multivariate data, although the levels of success in depicting association vary widely.

Most students produced graphs rather than tables. This was understandable, given the wording of the Questions and the supply of graph paper to students. For about half the representations that addressed two or more variables, however, the representation—whether table or graph—reproduced the original data, showing no transnumeration but for the process of turning data into the chosen representational form. This prevented a clear depiction of association, with one exception: scatter graphs. The representational power and simplicity of a scatter graph lies in the fact that although it involves no transnumeration apart from graphing, the resulting graph *does* demonstrate association. This technique could be given greater and earlier emphasis in school data analysis work. The results also suggest that simple techniques, such as grouping and sorting, are powerful yet under-utilised. Students who sorted the data before listing it, for example, produced representations where the association was visible without great reader effort. These strategies could receive more explicit emphasis in teaching, and, in fact, the repertoire of techniques in the framework could be made available to students as an explicit list.

Some students whose representations did not clearly show the association still recognised its existence and could talk about how it might be seen in their depiction, for example, by seeing if one column is high when the corresponding column for the second variable is low. These students would benefit from seeing some effective representations, so that they can see how data analysis and reporting is enhanced by the use of other transnumerative techniques that produce better representations.

Finally, there is scope for much further research. Two particular areas of interest include the responses from other age groups (e.g., what happens when students have greater familiarity with sophisticated techniques?), and the outcomes when spreadsheet programs are used.

Acknowledgment.
Thanks to Jane Watson for useful discussions about the framework.

References

Board of Studies. (2000). *Curriculum and Standards Framework II: Mathematics*. Melbourne: Author.

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.

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*, 382-393.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education,* 33, 259-289.

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.

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, B. Perry, & M. Mitchelmore (Eds.), *Numeracy and beyond* (Proceedings of the 24th annual conference of the Mathematics Education Research Group of Australasia*,* pp. 378-385). Sydney: MERGA.

Nisbet, S. (2003). Organising and representing grouped data. 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. 562-569). 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.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, Connecticut: Graphics Press.

Watson, J. M., & Callingham, R. A. (1997). Data Cards: An introduction to higher order processes in data handling. *Teaching Statistics, 19*, 12-16.

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.