ASSESSING EARLY MATHEMATICAL DEVELOPMENT

Ray Godfrey and Carol Aubrey Canterbury

Christ Church University College

The work discussed here is part of an international study involving Dutch, Belgian, German, Greek, Finnish and Slovenian as well as English children. The project is co-ordinated by the University of Utrecht and employs the Utrecht Early Mathematical Competence Test. In the long term our interest lies equally in critiquing the test and the Dutch methodology and in drawing what conclusions may be drawn from the data. This paper looks only at a few aspects of the analysis which has been carried out in the few weeks since the full British data set has been available.

Background

Van de Rijt and Van Luit (1998) briefly presented the results of the first phase of an international study employing the Utrecht Early Mathematical Competence Test. Most of the paper was concerned with the technical merits of the test, but the following substantive comparison was made.

COUNTRY	MEAN SCORE	STANDARD DEVIATION
NETHERLANDS	26.2	7.7
BELGIUM	21.5	7.4
GERMANY	23.5	7.0
GREECE	20.0	7.4
UK	15.9	6.4
SLOVENIA	20.4	-

Table 1

It is absolutely clear that in fact the methodology of the study does not make such a crude comparison at all meaningful. However, when the full results are available later this year, comparisons resembling this one may well attract public attention.

The Test exists in three Forms. Forms A and B are discrete. Form C consists of 50% of the questions in A and 50% of those in B. Each Form consists of 40 questions, 5 on each of the topics listed in table 2.

In each country a sample of about 300 children at the beginning of schooling took the same Form of the Test three times over the course of a year. (In the case of the

UK sample these were roughly February-April 1998, June-July 1998 and February-March 1999.) The comparison shown above refers only to the first of these occasions. It takes no account of the fact that many of the UK children were under 5.

1	Comparison	including number and size words
2	Classification	including multiple criteria
3	Correspondence	one-to-one
4	Seriation	dealing with discrete and ordered entities
5	Counting Words	using number words flexibly and in sequence: 0 - 20 forwards and backwards
6	Structured Counting	allowing touch counting in varieties of arrangements
7	Resultative counting	excluding touch counting
8	General Number Knowledge	and 'problems'

Table 2

This paper presents part of our initial analysis of the British data on three levels: looking at the total score for each pupil in each; within these totals looking at the scores for each of the eight topics; looking at responses to individual questions. We have not yet carried out a reliability analysis. At total and subtotal level the data are currently taken at face value. The analysis is far from complete and is presented in order to draw suggestions and advice to assist in further analysis.

Total Scores

All the schools used were in Kent, a large authority with a wide range of schools, urban and rural, with roles varying from under 30 to around 600 and large differences in proportions of pupils eligible for free school meals or registered as having special educational needs. Efforts were made to ensure that the sample of schools was representative of the county in terms of affluence and academic achievement. A typical sample size within each school was 10 pupils. However, some schools were too small to provide 10 new pupils and some large schools provided 10 from each reception class. The total scores have been analysed through least squares linear regression using ML Win to look at differences between areas of the county, schools and classes simultaneously with those between individuals. At no stage in the analysis has the variation between parts of the county or between classes been worthy of note.

It is, of course, hardly worth considering the scores without taking age into account. Figure 1 shows a predictable pattern in the plot of total scores against age and suggests that the differences between scores in the different testing cycles might be simply part of the pattern of age dependency. Different versions of the regression model agreed that scores increased by about 0.92 with every month of age and passed through a level of 21.0 at the mean age of 5 years 5 months. However a rather more effective model (reduction in deviance 39, tested as χ^2 with $p < .0001$) suggests a central value of 20.3 and a monthly increase of 0.50 but with a drop down of 1.5 in the first cycle and a jump up of 3.8 in the third cycle. The same effectiveness could not be produced by making the effects of age non-linear. Possible interpretations of this include maturation through qualitative shifts, the effects of the school curriculum and a familiarity effect or retesting.

It is worth noting that no differences were found between boys and girls in terms of the level of scores, which is consistent with the findings of Van de Rijt and Van Luit; but there are some hints that boys are more variable (or less predictable) than girls.

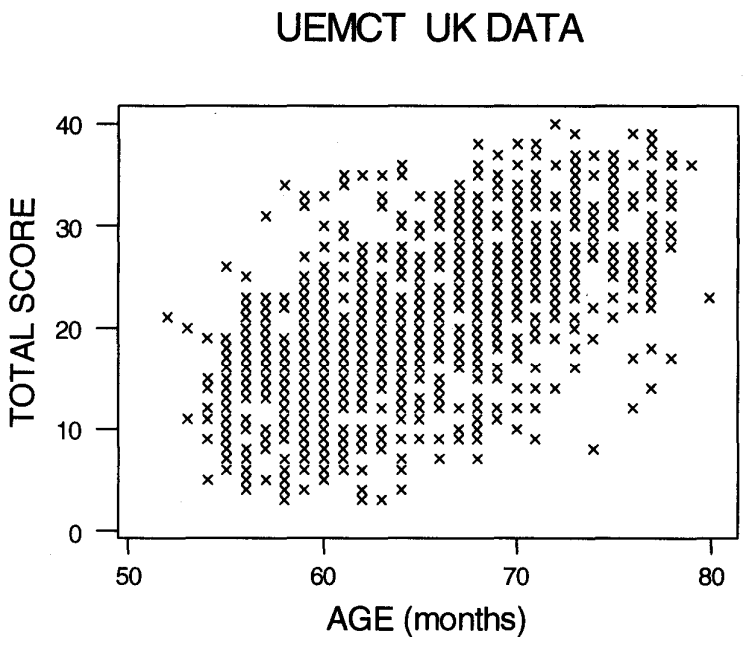

Figure 1

Topic scores

Scores on the eight topics for the most part show little difference between testing cycles 1 and 2 but a large difference between cycles 2 and 3. Figure 2 shows the example of the Correspondence topic. There is scarcely any difference between boys' and girls' scores in each cycle.

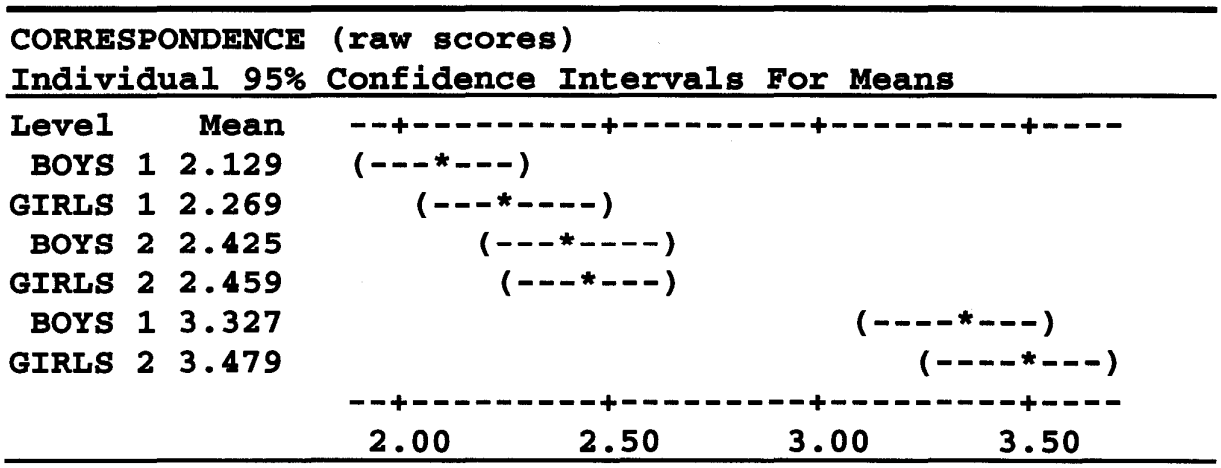

Figure 2

Analysis of variance for each of the topics, after allowing for Age as a covariant, consistently showed significant effects of Cycle in most cases, but no very impressive effects of sex (see Table 3). The most marked Cycle effect is that for Counting Words which is illustrated in Figure 3.

TOPIC	CYCLE EFFECT		SEX EFFECT	
	F	p	F	p
COMPARISON	15.33	0.000	2.04	0.153
CLASSIFICATION	1.14	0.320	0.11	0.742
CORRESPONDENCE	1.22	0.294	2.57	0.109
SERIATION	2.93	0.054	0.03	0.859
COUNTING WORDS	36.16	0.000	0.19	0.660
STRUCTURED COUNTING	15.73	0.000	1.63	0.202
RESULTATIVE COUNTING	3.98	0.019	0.12	0.728
GENERAL NUMBER KNOWLEDGE	7.90	0.000	0.27	0.601

Table 3

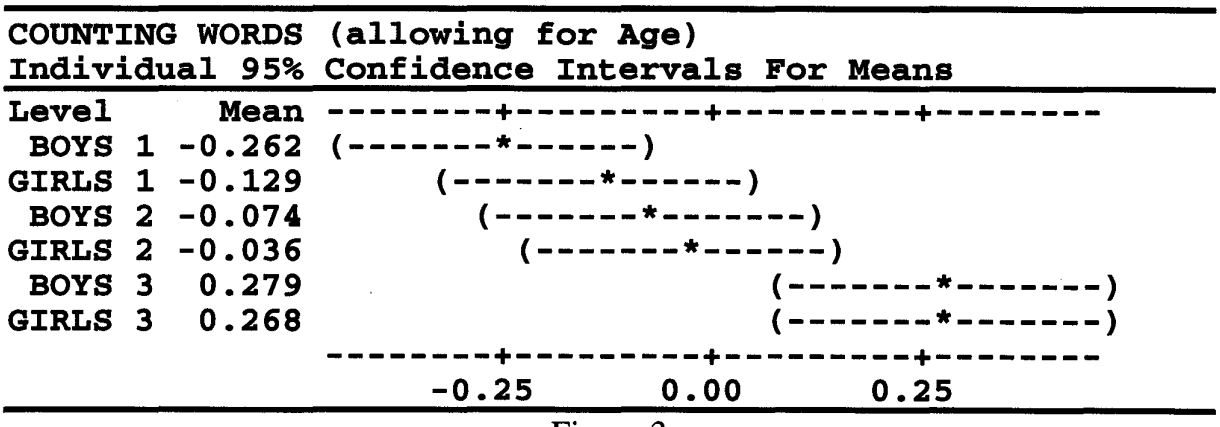

Figure 3

The appropriateness of the regression and ANOV A modelling described above is supported by inspection of the residuals in each case.

Individual Items

As a preliminary step towards analysing individual items, the facilities of each question in each cycle of testing have been plotted along with facility levels deduced from information given by Van de Rijt and Van Luit about their pilot sample. Figure 4 shows the graph for Classification Questions.

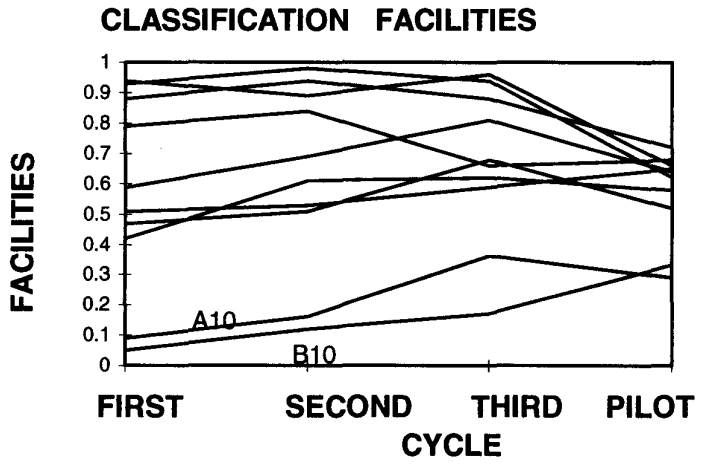

Figure 4

The two questions which have far lower facilities than others are AIO and BIO. AIO is based on a copy of the picture shown in Figure 5. The prompt is: *Here you see an apple with a stalk, without a leaf and a little worm coming out of the apple. Point out all the apples which are exactly the same.* The difficulty here seems likely to be linguistic rather than mathematical, insofar as such a distinction can be made.

Other details which look interesting include the complexity of the changes in facility of Comparison items involving words such as "fatter" and "thicker" and the difficulty of specific items

Facilities for counting questions are more evenly spread. Unsurprisingly, much seems to depend on the actual numbers involved in the questions.

Performance in some questions declines with time. Informal observations by testers suggest that this is because, with familiarity, some children were not listening carefully to questions.

Under Seriation questions about ranking or ordering objects in pictures proved especially difficult. Again it is attractive to see much of this difficulty as due to linguistic complexity rather than mathematical understanding.

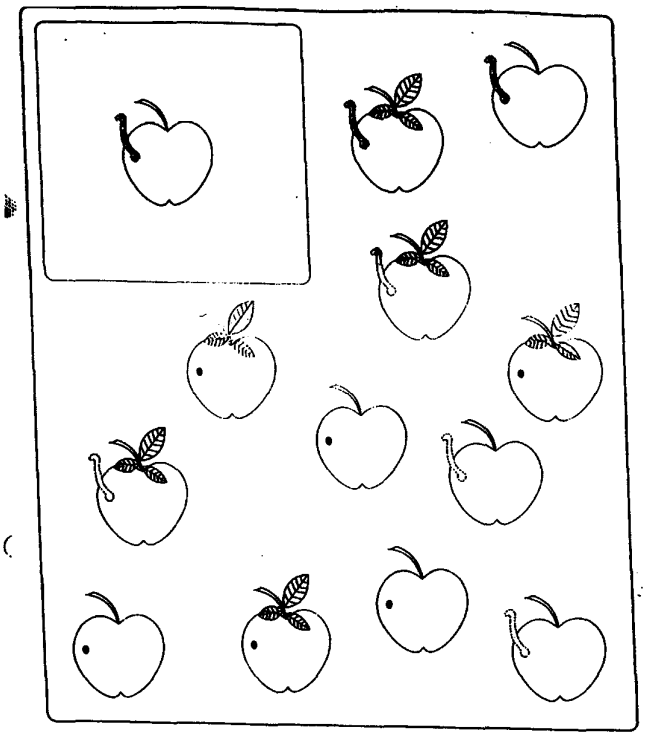

Figure 5

There is much further analysis to do. It is necessary to examine how far the differences between schools' test scores can be explained in terms of other numerical data collected on the schools or in terms of qualitative comments from the schools at the time of feedback to them.

Scores for individual pupils need to be compared with schools' Baseline Assessments and, eventually, with their scores in National Assessments.

In August the full international data set should become available$_1$ allowing more fruitful comparisons between the countries involved.

Bibliography

Van de Rijt, B.A.M. and Van Luit J.E.H (1998) "Development of Early Numeracy in Europe" presented at EeER 98 Ljubljana, Slovenia