

CHAPTER 6 DEVELOPMENT OF TEST QUESTIONS

6.0 INTRODUCTION

I will show the development of the test questions in a chronological manner. Examples of each 'type' of question which had been posed to various subjects will be given, the rationale for including them when this is not obvious, and comments as to their usefulness/validity will be made.

6.1 PRELIMINARY DEVELOPMENT OF TEST QUESTIONS

An 'Adult Test' had been developed for the MSc Research Project and typical questions of two types are given below:

13. Guess the number of film theatres in the U.S.S.R.?
A. 1,456 B. 14,560 C. 145,600
21. How many X's can you write down in 15 seconds? Guess, then test.

Question 13 was taken from a test developed by Raidt [1982]. She claimed that one could estimate the answer to this problem by using one's general knowledge about the number of cities in the U.S.S.R. of a given population and therefore, predict the correct order of magnitude of film theatres. These test questions were shown to the Hillcroft volunteers requesting their views on the usefulness of the questions. They did not consider Question 13 to be a valid question because questions similar to it rely upon trivial knowledge (the number of cities etc.). They especially felt that pupils would not find them interesting (and I agreed) and they voiced the opinion that pupils might just guess the middle answer of the three choices. I shall discuss multiple choice answers more fully in the next chapter. They also thought the subject could "cheat" to achieve the answers to Question 21 thus making this type of question unsatisfactory. The volunteers suggested that all the information necessary to answer the question must be contained in the question itself. They also indicated that computational estimation questions should involve relatively straightforward arithmetical operations and should be embodied in everyday situations. I found these remarks useful and decided to develop two 'test papers' which would use this information and involve areas of

interest suggested by the volunteers. The first eight questions cited below are various computational estimation tasks and the next six cited are quantitative estimation tasks.

1. Is $35 + 8$ more than or less than 40?

8. The number of students in a certain lecture were:

Monday	Tuesday	Wednesday	Thursday	Friday
498	506	587	412	475

What was the total number of students attending the lectures?

9. Can you buy 4 LPs costing £3.65 each if you had £14?

11. The Government spent the following amounts on four projects:

Project A	£11 954 164	Project B	£1 126 005
Project C	£ 4 170 522	Project D	£ 750 572

To the nearest million, how many millions were spent on these projects?

12. If there are 5280 feet in one mile, about how many feet are there in 7.189 miles?

a) 35 b) 350 c) 3500 d) 35000 e) 350000

13. How do you know that:

9×23 is less than 270?

$35/60$ is more than $1/2$?

It is less than 2 hours from 10:35 to 12:15?

$28 + 18 + 28$ is less than 80?

14. What is 25% profit on £320?

£8, £80, or £800

15. If lino costs £4.50 per roll, how much will it cost to lino a room measuring 10 yards by 20 yards if each roll is 1 yard wide and 10 yards long?

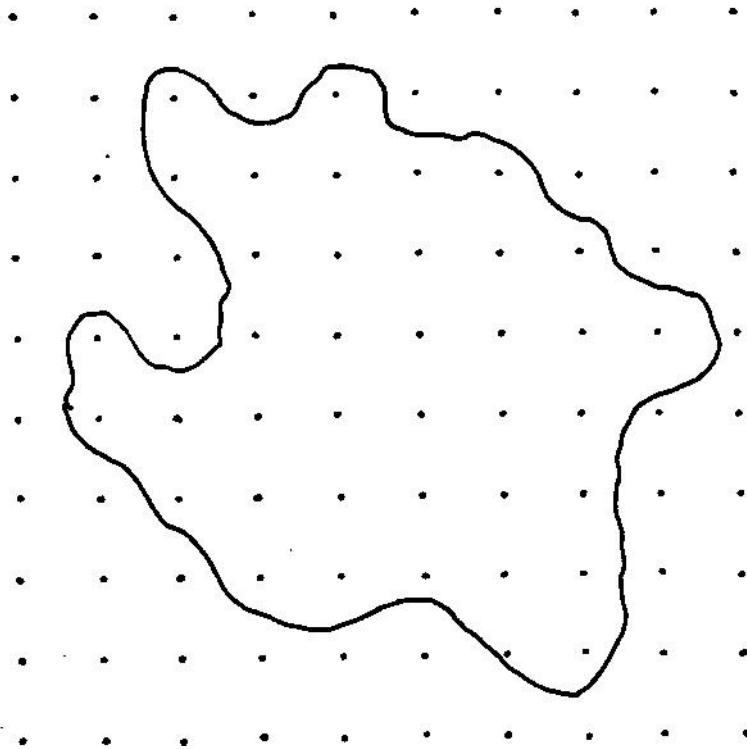
It should be obvious that some questions were designed to ascertain whether pupils would be able to attend to the significant figures in the numbers and to ignore numbers which were comparatively unimportant. Question 8 involves 15 digits but provided the estimator 'rounds' the 5-3 figure numbers to 3 - 500's and 1 - 400 and 1 - 600 or more simply to 5 - 500's roughly, an estimate of 2500 (a very accurate estimate) is obtainable ignoring most of the digits.

16. How many 10p pieces stacked up will be as high as a 10p piece on end?

17. What is the thickness of the lead in a pencil?
18. What is the room temperature?
19. What is the outside temperature?
20. How many dots are inside the closed curve in the figure shown?
(Figure 6.1 below) There are a total of 100 dots in all.

FIGURE 6.1

HOW MANY DOTS INSIDE THE CURVED LINE?



21. It is claimed that a shower uses less water than a bath. How would you attempt to prove or disprove this claim if the bath and shower were in different parts of the house?

The adult volunteers considered this to be a better test with some exceptions. Questions which required knowledge of certain terms, e.g. percentages, fractions, were considered inappropriate for pupils because if they had not been exposed to these terms in the recent past, they might

not remember their meaning. Also, questions which could be solved by prior information, (e.g. the thickness of a pencil lead is often printed on the pencil) were deemed invalid because they were not testing the skill of estimation. The reading ability of the subject during these tests interfered with his/her ability to estimate inasmuch as the time required to read and understand the question often impeded his/her progress. Consequently, the decision was made that any school test would need to be verbal, possibly utilising an overhead projector to give numerical data. Also, all questions would be timed to try to 'force' the pupils to estimate rather than count or calculate algorithmically.

Other questions were omitted and it may be useful to identify the reasons for the omission of one such question. Although most of the adults found the 'shower/bath' question interesting for themselves (several said they often had thought about it), they did not think it would be interesting to pupils. They thought that very few of them would have enough experience in household matters to make sensible suggestions and, therefore, it was omitted.

The temperature questions produced an interesting response. Inside temperatures were given in Fahrenheit while outside temperatures were given in Celsius. I think a possible reason for this is that weather reports often give the temperature in Celsius but the adults were accustomed to setting interior thermostats in Fahrenheit. The fact that Britain, at present, is operating a dual system of units complicates the task of setting questions because when the subject is asked for an estimate, s/he may not be certain which system of units is required. I decided that test questions for pupils would be posed in terms of the metric system. This decision was taken after a great deal of deliberation and consultation but was due, primarily, to the fact that the teachers (who volunteered to conduct the tests) believed it would be a retrogressive step if the Imperial system were used.

At this stage in the research, I decided that sufficient experience in interview technique had been achieved and a set of questions was taking form to enable me to begin testing and interviewing pupils, which was the main preliminary aim of this part of the research.

6.2 PILOT TEST QUESTIONS FOR SCHOOLS

The questions to be used with the pupils, as with the adults, had to be of interest to them and to address a specific estimation skill e.g. computational estimation of an addition problem. I consider it vital that the question needs to be such that the pupils could identify with the possibility that they could actually be required to estimate the answers to the questions and that the questions were ones which they could be expected to encounter in their day-to-day lives. Regardless of my wishes, I am certain that some of the final questions were not of interest to some of the pupils but oral feedback during the Pilot Testing gave me some assurance that the questions asked were generally interesting to the pupils and, I hope, their interest would give them some incentive to give estimates to the best of their ability. In the previous stage of the research, the complication of using the Bright [1976] model (for classifying quantitative estimation tasks as described in Chapter 2) was not deemed necessary due to the speculative nature of the research. However, for this series of tests, it was used to establish the area of estimation to be investigated and I decided to concentrate on Class A tasks which require an estimate of the measure when the object and unit were given in the question. Class B type questions were not included as this class is much more subjective and not easily analyzed. The reader will find the 'category' of quantitative estimation questions in the following discussion indicated by "A", "B", "C" or "D" as defined below:

"A" the object was present, the unit present

"B" the object was present, the unit absent

"C" the object was absent, the unit present

"D" the object was absent, the unit absent.

Many multiple choice questions were available and were used in other studies and the ease of result processing makes them attractive. I chose distracters to computational estimation tasks which could be obtained by incorrect arithmetic manipulations or order of magnitude errors. The quantitative estimation task distracters were more difficult to determine and integer values were selected which were, roughly, in error by between 30% and 70% with one exception which was double the correct length i.e. an error of 100%. Subsequent to conducting the tests, I found several reasons that multiple choice options may not be appropriate for

estimation tests. One obvious one is that it limits the subject and may promote a form of random guessing. Further comment will be made about multiple-choice questions in Chapter 7.

Questions which were found to be useful in the adult tests will not be repeated at this time except when pupils showed them to be inappropriate. Secondary school classes 'tested' during the pilot testing programme were taken from across the ability range and the pupils were, in all cases, encouraged 'to have a go' and not to try to get the 'right answer'. The first Pilot Test was conducted with three classes in the secondary modern girls' school where I was teaching and a few examples of the oral questions (with categories) are given below:

Write down a number which would make each of the following statements true.

1. To get more than 11, I need to add ___ onto 4.
17. How long are the following items in cm.
 - a) the pencil? "B"
 - b) the string? "B"
23. Roughly, how many tiles would cover:
 - a) the table? "A"
 - b) the wall of the classroom? "A"
24. Roughly, how many blocks would:
 - a) fit into the box? "A"
 - b) fit under the table? "A"
 - c) fill the entire classroom? "A"

The following questions were shown on the OHP:

25. $19.8 \times 42.3 =$
A. 83.754 B. 837.54 C. 8375.4 D. 8.3754
26. $173.6 + 243.2 - 51.6 =$
A. 443.6 B. 36.5 C. 365.2 D. 44.3

Questions similar to 1 confused and irritated some of the pupils and I discovered that they were concentrating on the 'words' and not the numbers resulting in confusion and some pupils claimed each question 'sounded' the same. The questions on lengths, areas, and volumes involving relatively small numbers (e.g. How many tiles would cover the

table?) were particularly popular with the pupils who said that they were "more like real life". The test was modified to eliminate the problem encountered with Question 1 and to include more questions involving quantity.

Questions which required detailed answers (How do you know that 9×23 is less than 270?) were deleted because data handling would become unmanageable. This was obvious when pupils' responses included comments such as "Cos it is!". These questions were more suited to an interview situation. The second Pilot Test was given by colleagues to pupils in two schools. Some questions (see 1 & 2 below) were included to test numerical estimation. More 'everyday' problems were included replacing those previously found unacceptable. Some questions from the second Pilot Test are shown below:

1. How many shapes on the OHP? (Scattered in groups of 3)
2. How many shapes on the OHP? (Scattered in groups of 5)

The number of newspapers delivered by a certain girl on each day of five days will be given on the OHP. Roughly what was the total number of newspapers delivered on the five days?

	Monday	Tuesday	Wednesday	Thursday	Friday
6.	23	6	34	59	5
8.	57	19	93	21	6

Sue, Jill, Sharon, and Beth collect stamps. The numbers in each girl's collection will be shown on the screen. What is the total of their stamp collection approximately?

	Sue	Jill	Sharon	Beth
9.	382	224	310	175
11.	10	359	1456	5

A 1 m. stick should be shown to the pupils.

28. What is the height to the nearest metre of:
 - a) the flagpole? "A"
 - b) the assembly hall? "C"
30. How many seconds have you been alive?
31. How many words are there in the Collins Dictionary?

Question 11 (and others) included numbers which can be 'ignored' when one is estimating i.e. Sue and Beth do not contribute many stamps to the total. I was interested in determining whether the pupils would ignore these 'inconsequential' numbers. I have had considerable classroom experience which has shown that pupils will add these numbers (in the wrong column) thus obtaining unreasonable answers. Question 28 was included to test the effect that the pupil's height had upon his/her estimation of heights. Obviously, taller pupils will see the height of the objects from a different perspective than shorter pupils. I thought this might prove interesting. Teachers conducting the tests were asked to make a special note of pupils of extremes in height for each group. Questions 30 and 31 were written to try to discover what the pupils believed to be a 'large' number as I became interested in this question. During the interviews with the primary pupils, it became obvious that younger pupils believe that 100 is a 'large' number while secondary pupils will consider 1000 or more to be a 'large' number. These questions were not wholly successful as answers such as "zillions" were recorded. The problem with this question became one of vocabulary. Millions, billions, trillions and all the possible combinations of these do not give any true indication of the pupils' opinion of the size of the numbers. Consequently, it was decided that the 'large number' question would have to be a multiple-choice question and as the 'coin' problem had been popular with the pupils, the following question was written.

A stack of 1 million 10p coins would be as tall as:

- A. A very tall man (2.3m.)
- B. A tall flagpole (23m.)
- C. A tall building (230m.)
- D. A mountain (2300m.)
- E. A plane at its highest altitude (23000m.)
- F. The orbit of a satellite (230000m.)

A local primary school headmistress agreed to allow me to 'test' some pupils and two vital flaws in the test were discovered for the primary age range. All questions for these pupils must deal only with integer values e.g. numbers should not contain decimal points and the time period that they could be expected to concentrate on the task was less than the secondary pupils. Consequently, a separate and shorter primary test

would need to be developed, albeit with a core element common with the secondary test.

The second Pilot Test results were analyzed with some of the techniques used for the final tests described in the next chapter. The results of this analysis showed tendencies and the following hypotheses were made:

1. If the object is present, absence of the unit does not affect results,
2. Pupils correctly ignore inconsequential numbers,
3. Pupils' heights may be of importance in estimating heights.

When pupils produce 'poor' estimates, they tend to:

4. Underestimate long horizontal lengths,
5. Overestimate heights,
6. Underestimate areas,
7. Underestimate very large numbers.

Hypothesis 1 was indicated by the fact that many pupils were able to estimate reasonably well lengths which they could see although they were not shown the metric rule. Most pupils appeared to ignore Sue and Beth's contribution in Question 11 resulting in Hypothesis 2 and, Hypothesis 3 came from the fact that the younger (and consequently shorter) pupils were less successful at estimating heights. Obviously, this hypothesis could be reworded in terms of age.

Hypotheses 4 - 7 were much more obvious from the data as unsuccessful estimates gave clearer evidence than successful ones. These areas of difficulty are much more interesting to explore and the final test data provides considerable substantiation for Hypotheses 6 & 7. Hypotheses 4 & 5 involve units of measurement and the final tests provide some clues as to the reasons for this type of estimation to cause difficulties for the pupils. The last four hypotheses could also be reworded for a minority who were at the other extreme of the scale, i.e. vastly over-estimating areas.

6.3 FINAL TESTS

Final Tests were written to test the above hypotheses by writing questions to augment previous questions which addressed each area of interest and the Final Tests are shown in Appendix I. A letter (see Appendix II) was sent to every school in the London Borough of Sutton to enlist support for this project and five primary and six secondary schools volunteered to take part in the testing programme. Instructions for the conduct of the tests by colleagues were needed and a script (see Appendix III) was written in an attempt to achieve a common approach by the teachers administering the tests. An example of each new type of question was to be shown on the OHP for the pupils to understand what was required. To ensure that the pupils' responses could be understood, they were asked to write the digits from 0-9 at the top of their answer sheets. Reys & Bestgen [1981, p124] emphasize that "The amount of time allocated for each question must be controlled carefully". I heeded this advice and emphasized to the teachers the need for strict control of timing. Questions 'flashed' on an OHP screen were shown for a short time (10 seconds) to reduce the chance of pupils calculating or counting rather than estimating while other questions required the teacher to show an object to the class for a specified period of time. The tests were to be completed within 30 minutes by the secondary pupils and 20 minutes by the primary pupils. The Primary Test was given to 455 pupils in 5 schools and the Secondary Test was given to 764 pupils in 6 schools in the Summer term of 1985. A discussion of the results of these tests can be found in the next chapter.

