Leading education and social research
Institute of Education
University of London

Millennium Cohort Study Data Note 2013/1

Interpreting Test Scores

Roxanne Connelly

September 2013

Centre for Longitudinal Studies
Following lives from birth and through the adult years
www.cls.ioe.ac.uk

CLS is an ESRC Resource Centre based at the Institute of Education, University of London

Millennium Cohort Study Data Note: Interpreting Test Scores

Roxanne Connelly

September 2013

Acknowledgments: The author would like to thank Professor Lucinda Platt for her helpful advice and comments on this data note.

First published in September 2013 by the
Centre for Longitudinal Studies
Institute of Education, University of London
20 Bedford Way
London WC1H 0AL
www.cls.ioe.ac.uk

ISBN 978-1-906929-69-5

The Centre for Longitudinal Studies (CLS) is an ESRC Resource Centre based at the Institution of Education. It provides support and facilities for those using the three internationally-renowned birth cohort studies: the National Child Development Study (1958), the 1970 British Cohort Study and the Millennium Cohort Study (2000). CLS conducts research using the birth cohort study data, with a special interest in family life and parenting, family economics, youth life course transitions and basic skills.

This document is available in alternative formats.
Please contact the Centre for Longitudinal Studies.
tel: +44 (0)20 7612 6875
email: clsfeedback@cls.ioe.ac.uk

Contents

1. Introduction

One particularly valuable element of the Millennium Cohort Study is the standardised tests undertaken by the cohort members to assess their cognitive skills and educational attainment (see for example Kelley *et al.*, 2009; Kiernan and Mensah, 2009; Schoon *et al.*, 2012; Sullivan *et al.*, 2013). The interpretation of the test results may, however, appear complex. The results of each test comprise a sequence of variables deposited with the main MCS datasets for each sweep, and the nature of these variables can be difficult to fully understand without reference to the original test materials and testing procedures. This data note presents an overview of the MCS variables covering performance on standardised tests in sweep two (age 3), sweep three (age 5) and sweep four (age7) of the survey and aims to aid the analyst in successfully utilising these measures in their research. Three suites of tests are described: The British Ability Scales (used at sweeps 2, 3 and 4), The Bracken School Readiness Assessment and the NFER Progress in Maths Test.

2. British Ability Scales II (BAS II)

The British Ability Scales II (BAS II) is a battery of twelve core sub-tests of cognitive ability and educational achievement (Elliott *et al.*, 1996). This battery of sub-tests is suitable for children aged from two years and six months (2:6) to seventeen years and eleven months (17:11). The BAS II has demonstrated construct validity as a measure of cognitive ability (Elliott, 1997; Elliott *et al.*, 1997) and high test-retest reliability (Elliott *et al.*, 1997). The BAS II is also considered to be compatible with current psychological practice (Hill, 2005).

It should be noted that the BAS II comprises assessments of both cognitive ability and educational achievement. The ability sub-tests are designed to measure abilities which are important for learning and educational attainment such as reasoning, perception and memory. The educational achievement sub-tests are designed to provide standardised measures of basic literacy and numeracy skills. The achievement tests are included in the BAS II in order to provide a standardised basis for comparing the educational attainment of a child and their level of cognitive ability. Although a child's performance on the educational achievement sub-tests should correlate with their performance on the cognitive ability sub-tests, the educational achievement sub-tests are not designed as measures of cognitive ability (Elliott *et al.*, 1996).

One of the particularly beneficial features of the BAS II is its flexibility; the core sub-tests of the battery are individually interpretable. In order to assess a child's level of performance the child need not complete all of the tests in the battery (Elliott *et al.*, 1997). The BAS II is therefore particularly suitable for the collection of data within a time-restricted survey setting: the MCS cohort members have only completed four of the BAS II sub-tests. The sub-tests undertaken by the MCS cohort members at sweeps two, three and four of the survey are described in Table 1.

Table 1: The BAS II sub-tests undertaken at MCS2, MCS3 and MCS4.

Sub-Test	MCS2 (Age 3)	MCS3 (Age 5)	MCS4 (Age 7)	Task	Ability/Process
Ability Scales					
Naming Vocabulary	✓	✓		*The child is shown a series of pictures of objects and is asked to name them.*	Expressive Verbal Ability
Pattern Construction		✓	✓	*The child is asked to replicate a design using patterned squares.*	Spatial Problem Solving
Picture Similarities		✓		*The child is shown a row of four pictures and is asked to identify a further congruent picture.*	Non-Verbal Reasoning
Achievement Scales					
Word Reading[1]			✓	*The child is asked to read a series of words presented on a card.*	Educational Knowledge of Reading

Note: Full details of the BAS II sub-tests, their design and their theoretical basis are provided in the BAS II Technical Manual (Elliott et al., 1997).

2.1. Test scores

The cohort member's score on each sub-test completed is presented in three forms: Raw Scores, Ability Scores and Standardised Scores. The variable names for these scores are presented in Table 2.

[1] The parents of children living in Wales were asked to select either an English reading test (BAS II Word Reading) or a Welsh reading test (the '*Our Adventures*' section of the '*All Wales Reading Test*') for their child. The '*Our Adventures*' test required children to identify the correct word to complete a sentence and comprised 58 items. The children completed all 58 items or were stopped after 30 minutes. The '*Our Adventures*' test was undertaken by 139 cohort members. The results for this test are available for each test item (dcoq0100 to dcoq5800) and the analyst must add the number of correct responses to produce an overall raw score. Age-adjusted scores are not provided. The '*Our Adventures*' test scores and the BAS II Word Reading Test are not directly comparable. When using the BAS II Word Reading test results, analysts should bear in mind that a non-random selection of 139 Welsh cohort members did not complete this test.

Table 2: The variable names for the test scores of the BAS II sub-tests undertaken at MCS2, MCS3 and MCS4.

Survey (Survey Number)	Sub-Test	Raw Score	Ability Score	Standardised Score
MCS2 (SN 5350)	Naming Vocabulary	bdbasr00	bdbasa00	bdbast00
MCS3 (SN 5795)	Naming Vocabulary	ccnsco00	cdnvabil	cdnvtscr
	Pattern Construction	cccsco00	cdpcabil	cdpctscr
	Picture Similarities	ccpsco00	cdpsabil	cdpstscr
MCS4 (SN 6411)	Pattern Construction	dctots00	dcpcab00	dcpcts00
	Word Reading	dcwrsc00	dcwrab00	dcwrsd00

2.1.1. Raw Scores

Raw scores are simply the number of correct answers the child gave in each test, however the simplicity of these scores is misleading. In the design of both the MCS and the BAS II great thought was put into the child's experience of the testing process and the time taken to administer the tests. As a result, the cohort members do not all complete the same set of test items. The aim is to present the child with the test items most suitable for their age and ability, excluding items which are likely to be either too easy or too difficult.

The BAS II sub-test items comprise of a number of questions or tasks of increasing difficulty, the children all complete an initial set of items. Based on performance on this initial set of items the interviewer will progress by either: stopping at a predetermined point if the child has scored a sufficient number of correct and incorrect answers to determine their ability, progressing to more difficult items if the child has not found the initial set of items sufficiently challenging, or routing back to more simple items if the child has found the initial set too challenging.

For example, in the Pattern Construction sub-test completed in MCS4 (age seven) there are 23 items in the total test, however the cohort members do not necessarily complete all of these items. Examining the test items completed we can see that item 8 was the simplest question completed by 96% of the cohort members, however 4% of cohort members were routed back to start at item 1. The final item presented to the majority of the cohort members (95%) was item 16, whereas for 4% of the sample it was more appropriate to stop the test earlier, at item 13, and for a small number of cohort members the test was continued up to item 20 or 23.

This testing procedure protects the self-esteem and motivation of the child, and also reduces the overall time required to complete the tests. However, as the cohort members complete sets of items of varied degrees of difficulty their raw scores cannot be directly compared and are not intrinsically meaningful. It cannot be assumed that a one point increase in a cohort member's raw score holds the equivalent meaning across the whole sample. For some cohort members a one point increase in their raw score will refer to the successful

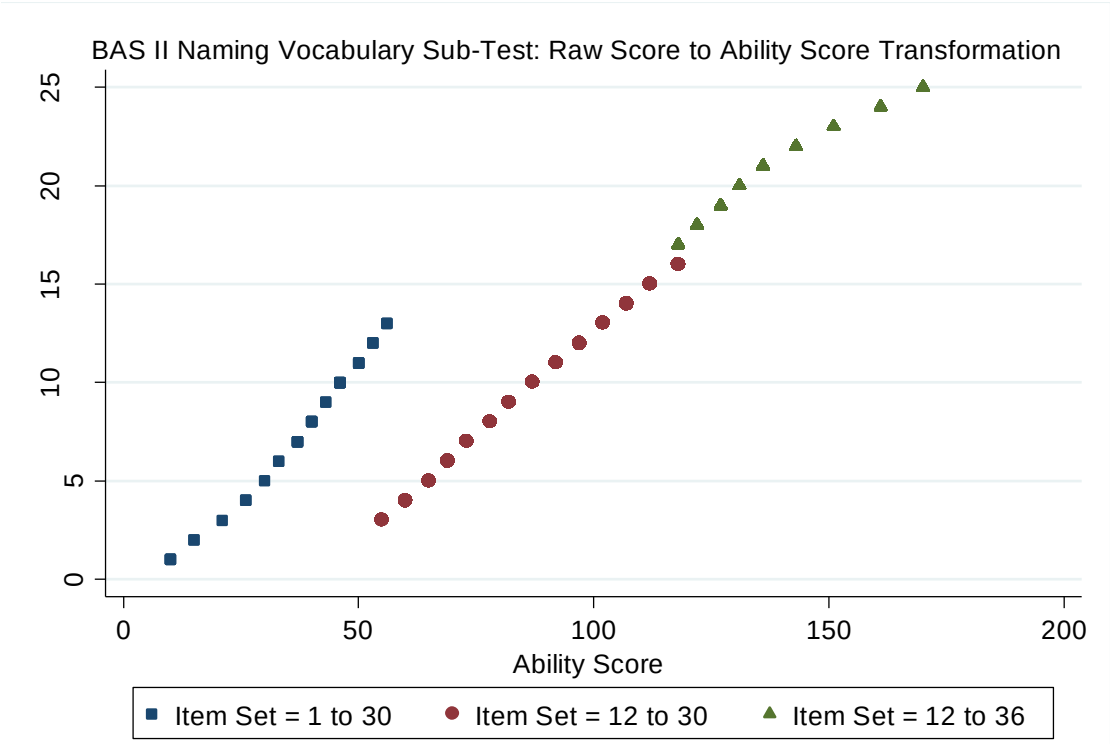
completion of a simple item, yet for other cohort members a one point increase in their raw score will refer to the successful completion of a difficult item.

2.1.2. Ability Scores

To remedy the problem of comparability of test scores across different sets of items, raw scores are converted into ability scores. Ability scores are a transformation of the raw scores, taking into account the specific set of items which the cohort members were presented with. This transformation is made through the use of item response theory, specifically the Rasch model (Rasch, 1960; Rasch, 1961). The Rasch technique defines a theoretical model which describes the probability of successful completion of a set of test items, based on the difficulty of the items and the ability of the child. This model can then be used to predict an individual's ability based on the difficulty of the items which they were able to successfully complete. Look-up tables for the transformation from raw scores to ability scores are provided in the BAS II testing materials (see Elliott *et al.*, 1996).

If we consider the test scores for the Naming Vocabulary sub-test completed in MCS3, we can see that cohort members completed one of three sets of items (see Figure 1). Those cohort members who completed items 1 to 30 (blue squares) completed more simple items than those cohort members who completed items 12 to 30 (red circles). Based on the Rasch model, a child with a raw score of 5 on the item set 1 to 30 is deemed to have an ability score of 30; however a child with a raw score of 5 on the item set 12 to 30 is deemed to have an ability score of 65. The most able children completed items 12 to 36 which provided them with the opportunity to gain a higher score. Although ability test scores overcome the problem of comparing the raw scores of cohort members who have completed different test items, they still have analytical weaknesses. Ability test scores are not based on a truly continuous scale as not all ability test scores can be obtained, and the upper limit of the ability test scores varies between the subtests. Just like raw scores, ability test scores cannot be meaningfully compared between tests.

Figure 1: The relationship between Raw Scores and Ability Scores in the Naming Vocabulary Sub-Test (MCS3).



Note: Millennium Cohort Study Sweep Three (SN 5795)

2.1.3. Standardised Scores

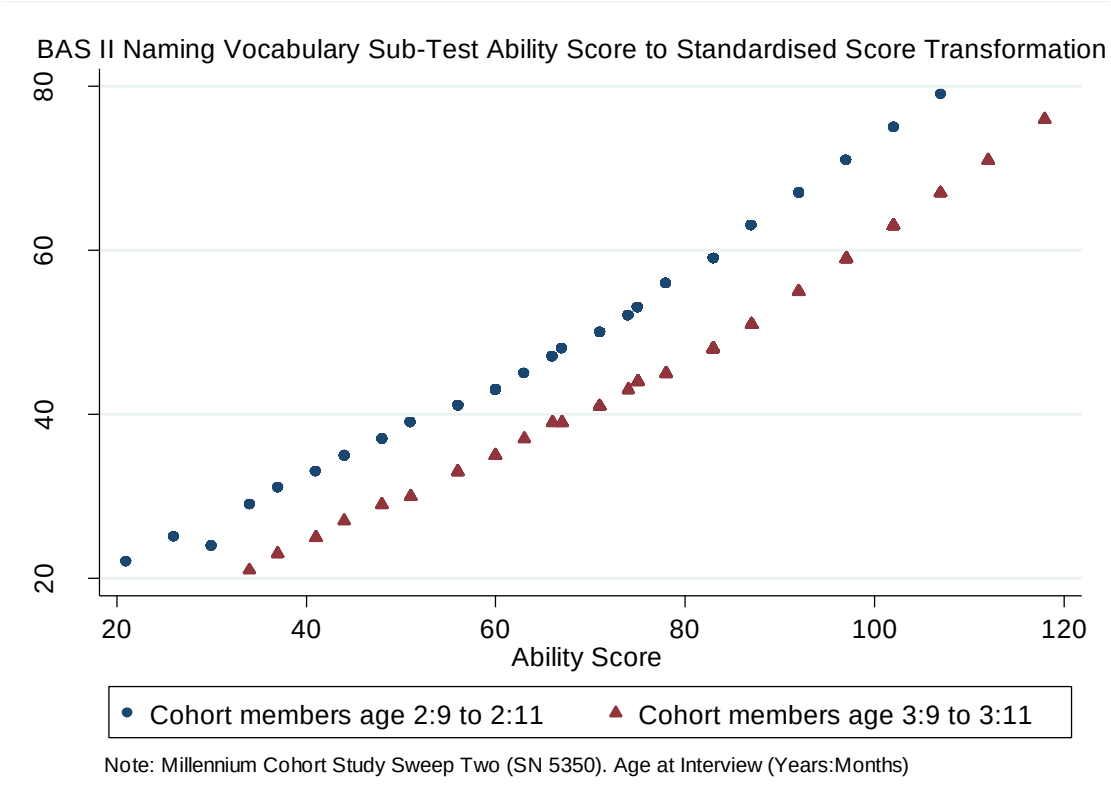
A further weakness of the ability test scores, described above, is that they are not adjusted for age. Based on these non-adjusted scores, older children are likely to gain higher scores due to their more advanced stage of cognitive development and greater educational experience, rather than their ability. In the MCS this is particularly important as the cohort members are born throughout the year and may also be interviewed at different ages. For most analytical purposes, researchers require a test score which is adjusted to take account of how a child is performing on the sub-tests in relation to other children of the same age, standardised test scores provide these age-adjusted measures.

Look-up tables for the conversion of ability test scores to standardised scores are provided in the BAS II manual (Elliott *et al.*, 1996). This transformation is made with reference to a norming sample of 1,689 children. This sample aimed to be representative of the UK national population in 1995; the demographic characteristics of this sample were carefully selected in relation to type of school attended, region of residence, free school meal entitlement, gender, parental education and ethnicity. Norms are provided for children in three month age groups. By comparing a child's performance with the performance of children of similar age, the ability scores of younger children result in higher standardised scores compared to older children who achieved the same ability scores. Figure 2 shows the relationship between the standardised and ability scores of the cohort members in the MCS2 Naming Vocabulary sub-test, we can see that the younger children (blue circles) achieve

higher standardised scores than the older children (red triangles) even when they have the same ability score.

As the standardised scores have been adjusted for both item difficulty and age, we can use these scores to compare the performance of younger and older cohort members on a more level playing field. However, it should be noted that age adjustment is made within three month age bands and there may still be variation in the cognitive development of the cohort members within these age groups. It may be appropriate, therefore, to include the cohort members' age in months in multivariate analyses to provide additional control for age.

Figure 2: The relationship between Ability Scores and Standardised Scores in the Naming Vocabulary Test (MCS3) for older and younger children.



Note: Millennium Cohort Study Sweep Two (SN 5350). Age at Interview (Years:Months)

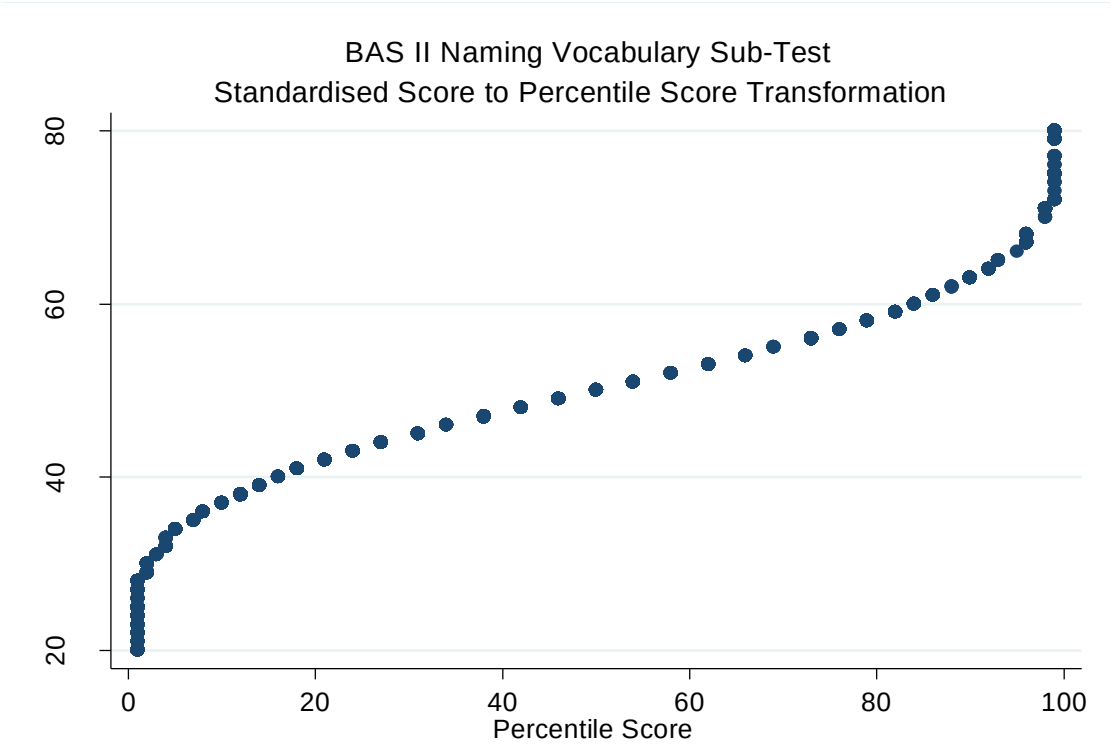
The standardised scores for each of the BAS II ability sub-tests have a mean of 50 and standard deviation of 10, and the scores are bounded between 10 and 80. A child whose standardised ability score is equal to the norming sample will have a score of 50, a child with a score of 40 has an ability score one standard deviation below the mean score of the norming sample, and a child with a score of 60 has an ability score that is one standard deviation above the norming sample. The exception is the BAS II achievement sub-tests, such as the Word Reading test undertaken at MCS4 (age seven): these standardised scores have mean of 100 and a standard deviation of 15. The standardised scores for the achievement scales are provided in this way to facilitate comparison with other achievement test scores, which are generally scaled in this manner (Elliott *et al.*, 1996).

2.1.4. Percentile Scores

In the MCS2 dataset (SN 5350) percentile scores are also provided for performance on the BAS II Naming Vocabulary sub-test (bdbasp00). Percentile scores adjust for the difficulty of the test items and the cohort member's age, in the same manner as standardised scores. The percentile scores indicate the percentage of individuals in the norm group who scored below the level of the cohort member. For example, a cohort member with a standardised score of 50 has a percentile score of 50 as 50% of the norm sample scored below this level. However a cohort member with a standardised score of 71 has a percentile score of 98 as 98% of the norm sample performed below this level.

Percentile scores may seem like an intuitive way to present and analyse the cohort member's performance on the ability tests, however one should note that percentile scores are 'ordinal' rather than 'numerical' measures and therefore percentile units are not constant across the entire distribution of scores. As can be seen in Figure 3, an increase of a few points in standardised scores which fall in the middle of the distribution result in larger changes in percentile scores than an increase of the same magnitude which occurs in the tails of the distribution. The difference between two cohort member's percentile scores will vary depending on where in the distribution their standardised scores fall, which may imply that the same increase in standardised score refers to a larger difference in performance in the middle of the distribution, than in the tails, which is not the case. Therefore, percentile scores can present misleading indications of the difference between the performance of cohort members, or of the changes in cohort members' performance over time.

Figure 3: The relationship between Standardised Scores and Percentile Scores in the Naming Vocabulary Test (MCS2).



Note: Millennium Cohort Study Sweep Two (SN5350).

2.1.5. Age-Equivalent Scores

Although not provided in the MCS datasets, the MCS *'A Guide to the Datasets'* document provides look-up tables to produce age-equivalent scores for the BAS II sub-tests and the Bracken School Readiness test, described below (Hansen, 2012, p. 58). Age-equivalent scores indicate the age at which the cohort member's standardised score is the median standardised score for the norming sample. These scores provide only very rough estimates of the child's ability and should be treated with caution. Elliott *et al.* (1996) state that these scores are most suitable for describing the performance of children with severe learning difficulties, who achieve standardised scores commensurate with a child of a much younger chronological age. In psychological practice age-equivalent scores are generally only used to express a child's level of performance to parents in a simple manner and have major limitations as an analytical measure.

Age-equivalent scores can act to disguise the true nature of a child's performance. Although a child may achieve an age-equivalent score at a level higher or lower than their chronological age, this level of performance may be well within normal bounds of performance for a child of that age (Maloney and Larrivee, 2007). Age-equivalent scores can also encourage typological thinking and the categorisation of "normal" performance, when in practice we would expect a degree of variation in "normal" performance at a given age (Salvia *et al.*, 2006). The use of age-equivalent scores may also lead to extrapolations beyond the age bounds for which the test is developed, and age-equivalent scores may refer to ages outwith the norming sample (Salvia *et al.*, 2006). Furthermore, these scores imply

equivalence in the changes in test performance across the child's lifespan, although we would expect different rates of test score improvement at different stages of the child's development (Maloney *et al.*, 2007). Overall, age-equivalent scores result in the loss of information from the standardised scores on which they are based. A child's level of performance in relation to the overall distribution can be more effectively determined by citing the standard deviation of standardised scores.

2.2. Composite Scores of Ability

With multiple measures of cognitive ability over the course of the survey analysts may seek to reduce this information into a single measure for the purposes of their research. The BAS II subtests were designed to be combined to produce a General Conceptual Ability Score (GCA) through the summation of sub-test scores. If a sub-test score is missing the BAS II manual states that the remaining scores can still be used to produce a GCA score by pro-rating the sub-test scores which are available. However, in the MCS there are only one to two ability sub-tests completed at each survey sweep (see Table 1), therefore the number of standardised test scores is too low to follow the BAS II prescribed method. In order to produce composite scores analysts have generally employed Principal Components Analysis (PCA), this technique can be used to reduce a number of correlated variables into a smaller number of composite variables (i.e. principal components). The cohort members' scores on the standardised ability sub-tests are significantly correlated, hence these variables are suitable for PCA. An example of the use of PCA to derive a single measure can be found in Jones *et al.* (2008, p. 118). Further data reduction techniques could also be utilised in this setting, such as Multivariate Regression (Haase, 2011) or Structural Equation Modelling (Acock, 2013).

3. Bracken School Readiness Assessment - Revised (BSRA-R)

At MCS2 (age three) the cohort members completed the Bracken School Readiness Assessment-Revised (BSRA-R), this assessment is one element of the Bracken Basic Concept Scale-Revised (Bracken, 1998). The BSRA-R is used to assess the 'readiness' of a child for formal education by testing their knowledge and understanding of basic concepts (Bracken, 1998). Basic concepts are defined as aspects of children's knowledge which are taught by parents and pre-school teachers to prepare a child for formal education (e.g. numbers, letters, shapes), and upon which further knowledge builds. The acquisition of basic concept knowledge and skills is important for a child's future educational attainment (Breen, 1985; Sterner and McCallum, 1988; Zucker and Riordan, 1990). The cohort members completed all six sub-tests of the BSRA-R at MCS2 when they were around three years of age. The BSRA-R sub-tests are described in Table 3.

Table 3: The BSRA-R sub-tests undertaken at MCS2.

Sub-Test	Task
Colours	*The child is asked to name basic colours from a picture.*
Letters	*The child is asked to name lower and upper case letters from a picture.*
Numbers/Counting	*The child is asked to name numbers from a picture and assign a number value to a set of objects (involves counting skills and number knowledge).*
Sizes	*The child's knowledge of sizes (e.g. tall, long, big, small, thick) is assessed using a series of pictures.*
Comparisons	*The child's ability to match and differentiate objects is assessed using pictures.*
Shapes	*The child's ability to identify one-dimensional (e.g. curve, angle), two-dimensional (e.g. square, triangle), and three dimensional (e.g. cube, pyramid) shapes is assessed.*

Note: Full details of the BSRA-R sub-tests, their design and their theoretical basis are provided in the BSRA-R examiner's manual (Bracken, 1998).

3.3. Test Scores

The BSRA-R results are presented as series of variables, both for the individual sub-tests and for composite scores for the entire assessment. The variable names for these scores are presented in Table 4.

Table 4: The variable names for the sub-test scores of the BSRA-R (SN 5350).

	Raw Score	% Mastery	Standardised Score	Percentile	Normative Classification
Colours	bdcosc00	bdcmas00			
Letters	bdlesc00	bdlmas00			
Numbers	bdnosc00	bdnmas00			
Sizes	bdszsc00	bdsmas00			
Comparisons	bdcmsc00	bdomas00			
Shapes	bdshsc00	bdhmas00			
Composite Score	bdbsrc00	bdsrcm00	bdsrcs00	bdsrcp00	bdsrcn00

3.3.1. Raw Scores

The BSRA-R raw scores are the number of correct answers the cohort members attained. Unlike the BAS II the BSRA-R does not present different items to the children and therefore item-difficulty adjustment is not required. Raw scores are provided for each individual sub-test, and are added together to provide a composite raw score, known as the School Readiness Composite (Bracken, 1998). All six subtests are designed to measure "readiness" concepts which a child should ordinarily have mastered before they commence formal education. However Bracken (1998, p. 48) notes that this score is not necessarily a complete measure of all the concept knowledge which will be necessary for a child to succeed in school.

3.3.2. Percent Mastery

It should be noted that the BSRA-R sub-tests have different maximum scores (e.g. 10 in the Comparisons sub-test and 20 in the Shapes sub-test).To account for the different maximum possible scores in each test, percent mastery scores are also provided. These scores represent the raw score of the cohort member relative to the maximum possible score in each subtest. Percent mastery scores are provided for each of the six sub-tests and also as a composite score. Percent mastery scores provide the basis upon which performance on each sub-test can be compared.

3.3.3. Standardised Scores

The mean raw scores on the BSRA-R tests increase with the cohort member's age, due to the average accumulated growth in concept knowledge over time. In most circumstances the analyst requires comparable estimates of the cohort member's performance, especially as the MCS children are born throughout the year. The composite raw score is used to produce an age-adjusted standardised composite score. Standardised scores are available for the composite score only and are not available for the individual sub-tests.

Look-up tables for the conversion of raw composite scores to standardised scores are provided in the Bracken examiner's manual (Bracken, 1998). This transformation is made with reference to a norming sample of over 1,100 children between ages two years and six months (2:6) and eight years (8:0). The norming sample was selected to be representative of U.S. population in 1995 based on age, gender, ethnicity, region of residence and parental education. Norms are provided for children in three month age groups. The standardised scores have a mean of 100 and a standard deviation of 15. The standardised scores provide the basis to compare the BSRA-R performance of cohort members of different ages. However, as in the case of the BAS II standardised scores described above, additional accuracy may be gained by also controlling for the cohort members' age in months, as there may be variation within the three-month norming age groups.

Percentile scores are also provided based on the BSRA-R standardised scores, see section 2.1.4. for a discussion of percentile scores. The BSRA-R also includes a 'Normative Classification' variable, this variable places the cohort members into a categorical grouping based on their standardised composite score (see Table 5). Although the categorisation of

the test scores can seem appealing, categorisation is gained at a cost of a loss of the detailed information provided by the standardised scores (see Altman, 2006).

Table 5: The Normative Classification variable of the BSRA-R (bdsrcn00 in SN 5350).

Normative Classification	n (%)	Mean Standardised Composite Score (SD)	Minimum Standardised Composite Score	Maximum Standardised Composite Score
Very Delayed	310 (2)	66 (2)	56	69
Delayed	1,733 (12)	79 (4)	70	84
Average	8,582 (61)	101 (8)	85	115
Advanced	2,784 (20)	122 (4)	116	130
Very Advanced	630 (5)	136 (4)	131	149

4. NFER Progress in Maths Test (MCS Edition)

At MCS4 (age seven) the cohort members completed a shortened version of the National Foundation for Education Research (NFER) standard Progress in Maths (PiM) test. The progress in Maths (PiM) test assesses a child's mathematical skills and knowledge by asking them to complete a series of calculations in a paper and pencil exercise. The test is read aloud to the child and covers topics such as numbers, shapes, measurement and data handling. In order to complete a maths test within the restricted interview time an MCS version of the original PiM test was developed, this version of the test required the cohort members to complete fewer items than the full PiM test involves.

4.4. Test Scores

The cohort member's score on the PiM test is presented in three forms: total raw score (mtotscor in SN6411), scaled raw score (maths7scale) and standardised score (maths7sas). Currently only the total raw score is available in the MCS4 datasets deposited with the UK data service. The scaled raw score and standardised score will be available with the next MCS4 deposit, and are currently available on request[2].

4.5. Raw Scores

The total raw score is simply the number of correct answers the child gave on the test. There are 20 possible test items, although for 5 of the items the child has the possibility of scoring 2 'marks' for a single question. However, as in the case of British Ability Scale scores described above, the simplicity of this score can be misleading. In the MCS version of PiM

[2] Please check the most recently deposited MCS4 data or contact cls@ioe.ac.uk.

test children were routed throughout three possible sets of test items to reduce the time taken to complete the test. All children completed an initial set of items (items 1 to 7) and were then presented with an 'easy', 'medium' or 'hard' set of items depending on their performance on the initial set. The routing system for the PiM test is described in Table 6. As cohort members completed items of varying difficulty, it cannot be assumed that a one point increase in the total raw score will have the same meaning for each child.

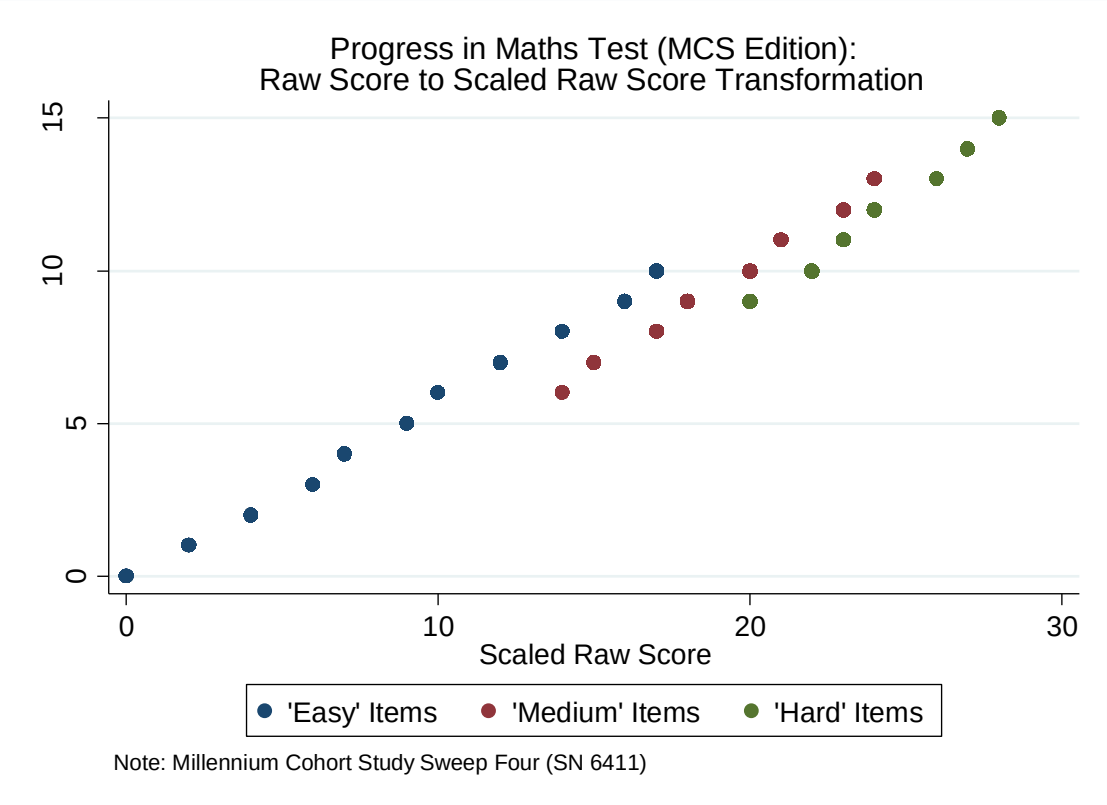
Table 6: The system of routing for the PiM test (MCS Version).

Total Score for Items 1 to 7	Further Questions		% cohort members
0 to 5	8 to 12	(Easy)	34
6 to 8	13 to 16	(Medium)	44
9 to 10	17 to 20	(Hard)	22

4.6. Scaled Raw Scores

To allow for comparability between the test scores of children who completed different items, total raw scores are converted to scaled raw scores. This transformation is made using Item Response Theory, and the Rash model described in section 2.1.2. This transformation takes into account the specific set of items completed and is illustrated in Figure 4. Looking at the cohort members with a total raw score of 10, it can be seen that those cohort members who completed the initial set of items followed by the easiest items (blue circles) are given a lower scaled raw score than cohort members who also had a total raw score of 10 but completed the 'medium' (red circles) or 'difficult' questions (green circles). Although scaled raw scores overcome the problem of comparability between children who completed different test items, these items are not adjusted for the child's age.

Figure 4: The relationship between Total Raw Scores and Scaled Raw Scores in the Progress in Maths Test (MCS Edition).



Note: Millennium Cohort Study Sweep Four (SN 6411)

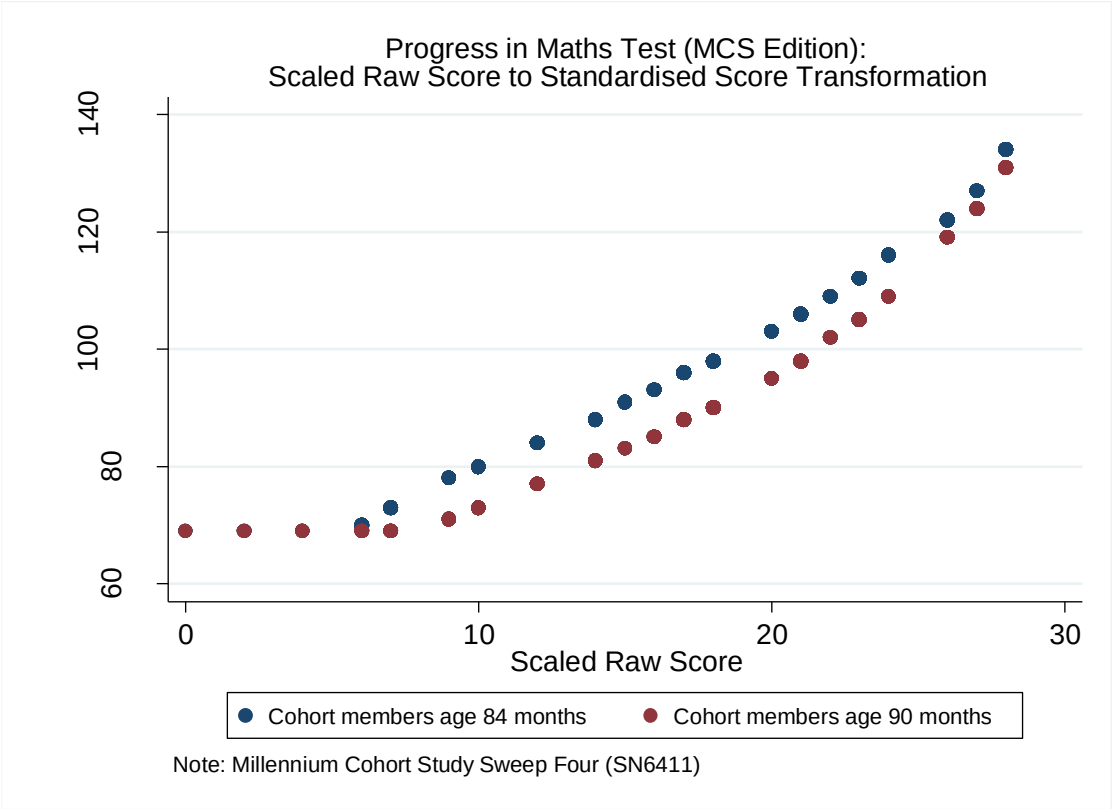
4.7. Standardised Scores

Older children may have more experience with mathematical concepts and tasks and have an unfair advantage in this test, which should be taken into consideration. The full PiM test was standardised with reference to a UK norming sample in 2004, however a norming sample is not available specifically for the shortened MCS version of the test. Standardised scores for the MCS edition of the PiM test are estimated based on the reference sample that completed the full PiM test.

By comparing a child's performance with the performance of this reference sample, the scaled raw scores of younger children result in higher standardised scores compared to older children who achieved the same scaled raw score. Figure 5 illustrates the relationship between the standardised and scaled raw scores, it can be seen that younger children (blue circles) achieve higher standardised scores than older children (red circles) even when they have the same scaled raw score. Standardised PiM scores have been adjusted for both the difficulty of items completed and the child's age in months; these scores can therefore be used to compare the performance of younger and older children[3].

[3] At the lower end of the distribution shown in Figure 5 we can see that the scores seem to reach a plateau. It is common for standardised tests to be truncated at the ends of the distribution in order to avoid making meaningless distinctions between these extreme scores. The British Ability Scale Standardised scores are also truncated.

Figure 5: The relationship between Scaled Raw Scores and Standardised Scores in the Progress in Maths Test (MCS Edition).



Note: Millennium Cohort Study Sweep Four (SN6411)

References

Acock, A. C. (2013). *Discovering Structural Equation Modeling Using Stata*. College Station: Stata Press.

Altman, D. (2006). 'The cost of dichotomising continuous variables'. *British Medical Journal,* 332 (7549)*,* 1080.

Bracken, B. (1998). *Bracken Basic Concept Scale Revised: Examiner's Manual*. London: The Psychological Corporation.

Breen, M. J. (1985). 'Concurrent validity of the Bracken Basic Concept Scale'. *Journal of Psychoeducational Assessment,* 3*,* 37-44.

Elliott, C. D. (1997). 'The Differential Ability Scales'. In D. P. Flanagan, J. L. Genshaft and P. L. Harrison (Eds), *Contemporary Intellectual Assessment: Theories, Tests, Issues* (pp. 183-208). New York: Guilford Press.

Elliott, C. D., Smith, P. and McCulloch, K. (1996). *British Ability Scales Second Edition (BAS II). Administration and Scoring Manual.* London: Nelson.

Elliott, C. D., Smith, P. and McCulloch, K. (1997). *British Ability Scales Second Edition (BAS II). Technical Manual*. London: Nelson.

Haase, R. (2011). *Multivariate General Linear Models*. London: Sage.

Hansen, K. (Ed.) (2012), *Millennium Cohort Study. First, Second, Third and Fourth Surveys. A Guide to the Datasets.* (7th ed.). London: Centre for Longitudinal Studies, Institute of Education.

Hill, V. (2005). 'Through the Past Darkly: A Review of the British Ability Scales Second Edition'. *Child and Adolescent Mental Health,* 10 (2)*,* 87-98.

Jones, E. and Schoon, I. (2008). 'Child Cognition and Behaviour'. In K. Hansen and H. Joshi (Eds), *Millennium Cohort Study. Third Survey: A User's Guide to Initial Findings*. Institute of Education, University of London: Centre for Longitudinal Studies.

Kelley, Y., Sacker, A., Gray, R., Kelly, J., Wolke, D. and Quigley, M. (2009). 'Light drinking in pregnancy, a risk for behavioural problems and cognitive deficits at 3 years of age?'. *International Journal of Epidemiology,* 38 (1)*,* 129-140.

Kiernan, K. and Mensah, F. (2009). 'Poverty, Maternal Depression, Family Status and Children's Cognitive and Behavioural Development in Early Childhood: A Longitudinal Study'. *Journal of Social Policy,* 38*,* 569-588.

Maloney, E. and Larrivee, L. (2007). 'Limitations of Age-Equivalent Scores in Reporting the Results of Norn-Referenced Tests'. *Contemporary Issues in Communication Science and Disorders,* 34*,* 86-93.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). 'On general laws and the meaning of measurement in psychology '. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistica and Probability,* 321-334.

Salvia, J., Ysseldyke, J. and Bolt, S. (2006). *Assessment: In special and inclusive education*. Boston: Houghton Mifflin.

Schoon, I., Jones, E., Cheng, H. and Maughan, B. (2012). 'Family hardship, family instability, and cognitive development'. *Journal of Epidemiology and Community Health,* 66*,* 716-722.

Sterner, A. G. and McCallum, R. S. (1988). 'Relationship of the Gesell Developmental Exam and the Bracken Basic Concept Scale to academic achievement'. *Journal of School Psychology,* 26*,* 297-300.

Sullivan, A., Ketende, S. and Joshi, H. (2013). 'Social Class and Inequalities in Early Cognitive Scores'. *Sociology,* 29*,* Online Preview.

Zucker, S. and Riordan, J. (1990). 'One-year predictive validity of new and revised conceptual measurement'. *Journal of Psychoeducational Assessment,* 8*,* 4-8.

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@ioe.ac.uk
Web www.cls.ioe.ac.uk