UNDERSTANDING DATA TABULATION AND REPRESENTATION

Chris Reading

Centre for Cognition Research in Learning and Teaching

University of New England, Armidale, Australia

Statistics has received increased recognition in mathematics school curriculums in Australia partially due to the strand status assigned to statistics (within Chance and Data) in A National Statement in Mathematics for Australian Schools. Consequently, research has focused on considering what 'statistical thinking' really means. To assist teachers to plan and assess the teaching of statistical concepts more needs to be known about students' statistical understanding. This paper takes up the theme by considering students' responses to two open ended tasks, one of which presents the data in table form and the other graphically. Both tasks require students to describe what they understand by the data representation. A developmental sequence of eight levels was identified and the responses to the two different data presentations were analysed. The SOLO Taxonomy was used as the theoretical framework to assist this process.

Introduction

More statistical ideas are being incorporated into mathematical syllabuses across Australia but poor awareness of students' statistical understanding on the part of teachers (Watson, 1998) may well be a contributing factor to the poor treatment of statistical components in the curriculum. Truran (1997) identified lack of knowledge of statistical understanding as a concern when trainee teachers had difficulty in interpreting and applying concepts in the Chance and Data strand to create well structured sequences of lessons. In encouraging teachers to give students a chance to show what they can do statistically Shaughnessy (1997) stressed the need for research into students' thinking about chance and data.

A recent development in investigating student understanding has been the use of the SOLO Taxonomy (Biggs & Collis, 1982) as a framework, in both probability and data handling. SOLO levels have been used to classify student responses concerning uncertainty (Moritz, Watson & Collis, 1996), data representation (Chick & Watson, 1998), data reduction (Reading & Pegg, 1996) and data interpretation (Reading, 1998). This paper contributes by exploring students' responses to questions concerning the understanding of data tabulation and representation, using the SOLO Taxonomy as the theoretical framework.

The SOLO Taxonomy

Detailed descriptions of the SOLO Taxonomy can be found elsewhere (see for example, Biggs & Collis, 1991; Pegg, 1992). The model, which allows students' responses to be categorised, consists of five modes of functioning, with levels of achievement identifiable within each of these modes. The two modes relevant to the research being reported are the ikonic mode (making use of imaging and imagination) and the concrete symbolic mode (operating with second order symbol systems such as written language). Although these modes are similar to Piagetian stages, an important difference is that with the SOLO Taxonomy earlier modes are not seen to be replaced by subsequent modes and in fact are often being used to support growth in the later modes.

A series of levels have been identified within each of these modes, three of which are relevant to the this report. These are: unistructural - with focus on one aspect, multistructural - with focus on several unrelated aspects and relational - with focus on several aspects in which inter-relationships are identified. These three levels form a cycle of growth which recurs within modes and in different modes. Within a mode the relational level response in one cycle is similar to, but not as concise as, the unistructural response in the next. A similar cycle of levels is identified in different modes but the nature of the element on which the cycle is based is different. This taxonomy is particularly useful because of the depth of analysis which can be achieved when interpreting students' responses.

Research Design

One hundred and eighty secondary students, selected randomly over gender (male, female), mathematical ability (low, middle, high) and academic years (7 to 12) were tested on a range of statistical questions. This paper reports on the responses to a two part question concerning the understanding of data tabulation and representation, an important step in the process of data analysis. The question was not testing the ability to arrange data into a table or a graph, but aimed at presenting students with some data and allowing them to describe what information they were able to gather from the representation. Part I of the question presented the data in a table, while in Part II the data presentation was graphical. The two parts were used in the question to investigate whether the form of data presentation influenced student understanding. The open-ended question allowed the student to respond with as much information as he or she felt was necessary.

Analysis of Responses to Part I

The question, as presented to students, is shown in Figure 1. Investigation of student responses showed that it was possible to divide the responses into a number of levels based on the statistical quality of the answer given. Three major groupings of the levels were identified based on the depth to which the response indicated the ability of the student to understand the representation of the data.

Part I Question

A class teacher wanted students to practice collecting data. One Year 8 student decided to collect data concerning the number of ice creams that she ate during a week for a seven week period. The table the student came up with is given below.

Week 1	3
Week 2	5
Week 3	7
Week 4	4
Week 5	2
Week 6	7
Week 7	5

What does the table tell you ?

Figure 1

First Group (No Data Use)

Responses in the first group dealt with only the requirements of the question and three broad levels, coded as 0, 1 and 2 were observed. These responses attempt to rationalise the requirements of the question but appear to make no use of the data when formulating the response.

Level 0 These responses indicate that the question has not been considered or the requirements were not clear. For example:

(7201) *Its like a graph.*

Level 1 These responses indicate that not all aspects of the question have been considered sufficiently to produce an answer. Usually, some key fact from the wording of the question is reproduced in the response. For example:

(7111) *The table tells us about the ice cream eating habits of a year 8 girl.*

Level 2 These responses indicate that all aspects of the question have been considered and a reasonable answer attempted but still no use is made of the data. For example:

(8213) *The table tells you in column one: what week it was and in column two: how many ice creams were eaten in that week.*

Second Group (Data Item Use)

Responses in the second group are concerned with attempting to understand the data, with three levels, coded as 3, 4 and 5, being observed. However, attempts to describe the data are produced in non-statistical terms.

Level 3 These responses indicate that, although students have considered the data, focus is directed back to the key facts in the question usually indicating in some way that weekly data is available. For example:

(11105) *It tells me how many ice creams she ate in 7 weeks, how many she ate each week.*

Level 4 These responses indicate an awareness that features of the data need to be mentioned in the answer. However, restricted experiences at data description result in the information in the table being quoted verbatim. For example:

(12207) *The table tells me that for week 1 the student ate 3 ice creams, in week 2 she ate 5 ice creams, week 3 she ate 7 ice creams etc.*

At this stage there is a divergence of the responses into two distinct paths which appear to develop at seemingly parallel rates. These are labelled:

Path A for responses which describe statistical features of the data

Path B for responses which make judgements about the data.

Level 5 These responses describe the data by making a simple observation. They suggest readiness to engage in data description but a lack of experience and appropriate tools to produce a statistically sophisticated response. For example:

Path A (11111) *Some weeks she ate more than other weeks.*

Path B (8108) *That the girl is very unhealthy.*

Third Group (Data Feature Use)

The final group of responses indicate a readiness to describe the information contained in the data in a more acceptable statistical form. Only two levels of responses, coded as 6 and 7, were identified. Both levels are split into A and B paths, with Level 7 also having some responses incorporating elements from both paths.

Level 6 These responses indicate the use of data from the table to make one detailed observation. They show more sophistication than those at Level 5, linking the observations to features of the data, rather than making broad statements. For example:

Path A (10109) *The table tells me that the amount of ice creams eaten varies from 2 - 7 over the 7 weeks.*

Path B (9201) *The table tells you that she likes ice cream for a couple of weeks then she gets back into them again.*

Level 7 These responses indicate a more in-depth understanding by presenting more than one observation related to the data. For example:

Path A (9112) *She ate the most ice creams in week 3 and 6. And the least in week 5. On average she ate 4.7 or 5 ice creams a week.*

Path B (12212) *The student likes ice cream or it is summer and she wants to keep cool.*

Some responses showed features of both Path A and Path B,

(10212) *The table tells you that the girl ate 33 ice creams in 7 weeks and that she must have liked* *ice creams.*

A third level in this group was not found but it is anticipated that such a level may contain responses which not only mention statistic(s) and judgement(s) but use the statistics as evidence for the judgements made.

The results, arranged by academic year, appear in Table 1 and a number of interesting points become apparent. First, there are only eleven students (18%) from the two senior years whose responses fall within the first group, compared to twenty five (42%) from Year 7 and 8. Second, there is a larger number of senior students compared to junior, noticeably twelve in Year 12, whose responses were coded as Level 7. Third, there is a large bulge in most years at Level 2. Fourth, there appears to be a larger number of responses in the last level of each of the first two groups (that is Levels 2 and 5), than in the previous two levels of the group. This is more noticeable in the junior years. Last, there appears to be a balance in the number of students whose responses reflect Path A and Path B.

Table 1

Response Level and Path by Academic Year for Part I Question

Level	**Year**												**Total**
	7		**8**		**9**		**10**		**11**		**12**		
0	1		0		0		0		0		0		**1**
1	4		1		0		1		0		2		**8**
2	4		5		10		11		6		3		**49**
3	14		3		3		2		7		2		**20**
4	0		5		0		3		0		2		**10**
	A	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	
5	1	5	3	6	4	3	2	4	4	2	3	3	**40**
6	1	0	3	2	1	4	2	2	2	3	1	2	**23**
7	1	0	1	0	3	1	1	0	5	0	4	3	
	0		1		1		2		1		5		**29**
Total	**30**		**30**		**30**		**30**		**30**		**30**		**180**

These results suggest that, for the process of understanding data presentation, the level of response improves progressively with academic year, although, the bulge at Level 2 suggests that many students have difficulty actually describing the data at all. Further, there appears to be no particular preference for descriptions using statistical features or judgemental observations.

Analysis of Responses to Part II

The second part, II, of the question is shown in Figure 2 (next page). Answering this question meant that students needed to be able to read the graph before describing the data. Examples of responses to Part II are not given because they are similar in form to those given for Part I and were coded into similar levels. The results, arranged by academic year, are presented in Table 2 and some interesting features emerge. First, while twenty one students (35%) from the two senior years gave responses in the first group, there were thirty four (57%) from the Years 7 and 8. Second, Level 7 contained thirteen Year 11 and 12 students but only three Year 7 and 8 students.

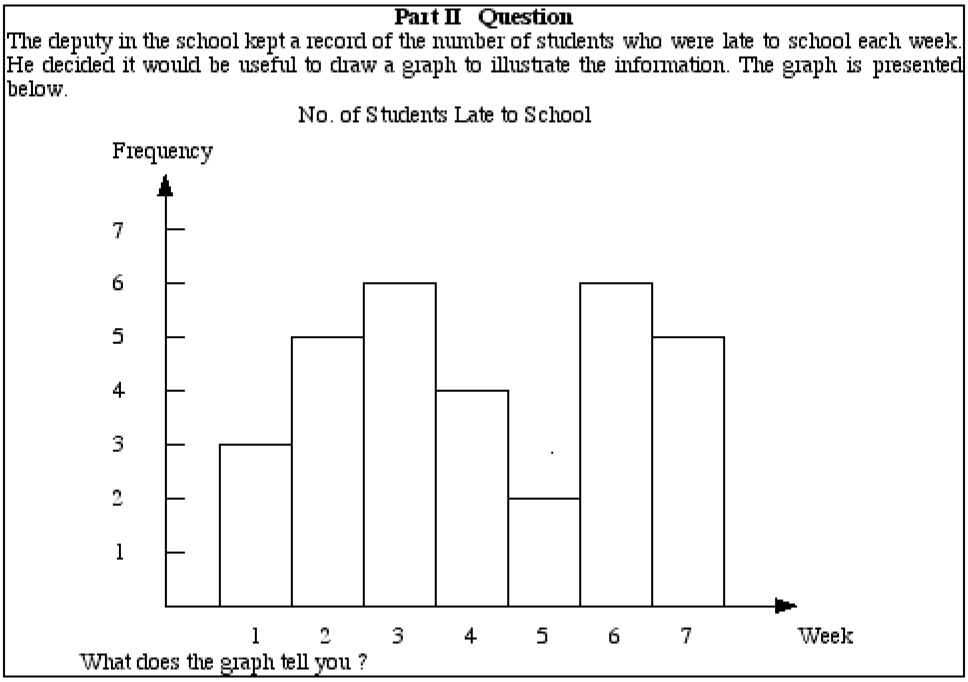

Figure 2

Third, there appears to be a larger number of responses in the last level of each of the first two groups (Levels 2 and 5). Last, there are almost twice as many students coded in Path A as Path B, with Path A more popular in all but Year 7 and Year 10.

Table 2

Response Level and Path by Academic Year for Part II

Level	Year												Total
	7		8		9		10		11		12		
0	0		0		0		0		0		0		0
1	8		7		1		0		2		3		21
2	11		8		11		14		12		4		60
3	2		1		3		4		4		1		15
4	0		0		0		1		0		2		3
	A	B	A	B	A	B	A	B	A	B	A	B	
5	0	7	3	3	5	3	0	7	4	2	7	3	44
6	1	0	6	0	5	0	4	0	1	0	2	0	19
7	1	0	1	0	0	2	0	0	4	0	6	2	
	0		1		0		0		1		0		18
Total	30		30		30		30		30		30		180

These observations suggest a slight improvement in the quality of responses with increasing academic year and a preference for describing the data using statistical features rather than judgements. Difficulties were experienced making the step to actually describe the data and also in the step to describe the data in more statistical terms.

Comparison of Part I and Part II

The framework developed appears to be adequate for explaining students' understanding, as far as the basic description of data presentations is concerned. There is a slight upward shift in the trend of the responses over the academic years as would be expected. All but one student felt that he or she understood the question sufficiently to attempt an answer and those few who had the most problems responding (Level 1), including misinterpreting the graph, were mostly in Years 7 and 8.

Testing the hypothesis that the level of response is independent of the part of the question yielded c2 = 13.53 (6 d.f) which is significant ($p < 0.05$) and indicates that the level of a response is associated with the part of the question being answered. Many more responses than expected were coded at the lower levels in Part II, while Part I had more responses than expected in the uppermost levels. This suggests that students exhibit a higher level of understanding when the information is presented in a table rather than as a graph. Considering the number of responses that were graded into each path for Parts I and II, a test of independence resulted in c2 = 2.39 (1 d.f.) which is not significant ($p > 0.10$) suggesting that the choice of path is not associated with the part of the question being answered. This means that the path, A or B, used by a student in Part I does not indicate which path will be used in Part II. For example, giving a judgemental response when describing tabulated data, does not necessarily mean a judgemental response will be given when describing graphed data.

Three noticeable trends emerge when comparing the results of the analysis of Part I and Part II. First, there are more Year 7 and 8 students in the first group than there are Year 11 and 12 students. Second, there are more senior students than junior students in Level 7. Third, there is a large bulge in the numbers at Level 2 in most years. So, irrespective of the form of data presentation there is a general trend for a slight increase in the level of performance of the students over the six academic years and an unexpectedly large number of students repeating key facts from the question and not using the data in any way in the description. The better quality responses from Years 11 and 12 could mean that by this stage most students are ready to describe the data in more detail, despite the fact that little statistical work is actually undertaken in their curriculum studies.

The comparison of results also highlights three differences. First, the number of Year 11 and 12 students in group one is much larger for Part II (21) than for Part I (11). Second, the overall number of responses at Level 7 is much larger for Part I, (29), than for Part II (18). Lastly, similar numbers of responses are categorised as Path A or Path B in all years for Part I but most years show a predominance of Path A type responses for Part II. These differences suggest that although there is an increase in level with academic year, the overall range of performance is better when data are presented in table form than as a graph. Also, that when presented with data in graphical form younger students are more likely to use judgements in their descriptions while older students are more likely to use statistics.

SOLO Taxonomy Framework

The SOLO Taxonomy is now used, along with the groups of levels described earlier, to develop a framework which can be used to assist with the interpretation of student responses. The first group of responses exhibit ikonic features, while the second and third groups represent two different cycles in the concrete symbolic (CS) mode.

Ikonic mode responses show no evidence of linking the required task with any sort of symbolic representation. Level 1 responses were coded as a mixture of unistructural (U) and multistructural (M) responses, while Level 2 responses were relational (R).

Second and third group responses have been able to link the concepts in the question to concrete experience. The answer, which suggests that the question has been understood, links directly to aspects of the data. These responses are in the CS mode, with two cycles, of U, M and R levels.

The first cycle involves appreciating that it is possible to describe data. The elements in the first cycle are the actual pieces of data themselves (data items). A relational response in the first cycle is not achieved until the student is able to consider all data items as a functioning set, and the data items as capable of being described in another form. The U, M and R levels in this cycle correspond to the Levels 3, 4 and 5 described earlier, with the R level showing the split into two parallel paths.

The second cycle involves showing an appreciation for the need to refer to the features or behaviour of the data as part of the description. The elements in the second cycle are the various features (or properties) of the data which a statistical description could include. A relational response in the second cycle is not achieved until the student is able to present both statistical facts and judgements, and also some relation between them. In this cycle, the U and M levels correspond to the Levels 6 and 7 as outlined earlier. These two levels still contain the separate A and B processing paths, while some M level responses show evidence of elements from both paths. No responses were observed at the R level.

In summary, the main feature which distinguishes the concrete-symbolic mode responses from those in the ikonic mode is the retrieval of facts from the recorded data. Ikonic mode responses go no further than recognising from the question the variables which are being measured. CS mode responses show that the data items have been considered. Within this mode, the first cycle responses suggest that the data items have been considered as separate items while the second cycle responses indicate an overview of the data, in the form of a statistic or judgement.

Conclusion

Three major findings have evolved as a result of this study. First, presenting data in graphical form alters the way students describe data, as compared to tabular presentation. The differences in approach to data description include an overall lower level of understanding and a greater likelihood to discuss statistics rather than make judgemental comments.

Next, the three broad grouping identified, namely, No Data Use, Data Item Use and Data Feature Use, assist in determining the stage a student has reached in understanding data representation and tabulation. No Data Use responses are dealing with aspects of the question and not the data, while in the other two groups use is made of the data, Data Item Use responses in a less statistically sophisticated fashion than the Data Feature Use responses. These groupings offer teachers a means to follow better student thinking when planning lesson sequences within the curriculum and assessing specific student outcomes.

Last, the groups of levels identified can be categorised as cycles of U-M-R levels, based on the SOLO taxonomy. The No Data Use group is a U-M-R cycle in the ikonic mode where the elements of focus are the facts in the question. The other two groups represent two U-M-R cycles in the CS mode. The elements of focus in the first cycle, the Data Item Use group, are the actual data items while the focus in the second cycle, the Data Feature Use group, is on the various features (or properties) of the data. The identified levels in the CS mode are consistent with the U, M and R levels described in more general terms by Chick and Watson (1998, p. 156). Using this framework to assess responses, teachers can gain a greater awareness of students' understanding which will allow them to better prepare lessons based on the curriculum and to assess what students really know, understand and can do.

References

Australian Education Council (AEC) (1991) A National Statement on Mathematics for Australian Schools, Carlton, Victoria: Curriculum Corporation.

Biggs, J. and Collis, K. (1982) Evaluating the Quality of learning: the SOLO Taxonomy. New York: Academic Press.

Biggs, J. and Collis, K. (1991) Multimodal learning and the quality of intelligent behaviour. In H. Rowe (ed.), Intelligence, Reconceptualization and Measurement, New Jersey: Laurence Erlbaum Assoc, 57-76.

Chick, H. and Watson, J. (1998) Showing and telling: primary students' outcomes in data representation and interpretation. In C. Kanes, G. Merrilyn, and E. Warren, (eds) Proceedings of the Twenty First Annual Conference of the Mathematics Education Research Group of Australasia, Gold Cast, Australia: Griffith Uni Print, 153-160.

Moritz, J., Watson, J. and Collis, K. (1996) Odds: Chance measurement in Three Contexts. In P. Clarkson, (ed.) Proceedings of the Nineteenth Conference of the Mathematics Education Research Group of Australasia, Melbourne: Deakin University Press, 390-397.

Pegg, J. (1992) Assessing students' understanding at the primary and secondary level in the mathematical sciences. In J. Inzard & M. Stephens (eds), Reshaping Assessment practice: Assessment in the Mathematical Sciences Under Challenge, Melbourne: Australian Council of Educational Research, 368-385.

Reading, C. (1998) Reactions to Data: Students' Understanding of Data Interpretation. In L.

Pereira-Mendoza, L. Kea, T. Kee, and W-K. Wong, (eds) Proceedings of the Fifth International Conference on Teaching of Statistics, Singapore, ISI Permanent Office, Netherlands, 1427-1434.

Reading, C. and Pegg, J. (1996) Exploring Understanding of Data Reduction. In L. Puig, and A. Gutierrez, (eds) Proceedings of the 20th Conference of the International group for the Psychology of Mathematics Education, Valencia, Spain, 4, 187-194.

Shaughnessy, M. (1997) Missed Opportunities in Research on the Teaching and Learning of Data and Chance. In F. Biddulph and K. Carr, (eds) Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia, Rotorua, N. Z.: Univ. of Waikato, 6-22.

Truran, K. (1997) Beliefs About Teaching Stochastics Held By Primary Pre-Service Teaching Students. In F. Biddulph and K. Carr, (eds) Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia, Rotorua, N. Z.: Univ. of Waikato, 538-545.

Watson, J. (1998). Professional Development for Teachers of Probability and Statistics: Into an Era of Technology. International Statistical Review, 66 (3), December, 271-289.