# MATHEMATICS ASSESSMENT FOR LEARNING AND TEACHING: AN OVERVIEW OF THE AGE STANDARDISATION MODEL AGES 5-14

Julian Williams, Lawrence Wo and Sarah Lewis

School of Education, University of Manchester

*The MaLT project, by the University of Manchester with Hodder Murray, researched and developed a new set of standardised diagnostic assessment. Data was collected in February and March 2005 from a nationally representative sample of 12591 pupils aged 5–14 from 111 schools. The tests for each year group were vertically equated using Rasch methodology and the scores were then age-standardised and matched against National Curriculum levels using teacher assessment. Single year and three year sub-sample scores enabled comparisons to be made between maturation (within-year) and year-group (between year) effects: progress appears to decline with age, and plateau at Key Stage 3, and particular features at year 2 and year 6 are notable. We suggest these findings have implications for policy.*

## BACKGROUND AND CONTEXT

In a continuing effort (see Williams & Ryan, 2000) to develop diagnostic assessment as an important part of formative assessment while continuing to provide measures of performance and progress that serve summative purposes, we developed a new set of age-standardised diagnostic assessment materials for ages 5-14 (Mathematics Assessment for Learning and Teaching, with Hodder Murray). In this paper we report the age-standardised profiles age 5-14.

Children's cognitive capabilities develop rapidly during their early years in primary school and these individual differences due to age make comparisons between pupils' performance difficult. The majority of previous research has found that birth-date effect exists for younger pupils and some research has also found that this continues through secondary school (Bell and Daniels, 1990). Pupils' mathematics performance relative to national norms can however, be determined through comparisons between individuals' test scores and national test score norms, taking account of age using of age-standardised scores. Previous work (e.g. Schagen, 1990) has combined age-standardisation and vertical equation to provide age standardisation throughout the range 5-14.

Our study has followed this methodology, using a large 2005 dataset, and reports new findings (i) on the plateau of progress at Key Stage 3, and (ii) on the difference between the effect of maturation and year-group, which might be accounted for by differential curriculum exposure. Our analysis reveals some interesting anomalies with implications for future research and policy.

## METHODOLOGY: SAMPLE, EQUATING AND AGE-STANDARDISATION

Schools were invited to participate in the project in November and December 2004 and those which expressed interest identified classes which could take part. A pre-test for this project was carried out in April-May 2004 in order to ensure that the mathematical items for each paper were appropriate. The responses were analysed and only a few changes were made to the papers for the standardisation procedure.

From these schools, a nationally representative sample was created and stratified by geography, school type/status and pupil attainment for the standardisation procedure in February and March 2005. A total of 12591 pupils aged 5 – 14 (minimum 1015 and maximum 1390 per school year group) from 120 schools in England and Wales participated in the project (see table 1).

| Year | R | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------|------|------|------|------|------|------|------|------|------|------|
| Male | 496 | 669 | 752 | 710 | 666 | 685 | 671 | 599 | 552 | 590 |
| Female | 513 | 680 | 634 | 645 | 572 | 614 | 641 | 617 | 608 | 645 |
| not known | 6 | 8 | 4 | 3 | 0 | 1 | 0 | 6 | 1 | 3 |
| Total | 1015 | 1357 | 1390 | 1358 | 1238 | 1300 | 1312 | 1222 | 1161 | 1238 |

**Table 1: Number of pupils per year group in the sample**

The test items were assembled into papers (30 mark points for Reception to Year 2 and 45 mark points for Years 3 to 9) with no common items between papers. Test equating data was achieved by some pupils sitting two standardisation papers, and by data from the pre-test in which pupils sat approximately half a paper of the following year, thus:

| | R/Y1 | Y1/2 | Y2/3 | Y3/4 | Y4/5 | Y5/6 | Y6/7 | Y7/8 | Y8/9 |
|------|------|------|------|------|------|------|------|------|------|
| **Pre-test** | 274 | 355 | 209 | 290 | 289 | 326 | 236 | 239 | 390[1] |
| **Standardising** | 87 | 74 | 79 | 118 | 63 | 31 | 144 | 100 | 75 |

**Table 2: Number of pupils providing anchoring**

Validation included analyses at the pre-test and main test stages (including subgroup DIF etc) that led us to believe that the construct, scale and the vertical equation was safe (see Ryan and Williams, 2005). However, it is important to be aware that vertical equation becomes progressively relatively invalid as the curriculum changes: we view the comparison of 3 years of data as reasonably safe.

Age-standardised scores for each paper were calculated using (up to) three years' item-level data, i.e. the target year and the years above and below (e.g. in calculating age-standardised scores for Year 6, the dataset also included Years 5 and 7). Pupil abilities were calculated from the Rasch analysis, and by using parallel linear regression, we modelled the variation of ability with age (at the month level), at different percentile points. The standardised scores were norm-referenced with average 100 and standard deviation 15 for any age on any test paper. This method for

age standardisation of test score has been used previously on a range of different datasets giving continually reasonable and consistent results (Schagen, 1990).

## RESULTS AND DISCUSSION

Pupils' vertically-equated 'ability' was modelled against age using (i) the 10th percentile, lower quartile, median, upper quartile, and 90th percentile abilities for each month (shown by points in Fig 2.); (ii) the linear equation for each years data separately (shown by line segments); and (iii) the smoothed quintic curves through the whole equated data set (shown by the curves of best fit). The slopes begin at about 1.5 logits per year in Reception and plateau at about 140 months, with a final slight upturn in year 9 (to be discussed later).
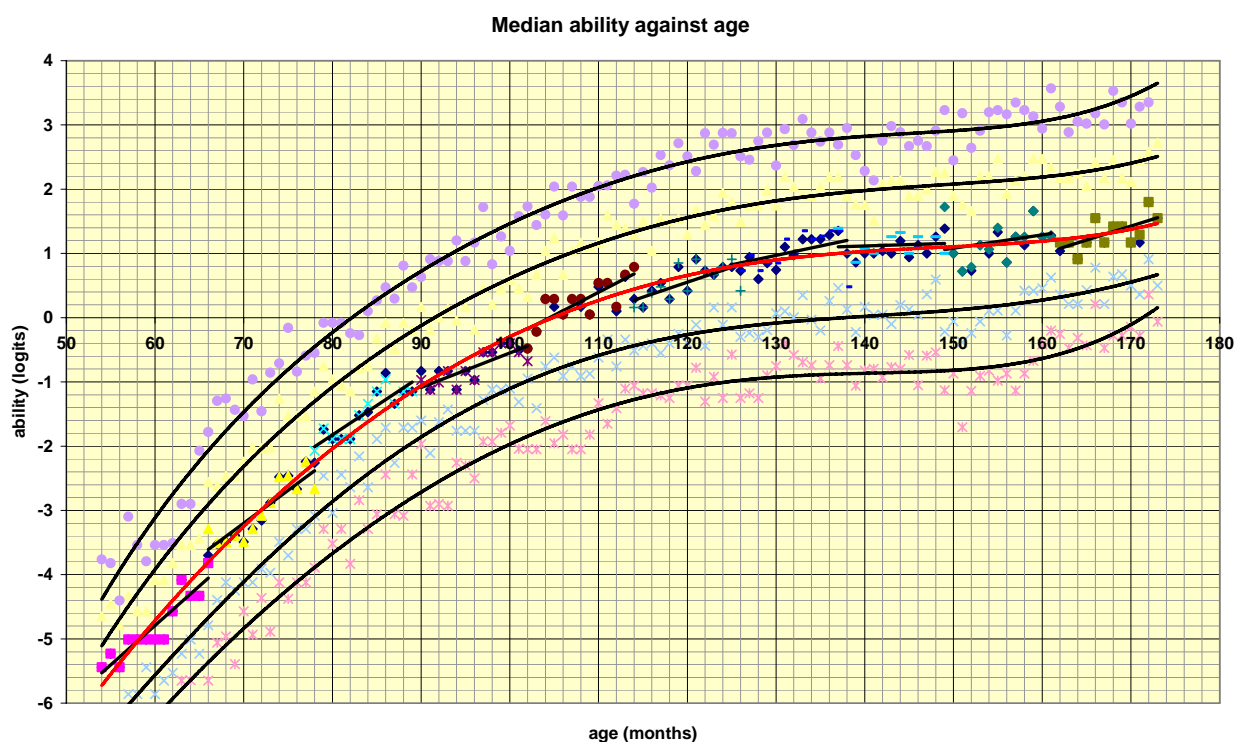


**Figure 2: Graphs of all pupils' 10th percentile, lower quartile, median, upper quartile, and 90th percentile ability (logits) for ages 54 to 173 months**

Inferences from this plateau should be drawn only with due care: maintaining the same logit might actually indicate some progress, if the pupils must learn a lot more curriculum material to achieve the same logit on a higher paper for instance (or vice versa). This would certainly be an issue in equating a performance of, say, 1 logit on the year 3 paper (typical of a 94 month old at the 90th percentile level) with the same logit attained by a median scorer on year 7 at about 145 months old, since the score of one logit, although vertically equated through performances on the intervening, equated papers at years 4, 5, 6, might actually represent a broader competence at the older paper which assesses a broader range of competence and knowledge. The point is that equating of groups' scores is always done through common items that are performed equally well by both groups, and cannot be based on items that are only

accessible to one group. However, the equating of years 7 and 8 is quite safe in this regard, so the apparent plateau of the graph there really is a valid representation of pupil performance.

Looking more closely at the median abilities (at each age) for years 3, 4 and 5 separately (see the three separate line segments in Fig 3.), and comparing them with the age-standardised line for the 3-year vertically equated set union of these three years (the longer line segment spanning the full age-range) we find discrete 'discontinuities of approximately 0.4 logit and 0.1 logits respectively at the junctions at age 102 months and 114 months respectively (the next discontinuity at age 126 months is back up to 0.6 logits).
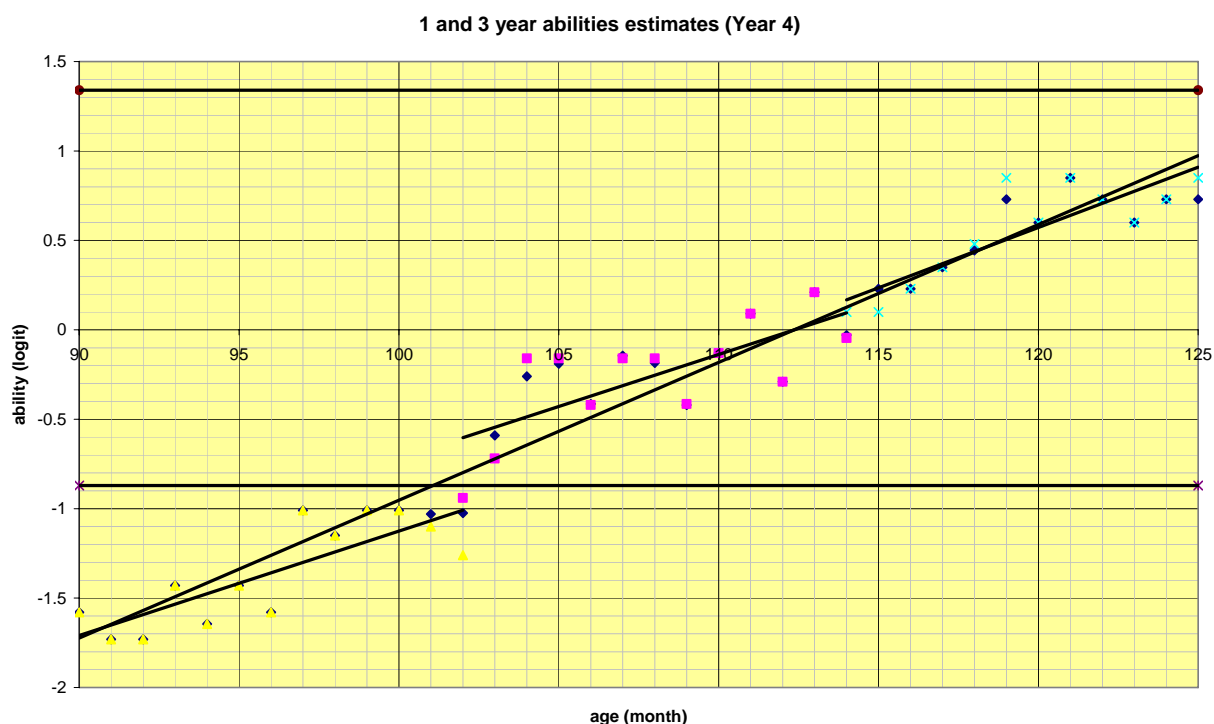


**Figure 3: Graph of the median pupils' abilities (logits) and age in years 3, 4 and 5 (separately, and collectively)**

We interpret the jump between one year's median ability and the next as the difference between the median score for children of the same age in adjacent year groups as a 'year-group effect', i.e. as related to the fact that the child in the older year groups has been exposed to a different history of curriculum and instruction (and perhaps significantly a different relation with a different peer group).

The positive year-group effect can also be clearly be seen for the Years 2 and 6, which are also the two years in which National Tests (SATs) take place and is demonstrated in the two graphs below (figures 4 and 5).
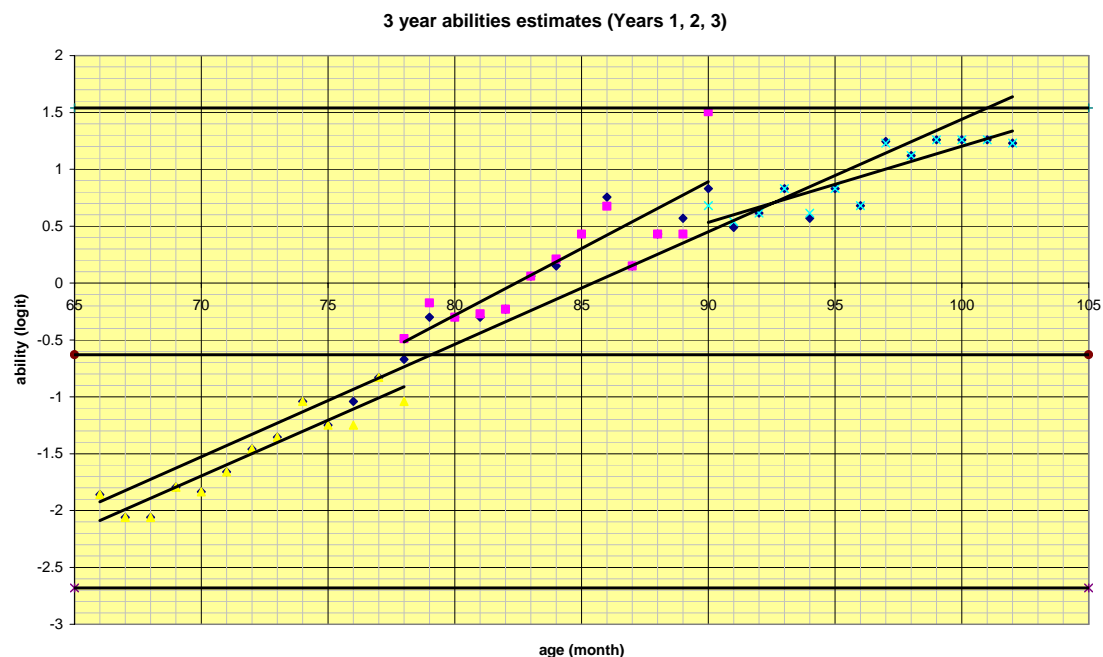
**3 year abilities estimates (Years 1, 2, 3)**



**Figure 4: Graph of pupils' abilities (logits) and age in years 1, 2 and 3**

Further, the findings for pupils' abilities in Years 2 and 6 also suggest the effect of experiencing preparation for the National tests (SATs), giving rise to an increase in abilities in these two years and a clear drop in abilities scores in years 3 and 7. The progress graphs (Fig 2.) curvature at year 9 also may be considered to add weight to this interpretation.
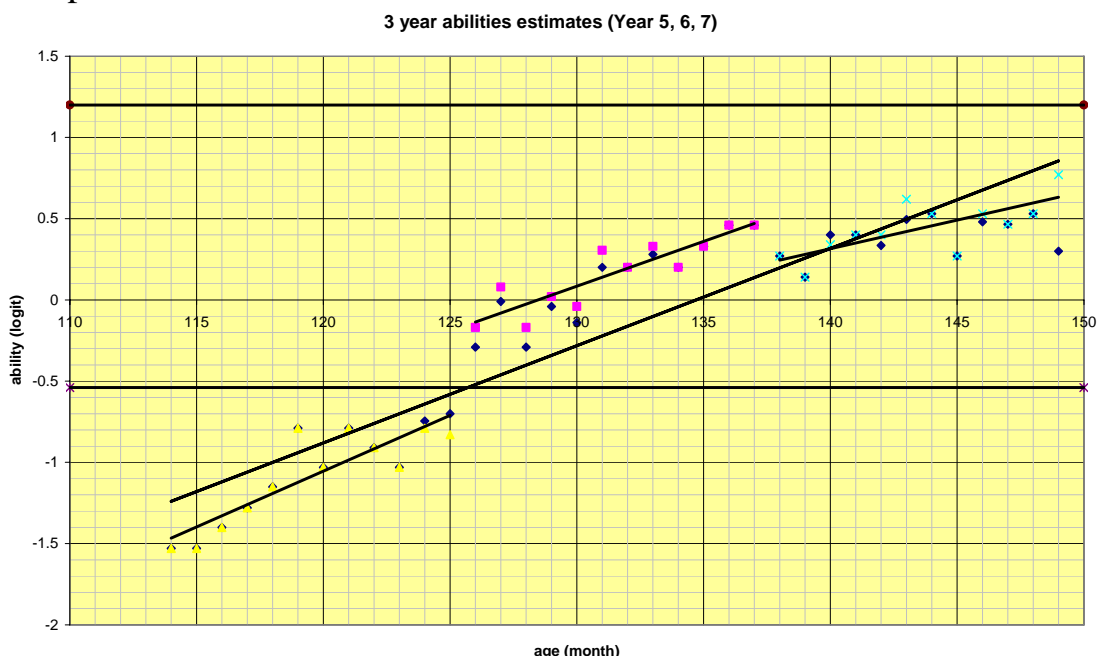
**3 year abilities estimates (Year 5, 6, 7)**



**Figure 5: Graph of pupils' abilities (logits) and age in years 5, 6 and 7**

These results are also suggestive for value-added researchers. If this is a general effect then differences in 'value-added' within school years and across school years should be interpreted separately and with caution. Calculating age-standardised value-added from a table based on the three-year graph in Figure 6 would be likely to

severely under-estimate the actual progress made by a Reception child during a period in their Reception year. Meta-analysis of value added also may be threatened by such an effect.

Figure 6 highlights the Reception year as one with an especially severe year-group effect. The interpretation might be (i) connected with the varied effect of diverse school starting times of Reception/year one children, and (ii) the effect of early years' induction into the social practices of teacher assessment.
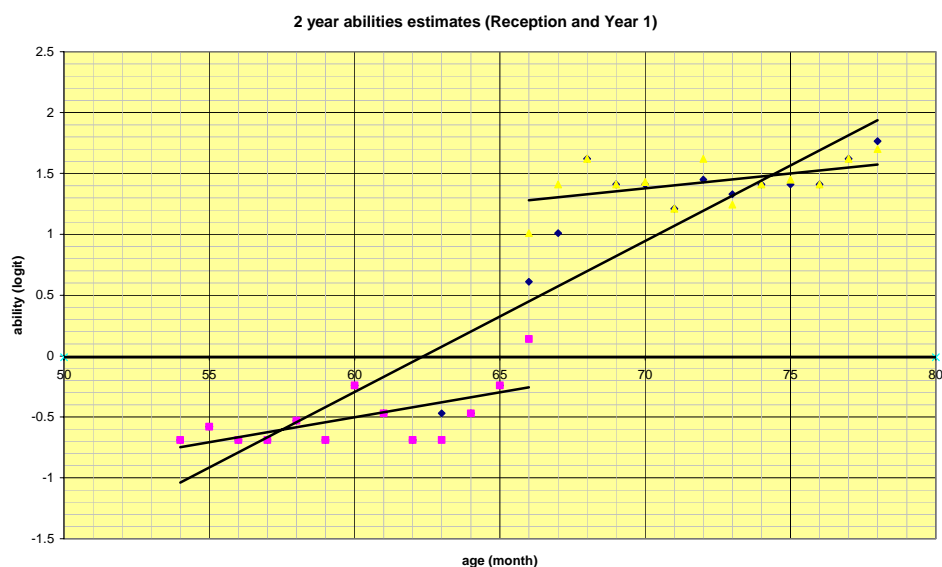


**Figure 6: Graph of pupil abilities and age in Reception and Year 1**

## CONCLUSIONS

We have begun to describe and account for children's progress at various ages 5-15 using vertically equated, age-standardised tests of performance in mathematics: the plateau, the curriculum-maturation distinction, and the year 2/6/9 effect all have policy implications.

## NOTES

[1] Year 8/9 includes those who sat the Y9 anchor paper which consisted of the harder Y8 items

## REFERENCES

Bell, J.F. and Daniels, S. (1990) Are Summer-born Children Disadvantaged? The Birthdate Effect in Education, *Oxford Review of Education 16 (1)*, 67-80.

Schagen, I.P. (1990) A method for the age standardisation of test scores, *Applied Psychological Measurement 14 (4)*, 387-393.

Williams, J. S. and Ryan, J. T. (2000). National testing and the improvement of classroom teaching: can they coexist? *British Educational Research Journal 26 (1)*, 49-73.