A Longitudinal Study of Pupils' Probability Concepts

David Green - Leicestershire, England

1. Introduction

In 1986 the author tested 1000 English school pupils aged 7-11 years. The two tests used comprised 11 questions on "Randomness" and 12 questions on "Comparison of Odds". At ICOTS II the author reported on some of the "Randomness" questions (Green, 1986) and subsequently has contributed a paper for the UNESCO book *Studies in Mathematics Education Volume 7* (Green, 1989).

In 1990 the author attempted to retest some of the same children who had been tested four years earlier. 305 children were successfully located and retested. The two 1990 tests used comprised 19 "Randomness" questions and 12 "Comparison of Odds" questions, of which 4 "Randomness" questions were common to both studies as were 11 "Comparison of Odds" questions. In both studies the tests were carried out in the period April-June (1986 and 1990). A 70 page monograph has been prepared detailing many test results from these two studies (Green, 1990).

This paper reports on one of the "Randomness" questions and on a set of ten "Comparison of Odds" questions, all of which were common to both studies, and for which results have not hitherto been published.

2. The sample and testing procedure

Of approximately 500 pupils tested (in 1990), 305 were common to both 1990 and 1986 studies. For reasons of limited access at one large school, no pupils in Years 1 and 3 were tested there but all 200 pupils in Year 2 were. This explains the much larger size of the Year 2 sample. It was not feasible to test any "Year 4" pupils. Basic characteristics of the 1990 sample are shown in Table 1.

TABLE 1(a)
Basic details of sample

School Year	Age (in 1990)	Boys	Girls	Total	Mean Age (in 1990)
1	11-12	36	27	63	12.4 yrs
2	12-13	98	95	193	13.2 yrs
3	13-14	27	22	49	14.2 yrs
		161	144	305	

TABLE 1(b)
Distribution of general reasoning ability (assessed 1990)

Ability Grade	Boys	Girls	Total
A	33	25	58
B	30	43	73
C	47	37	84
D	31	27	58
E	20	12	32
	161	144	305

All testing was done on a class basis using prepared printed tests. The questions were read out to the pupils. The total time taken was about 50 minutes. The Ability Grade was supplied by the teachers of the pupils on a scale A (top) to E (bottom).

3. Randomness question

The first question asked pupils to generate a pseudo-random sequence of 50 Hs and Ts to simulate tossing a coin (see Figure 1).

1. Pretend you are tossing a coin *properly* 50 times. Put H or T in each box.

Start

□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□

Finish

FIGURE 1
Coin tossing simulation item

Various analyses of the generated sequences have been made. There are three basic aspects to this which have been identified: Frequency, Independence, and Consistency.

Frequency analysis: Do pupils on average have as many Heads as Tails? Do their results conform to a binomial model? Table 2 and Figure 2 supply the answers.

TABLE 2
Frequency of heads in 50 mentally simulated coin tosses

	All Pupils (N = 303)		Non-extremist Pupils (N = 280)	
	1986	1990	1986	1990
Sex				
Boys	24.7	24.0	24.8	25.0
Girls	24.5	25.2	24.5	25.0
Ability				
A	24.2	24.8	24.2	24.8
B	24.5	25.0	24.5	25.0
C	24.9	25.0	25.0	25.0
D	24.6	24.9	24.6	25.1
E	25.0	26.0	25.0	25.3

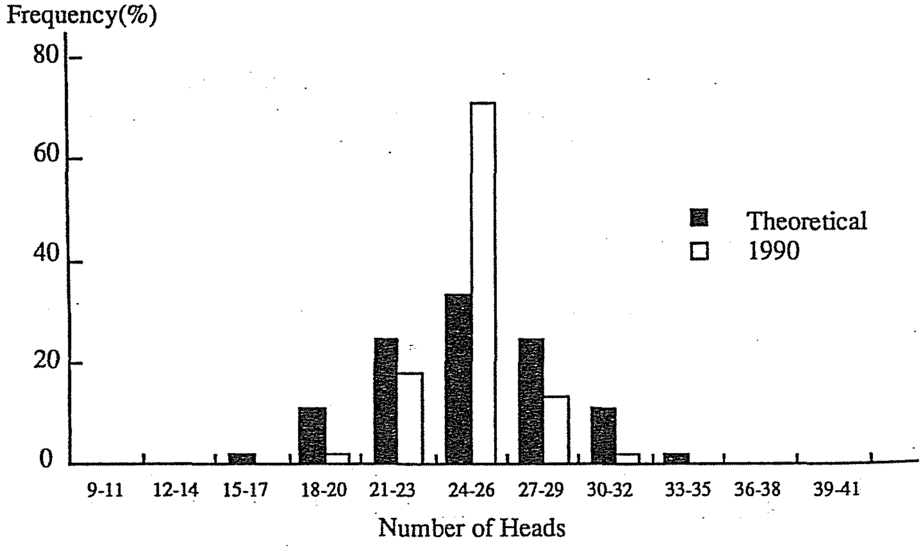

FIGURE 2
Distribution of number of heads in 50 tosses (1990)

Table 2 (left side) includes all pupils who answered this question. Table 2 (right side) excludes 23 pupils on the grounds that their sequences were extreme - either because they alternated H and T all the way through or because they repeated one outcome at least 25 times.

It is quite clear from Table 2 that there is no significant bias towards H or T, as can be checked formally with a z-test. The means are remarkably close to the theoretical 25. On the other hand, the numbers of responses falling into the middle group is in both years too large to be attributable to chance.

Independence analysis: First we examine the "extremists", who fall into two camps - the alternators and the replicators. The "longest run" statistic can be used to identify them. If the longest run (lr) is 1 then the sequence is HTHTH ... (or THTH ...), so lr = 1 determines an alternator. If the longest run is 25 or more (this is an arbitrary number, any number above 10 would seem suitable), then we have a replicator. The figure of 25 is particularly appropriate here as the question required two rows of 25 to be filled. The breakdown of the extremists is shown in Table 3.

TABLE 3
The extremist pupils - alternators and replicators

	Alternators (lr = 1)		Replicators (lr ≥ 25)	
	1986	1990	1986	1990
Sex				
Boys	2	10	-	1
Girls	3	8	1	4
Ability				
A	-	2	-	-
B	-	1	-	1
C	-	4	-	-
D	1	5	1	3
E	-	6	-	1

With the relatively small numbers of cases it would be unwise to assume that the increase in the incidence of extremism over the four years is genuine and significant. Of the 18 alternators in 1990 only one had been an alternator in 1986 (rather surprisingly, the majority had lr = 3 or lr = 4 in 1986).

The above discussion about extremists is really about a lack of independence of one outcome in relation to the previous outcome. Thus the number of runs (nr) in the sequence is an important statistic. Alternators have nr = 50 (the maximum possible) and replicators (if they persist) have nr = 1. The statistics lr (length of longest run) and nr (number of runs) are, of course, related but are not equivalent, and whereas lr is an easy measure whereby to identify extremists, it is rather more volatile and prone to irrelevant fluctuation than the much more stable nr.

For our sample the mean length of longest run was close to 4.0 for all abilities and for both sexes and across the ages, with no significant change occurring over the four year period. Table 4 shows the results for nr (for the non-extremists only).

TABLE 4
Mean number of runs (nr) for non-extremists

	1986 (N = 280)	1990 (N = 298)
Sex		
Boys	29.7	29.7
Girls	30.5	30.8
Ability		
A	29.0	30.2
B	29.1	29.6
C	31.1	30.5
D	30.4	30.3
E	30.9	31.4

From Table 4 it can be seen that all groups have too many runs, the theoretical mean being 26. Calculation of standardised statistics confirms the existence of a significant and persistent bias. These results closely match those found by Ruma Falk. It has been widely reported by psychologists that adults generate too many runs and have too few longer runs.

Consistency analysis: As the question (Figure 1) required two rows of 25 outcomes to be simulated, it was natural to examine the two rows separately and compare them. This would indicate whether the pupil was consistent (or possibly compensatory). What was found showed a high degree of consistency. Table 5 provides some basic statistics.

TABLE 5
Numbers of heads in each row

	1986		1990	
	Row 1	Row 2	Row 1	Row 2
mean	12.4	12.2	12.6	12.5
s.d.	1.6	1.7	2.1	1.9

The mean difference of the numbers of heads between the two rows is very close to zero (its expected value), but the standard deviations of the differences are found to be well below their expected values. This shows that the pupils are sticking too closely to the even split between H and T - the variation due to randomness is under-represented.

4. Comparison of odds

Twelve questions were used, ten of which were common to both studies. The ten questions reported here all took the form of the example shown in Figure 3. Four versions of the questions were used, with a test design which alternated which half of the questions were on which page, and which was accompanied by a diagram.

1. Two boxes have in them some white balls and some black balls.
You must pick a black ball to win a prize.
The boxes are shaken up and you cannot see inside.

Box A has 2 black balls and 1 white ball A ● ● ○

Box B has 3 black balls and 1 white ball B ● ● ● ○

Which box gives a better chance of picking a black ball?
Tick one answer:

Both give the same chance □ A is better for black □

B is better for black □ Don't know □

FIGURE 3
Comparison of Odds question

TABLE 6
Comparison of odds questions details and facilities

Question	(A) B:W	(B) B:W	Correct Answer	Percentage Correct 1986	 1990
1	2:1	3:1	B	72	85
2	5:2	5:3	A	54	77
3	2:2	4:4	=	26	64
5	3:1	6:2	=	7	25
6	10:5	3:1	B	30	54
7	8:1	9:1	B	70	82
8	5:4	5:1	B	52	83
9	2:6	1:3	=	11	33
11	1:1	6:6	=	22	56
12	6:2	3:1	=	6	28

The complete set of questions is summarised in Table 6. In all cases pupils were asked to decide which was better for black. The results in Table 6 show the dramatic improvement over the four year period.

Analysis of the results by version suggests that neither order nor presence of a diagram is a significant factor. This may seem surprising - and may in part be because the actual form of diagram used was not helpful.

It will be noted that Questions 5, 9, and 12 are conceptually really all the same and yet the percentages correct vary substantially. Questions 5 and 12 are close, but Question 9 with the B:W ratios reversed is rather easier. It is just those three questions which the younger children (1986 study) found most difficult, with percentages correct very low indeed.

A "C-score" was calculated for each pupil, being the number of Comparison of Odds questions correctly answered, yielding a score from 0 to 10. The mean C-score overall rose from 3.5 (s.d. = 1.9) in 1986 to 5.9 (s.d. = 2.5) in 1990. The boys slightly outperformed the girls, with means of 3.7 in 1986 and 6.1 in 1990, as compared with the girls' 3.3 and 5.6. Also, there was a clear relationship with Ability, the mean C-score increasing from 2.3 to 4.4 over grades E-A in 1986, and from 3.3 to 8.2 in 1990. The correlation between C-scores in the two studies was +0.41 ($p < 0.001$) which is fairly high when it is realised that for the 1986 study the scores were compressed and no doubt subject to random guessing in some cases.

It is well-known that girls tend to outperform boys at arithmetic at younger ages but that boys catch up and surpass the girls. Table 7 shows the results for the "Comparison of Odds" questions for 1986 and 1990.

TABLE 7

Comparison of odds percentages correct by sex

Question	1986		1990	
	(ages 7-10)		(ages 11-14)	
	Boys	Girls	Boys	Girls
1	71	75	88	85
2	60	50	81	76
3	34	19	68	62
5	10	4	31	20
6	27	35	56	55
7	68	75	83	81
8	58	67	86	81
9	13	11	39	28
11	28	16	59	54
12	8	4	31	24

Whereas the 1986 results are mixed, the 1990 results without exception show that boys have a higher percentage correct (albeit marginally so in some cases!).

The analysis of questions by ability is particularly interesting for the three hardest questions (Questions 5, 9, and 12). These seem to discriminate markedly between the Grade A pupils and the others. An analysis is presented in Table 8. The ten questions have been analysed into one variable or two variable types (v1 or v2) depending on whether or not both the Black and White numbers vary. Also, they have been allocated to one of three difficulty levels depending on whether two corresponding numbers (i.e. both of one colour or both in one bag) are the same (d1), or differ but the numbers are small and the ratios therefore simple (d2), or differ and the ratios are difficult (d3). The results are included in Table 8 which lists the questions in order of increasing difficulty as judged by the overall facilities. It is clear that it is in the most difficult questions where the most able make greatest relative progress - whereas the least able make no improvement at all. Indeed, their performance may even decline, suggesting that whereas in their earlier years they have "no idea", later on they have "the wrong idea"! What might be called a false intuition (Fischbein, 1989) is beginning to develop. It is also apparent that problems which have *equality* of the ratios are more difficult than would seem "logical" (compare Questions 5 and 6 for example).

TABLE 8
Comparison of odds questions improvement in facility for grade E and grade A ability pupils : questions ranked in order of difficulty

			% Correct		Improvement	
Question	Type	Answer	1986	1990	Grade E	Grade A
1	v1 d1	B	72	85	+17	+14
8	v1 d1	B	52	83	+9	+15
7	v1 d1	B	70	82	+26	+14
2	v1 d2	A	54	77	+17	+10
3	v2 d1	=	26	64	+10	+33
11	v2 d1	=	22	56	+10	+43
6	v2 d3	B	30	54	+28	+14
9	v2 d2	=	11	33	+12	+50
12	v2 d2	=	6	28	-4	+58
5	v2 d2	=	7	25	-7	+54

5. Conclusion

This study has shown that English school pupils display differing developmental patterns in a test of randomness and in a test of comparison of odds. In simulating coin tossing, pupils are highly accurate in reflecting equal probability of Heads and Tails and produce sequences whose first and second halves are highly consistent. Indeed, they are *too* consistent to reflect random variation! Pupils, like adults, have too many too short runs. In the four year period of this study little change arose in these traits. Sex and ability were of little consequence. By contrast, in the Comparison of Odds questions, dramatic improvement occurs for all ability groups, and both sexes. Ability and, to a lesser extent, sex, are significant factors.

It seems natural to believe that the two types of questions discussed here are both legitimate topics for research into probability concepts. Clearly, they are not closely related to each other, however. Comparison of odds relying on the ratio concept is the subject of fairly explicit teaching in schools whereas the study of random sequences is not. Would matters be different if it were? Would there be found the dramatic range of performance linked to general reasoning ability as is evident with comparison of odds? The writer's very limited investigation into whether experience of a computer simulation game is effective (Green, 1990b) suggests otherwise. Clearly this is an area worthy of a fuller programme of research.

References

Fischbein, E (1989) *Intuition in Science and Mathematics.* D Reidel.

Green, D R (1986) Children's understanding of randomness. In: R Davidson and J Swift (eds) *Proceedings of the Second International Conference on Teaching Statistics.* Victoria, British Columbia.

Green, D R (1989) School pupils' understanding of randomness. In: R Morris (ed) *Studies in Mathematics Education* (Vol 7), Chapter 2, pp.27-39. UNESCO.

Green, D R (1990a) *Probability Tests Report.* Mathematical Sciences Report No A132). Available from the author at: Department of Mathematical Sciences, University, Loughborough, LE11, 3TU, UK. Price £4.00 (overseas £5.00 or US$10.00) including postage.

Green, D R (1990b) Using computer simulation to develop statistical concepts. *Teaching Mathematics and Its Applications* 9(2), 58-62.