UNDERSTANDING DATA COLLECTION

Chris Reading
Centre for Cognition Research in Learning and Teaching
University of New England, Armidale, Australia

Recent changes in the research agenda, fuelled by curricula changes, have focused on considering what 'statistical thinking' really means. To assist educators in both curriculum design and assessment more needs to be known about students' statistical understanding. This paper takes up the theme by considering students' responses to two open-ended tasks, based on scenarios involving data collection. The first task requires a suitable data collection method to be suggested, while the second task suggests the method but asks for implementation details. In both, a justification for the answer is elicited. A developmental sequence of nine levels was identified and the responses to the two data collection questions were analysed. The SOLO Taxonomy was used as the theoretical framework to assist this process.

Introduction

As more researchers focus on students' statistical understanding, some aspects of statistics are being more thoroughly researched than others. Just as the understanding of simple probability has been identified as critical to the statistical process, so too is a basic understanding of data collection. Too often students are given data to work with or told how to collect data, rather than experiences which involve data collection decisions. Shaughnessy (1997) advocates encouraging teachers to give students a chance to show what they *can* do statistically. This should include making decisions about data collection and more research into students' understanding of data collection is needed.

The SOLO Taxonomy (Biggs & Collis, 1982) is being increasingly used as a framework in both probability and data handling. SOLO levels have been used to classify student responses concerning; data representation (Reading, 1999; Chick & Watson, 1998), data reduction (Reading & Pegg, 1996), data interpretation (Reading, 1998) and uncertainty (Moritz, Watson & Collis, 1996). This paper explores students' responses to questions concerning the understanding of data collection, 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). The model, which allows a deep analysis during categorisation of students' responses, 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). The three relevant levels identified within each of these modes 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. A cycle of growth forms as the three levels recur within the modes, with the relational level response in one cycle similar to, but not as concise as, the unistructural response in the next. Different cycles of levels are identified by the nature of the element on which the cycle is based.

Research Design

One hundred and eighty secondary students, selected randomly over gender, mathematical ability and academic years were tested on a range of statistical questions. This paper reports on the responses to a two part question which presented short scenarios to students and then asked about some aspect of the related data collection. The question was open-ended and students were asked to explain the reasons for their answers. Part I of the question sought students' ideas on method of collecting data, while Part II was more specific, the method was given and implementation details were sought. For a more detailed discussion of the analysis of responses to these questions see Reading (1966).

Analysis of Responses to Part I

The question as presented to students is shown in Figure 1. Based on the depth to which the response indicated the ability of the student to understand the collection of the data, three major groupings of the levels were identified.

PART I Question

Radio stations have their own way of working out the most popular song on the radio and they often produce Top 40 charts. Imagine that you have been asked to do this independently of the radio station and answer the following questions :
(i) Describe the best way to find out what the most popular songs are on the local radio station.
(ii) Why did you decide to find out this way?

Figure 1

First Group (No Method Suggested)

Observed responses, which were coded into two broad levels (1 and 2), attempt to rationalize the requirements of the question but show no real concern about actual data collection.

Level 1 These responses do not fully address the question, suggesting the use of data that have already been collected rather than collecting data to address the problem. For example a Year 7 student wrote:
(i) *You could ring up the radio station, ask someone that works there, go in and ask them.*
(ii) *It was the first thing that came into mind. It would be easier than worrying about it.*

Level 2 These responses indicate that all aspects of the question have been considered, but a suitable explanation as to why the answer was chosen was not

given. For example a Year 12 student gave a reason to collect data rather than for the method chosen:
(i) *Watch programs such as Rage or Video Smash Hits then maybe listen to the radio to see if it is right, or just ring up the station and ask.*
(ii) *So that I know what to expect if they ever play a song on the radio.*

Second Group (Concern with Physical Aspects of Data Collection)

Responses in the second group (coded as Levels 3, 4 and 5) are concerned with rationalizing the method of data collection. These responses attempt to describe suitable data collection, but are mainly concerned with physical aspects collection such as, the time or cost involved. There is no evidence of concern for the quality of the resulting sample.

Level 3 These responses indicate that, in attempting to justify the suggested method of data collection, focus was directed back to the question and not to any specific aspect of the collection of the data. For example a Year 7 student wrote:
(i) *Have a piece of paper sent to all houses, get them to write their favourite songs on them and return them to the radio station.*
(ii) *I decided this way because I think it would be a good idea.*

Level 4 These responses give reasons, with an explanation for the method chosen, which focus on physical aspects of the collection process. There is no real concern for the accuracy of the resulting sample. For example a year 9 student wrote:
(i) *Have a phone in census or a questionnaire that is put through the public for their forty favourite songs.*
(ii) *I decided to find out this way as you can get a larger amount of information in a relatively short amount of time.*

Level 5 These responses indicate that concern for the physical aspects of the data collection have been rationalized. However, the only concern for the quality of the resulting sample is that data have been collected in such a way that the sample is fair or accurate with no indication as to how this is to be achieved. For example a Year 8 student wrote:
(i) *By finding what music is bought as singles most at the music stores.*
(ii) *Because it is the most accurate way of finding out this.*
and a Year 12 student wrote:
(i) *Have people ring in and vote for their favourite song. The song that is most popular will be no. 1, the second most popular no. 2 and so on.*
(ii) *To give everybody an equal chance of giving their opinion of their favourite song.*

Third Group (Concern with Quality or Accuracy of Resulting Data)

The final group of responses (coded as Levels 6 and 7) indicate concern for the quality or accuracy of the data in the resulting sample.

Level 6 These responses indicate the need for sample selection to be arranged so as to produce a range of data in the sample, based on one variable. For example a year 12 student used 'time' as the variable:
(i) *Do a random survey on the radio turning it on at different times of the day at different intervals noting the songs that are being played.*
(ii) *Because it gives you a less bias opinion and view on the popularity of certain songs. You can get a wider census area, making the results more realistic.*

Level 7 These responses indicate that selection of the sample has been based on more than one variable in an attempt to improve the range of the responses which are collected. For example a Year 12 student used 'age' and 'background':
(i) *Collect group of people of varying ages and background who listen to the radio and ask them their favourite songs.*
(ii) *Not biased to any group of people and asks people who are interested in music because they listen to the radio.*

The results, arranged by academic year in Table 1, illustrate a number of interesting points. First, there are only three students (5%) from Years 11 and 12 whose responses fall within the first group (Levels 1 and 2), whereas in Years 7 and 8 there are nine students (15%), in this group. Second, there are only three students, all in Year 12, whose responses were coded as Level 7. Last, there was an overall bulge at Levels 3 and 4. This bulge is consistent in all years, except Year 12 where the bulge shifts to Levels 4 and 5.

Table 1 - Response Level by Academic Year for Part I

Level	Year						Total
	7	**8**	**9**	**10**	**11**	**12**	
1	5	1	1	0	2	0	**9**
2	1	2	3	5	0	1	**12**
3	10	9	9	6	10	5	**49**
4	10	14	9	12	12	9	**66**
5	2	3	4	5	5	7	**25**
6	2	1	4	1	1	5	**16**
7	2	0	0	0	0	3	**3**
Total	**30**	**30**	**30**	**30**	**30**	**30**	**180**

These results suggest that, when dealing with data collection, the level of response improves progressively with academic year, although, the bulge at Levels 3 and 4 suggests that many students are more concerned with sample selection based on physical aspects rather than quality of the data.

Analysis of Responses to Part II

Answering this question (Figure 2) meant that students did not need to be concerned about the method of collection, the actual details of the sample were required.

A similar hierarchy of levels of response was observed for Part II so examples of responses for Levels 1 to 7 are not included. However, two other Levels 0 and 8, were observed.

Part II Question

There are often surveys of the community to see what T.V. programs they like to watch. The editor of the school magazine is interested in writing an article about the viewing habits of the students at A.H.S. and asked you to find out the information.
(i) You are only able to ask 30 students from the school. Which students would you select to ask ? (Don't use names)
(ii) Why would you select these students ?

Figure 2

Level 0 These responses (in the First Group before Level 1) indicate that no attempt at all has been made to answer the question.

Level 8 This response (in the Third Group after Level 7) indicates that the selection of the sample, based on a number of variables, also takes into account the composition of the population from which it is drawn. This is a more thorough attempt to make the sample representative. This unusually good response from a very insightful Year 8 student, who achieved at Level 6 in Part I was:
(i) *I would try to select a broad spectrum of the populous, taking into account age and social groupings. I would keep the divisions proportionate to what they are in the school environment e.g. there are 80 people in one social group and 20 in another therefore i would take 4 people randomly from group 1 and 1 person from group 2.*
(ii) *I would use this method to be sure of getting the full range of viewing habits within the school, but so as not to overestimate the statistical effects of minority groups.*

Three interesting points arise from the results, arranged by academic year in Table 2. First, there are no students from the two senior years whose responses fall within the first group (Levels 0, 1 and 2), whereas in Years 7 and 8 there are a number of students, eight (13%).

Table 2 - Response Level by Academic Year for Part II

Level	Year						Total
	7	**8**	**9**	**10**	**11**	**12**	
0	1	0	0	0	0	0	**1**
1	1	0	0	0	0	0	**1**
2	3	3	0	3	0	0	**9**
3	1	0	0	0	1	0	**2**
4	12	7	8	3	1	3	**34**
5	7	9	10	5	7	11	**49**
6	2	7	10	13	13	6	**51**
7	3	3	2	6	8	10	**32**

8	0	1	0	0	0	0	**1**
Total	**30**	**30**	**30**	**30**	**30**	**30**	**180**

Second, there are only seven (12%) students from Year 7 and 8 whose responses were coded as Level 7 or 8 compared to eighteen (30%) Year 11 and 12 students. Last, there is an overall bulge which varies from year to year, ranging from Levels 4 to 5 for Year 7 through Levels 4 to 6 for Year 9 to Levels 5 to 7 for Years 11 and 12.

These observations suggest an improvement in the quality of responses with increasing academic year and the ability to show more concern with the accuracy of the sample when the method of collection is specified.

Comparison of Part I and Part II

The framework developed appears to be adequate for explaining students' understanding of the basic concepts of data collection. An upwards shift in the level of response with increasing academic year is more pronounced in Part II than Part I. There is an association between the level of the coding and the part of the question being answered ($\chi^2 = 109.5$, 6 d.f., is very significant, $p < 0.001$). Far less responses are coded into the Levels 0 to 3, and far more responses in the Levels 6 to 8, for Part II of the question than for Part I. Even the large bulge of responses at Levels 3 and 4 for Part I has shifted up to Levels 4, 5 and 6 for Part II.

This suggests that, given a sampling scenario, students are able to discuss the practicalities of the collection of data, but find it difficult to rationalize this sufficiently to consider the consequences of the sample choice on the data produced. However, once students have been prompted with some information about the sample and are able to concentrate on which members of the population to choose, consideration is given to the aspects of selection which affect the sample selected and hence the quality of the data collected.

The significantly higher level attained in Part II suggests that, with prompting as to the physical details of the sample, students were able to demonstrate a deeper understanding taking into consideration the variables which might possibly affect the resulting data.

SOLO Taxonomy Framework

The levels established for the classification of the responses, along with the structure of the SOLO Taxonomy, were used to create a framework which could be used in future to code student responses to data collection questions. The first group of three levels exhibit ikonic features while the second and third groups represent two different cycles in the concrete-symbolic (CS) mode.

Ikonic responses suggest that the required task could not be linked with any sort of symbolic representation. Level 1 responses were coded as a mixture of unistructural (U) and multistructural (M) responses. As the task had not been addressed, it is difficult to determine how many visual cues from the question are in focus or what

personal beliefs and experiences have been drawn upon, without further investigation. Level 2 responses correspond to the relational level within this mode.

The responses in the second and third groups have been able to link the concepts in the question to concrete experience. The answers include reasons linked directly to the practical aspects of data collection or to concerns about the accuracy of the sample. These responses are in the CS mode with two cycles of U, M and R levels.

The first cycle involves consideration of physical aspects of the data collection. The elements in the first cycle are the practicalities which need to be taken into consideration when data are to be collected. Typical considerations are the number in the sample and the type of data to be collected as influenced by things, such as, the cost and time involved. A relational response in the first cycle is not achieved until the student is able to consider all physical considerations as a functioning set, and hence come to the realization that more needs to be considered. The U, M and R levels in this cycle correspond to the Levels 3, 4 and 5 identified earlier.

The second cycle involves appreciating that the method of selection of the sample influences the quality of the responses. Reasons given in responses now indicate that some attention has been focused on ensuring that the data collected presents a suitable range of opinions. The elements in the second cycle are the various variables that may be used in the selection process to ensure an accurate sample. A relational response in the second cycle is not achieved until a variety of variables have been considered as concern is centred on making the sample as representative of the population as possible. In this cycle, the U, M and R levels correspond to the Levels 6, 7 and 8, as outlined earlier.

The main feature which distinguishes the concrete-symbolic mode responses from the ikonic mode responses is evidence of the recognition that data need to be collected to address an issue. Ikonic mode responses either suggest using information collected by others or consider a personal judgement sufficient. CS mode responses discuss one or more aspects of the data collection process. Within this mode, the first cycle responses are only concerned with physical aspects of data collection, while second cycle responses consider the influence of variables related to the method of selection on the quality of the data.

Conclusion

Three major findings have evolved as a result of this study. First, students are better able to consider variables influencing the selection of a sample when the physical aspects of the data collection process have already been resolved for them.

Next, the three broad groupings identified, namely, *No Method*, *Concern with Physical Aspects* and *Concern with Accuracy* assist in determining the stage a student has reached in understanding data collection. *No Method* responses are addressing the issue but not by collecting data, while the other two groups deal with data collection, *Concern with Physical Aspects* responses in a less statistically sophisticated fashion than the *Concern with Accuracy* responses. These groupings

offer educators a means to better follow student thinking when planning lesson sequences within the curriculum and assessing specific student outcomes.

Last, the groups of levels identified can be categorized as cycles of U-M-R levels, based on the SOLO taxonomy. The *No Method* 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 *Concern with Physical Aspects* group, are the physical aspects of collection while the focus elemens in the second cycle, the *Concern with Accuracy* group, are the various variables which could affect the accuracy of the sample. Educators could make use of the suggested framework when considering the quality of responses to gain a greater awareness of what students really know, understand and can do.

References

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.

Reading, C. (1996) *An Investigation of Students' Understanding of Statistics*, Unpublished Doctoral Dissertation, University of New England, Armidale, Australia.

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. (1999) Understanding Data Tabulation and Representation. In O. Zaslavsky, (ed.) *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education,* Haifa, Israel, **4**, 4-97-4-104.

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.