Transnumeration and the Art of Data Representation

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

First year university students were provided with two sets of data and told about messages or conclusions that could be drawn from each set. The students were then asked to represent the data so that the representation effectively conveyed the message. These tasks required students to "transnumerate" the data, or, in other words, reorganise data to better understand them. Four phases of transnumeration were identified: recognising the message, choosing the representation, transforming the data, and representing the transformed data. Although many students were able to produce effective representations, other students had difficulties with aspects of the transnumerative process, many at the basic numeracy level.

Instances of the problematic graphical representation of data abound. Some problems arise from a deliberate attempt to misrepresent data (e.g., using axes that do not start from 0), some from a failure to understand what kinds of representation are appropriate for different types of data (e.g., the use of pie charts to display data that do not represent quantities expressed as proportions), and some from bad choices about how to effectively use a graph to tell the "story" of the data. This final cause is the focus of the current study.

The issue of choosing how to take a data set and use it effectively to tell the story of the message contained within it is a critical component of statistical thinking. The term *transnumeration* was introduced by Wild and Pfannkuch (1999) for the process of changing data representations to better understand the data. Pfannkuch and Rubick (2002) identify three specific instances of transnumeration in statistical thinking: (i) making measurements that capture characteristics of the situation of interest; (ii) transforming raw data into other representations—such as sorted data, graphs, tables, and summary statistics—in order to search for meaning in the data; and (iii) communicating the meaning in the data to others.

In a recent study Pfannkuch, Rubick, and Yoon (2002) considered the types of "transnumeration thinking" exhibited by students investigating the raw data given in the Data Cards Protocol of Watson, Collis, Callingham, and Moritz (1995). Since students were working with raw data, a key transnumerative process was to classify the data, prior to summarising and displaying the data. In the studies of Pfannkuch et al., as well as the earlier work of Watson and her colleagues (e.g., Chick, 2000; Chick & Watson, 2001; Watson et al., 1995), students had to make decisions about how to *represent* the data. Students' facility with this task depended on what aspects were considered and what representation techniques were in their repertoire of statistical tools. Chick and Watson (2001) suggest that the transnumerative process of *interpreting* data may be easier than the process of finding an appropriate representation. Chick (2000) also highlights that many students do not use statistical representations, graphical or otherwise, to support the claims they make about what the data are showing.

Choosing how best to represent data has been shown to be problematic by a number of studies. Li and Shen (1992) provide examples of graphs produced by students that do not effectively convey the story in the data, because of poor choices about the kind of representation to use or poor implementations of an appropriate representation. Chick

(2000) also gives an example of a representation that is inappropriate for the type of data. Typical problems include incorrect choice of graph type, inappropriate scales on axes, graphs that present two incomparable quantities in ways implying comparison is possible, misunderstanding of basic concepts such as percentage and proportion, conveying all information when some form of data reduction is more appropriate, and not using graph conventions such as having the independent variable on the horizontal axis.

The Data Cards Protocol studies referred to above gave students a data set but did not tell students any of the messages in the data before getting students to produce their representations. In contrast, Moritz (2000) gave primary students messages about data, though without giving explicit data sets, and asked the students to represent these messages graphically. For example, he asked students to draw a graph showing that people grow taller as they get older. To achieve success, students had to think quite carefully about how best to depict that message graphically, but with no explicit data.

The focus of this paper is to consider some of the transnumerative processes required in order to make appropriate representations, particularly graphical ones, to convey the information contained in previously collected data (the transnumerative processes associated with the collection of data will not be considered). In addition, some of the difficulties encountered by students will be identified.

Method

The participants in this study were first year university students enrolled in a mathematics service subject intended for those from the arts, science, commerce, and medicine faculties undertaking courses requiring only basic calculus. The students had passed Year 11 and 12 mathematics subjects involving study of functions, and some basic statistical ideas such as collecting, organising, displaying, and interpreting data, as well as calculating straightforward summary statistics. The university mathematics subject undertaken did not have statistics as a focus, but was intended, in part, to give students an appreciation of how mathematics is used to model and understand real-world phenomena.

Two assignment questions provided data for the study. The first—the Down's Syndrome Question—arose from one student's response to an earlier assignment question that had asked students to find some data in a book or magazine, discuss the data's representation, and then represent the data using an alternative approach. Figure 1 shows the resulting Down's Syndrome Question given to students in this study.

The second question—the Exam Success Question—originated with the subject's previous lecturer, Glen McPherson (personal communication, 1997). It gave students the raw data shown in Table 1, and asked them to present the data in such a way as to convince readers that attending tutorials results in a greater likelihood of exam success.

For both of these problems transnumerational activity is required. In the Down's Syndrome Question the request to transnumerate is explicit, with a particular outcome specified. In the Exam Success Question the task is more open-ended, with students able to make their own decisions about how to treat and display the data.

The questions were given on separate take-home assignments. A total of 41 students responded to the Down's Syndrome Question and 46 responded to the Exam Success Question. Their responses were analysed to determine typical problems and successes.

The following data were found by a student and the accompanying well-labelled graph was produced. One drawback with the graph is that as age increases the graph is *decreasing*, so that the visual impression is that risk reduces with age. Produce a graph which is not only correct but actually *shows* that the risk increases with age.

Maternal age	Risk of Down's Syndrome
Below 30	1 in 1000
30	1 in 880
32	1 in 720
34	1 in 460
36	1 in 280
38	1 in 180
40	1 in 100
42	1 in 70
44	1 in 40
46	1 in 25
48	1 in 15

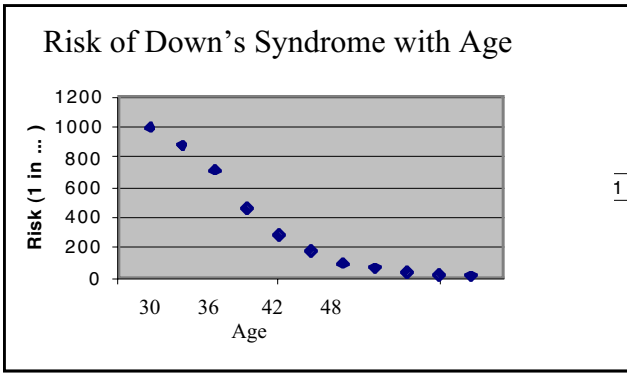

Figure 1. Down's Syndrome Question.

Table 1
Number of Students Passing and Failing First Year Mathematics

Number of tutorials attended	Performance on exam	
	Pass	Fail
> 20	137	34
≤ 20	38	36

Results

The Down's Syndrome Question

Of the 41 students completing this question, 28 were successful in producing an acceptable representation that showed risk increasing with age. Half of these students converted the stated risk values into percentage values, whereas another ten equivalently determined the "risk in 1000" or stated how many times more likely Down's syndrome was to occur compared with age 30. The remaining three used the word "probability" and calculated a decimal value for the risk.

For the remaining students, the most common difficulty arose from treating the risk ratios as fractions. Eight students expressed the risk values in fraction form (e.g., 1/720), and placed these fractions on the y-axis, with 1/1000 closest to the origin, and then 1/900, 1/800, and so on, at equally spaced positions up the axis (for a slight variant of this, see Figure 2a). Instead of positioning 1/720 at the correct position between 1/700 and 1/800, it was positioned as if it was 720 between 700 and 800. Some of these students also had difficulty deciding what should be at the top end of the axis: one had 1/1 as the next value after 1/100, whereas the student who produced Figure 2a did not actually write axis values

above 100 although the graph has data points in that region. Since the scale on the *y*-axis has been distorted, the way in which risk increases with age is thus distorted too.

Two students achieved the goal of having an increasing graph by reversing the conventional allocations of variables to the axes, placing maternal age on the *y*-axis and risk on the *x*-axis. As they did not transform the risk values in any way (e.g., into percentages, decimals, or fractions), in order to achieve the increasing effect it was necessary for them to reverse the order of the risk values. In addition, one of these students made no attempt to scale these values, instead spacing the denominators of the 11 risk values equally along the *x*-axis. The result, shown in Figure 2b, is a linear graph.

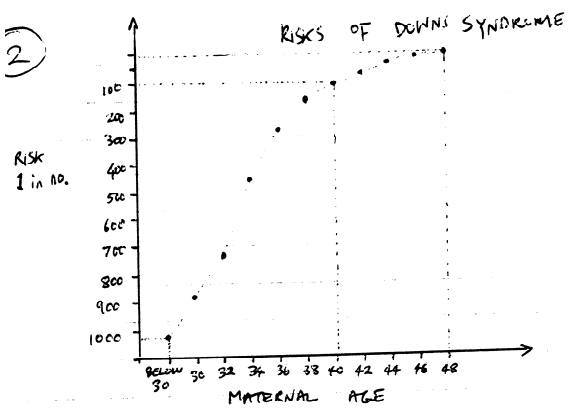

Figure 2a. Inappropriate treatment of the risk ratios on the *y*-axis.

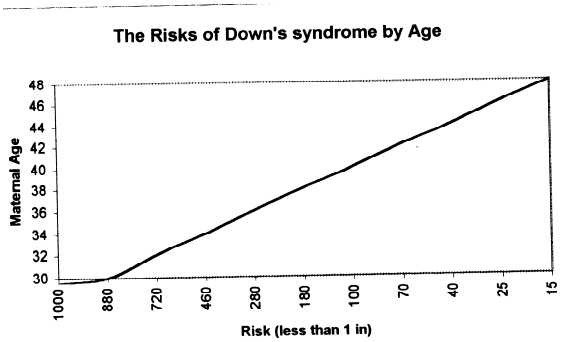

Figure 2b. Reversal of conventional use of axes, together with inappropriate treatment of risk ratios.

Three other transnumeration difficulties are worth mentioning. One student converted the risk values to percentages, and used these values on an axis labelled "probability", which resulted in probability values apparently greater than 1. Another student obtained decimal values for risk, but then had trouble locating these positions on an axis as the values were expressed in exponential notation. The axis had equally spaced tick marks, which were labelled 2×10^{-3}, 4×10^{-3}, 6×10^{-3}, 8×10^{-3}, 2×10^{-2}, 4×10^{-2}, 6×10^{-2}, and 8×10^{-2}, thus resulting in a non-linear jump between 8×10^{-3} and 2×10^{-2}. Finally, the most unusual unsuccessful example of transnumeration came from a student who obtained new risk values by subtracting the denominators of the old values from 1100 (so that, for example, 1 in 880 was assigned the risk value of 220). The resulting graph had these new risk values—which get bigger with age—on the *y*-axis. The student went further, however, calculating a new percentage by comparing the new risk value with 1100, and concluding that (in this case) "at age 30 there is a 20% chance of getting Down's Syndrome". Only this particular percentage value was actually calculated and it was not related to anything on the graph. Furthermore, the student did not indicate a reason for the choice of 1100.

The Exam Success Question

Table 2 shows the range of approaches taken by students for the Exam Success Question. Of the two-thirds of students who produced a successful representation, all but one did so by converting the numbers in Table 1 to percentage pass and fail rates for each

of the two groups (those attending more than 20 tutorials and those attending 20 or less). Success with this computational transnumeration was followed by a variety of approaches to transnumerating for the purposes of presenting the results: 18 students used some type of bar graph, six students used pie charts, five presented the results in a table, and one made a written poster/advertisement stating the pass rate for the high attendance group and the fail rate for the low attendance group. The final student used a different approach, producing a graph with the *y*-axis representing the number of students passing versus number of students in the group on the *x*-axis, and represented the performance of the two groups via two lines whose slopes were equal to the respective pass rates.

Table 2
Student Responses to the Exam Success Question

Type of response (the "groups" are students attending >20 tutorials, and those attending ≤20)	No. of students
Appropriate responses TOTAL	31
Two bars, showing percentage pass rate for the two groups	4
Two bars, each 100% tall, showing pass and fail rates for the two groups	3
Two pairs of bars, each pair showing pass and fail percentages for one group	8
Two pairs of bars, one pair showing percentage pass rates for the two groups and the other showing percentage fail rates	2
Four bars, giving pass and fail rates for the two groups with percentage pass rates oriented above the horizontal, and fail rates oriented below	1
Two pie charts, one for each group, showing percentage pass/fail rates	6*
Tabular representation of percentage pass/fail rates for each group	3
Tabular representation of percentage pass rates only	2
Poster highlighting percentage pass rate for high attendance and percentage fail rate for low attendance	1
Graph showing two lines, one for each group; points on the line indicate number of students passing for number of students in the group	1
Inappropriate responses TOTAL	15
Two pairs of bars, each pair showing pass and fail numbers for one group, *but using absolute numbers*	4
Two bars, each divided in two, showing pass and fail rates for the two groups, *but using absolute numbers*	2
Some form of bar graph or pair of bar graphs (as above), *but using percentages determined relative to the* total *number of students*	7
Pie chart containing the four categories (pass/fail for each of >20/≤20), *using percentages relative to the* total *number of students*	1
Graph showing 3 lines, one for each group and for the total, *based on pass percentages relative to the* total *number of students; points on the line indicate number of students passing for number of students overall*	1

* One of these graphs was hand-drawn and poorly displayed 20% as much less than 1/5 of the graph

There were two main causes for unsuccessful transnumeration. Six students worked with the absolute numbers given in Table 1, failing to appreciate that the message in the data is best conveyed by considering proportions. Another nine students appreciated the proportion idea, but could not correctly identify the most appropriate whole. These students used the total number of people in the first year mathematics class as the whole, and so the comparisons made in the resulting graphs were equivalent to those done by students using absolute numbers.

Discussion

From these examples it appears that, in order to obtain the desired end representation, students must go through four phases of transnumeration. The first is to decide what message to convey from the data. In this study, this was determined for the students and so cannot be studied closely. The second step is to determine what sort of representation is required (in the case of the Down's Syndrome Question the nature of the final representation was specifically requested); this then informs the next phase, which is the choice of computational transformation to apply to the data. It is conceivable that these two processes could be carried out in reverse order, too, or even simultaneously; the data in the current study do not allow us to determine the approach taken by the students here. Finally, there is the process of using the transformed data in the representation. We shall discuss the final three phases in turn, informed by the data in the study.

Choice of Representation

As suggested above, in the case of the Down's Syndrome Question the type of representation was suggested to students in the question itself. This request was a consequence of the author already having a sense of "the story the data are telling", namely that the risk of giving birth to a child with Down's syndrome increases with maternal age. However, note what prompted the author to prepare the Down's Syndrome Question: it was based on a student's earlier less than ideal choice of how to represent the data (see Figure 1). The choice made by this student may have arisen as a consequence of not appreciating that the ideal way to communicate the idea of increasing risk—which was understood by the student—is to have a graph that is increasing. Another possibility is that the student wanted to use the given data directly (i.e., expressed exactly as in the data table) rather than transform it.

For the Exam Success Question, the students in this study had to make a choice of representation. Virtually all students appreciated that two groups of people were being compared and that a representation that allowed a comparison would be required. There was nothing intrinsically wrong with the representation type (bar graph, table, pie chart) chosen by the students in this study, although a couple of representations could have been improved to allow more obvious comparison between groups (e.g., one student, who elected to use a pie chart and erroneously used the total number of people as the whole, consequently had the four outcomes—pass and fail for each of the two groups—as sectors of the pie, making it more difficult to compare the results of the two groups). The main problems, however, arose with the way that the data were transformed for use in the representation, as discussed in the next section.

Transforming the Data

For both questions an effective representation can only be obtained if the data are transformed in an appropriate way. For the Down's Syndrome Question the conventional approach, carried out by most of the students, is to treat the risk values as fraction, decimal, or percentage values. Those students who elected to keep the data expressed as a risk value (e.g., 1 in 880) then had to do something unusual (and usually incorrect) when representing the data. The student who subtracted the risk denominators from 1100 was successful in producing a graph showing risk increasing with age, but the quantity so calculated was given no meaning despite the student still calling it "risk".

For the Exam Success Question the critical idea is proportion, and in particular proportion with respect to the correct wholes. It was apparent from these results that some students struggle with proportional reasoning and do not know when to use it. It would be interesting to know if students' performance would differ if, for example, the data given in the Exam Success Question indicates a greater number of students failing from the group who attended more than 20 tutorials. An absolute comparison would then show more people failing from the high tutorial attendance group compared with the number failing from the low attendance group. Such a comparison would appear to contradict the true message in the data and may encourage students to revise their transformation.

Representing the Data

Difficulties with actually representing the transformed data were especially apparent in the Down's Syndrome Question. Some of the problems were associated with basic numeracy skills, whereas others were associated with graph conventions. These are discussed below.

Number. One area that caused difficulty was associated with an understanding of basic number properties. Even though an appropriate data transformation may be made, if there are difficulties in interpreting the corresponding numerical values then the representation will be flawed. In the current study the most significant misconception about number, exhibited by nearly 20% of students, was failing to recognise the reciprocal and therefore non-linear relationship between unit fractions and their denominators. This had the consequence of distorting the representation. Another difficulty concerned the relative positions of small decimal values on the number line, perhaps because of the exponential notation.

Appropriate axis and scale choice. If students choose a graphical type of representation (e.g., a bar graph or a pie chart), then the final display may be unsuccessful if inappropriate choices of axis or scale are made, or proportions are misrepresented. This was found in the study of Li and Shen (1992). In the current study, two students reversed the variables on the axes from the conventional dependent variable versus independent variable. In addition, picking an appropriate scale was problematic for one of these students. Another student's otherwise appropriate choice of pie chart, together with appropriate transformation of data, ended up being flawed in the final representation because of a failure to actually show 20% as a 20% sector (this may also be regarded as a basic numeracy difficulty).

Conclusions

This study suggests that successful transnumeration for the purposes of representing data depends on four linked processes. Three of the phases of transnumeration seem particularly intertwined, namely identifying the message in the data, choice of representation, and the process of transforming the data. Success with the latter two tasks may well rely on success with the first. Without a sense of the message it is hard to choose what transnumerative process to use. Furthermore, evidence from other studies (e.g., Chick, 2000) suggests that some students do not understand that transnumeration is actually necessary because it allows the message to be conveyed *with clear evidence*. In considering the fourth phase, in which the transformed data are represented in the chosen way, it must be noted that, unfortunately, even with appropriate transnumeration taking place in the earlier phases, the final representation may fail if there are any misconceptions associated with the actual representation process. The current study cannot provide insight into the interaction between these four phases, as it focuses on the final graphs or tables. The use of interviews in the future might give information about links between the transnumerative processes leading up to the production of the final representation.

It should be noted that in virtually all cases considered here and in Li and Shen (1992) the unsatisfactory representations did not seek to obscure the data, in that all the original data values can be retrieved from the representations (although extra information may be needed, such as the total number of people for the Exam Success Question). What is critical is that some transformations are better than others for "telling the story in the data". Some students are clearly able to make appropriate choices, whereas others struggle with aspects of the task. Given these findings, it would be useful to know how to help students learn how best to represent and transform data, and to appreciate the necessity of using statistics from the data as evidence for claims made about the data's messages.

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

Li, K. Y., & Shen, S. M. (1992). Students' weaknesses in statistical projects. *Teaching Statistics, 24*(1), 2–8.

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. 157–164). 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.

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