CHAPTER 3

DESIGN - STUDY 1

Several research questions emerged from the literature review in Chapter 2. Of interest in this first study were: do levels exist in students' understanding of statistics and is it possible to use SOLO to describe these levels? If levels are found, can they be described, are there cycles within the modes, and what are major differences between responses coded as ikonic and concrete-symbolic? Also, does the level of understanding of a student depend on his or her academic year, mathematical ability or gender? Study 1, the first of two studies, consists of a pilot study and a main study. The pilot study was used for item development and refinement of test procedure. The main study then investigates the research questions. The object of this chapter is to describe the development of the experimental design of the first study and methodological considerations incorporated into the study.

There are two main sections in this chapter. The first section, entitled **Pilot Study,** clarifies aspects of initial research used to refine the test instrument and procedure. The statistical areas which define the scope of the study are outlined also. The second section, **Main Study**, describes the experimental procedure for the research pertaining to the above areas of enquiry. As well as aspects of the testing and the context of the study, limitations and constraints are presented. The results of the main study are presented in Chapter 4 and an overview of the study is provided in Chapter 5.

Pilot Study

The Pilot study was necessary to develop a test instrument, as there was not a previously existing instrument which was suitable. The main purposes of the pilot study were to decide which items were more likely to be successful in determining student understanding, to trial the written paper in terms of readability and time requirement, and to give an early indication of whether ranking of responses was possible. The following discussion of the pilot study is divided into four sections. The first clarifies the statistical areas to be covered by the study. Next, the process of item development is outlined. Following on from this, the first trialling of the written test is discussed. Finally, the conclusions from the pilot study and their implications on the main study are outlined.

Aspects of the Research

Before deciding on the content of questions to be used for the study a decision was needed as to what aspects of statistics were to be covered by the research.

Holmes' definition of Statistics, as discussed in Chapter 1, was used as a basis for this decision. This definition had five major areas, Data Collection, Data Tabulation and Representation, Data Reduction, Probability, and Interpretation and Inference. Details of what is contained within each of these areas can be found in Chapter 1, where the definition is reproduced in detail. Only four of the main areas were decided as focus areas for the study. Probability was not included in the study because the considerations needed to be reduced to a manageable size, and much research had already been undertaken in the area of understanding in probability. Reference to some of this research can be found in Chapter 2.

To design a study to address every aspect of the four remaining areas would have been beyond the scope of this research. Hence important aspects of each focus area were selected for investigation. Within each aspect, the choice of stimulus items was influenced by the students' limited sphere of experience (dictated by the N.S.W. Syllabus).

Data Collection

The first focus area covers all aspects of collecting sample and census data. This includes various types of census, types of data and sampling techniques. The area for investigation was the process of deciding on the type of sample that should be collected and the reasons for collecting such a sample. The research asks what students view as the best methods of sampling to use when investigating phenomena?

Data Tabulation and Representation

The major considerations in the second focus area are basic forms in which data can be presented. For the present study, only simple one way tables and frequency histograms were used. The investigation concentrated on what interpretation students placed on data that were presented to them in either of these two forms. The research asks what students 'see' when confronted with data?

Data Reduction

A wide range of statistics are included in the third area. As well as measures of central tendency and dispersion, there is regression and correlation and other summarizing statistics such as indexes and birth/death rates. Most experiences of secondary students in this area are in measures of central tendency and dispersion so attention was restricted to these. The study concentrated on investigating what measures students considered most useful for reducing data. The research asks how students condense data into a more usable form?

Interpretation and Inference

The fourth, and final, area of interest involves detailed interpretation, such as, estimating populations and parameters and comparing data. Also, the detailed design of experiments to take into account such problems as estimation of errors and inference, including the identification of trends and projections. The area of focus is students' ability to consider trends in data and to make predictions using the data. The research asks what patterns students 'see' when confronted with data and how they use data to make predictions.

These four areas form the basis of focus for the study. The restrictions within each area were made to keep the study within a manageable range. Far more research is possible than the small amount covered by this study.

Item Development

As few researchers have studied students' understanding in statistics it was necessary to investigate what type of questions would be most suitable for the study. Those questions trialled, some unsuccessfully are presented for the readers' perusal.

To best investigate students' understanding it was necessary to design questions which reflected various areas of statistics and enabled students to provide as much detail as possible. The lack of any previous knowledge of the structure of levels meant that open-format questions, rather than closed-format, had to be used. Open-format questions have been used successfully in various fields to allow a greater range of student responses when investigating levels of learning. In their investigation of the use of statistical heuristics in everyday inductive reasoning Nisbett, Krantz, Jepson and Kunda (1983, p.356) had subjects respond to various open-format problems. The use of this type of question is not widespread, however, in the field of statistics.

Once the decision was made to use open-format questions it was necessary to determine the context of the questions. It was decided to make references in questions to practical situations with which the students would feel comfortable, for example, in the questions trialled, television viewing habits of students at the local school were used. Students are more likely to put their statistical skills to use on data with which they feel comfortable, such as local rainfall rather than shipping rates in a foreign port. Rogerson (1987) stressed that techniques should be applied in 'real-life' situations because research has shown that more successful statistical decisions are achieved when dealing with data that relate to familiar situations.

Questions were needed in each of the four statistical areas outlined above, that is, data collection, data tabulation and representation, data reduction and interpretation and

inference. During 1990 various questions were trialled on an informal basis with Year 10 and Year 7 classes. These two academic years were selected as a guide to minimum responses which may be expected of junior and senior students. The questions were given during the course of their normal mathematics lessons. The students were told that this was not a competitive situation and they were put under as little pressure as possible, in an attempt to get them to write as much as possible.

Of the questions that were trialled some proved to be more successful than others. A few questions were discarded as unsuitable and others were revised before being included in the pilot study. However, not all the questions in the pilot study were first trialled in this preliminary investigation. Nevertheless, it was felt that enough questions were used to obtain an indication of the type of questions that would be most suitable for the targeted age range. The following includes examples of the questions that were trialled, both successfully and unsuccessfully.

In the focus area of data collection the trialled question was on sample selection and is shown in Figure 3.1. This question was given to Year 7 and Year 10 and students responded well. There was enough variety in their responses to make analysis worthwhile and enough consistency to make grouping possible. This question was included in the written test but only with parts (i) and (ii). Due to the great variety of responses and the difficulty it would have created in coding, part (iii) was not included. Also, in this preliminary investigation, it was felt that the question had led the students too directly into the sampling technique. Hence, in the final written paper a more general sampling question was asked first and then the above question was given to probe more deeply into their ideas on sampling techniques.

We are interested in finding out the T.V. viewing habits of the students at A.H.S. We are only able to ask 30 students from the school.

(i) Which students would you select ? (Don't use names)

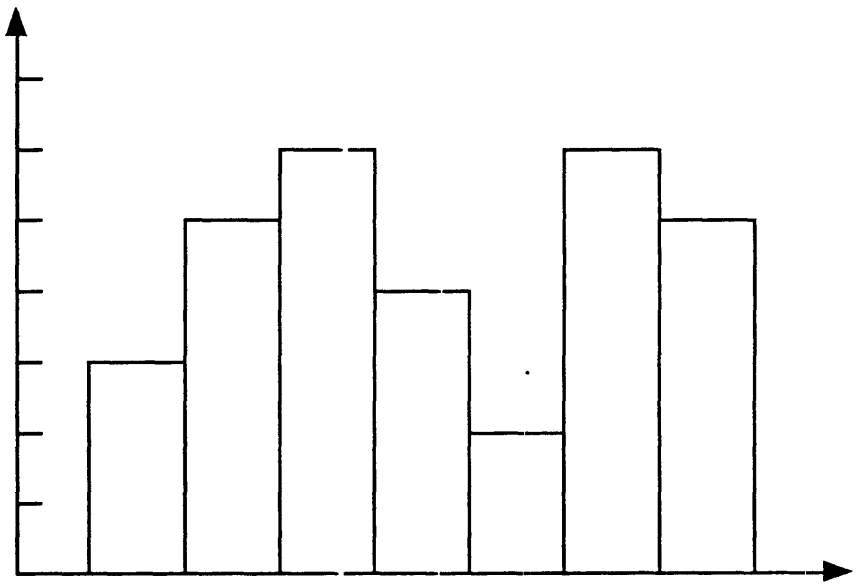
(ii) Why would you select these students ?

(iii) What questions would you ask them?

Figure 3.1 - Trialled Data Collection Question

In the focus area on data tabulation and representation, two questions were considered. One dealing with graphs and the other with information presented in table format. The first question is shown in Figure 3.2. Students took a long time deciding what to put as a heading and then mostly decided on answers which turned out to be very similar.

Most students selected for a heading some everyday event or some quantity from within their context of living. This reinforced the notion that the context of questions set should be limited to situations of relevance to the students. Then, when labelling the axes, many students selected some time scale for the horizontal and frequency for the vertical. The amount of time students took in doing this did not warrant the amount of useful information that was collected and hence, in the written test, both parts (i) and (ii) were removed. The students were simply presented with a labelled histogram and asked to write a story about what the graph told them. The results of the preliminary investigation were used as a guide for the heading and axes' labels and scales on the histogram for the test item.



(i) Write a heading.
(ii) Label the axes.
(iii) Write a story telling about what is in the graph.

Figure 3.2 - Trialled Data Tabulation and Representation Question - Graph

The second question presented information in table form and is shown in Figure 3.3. This question was only presented to Year 7, but again showed a useful variety of responses and so was included in the written test. It is worth noting that the word 'story' was removed from the question because of the fictional narratives supplied by some students. Instead students were asked to explain what the graph or table told them.

Research discussed earlier suggested that the form of presentation of data affected understanding, so in the present study questions were included in this focus area

and the final two, data reduction and interpretation and inference, in both tabular and graphical forms.

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

No. of icecreams that a Year 7 student ate during a week for a seven week period. Write a story about the information that this table tells you.

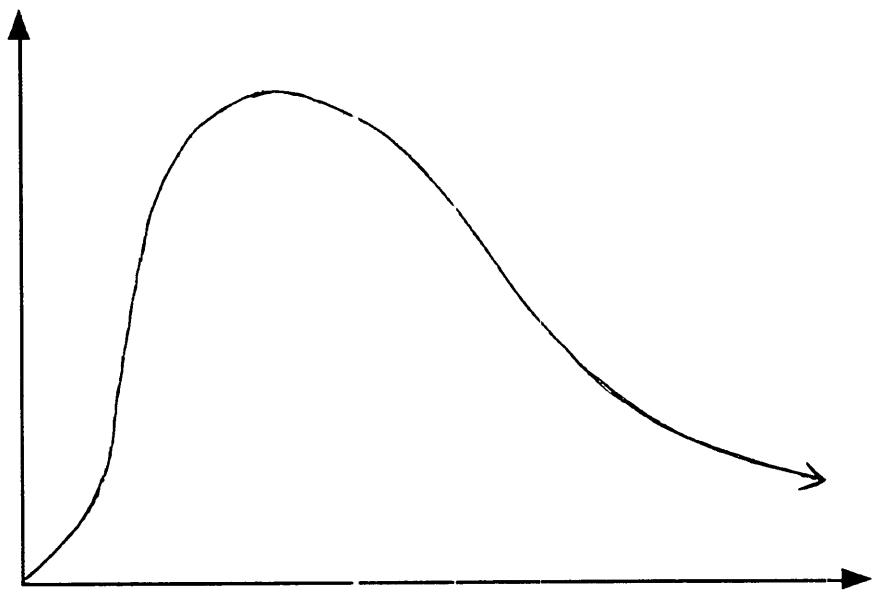
Figure 3.3 - Trialled Data Tabulation and Representation Question - Table

The first two questions trialled in the area of data reduction were not very successful and hence not included in the test. One question was on the average and the other on comparing the measures of central tendency graphically. The first question, concerning average, is presented in Figure 3.4. This question was presented to both Year 7 and Year 10 but did not give sufficient information on students' understanding of the mean to be included in the test. Apart from some attempts at a qualitative description, most students simply described a rote-learnt method for calculating the mean. There was one 'creative' student who said "I think I must have been the student who was away".

A friend missed a lesson on the topic 'average'. You have to ring up your friend and explain to him or her what an 'average' is. What would you say?

Figure 3.4 - Trialled Data Reduction Question - Average

The second question, shown in Figure 3.5, comparing the mean, mode and median, was only given to Year 10 as it was considered too hard for Year 7. Even so, it was finally not included as most students found it too difficult to answer, and just gave rote-learnt definitions of the three measures or attempts at a numerical estimate of them.



(a) Label the axes.

(b) Discuss the mode, mean and median.

Figure 3.5 - Trialled Data Reduction Question - Comparison of Measures of Central Tendency

The question presented in Figure 3.6, investigating the appropriateness of suggested measures of central tendency, was trialled more successfully. This question was presented to Year 7 and resulted in a variety of responses. However, many students took the easy option when responding by selecting one of the three suggested measures and giving a formal definition of that measure. To try to make this option not quite so obvious, when the question was included in the test, the choices of Sally, Sam and Sarah were not given and students were asked to selected a number (or numbers) themselves to describe the data and explain why they made that choice. It was felt that this gave more scope to the possibility of responses. As mentioned earlier, a similar analysis question was also included with the data in graphical form.

No questions were trialled in the focus area of interpretation and inference as it was felt that the information from trialling the other questions provided enough basis for designing a question in this area. However, before the questions were included in the written test for the pilot investigation, samples of responses were presented at some academic

gatherings to gauge whether responses warranted further investigation. During 1990 items were presented at a University of New England seminar for postgraduate students who were studying a course on levels of learning and the SOLO Taxonomy. Hierarchies of answers to two of the above questions were presented and sufficient interest was generated to suggest that it would be possible to ultimately divide these groups of responses into various levels of understanding. A selection of questions and ranked responses was also presented to a Mathematics Conference in 1992. Again reactions suggested that the grouping of responses was possible and a useful exercise.

The number of buses per day past the arboretum for a 3 week period was collected and then arranged in order

5	6	6	6	6	7	7
7	8	8	8	8	9	9
9	9	9	10	10	10	10

When asked to select a number which would best describe the data
Sally selected the mode which is 10
Sam selected the median which is 8
Sarah selected the mean which is 9
Who do you think best described the data and why?
OR if you think none of the three did a good job, select a suitable number to describe the data and explain why you selected it.

Figure 3.6 - Trialled Data Reduction Question - Best Measure of Central Tendency

The preliminary investigations indicated that it was possible to construct open-format statistical questions, and to group the responses and arrange them hierarchically in terms of the statistical understanding or thinking exhibited. Responses indicate that as statistical thinking developed so did the ability to appreciate and understand data and related statistics. These conclusions were encouraging and the next stage of the pilot study was designed to further refine the written test and the testing procedure, before the main study was begun.

Written Test

A four-question written test, designed to cover the four areas of study as outlined earlier in this chapter, was applied to a sample of students in order to refine both the instrumentation and the procedure. The main purposes of the pilot study were: to minimize any problems that may occur when the written test is given to a larger group in the main study; to check whether the responses would be suitable for coding; and to investigate the suitability of the template suggested in Chapter 2. Checks needed to be made on the readability of the test over all age groups and to make sure that the ideas were not too difficult. Also of interest was the time that was necessary for the test to be completed and whether two separate sittings would be needed to maximize concentration. Two aspects were important as far as suitability for coding were concerned. It was necessary to determine whether the questions designed would allow for a range of answers, and the responses were needed for use as a tentative framework for the coding of responses in the main study.

Sample

Only twelve students were included in the pilot study but they were purposively selected. Two students, one male and one female, were chosen from each of the six secondary school years. Students of average ability were randomly selected from the middle third of each academic year after students had been ranked according to mathematical ability. It was felt that students of average ability would suitably test the readability of the questions while at the same time being able to supply sufficient detail in their answers to determine whether coding was going to be possible.

Instrumentation

A written test was designed consisting of four questions covering the four areas of interest. Question 1 covered aspects of data collection. Part I was a general question about data collection and Part II asked for more specific details of a partially described sampling situation. Question 2 required data to be described. Part I presented the data in tabular form and Part II in graphical form. Question 3 requiring data to be summarized as statistics included two parts, with data in tabular and graphical form, respectively. Question 4 required students to interpret data and to make predictions. Again, Parts I and II presented data in tabular and graphical form, respectively. The written test is reproduced in full in Appendix A.

The questions were of an open-format style based on ideas tested in the preliminary investigations. Attempts were made to word the questions in such a way that

students would be encouraged to write as much as they wanted or were able. At the beginning of the test a warm-up section was included which asked students what statistics meant to them and what they expected to be learning about in statistics. This was designed to lead students more gently into the concept of the open-format question and also to find out how students viewed statistics.

Procedure

Some consideration was given to whether students would best complete the questions on an individual basis or in small groups. Some researchers have found that a group situation makes the student more willing to discuss ideas (Abele, 1983, p.827). This may have been helpful in the present research where it was hoped that students would write detailed answers. However, it was felt that with the input of other students, answers may indicate a higher level of understanding than that of which a particular student was capable. Hence, students were given the test to be completed on an individual basis.

Some researchers read out questions to the students in case of reading difficulties (Green, 1987(b), p.287) but this was not done for two reasons. First, the students were being tested as a group and this would have not allowed them to work at their own pace. Second, they were told that if they had any problems with the questions, as they were written, then they could ask about them. Any answers to such questions were carefully constructed so as to make the wording of the question clear to the student without giving them any special information which would help in their answering of the question.

Students were tested as one group in a classroom. A simple set of instructions was given to the students for completing the test and these are reproduced at the beginning of Appendix A. Points most stressed were:

(i) this was not a school test and they would not be receiving a mark for it;

(ii) there were no right or wrong answers and anything they wanted to write would be useful;

(iii) they should try to write something for every question;

(iv) they should not rush as there was plenty of time.

All students started the test together, being encouraged to answer the warm up question 'What is Statistics?'. Those who were reluctant to put pen to paper were given extra encouragement. From then on students were allowed to work at their own pace. Students were given blank paper on which to respond so that they felt no restrictions as to a minimum or maximum amount that needed to be answered for any particular question.

Testing was done during the morning teaching session as it was considered that students were fresher then, and more detailed accurate responses would be obtained. A double lesson, 90 minutes, was allocated for the test. Students were allowed to finish when

they were ready and to leave. Half way through the session, all students were given the option of stopping then and completing the test at another time or continuing to work on the test.

Data Analysis Plan

Responses from each part of each question were ranked according to the degree of statistical understanding that the answer exhibited. Such rankings have been attempted by other researchers. For example, Truran (1985) ranked interview responses to a probability question into nine categories, the titles of which imply increasing statistical understanding. The framework, with three broad categories as suggested at the end of Chapter 2, was used as a guide to the ranking of the responses. If a response contained elements which allowed it to be assigned to more than one category then it was coded into the highest ranked of those categories. These ranked responses were then compared to the groupings obtained earlier for all those questions which had also been included in the preliminary study. An independent observer also ranked the responses to check for reliability of coding.

Results

As far as the testing procedure was concerned, most aspects appeared to be suitable. When offered a chance to have a break after forty minutes all students elected to continue. The total time allowed for the test, 90 minutes, appeared to be adequate for most students. None of the students needed to ask questions to clarify the wording of the written test. Most students were reluctant to answer the warm-up question, although, all made an attempt with encouragement. This had a positive effect of increasing the students' confidence to continue on with the other open format questions.

When attempting to grade the responses, initially, the answer to part (i) of a question were coded separately to the part (ii) answer. However, discussion of the ranking with others confirmed that it would be more suitable to group together an answer (part(i)) with its reason (part(ii)) to best ascertain a student's understanding. Consequently, this was done for all the questions where students were asked to justify their answer. There was sufficient variety in the responses to allow them to be ranked. Brief descriptions of the various categories in the ranking were created but there were not sufficient responses, however, to allow accurate descriptions of groupings to be made. This would be a priority for Study 1 when there would be many more responses.

The template established out of the literature review proved to be useful as a framework on which to build the hierarchy except that the descriptions of the three categories

needed to be extended to cover aspects of data collection as well. The first category are those who are only able to reproduce information from the question or make use of preexisting information. The second category are those who consider features of the data and also those who are primarily concerned with physical aspects of the data collection process, such as, the cost and time involved. The third category are those who can explain how data can be used and those who are aware that the method of data collection influences the quality of the data collected, and hence the use which can be made of data. This revised template was considered suitable to use to rank the responses in the main study.

Conclusions

The results of the pilot study were used to make decisions about the procedure and the instrumentation for the main study. Due to the success of the procedure used, presentation and timing of the test remained as tested. Hence it was decided to allow the students to take the written test, without a break, during a 90 minute time slot and to give them encouragement in answering the warm-up question. The chance was given to ask about any questions where the wording was a concern. As students did not appear to have problems in reading the questions and creating answers, and as the responses were able to be ranked, the written test was not altered. The next section describes the design of the main body of research for Study 1.

Main Study

The preliminary investigations indicated that it was possible to construct open-format statistical questions. Also, it was possible to group the responses and arrange them hierarchically. The pilot study was used to refine the instrumentation and methodology. The main study was designed to attempt to provide enough responses to be able to construct an hierarchy and describe the various categories. This section begins with background information including the context for the study and some of the limitations and constraints which affected the research. Next, the main features of the design of the application of the written test are described. Finally, the details of the data analysis plan are presented, before some concluding remarks.

Background

To consider the background of the study it is necessary to consider both the context, and the limitations and constraints that this imposes.

Context

The research was carried out with secondary students and the content was governed by the New South Wales Syllabus. The statistical components of the syllabuses for Years 7 to 12 are included in Appendix B. In Years 7 and 8 students meet introductory statistics ideas. This involves early ideas about data collection and representation, and only a glimpse at analysis. These ideas are considered in more depth in Years 9 and 10, where there are three separate levels of study. In the General course, designed for the lower 20% of students, the statistics is contained in an optional lobe, while the statistical content of the Advanced (designed for the upper 35% of students) and Intermediate (designed for the middle 45% of students) courses are the same, not venturing beyond basic concepts. Students in Years 11 and 12, unfortunately, have little exposure to statistics. Students studying Mathematics in Society (a non-calculus based course) do gain more practical experience at using statistics, but meet few new concepts. The Mathematics in Practice (designed for low-ability senior students) course does not have any specified section on statistics but encourages the interpreting of graphs, and collecting and interpreting data within the framework of the various practical topics covered. While for those in the Calculus dominated 2, 3 and 4 unit courses, the closest topics come to covering statistical ideas is to study some probability concepts. These levels of exposure to different statistical skills need to be taken into consideration when the data are being analyzed.

The students tested were considered to be typical of secondary students in the New South Wales Education system for a number of reasons. They attend a government secondary school which is located in a reasonably large rural-urban city, whose main industry is education. The school is average in size having between 700 and 800 students. Although a number of its students do well academically, the school also aims to provide sufficient courses and interest for those students who are not high academic achievers and those who may otherwise have left school at Year 10 and entered the work force.

As the research was to be carried out entirely within one school, approval was only necessary from the Principal, rather than the Department of Education. Appreciating the importance of the research, he agreed provided any testing was not too disruptive to the students normal schedule of education. It was also necessary to seek written approval from the parents or guardians of each of the students involved. The letters of approval to both the Principal and parents appear in Appendix C.

Limitations and Constraints

There were a number of limitations and constraints imposed on the study. These include the right of the student not to be included in the study and the use of

classrooms and class time for testing. For each stage of the study, the sample to be tested was chosen according to certain criteria, to be outlined later. However, in each instance, students were free to choose whether they wished to be included in the study. No students were to be pressured into completing the test as it was felt that the type of responses being sought needed to have the full co-operation of the participants.

Normal classrooms were used for testing so that students would feel more comfortable. Occasionally this was a problem as there were external distractions such as noisy outdoor activities or change of lessons by other students. However, as students were given as much time as they needed to complete the test this was not considered to be too much of a problem. All testing was done during school hours, and mostly during class time as students were reluctant to forego their lunch period. Some senior students, however, may preferred to be tested at lunch time, rather than miss lessons.

These limitations and constraints, despite being present, were considered not to hinder the testing process to any great degree. The basic design of each component of the study is now outlined.

Written Test

The main purpose of the written test in this study was to investigate the hierarchy of responses. This investigation comprised a number of parts. Firstly, groupings of responses to each question were identified. Then, a hierarchy of understanding was developed which would accommodate the groups of responses for each of the questions.

Once established, this hierarchy was worth investigating to determine whether it was consistent within areas and whether there was a similar pattern of growth across areas. Also, it was worth investigating whether various factors influenced the levels of the hierarchy. Finally, the researcher wished to determine whether the SOLO Taxonomy could adequately describe the levels found in the hierarchy, ultimately, determining whether there were modes of functioning in statistics

Detailed discussion of the research questions for this study are posed at the end of Chapter 2. However, the analysis of responses in the pilot study also suggested that a comparison of the difficulties of the questions would be worthwhile.

Sample

A sample of 180 students was selected from the total school population, excluding those 12 who had been included in the pilot study. The sample was stratified across school year, mathematical ability and gender. From each school year, 30 students were selected of which 15 were male and 15 were female. To do this, each school year was

ranked according to recent performances in mathematics and this ranking was divided into three bands. From each of these three bands, labelled Low, Middle and High, 5 male and 5 female students were selected randomly.

As mentioned earlier, each student had a choice of whether he or she wished to take part in the research. The four students who declined were replaced by a student of the same gender and from a similar position in the ranking bands. The two students who could not take part due to lack of parental consent were also replaced in a similar fashion.

Each student was given a code containing academic Year (7 to 12), followed by sex (1 for female and 2 for male) and a two digit individual identifier (01 to 05 for low, 06 to 10 for middle and 11 to 15 for high mathematical ability). For example, the code 8213 represents a Year 8 male who is in the high band for mathematical ability. These codes are used when referring to particular responses in the descriptions in following chapters.

Instrumentation

Each student was given the written test, as refined in the pilot study.

Procedure

As for the pilot study, students were given the written test during class time and in a normal classroom. The same instructions as for the written test in the pilot study were used and encouragement to get started on the warm-up section was given. The 180 students were not all tested at once in a hall situation for two main reasons. First, this would not allow students who wish to ask questions the chance to do so easily. Second, the students may have felt under pressure to perform as the hall is the venue normally used for major examinations. Consequently, each of the six years was tested separately.

Students were not given the chance to stop, and continue at a later date, during the test but were allowed to do the test in two parts, by prearrangement, if their timetable was such that they needed to attend specific lessons.

Data Analysis Plan

The data analysis has both qualitative and quantitative aspects.

Qualitative Analysis

The qualitative analysis involved the description of the levels and groupings into which the responses were coded and a SOLO framework for the assessment of students'

understanding. Initially, the responses to each question were grouped. The template for statistical thinking, which came out of the literature review, was applied to produce broad groupings of responses. Within these groups, responses were then ranked according to the degree of sophistication of statistical understanding they demonstrated and the amount of support that their reasons gave to the answers. The criteria for decisions on ranking were based on similar ideas to those used by researchers in coding open-format statistical questions. Nisbett, Krantz, Jepson and Kunda (1983, p.356) based their coding of responses to open-format problems on the degree to which they reflected the use of statistical principles. When scoring student responses from 1 to 8, Swing and Peterson (1988, p.57) based their decisions on "the amount and strength of support given for an affirmative response to the question".

An independent judge, not related directly to the collection of the data but aware of the direction of the research, ranked a sample of the responses to check on coder reliability. Thirty responses were randomly selected for each of the four questions. These were coded using the descriptions of levels devised by the researcher and sample responses for each level. The resulting codings were then compared to the codings made by the researcher. The results of this check on coder reliability are presented with the results for each of the four questions in Chapter 4.

From the discussion of the SOLO Taxonomy in Chapter 2, a framework of general characteristics of cycles of levels within modes was constructed to apply to the results. The three modes of interest were the ikonic, concrete-symbolic and formal. The ikonic mode is imagery based and personal, with actions being made more abstract by internal representation of some form. The concrete-symbolic mode is the sophistication of the abstraction of actions to a symbol system that applies to the experienced world. The transition from ikonic to concrete-symbolic is a transition from a personal world to a real world. The formal mode is a much more abstract system which allows for hypothesizing about alternative ways of ordering the world. The transition from concrete-symbolic to formal is a transition from describing the real world to questioning it. Within each of these modes there is a level structure, the three main levels of interest being unistructural, multistructural and relational. At the unistructural level, focus is on the relevant domain and one aspect is picked up and dealt with. At the multistructural level a number of aspects are dealt with but they are not integrated, while at the relational level a number of parts are integrated to produce a coherent structure and meaning. The framework allows also for the possibility of a cycling of these levels within the modes, especially in the concrete-symbolic mode, as suggested in Chapter 2.

This framework was used as a basis for describing levels within modes along SOLO guidelines for each of the focus areas. The descriptions of the suggested features which could be used to distinguish the various levels and modes, along with references to

sample student responses, are given at the end of the analysis for each of the four focus areas in Chapter 4. Some of the quantitative analysis in Chapter 5 is used to substantiate the delineation of the various levels and cycles within the modes. The plan for this analysis along with other quantitative analysis is presented next.

Quantitative Analysis

The quantitative analysis of the responses was carried out in two steps. First, the rankings of the responses were analyzed for each of the four focus areas separately, with parts of questions within focus areas being compared. These results are discussed in Chapter 4. Second, results for all four questions were combined and analyzed to assess overall student understanding of statistics. This discussion follows in Chapter 5. The detail of each of these analyses is outlined below.

Individual question analysis involved similar procedures for Part I and Part II of each of the four questions for the separate focus areas. After the responses were ranked for each student, frequency tables were tabulated by academic year, mathematical ability and gender. To consider the possibility of a relationship between academic year and the rankings, trends were discussed generally because the cell numbers were too small to perform a Chi-squared Test on the Contingency Table. However, to investigate possible effects of mathematical ability and gender on the ranking it was possible to use Chi-squared Contingency Table Tests. Any level in the rank which had expected numbers which were too small were combined with an adjacent level. The chi-squared tests were performed separately on the responses to Part I and Part II of each question.

Next, the two parts of the question were compared. The numbers of responses ranked at each level were tabulated for the two parts of the question. A Chi-square Test of the Contingency Table was used to test whether performance was similar on the two parts of the question. Then, a qualitative analysis was used to compare the performances of Part I and Part II, for each of the three variables, academic year, mathematical ability and gender.

Following the general comparison of the two parts of the question, the individual student performances on the two parts were compared. The levels were paired for the two parts of each question for each student and a Chi-squared Test of the Contingency Table was used to decide whether the performance of the student on the two parts of the question was related. Where a relationship was found, it was investigated further using a Nonparametric Sign Test. This test was selected as it is not necessary to assume any underlying distribution of the observations. For each student the level for Part II was subtracted from the level for Part I and a +, 0 or - recorded, depending on the difference. Testing the hypothesis that the codings are the same for the two parts of the question is

equivalent to testing that the proportion of positives is 0.5. As the sample size, n = 180, was large the normal approximation to the binomial distribution was used. A significant result suggested that the level had changed and an increase or decrease in level is suggested by the greater number of positive or negative differences.

Combined data analysis, the results of which are presented in Chapter 5, was used to substantiate many of the general observations made in Chapter 4. The Rasch model was used to produce an estimate of statistical understanding for each student which incorporates the information from the questions covering all four focus areas. The concept of the Rasch model was first introduced for the analysis of dichotomously-scored responses in 1960. When data are fitted to the model, person parameters can be freed from the item difficulties and item parameters can be estimated independently of the calibrating sample. Since being first introduced, the Rasch model has been extended to analyze other polychotomous data such as counts, repeat trials and rating scales. More recently, Masters (1982) has developed the partial credit model which is applicable to the data in the present study.

The partial credit model is relevant to items where responses can be coded according to an increasing, or decreasing, degree of 'correctness'. Rather than apply the Rasch dichotomous model to each step in performance level, which assumes hierarchical independence and increasing difficulty with each step, Masters (1982, pp.157-158) suggested considering the individual difficulty of each successive step in the item. The probability of student n responding in category x of item i is given by

$$p_{nix} = \frac{\exp \sum_{j=1}^{x} (\beta_n - \delta_i - \tau_j)}{1 + \sum_{k=1}^{m} \exp \sum_{j=1}^{k} (\beta_n - \delta_i - \tau_j)}$$

where β_n is the ability of student n; δ_i is the overall difficulty of item i; $\tau_1, \tau_2, \tau_3, \ldots, \tau_m$ is a set of parameters associated with the transition between response categories; and m is the number of response categories provided for each item. Masters (1982, pp.163-166) used a maximum likelihood procedure to estimate the parameters in the model. Means and standard deviations of the infit (weighted) and outfit (unweighted) statistics allow a check on how well the data fit the proposed model. The advantage of the partial credit model over other models is that the parameters in the model are separable. This makes it possible to produce sufficient statistics for person ability and for step difficulty.

The QUEST Interactive Test Analysis System (Adams & Khoo, 1993) which is based on the preceding equation was used to process the data. This software includes the most recent developments in Rasch measurement theory for the analysis of test and questionnaire responses as well as traditional analysis procedures. Three measures were used to check on the suitability of the test as a measure of statistical understanding. First, the item consistency is Cronbach's alpha as a measure of internal consistency (p. 41) which was used to measure the extent to which items reflect the same underlying construct. Second, the fit statistics for question estimates and student estimates are weighted and unweighted residual-based statistics and are expressions of the fit of the item to the model (p. 28, p.85). They were used to check whether the data compared favourably with the model. Third, the infit mean square map shows the infit mean square for each item and two vertical dotted lines which enclose acceptable values. One line indicates a mnean-square which is 30% above the expected value and the other a mean square which is 30% below the expected value (p. 32). This was used to determine the level of parameter fit in the model for each question. The estimates produced by QUEST were used to create a number of statistics which then allowed analysis of various features of student performance and the hierarchy of levels used to code the responses.

To consider the influence of variables on student understanding an Analysis of Variance with three sources of variation, academic year, mathematical ability and gender was applied to the estimates. The Scheffe Test was then used to compare pairs of means for any significant effects. Question difficulty and the structure of the hierarchy were also investigated. Overall difficulty scores for all questions and tau values, measuring the degree of difficulty in moving from one level to the next, were calculated by the QUEST system. Graphs of tau values for each question were drawn to investigate which levels are the more difficult for students to step up to from a previous level. Level thresholds, the estimate needed by a student to have a 50% chance of having a response coded at that level, were also calculated by QUEST, except for first level responses. QUEST cannot calculate threshold values for the first level because the response data do not contain sufficient information to discriminate between subjects responding at that level. The only estimate that QUEST can make for the first level response is that it is less than the value for the second level. Graphs of the thresholds were drawn for each question to investigate the grouping of levels into cycles. Finally, mean estimates for students were compared to the level thresholds to determine where the 'average' secondary student may be found in terms of the framework suggested based on the SOLO Taxonomy.

Following on from the combined data analysis, Chapter 5 specifically addresses the research questions posed at the end of Chapter 2. The research in Study 1 was designed to investigate in depth students' understanding in statistics and overcome some shortcomings of previous research.

Conclusion

There are a number of features of Study 1 which set it apart from earlier research. Firstly, a large sample of students, 180, were tested. This allowed the random sampling to be stratified over academic year, mathematical ability and gender so that the influence of these factors could be considered. Second, the questions were unlike those used in most earlier research. An open-format question style was used to allow students the opportunity to explain fully their thoughts. Also, the questions used were set in a context that was realistic to the students and the content of the questions was spread over differing aspects of statistics. Third, the test was given to the students in their normal school environment and not in an experimental situation, so that they would feel more comfortable about responding. Finally, the levels from the various aspects of statistics were combined in the analysis by using the Rasch model to estimate overall understanding

All these factors combined to produce a study which attempts to give a broad view of students' understanding over the full spectrum of statistical aspects. The results, conclusions and implications of the analysis of the data from the main study are presented in Chapters 4 and 5. Following on from this, a second study was planned to check on the accuracy of the assessment tool devised and to consider the effect of a longitudinal factor on statistical understanding.