Standards & Testing Agency

Reception baseline comparability study

Results of the 2015 study

March 2016

Contents

Table of figures

Introduction

This document provides details of the comparability study on the reception baseline carried out in September to December 2015.

The report is aimed at a technical audience but will potentially also be of interest to those involved in assessment in the early years.

Background

In the Government response to the consultation on primary assessment and accountability in March 2014 (available from www.gov.uk/government/consultations/new-national-curriculum-primary-assessment-and-accountability), the Department for Education announced that it would use a reception baseline assessment as the starting point for measuring progress in primary schools.

The optional reception baseline would eventually become the only way of measuring progress and therefore schools were encouraged, though not mandated, to participate. A number of reception baseline assessments were to be approved to provide school choice and pupils were to be assessed within their first half-term in reception.

Following a procurement process and school recruitment phase (where suppliers had to achieve recruitment volume requirements), there were 3 approved products from which schools could choose, from the following organisations:

- Early Excellence (EE)
- Durham University's Centre for Evaluation and Monitoring (CEM)
- National Foundation for Educational Research (NFER)

The criteria that potential reception baseline assessments were required to meet meant that suppliers were allowed to create products that were of different formats and that measured different constructs. The assessment had to have an age-appropriate content domain that was suitable for the range of pupils' attainment at the start of reception. Further, the clear majority of the content domain was required to be linked to the learning and development requirements of the communication and language, literacy and mathematics areas of learning from the early years foundation stage and demonstrate a clear progression towards the key stage 1 (KS1) national curriculum in English and mathematics.

There have always been concerns that the assessment of 4-year-olds would not provide a sufficiently robust measure on which to base the primary school progress measure. The multiple-supplier approach adds an additional risk that the different assessments may not be comparable. Therefore a comparability study was designed to determine

whether the 3 products were, in fact, sufficiently comparable to be used in the accountability system.

What is meant by comparability?

It is not uncommon to want to make comparisons between assessments. However, the viability of the comparison depends on the nature of the assessments being compared and the way the assessments are linked (Linn, 1993). It is also important to consider the ultimate purpose to which the scores on the 3 assessments will be put (Elliot, 2013).

Linking is the mechanism by which the score comparisons can be made. Linn (1993) identifies 5 forms of linking and 3 of these are further explicated by Dorans (2000) and Holland (2007) and are explicated below.

Equating is the strongest form of linking, and has as its goal that the scores being equated are interchangeable. In order for this to be the case, the assessments to be equated need to be measuring the same thing in the same way for all pupils. In other words, the assessments should be considered parallel forms of each other, having been created from the same specification. It is self-evident that if there are 3 separate specifications equating is not attainable in the context of reception baseline assessments.

Scale aligning or calibration refers to the transformation of 2 different tests onto a common scale. According to Dorans (2000) the main point of scaling is to have the same rank ordering of scores in a group of people. One can scale dissimilar constructs or similar constructs, but the data collection method is important, as is the reliability of the test in question. In the context of reception baseline assessment comparisons, we expect similar levels of reliability and similar item difficulty in the population, leading to concordance. In order for scaling to be viable, the correlation between the 2 measures being compared must be high.

Prediction is the least restrictive and demanding. The point is to predict one score from another. It is not expected that the inverse of the score comparison leads to the same outcome as the original comparison. Prediction methods are not satisfactory ways of creating comparable scores (Holland, 2007).

Since the criteria that the reception baseline assessments needed to meet enables the products to assess different constructs, for this study we have defined comparability in relation to the scaling definition. However, given the way in which results from the reception baseline will be used for accountability which implies a strong degree of interchangeability between the different assessments, we need to meet strict conditions on the relationship between the scores on each assessment. As a result, although high correlations between the rank orders assigned by each assessment, which would be a sufficient condition for prediction, will be necessary, they won’t necessarily be sufficient in this case.

Comparability study design

The data collection method is critical to any linking process. It is important to build sufficient links to ensure there is enough data to make the comparisons required. As part of the study, teachers in selected schools administered 2 reception baseline assessments to their pupils. One assessment was the reception baseline selected by the school and the second was assigned to the school as part of the study.

There were 2 factors considered when designing the data collection method: how to ensure all reception baseline assessments were tested in pairs and to counter-balance the testing such that a school's selected baseline assessment was administered first in some groups, and second in other groups. We were also cognisant that the study had to be manageable and demonstrate value for money. The final design included 12 combinations and a minimum of 300 pupils per combination was required (3,600 pupils in total). The design is presented in appendix 1.

It is also important that the individual assessments were administered and scored correctly and that the assessments themselves were reliable. In order to meet the criteria to become an approved reception baseline, suppliers had to provide evidence in relation to the reliability of their assessments. Separate reliability statistics were not generated as part of this study.

All schools that participated in the study were given the same training for the reception baseline assessment that was assigned to them as for those that chose that reception baseline assessment. For the CEM and NFER reception baseline assessments, this training is in the form of written guidance and there are no concerns that the administration arrangements would be different for schools that chose and schools that were assigned these reception baseline assessments. For EE, 1 day of external training was required. For schools that chose EE as their reception baseline assessment, this training took place in the summer term 2015. However, given the timeline involved in this study, some of the study schools were trained in September 2015, immediately prior to administering the assessment. Although no concerns were raised by schools in the study, the close proximity of the training and the administration may have impacted on the ability of teachers to internalise the training.

The final point to note is that with a comparability study, it is ideal for the assessments to be administered as closely together as possible such that any differences in performance cannot be attributed to learning having taken place by the pupil in between the 2 assessments. The use of observational assessment in the EE reception baseline means that the time between assessments in these studies is likely to be longer than would be ideal for comparisons involving EE. This is likely to have reduced correlations and therefore associated statistics which means that true statistics are likely to have been higher were we able to conduct assessments at exactly the same time.

Sample

A trialling agency, the Scottish Qualifications Authority (SQA), was procured to recruit schools to participate in the study. The recruitment phase took place later than anticipated due to constraints external to SQA. However, they were able to mobilise a flexible recruitment strategy whereby they used a mixture of cold calling and web registration to engage schools at the end of the 2014 to 2015 academic year.

Participation in this trial was voluntary and relied on school goodwill to participate. Only 20% of schools that were written to by SQA chose to participate, which is a little lower than we normally see for this type of trial. The timing of this recruitment, in the last 2 weeks of term, is probably the reason for the difference.

STA provided the sample design as well as 3 samples from which to recruit schools. The 2 stratifiers of interest were KS1 to KS2 progress measure decile (sample split into 10 groups) and region (9 regions). Given the size of the sample and the number of groups per stratifier, it was not possible to have a fully representative sample where schools in each region were represented in each progress decile. However, SQA worked very hard to ensure a spread of region and progress decile across each test pairing.

Due to recruitment constraints it was also necessary to be flexible with some pairings such that a small number of schools were assigned a different baseline than originally planned. Because of how SQA recruited the schools, this change was not apparent to the schools and had no adverse effect on the sample.

Sample tables provided by SQA are presented in appendix 2 and show the achieved sample against the original sample design. In the end 122 schools were recruited (4,690 pupils).

In relation to the 12 combinations, the following numbers of pupils was achieved:

Test combination*	Count of schools	Number of pupils
CEM(1)_EE(2)	11	543
CEM(1)_NFER(2)	11	382
CEM(2)_EE(1)	10	463
CEM(2)_NFER(1)	13	414
EE(1)_CEM(2)	9	320
EE(1)_NER(2)	10	363
EE(2)_CEM(1)	9	332
EE(2)_NFER(1)	10	331
NFER(1)_CEM(2)	11	411
NFER(1)_EE(2)	8	394
NFER(2)_CEM(1)	11	299
NFER(2)_EE(1)	9	438
Min	**8**	**299**
Max	**13**	**543**
Total	**122**	**4,690**

Table 1: Counts of schools and pupil numbers by test combination

Some schools decided to administer more than 1 baseline assessment to their pupils without being recruited to do so by SQA. Data from all schools that had multiple assessments has been included in the analysis below.

This means that of the 5,261 pupils matched in the comparability study: 2,099 pupils were in the EE/CEM combinations; 1,384 pupils were in the NFER/CEM combinations; and 1,722 pupils were in the EE/NFER combinations. There were 56 pupils in 1 school that matched to all 3 assessments.

Method

The data from the 3 baseline assessment suppliers was matched together on the combination of school details and unique pupil number (UPN) and a check for duplicates was undertaken.

Frequencies, summary statistics and correlations between subscale and overall reported scores for all 3 baseline assessments were run on the pupils in the comparability study.

As a means of comparison, frequencies and summary statistics were also run on the overall file which included all pupils who participated in the reception baseline.

Equipercentile linking using LEGS (2004, Linking with Equivalent Groups or Single Group Design, freeware software made available by the Center for Advanced Studies in Measurement and Assessment, University of Iowa) was carried out. Initially this was done on the 3 pairs on the overall reported score.

Further equipercentile linking was undertaken to examine sub-population differences by gender as suggested by Dorans (2000). Differences in linked scores between the 2 genders indicates an absence of population invariance. This would provide further evidence that score comparability is inappropriate.

Dorans (2000) further recommends a content comparison between the 3 assessments as well as measures of reduction in uncertainty based on the Pearson correlations between the 3 assessments.

Assessment structures

Full details of the structure of each of the assessment are provided in appendix 3. A summary is provided below:

- The CEM baseline, BASE, is an on-screen computer adaptive test which focuses on literacy (76% of items approximately) and mathematics (24% of items approximately).
- The EE baseline, EExBA, is an observational assessment where teachers make a series of judgements about each child based on a set of assessment criteria. The judgements relate to communication and language (25.5%), literacy (17%), mathematics (17%), personal social and emotional development (13%), physical development (8.5%) and characteristics of effective learning (19%).
- The NFER baseline is a resource-based assessment with a mixture of tasks and observational checklists which focuses on communication and language (14%), literacy (48%) and numeracy (38%).

All 3 baseline assessment providers include communication, literacy and numeracy in their assessments, as would be expected. However, the proportion that these content areas take up within each assessment is quite different. For example, numeracy is reported to be 38% of the NFER assessment but only 17% of the EE assessment, and approximately 23% for the CEM assessment.

Analysis

Frequencies, summary statistics, correlations

All 3 baseline assessments report 3 scales, literacy, numeracy and an overall reported score.

The CEM literacy and numeracy subscale scores and the overall reported score were standardised using trial data such that it had a mean of 100 and a standard deviation of 15. In the comparability study dataset the scores ranged from 50 to 150.

The EE literacy and numeracy subscale scores range from 1 to 8 and the overall reported score is a raw score that can range from from 0 to 56. In the comparability study dataset the overall reported score ranged from 2-56.

The NFER literacy and numeracy subscale scores and the overall reported score were standardised using trial data such that it had a mean of 100 and a standard deviation of 15. In the comparability study dataset the scores ranged from 69 to 131.

Summary statistics are provided for each sample. These show that there were some minor differences between the samples; however, the internal correlations for each assessment are similar across samples indicating that any sampling issues were not significant.

EE/CEM sample

There were 2,155 matched pupils in the EE/CEM sample. The following tables provide summary statistics and correlations between relevant scales. All correlations are significant at the 0.01 level. The correlation between the CEM reported score and the EE reported score was 0.601 (in bold text below).

	EE literacy	EE numeracy	EE reported score	CEM literacy	CEM numeracy	CEM reported score
N	2,155	2,155	2,155	2,155	2,155	2,155
Mean	1.98	3.26	25.28	92.72	96.90	93.84
Median	2.00	3.00	25.00	92.00	98.00	94.00
Standard deviation	1.55	1.76	11.24	14.55	16.98	15.06

Table 2: Summary statistics for the EE/CEM sample

	EE literacy	EE numeracy	EE reported score	CEM literacy	CEM numeracy	CEM reported score
EE literacy	1	.526**	.685**	.617**	.536**	.621**
EE numeracy		1	.716**	.528**	.606**	.585**
EE score			1	.576**	.574**	**.601****
CEM literacy				1	.773**	.971**
CEM numeracy					1	.897**
CEM score						1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3: Correlation between the EE reported score and the CEM reported score

NFER/CEM sample

There were 1,440 matched pupils in the NFER/CEM sample. The following tables provide summary statistics and correlations between relevant scales. All correlations are significant at the 0.01 level. The correlation between the NFER overall score and the CEM reported score was 0.841 (in bold text below).

	NFER literacy	NFER numeracy	NFER overall score	CEM literacy	CEM numeracy	CEM reported score
N	1,440	1,440	1,440	1,440	1,440	1,440
Mean	92.10	90.91	91.22	93.97	97.33	94.84
Median	91.00	90.00	90.00	92.00	98.00	95.00
Standard deviation	13.65	13.32	13.36	14.82	17.01	15.36

Table 4: Summary statistics for the NFER/CEM sample

	NFER literacy	NFER numeracy	NFER overall score	CEM literacy	CEM numeracy	CEM reported score
NFER literacy	1	.765**	.960**	.779**	.725**	.796**
NFER numeracy		1	.913**	.723**	.812**	.789**
NFER score			1	.802**	.806**	**.841****
CEM literacy				1	.796**	.974**
CEM numeracy					1	.910**
CEM score						1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5: Correlation between the NFER reported score and the CEM reported score

EE/NFER sample

There were 1,778 matched pupils in the EE/NFER sample. . The following tables provide summary statistics and correlations between relevant scales. All correlations are significant at the 0.01 level. The correlation between the EE reported score and the NFER overall Score was 0.735 (in bold text below).

	EE literacy	EE numeracy	EE reported score	NFER literacy	NFER numeracy	NFER reported score
N	1,778	1,778	1,778	1,778	1,778	1,778
Mean	2.10	3.53	26.47	91.98	92.74	91.93
Median	2.00	4.00	26.00	91.000	92.00	91.00
Standard deviation	1.53	1.78	11.00	13.06	14.01	13.19

Table 6: Summary statistics for the EE/NFER sample

	EE literacy	EE numeracy	EE reported score	NFER literacy	NFER numeracy	NFER reported score
EE literacy	1	.574**	.662**	.713**	.558**	.692**
EE numeracy		1	.736**	.674**	.625**	.697**
EE score			1	.752**	.604**	**.735****
NFER literacy				1	.743**	.952**
NFER numeracy					1	.912**
NFER score						1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7: Correlation between the EE reported score and the NFER reported score

Reduction in uncertainty

Dorans (2000) noted that in order to support scaling, the correlations between pairs of assessments must be high. A coefficient of alienation (COA) was suggested in the form of:

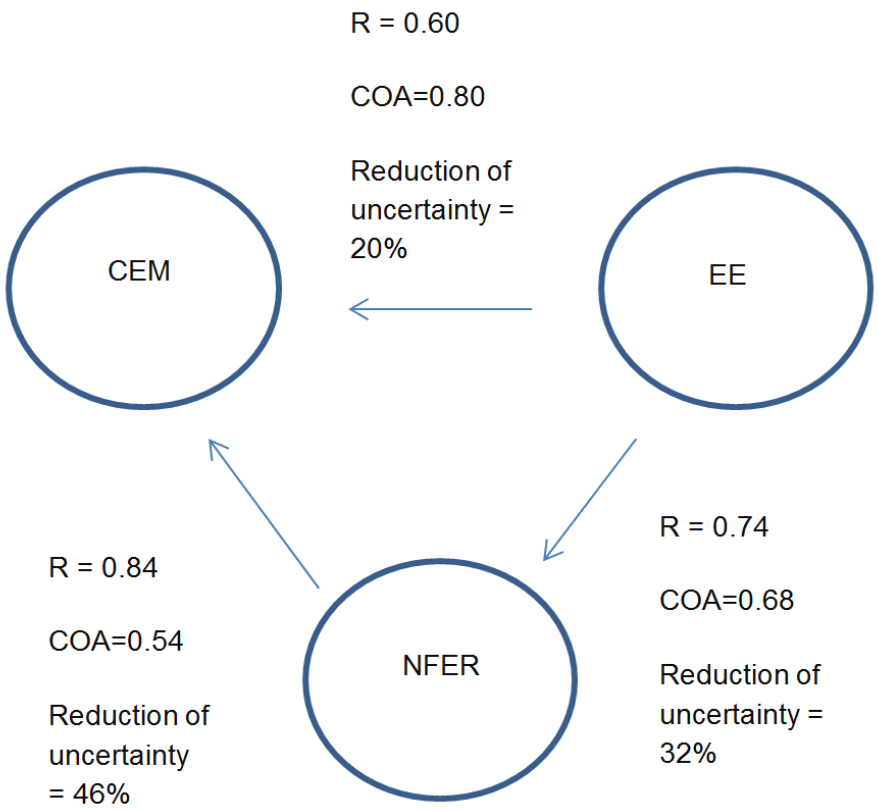
$$\text{COA} = \sqrt{(\mathbf{1} - \mathbf{r^2})} \quad \text{where r = Pearson correlation}$$

Dorans (2000) further suggests a definition of the reduction of uncertainty as:

$$\text{Reduction of uncertainty= 1-COA= } \mathbf{1} - \sqrt{(\mathbf{1} - \mathbf{r^2})}$$

Dorans (2000) recommends that if the reduction of uncertainty is less than 50%, concordance is unacceptable.

As can be seen by the diagram below, the most that uncertainty can be reduced is by 46% in the NFER/CEM pairing. This is not sufficient to suggest that scale alignment is appropriate. The best that could be done is including the reported score in a prediction equation for another reported score, but this does not fit the purpose of the reception baseline assessment comparisons.



Linking results

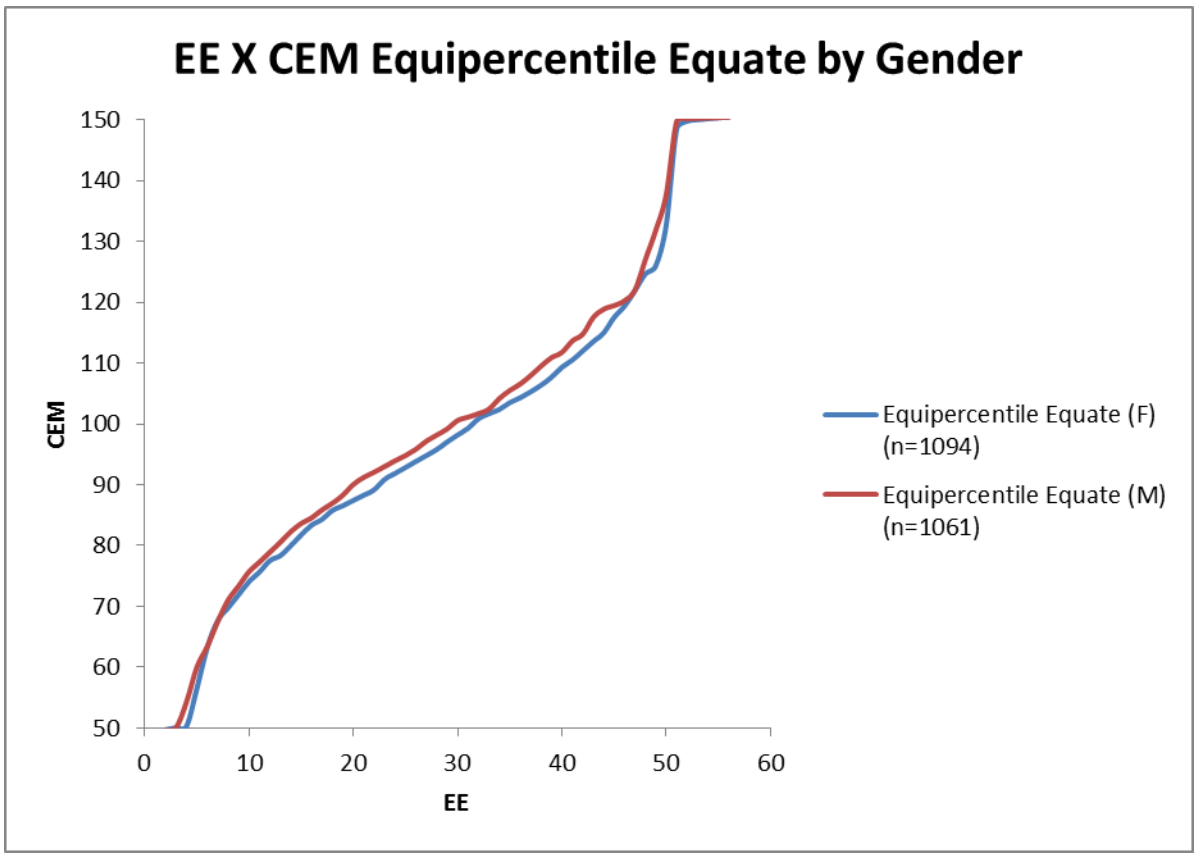
There are two graphs for each baseline assessment pairing presented. The first graph is the equipercentile link comparing girls and boys. Examining linked scores between girls and boys helps determine if there are differences between sub-populations. A standardised mean difference between girls and boys was calculated for each assessment in the pair and will be presented below.

The second graph is the overall equipercentile link with a standard error of the link 68% confidence interval.

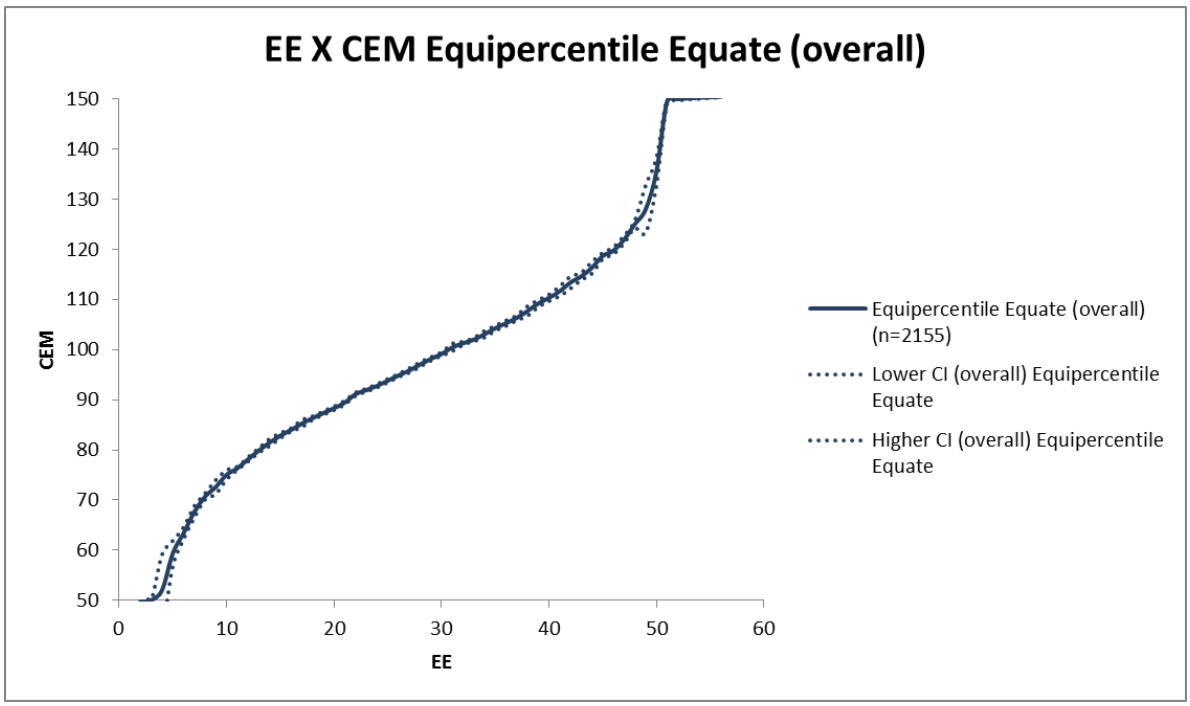
These graphs show that further investigation may be required as to whether the different assessments favour boys and girls differently. Although girls on average perform better than boys on all 3 reception baselines, the charts below indicate that there may be a further issue in relation to comparability as boys linked scores seem better on some assessments compared to others. These differences, if they exist, may be due to the nature of the assessments and the constructs being assessed which would be a problem given how results would be used.

EE/CEM sample

As can been seen in the following graph, at almost all points on the scale, boys have a higher linked score than girls. The standardised mean difference was 30% of the overall group standard deviation for EE and 16% for CEM. The average difference across the score range is nearly 2 CEM score points, the maximum difference is nearly 6 CEM score points.

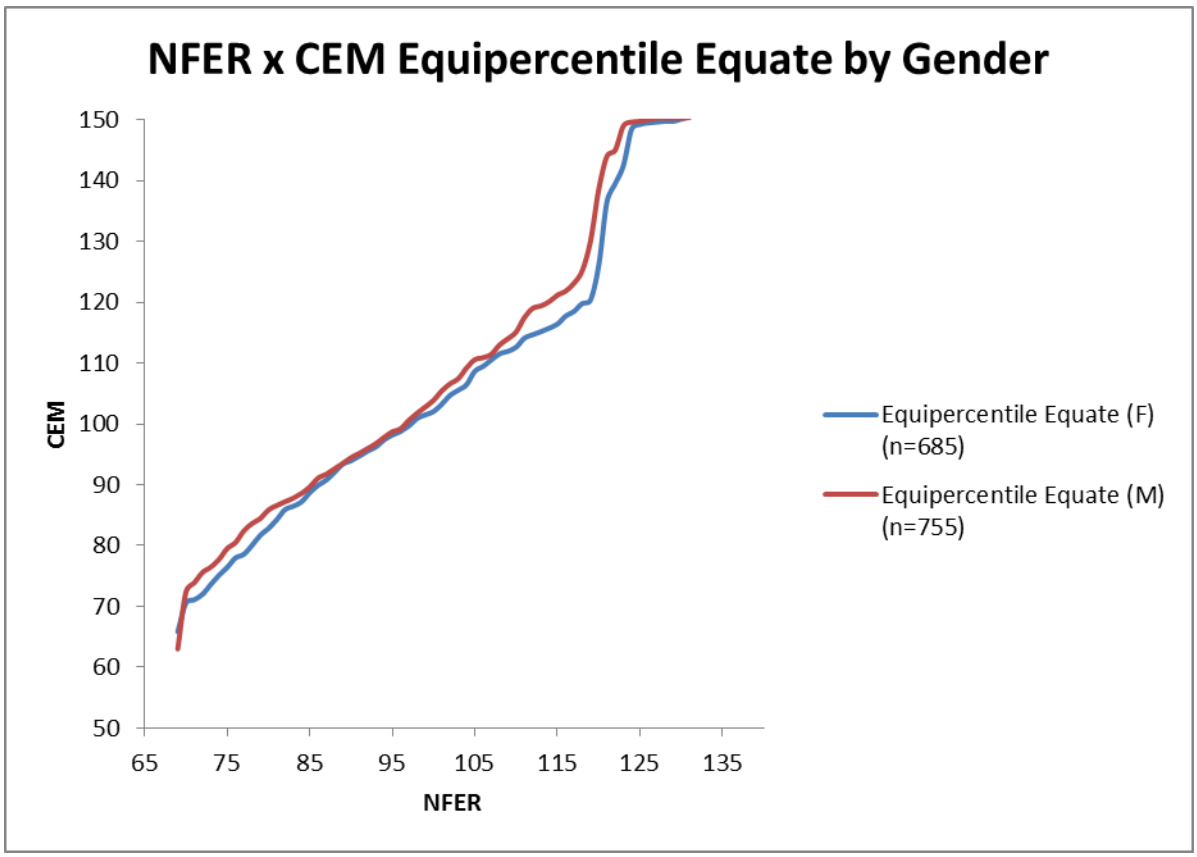


Below is the overall equipercentile link for EE/CEM. While the standard error of the link looks very close to the line, this is due to the difference in scale between the two assessments. The average width of the confidence interval based on the standard error of the link is 1.6, the maximum difference is 14.14.

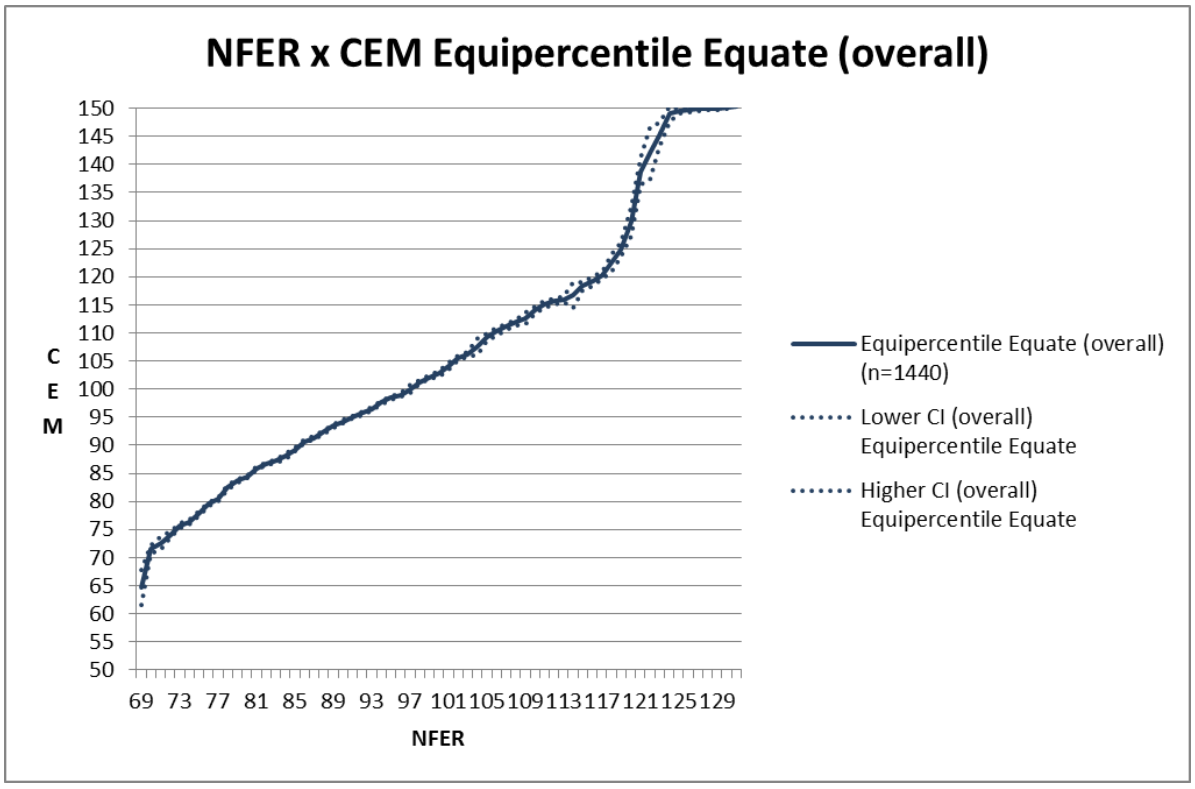


NFER/CEM sample

The differences between girls and boys in the NFER/CEM sample are different in that there is a small section of the middle of the scale where the difference between boys and girls is very small. Particularly at the top of the scale, boys have a higher linked score. The average difference across the score range is 2.3 CEM score points, the maximum difference is 12.5 CEM score points. The standardised mean difference for NFER was 34% of the overall group standard deviation; for CEM the standardised mean difference was 23% of the overall group standard deviation.

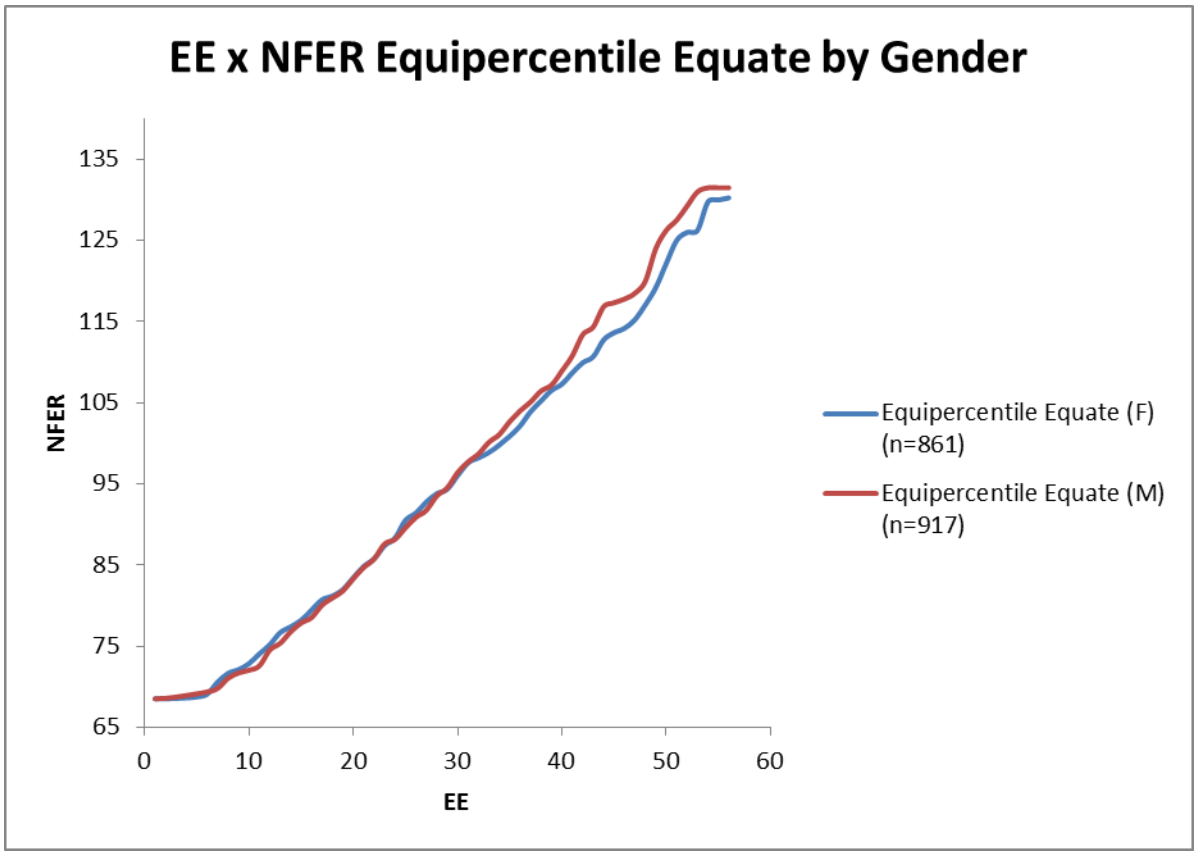


Below is the overall equipercentile link for NFER/CEM. The average difference in standard error of the link is 1.48, the maximum difference is 9.38.

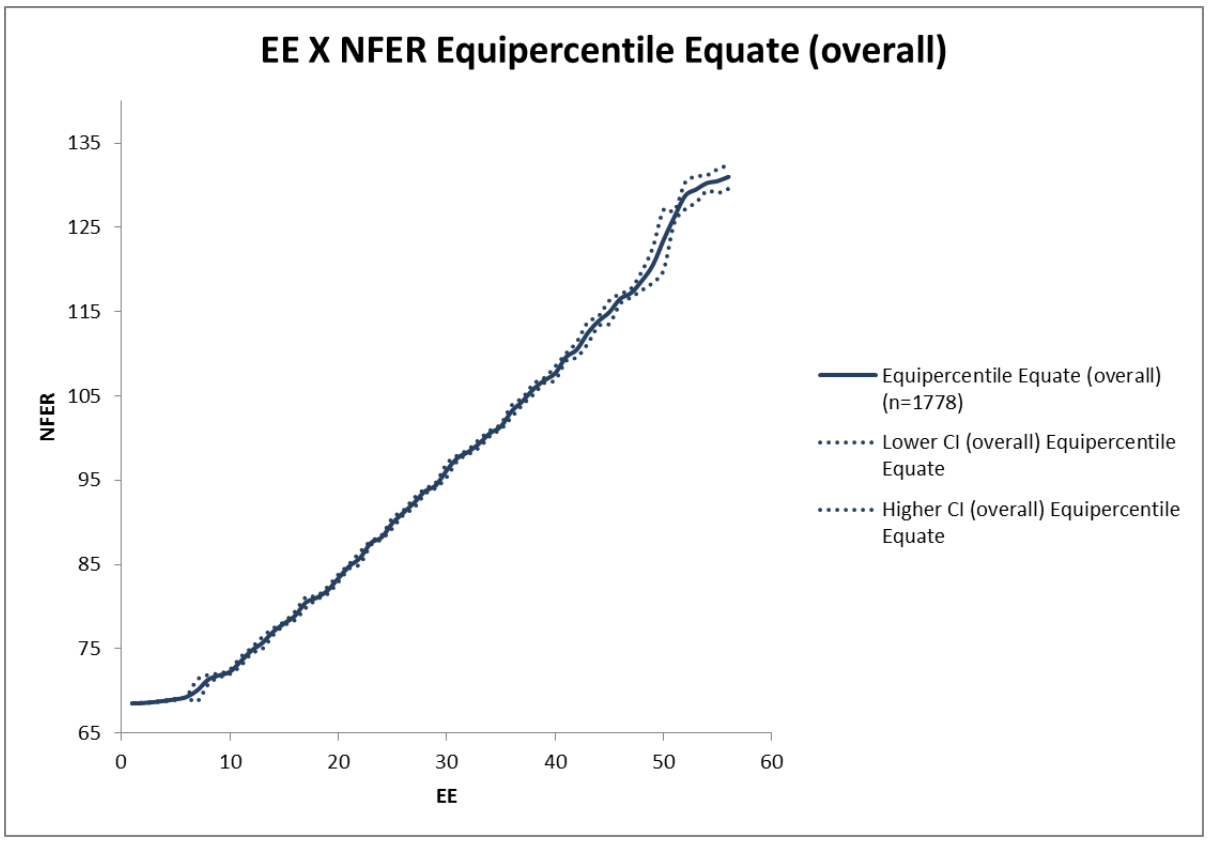


EE/NFER sample

The differences in linked scores between girls and boys in the EE/NFER sample are interesting in that for just over half of the scale there is very little difference between them whilst at the top of the scale boys have a higher linked score. The average difference is less than 1 NFER score point, and the maximum difference is just under 5 NFER score points. The standardised mean differences are very similar for both EE and NFER, at just over 25% of the overall group standard deviation.



Below is the overall equipercentile link for EE/NFER. The average difference in standard error of the link is 1.28, the maximum difference is 7.18.


EE X NFER Equipercentile Equate (overall)
NFER
135
125
115
105
95
85
75
65
0
10
20
30
40
50
60
EE
Equipercentile Equate (overall) (n=1778)
Lower CI (overall) Equipercentile Equate
Higher CI (overall) Equipercentile Equate

Conclusion

In summary, analysis was undertaken to examine the relationships between the 3 baseline suppliers' overall reported scores. All 3 products were designed to measure literacy and numeracy, but the specifications for the assessments differ. Equipercentile linking between the 3 pairs of reported scores show gender differences. These are not consistent across the score range and can be quite large in places, particularly for the NFER/CEM and the EE/NFER pairing.

While there were relatively high correlations, particularly between NFER and CEM, these were not sufficient to suggest that concordance is appropriate. It is likely that reported scores could prove useful in a prediction equation, but other background characteristics would be necessary in order to form a strong prediction. We have no evidence to suggest that the linkages that are possible are actually appropriate.

To illustrate what this means for the progress measure, we investigated the group of pupils to whom a pupil would be compared as having a similar starting point when they reached the end of key stage 2 (KS2).

The following example indicates what this would mean for the progress measure for a pupil using the NFER and CEM comparison, as the closest to demonstrating comparability, by virtue of the reduction of uncertainty presented earlier. In the progress calculation, a pupil's score on the KS2 tests is compared to the average score for pupils with the same starting point; in this case the same score on the reception baseline. Although we cannot know average scores on KS2 tests for this group of pupils for several years, we can look at whether the group of pupils against which a pupil will be compared will remain similar regardless of the reception baseline they took.

If we take a pupil who scored 90 on the NFER baseline, which is a common score in the middle of the distribution, this pupil would be compared to 40 other pupils in the progress measure who also scored 90. This same pupil on the CEM baseline scored 97 and would be compared to 40 other pupils in the progress measure who also scored 97. However, these two groups of 40 pupils to which the same pupil would be compared only had 3 other pupils in common.

This means that the average KS2 score to be used in the progress measure would be different depending on which baseline the pupil actually took. Although we cannot say how different, the range of scores on the CEM baseline for pupils who scored 90 on the NFER baseline is quite wide (from 82 to 120). Therefore, the CEM baseline does not agree with the NFER baseline that these pupils have a similar starting point, which is the requirement of the progress measure.

We therefore conclude from this study that there is insufficient comparability between the 3 reception baseline assessments to enable them to be used in the accountability system concurrently.

References

Dorans (2000). Distinctions among classes of linkages. RN-11. The College Board Office of Research and Development.

Elliot (2013). A guide to comparability terminology and methods. Research paper from Cambridge Assessment, www.cambridgeassessment.org.uk/insights/the-cambridge-approach-to-comparability.

Holland (2007). A framework and history for score linking. In Linking and Aligning Scores and Scales Dorans, N.J., Pommerich, M., and Holland, P.W. (Eds). Springer: New York.

Kolen, M.J, and Brennan, R.L. (2014). Test equating, scaling and linking: methods and practices, 3rd Edition. Springer: New York.

Linn (1993). Linking results of distinct assessments. Applied Measurement in Education, 6, 83-102.

Mislevy, R.J. (1992). Linking educational assessments: Concepts, issues, methods and prospects.

Newton, P., Baird, J. A., Goldstein, H., Patrick, H. and Tymms, P. (2007). Techniques for monitoring the comparability of examination standards. Qualifications and Curriculum Authority: London.

Appendix 1: Comparability study design

	CEM	EE	NFER	
1	P2	A1		School chose CEM and was assigned EE. EE administered first and CEM second
2	A2	P1		School chose EE and was assigned CEM. EE administered first and CEM second
3	P1	A2		School chose CEM and was assigned EE. CEM administered first and EE second
4	A1	P2		School chose EE and was assigned CEM. CEM administered first and EE second
5	A1		P2	School chose NFER and was assigned CEM. CEM administered first and NFER second
6	P1		A2	School chose CEM and was assigned NFER. CEM administered first and NFER second
7	A2		P1	School chose NFER and was assigned CEM. NFER administered first and CEM second
8	P2		A1	School chose CEM and was assigned NFER. NFER administered first and CEM second
9		P2	A1	School chose EE and was assigned NFER. NFER administered first and EE second
10		A2	P1	School chose NFER and was assigned EE. NFER administered first and EE second
11		P1	A2	School chose EE and was assigned NFER. EE administered first and NFER second
12		A1	P2	School chose NFER and was assigned EE. EE administered first and NFER second

Appendix 2: SQA samples tables

Provided on 6 January 2016, this table shows the distribution of schools by decile for the test combinations where schools had chosen CEM as their test provider.

Test Combination	Decile	Count of Schools	Actual %	Target %
CEM(1)_EE(2)	0	1	9%	9 to 10%
	1	1	9%	9 to 10%
	2	1	9%	9 to 10%
	3	2	18%	9 to 10%
	4	2	18%	9 to 10%
	5	1	9%	9 to 10%
	6	0	0%	9 to 10%
	7	0	0%	9 to 10%
	8	1	9%	9 to 10%
	9	1	9%	9 to 10%
	10	1	9%	9 to 10%
CEM(1)_EE(2) Total		**11**		
CEM(1)_NFER(2)	0	1	9%	9 to 10%
	1	1	9%	9 to 10%
	2	2	18%	9 to 10%
	3	0	0%	9 to 10%
	4	1	9%	9 to 10%
	5	2	18%	9 to 10%
	6	2	18%	9 to 10%
	7	0	0%	9 to 10%
	8	0	0%	9 to 10%
	9	1	9%	9 to 10%
	10	1	9%	9 to 10%
CEM(1)_NFER(2) Total		**11**		
CEM(2)_EE(1)	0	0	0%	9 to 10%
	1	2	20%	9 to 10%
	2	3	30%	9 to 10%
	3	2	20%	9 to 10%
	4	0	0%	9 to 10%
	5	0	0%	9 to 10%
	6	1	10%	9 to 10%
	7	1	10%	9 to 10%
	8	1	10%	9 to 10%
	9	0	0%	9 to 10%
	10	0	0%	9 to 10%
CEM(2)_EE(1) Total		**10**		
CEM(2)_NFER(1)	0	2	15%	9 to 10%
	1	1	8%	9 to 10%
	2	1	8%	9 to 10%
	3	0	0%	9 to 10%
	4	2	15%	9 to 10%
	5	1	8%	9 to 10%
	6	1	8%	9 to 10%
	7	1	8%	9 to 10%
	8	1	8%	9 to 10%
	9	2	15%	9 to 10%
	10	1	8%	9 to 10%
CEM(2)_NFER(1) Total		**13**		

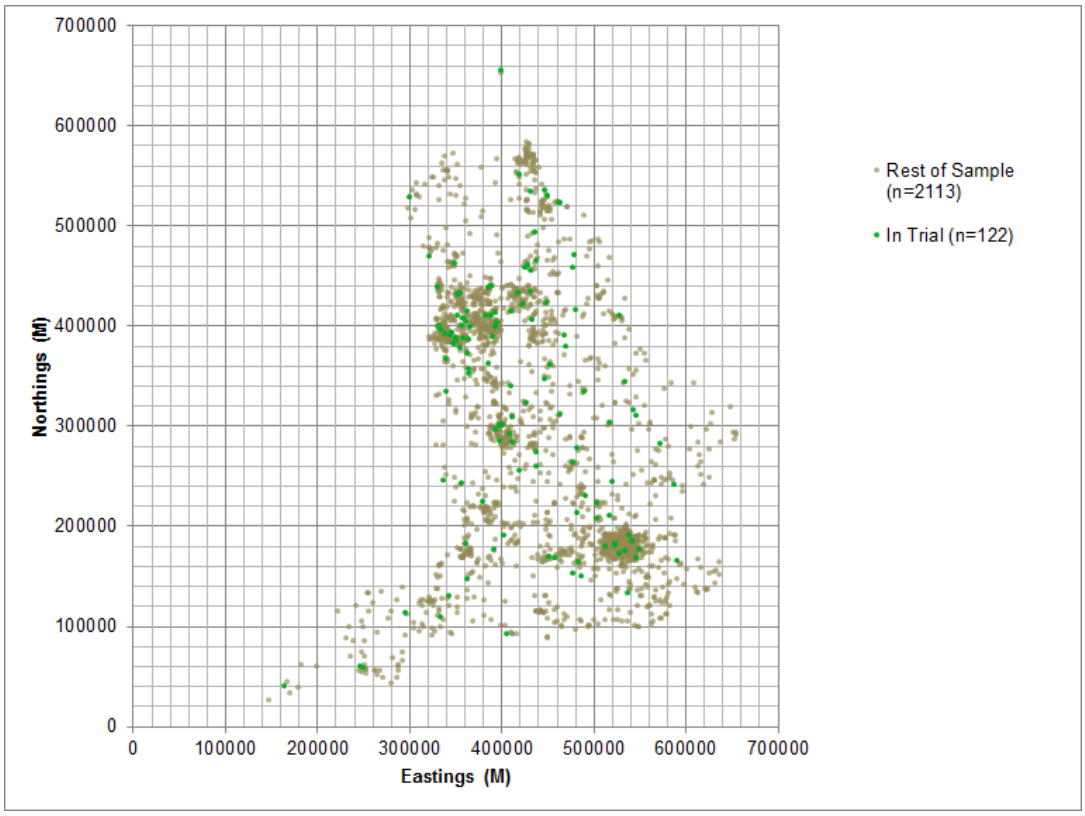
The distribution of schools by decile for the test combinations where schools had chosen EE as their test provider.

Test Combination	Decile	Count of Schools	Actual %	Target %
EE(1)_CEM(2)	0	1	11%	9 to 10%
	1	1	11%	9 to 10%
	2	2	22%	9 to 10%
	3	0	0%	9 to 10%
	4	0	0%	9 to 10%
	5	0	0%	9 to 10%
	6	2	22%	9 to 10%
	7	1	11%	9 to 10%
	8	1	11%	9 to 10%
	9	0	0%	9 to 10%
	10	1	11%	9 to 10%
EE(1)_CEM(2) Total		**9**		
EE(1)_NFER(2)	0	0	0%	9 to 10%
	1	2	20%	9 to 10%
	2	1	10%	9 to 10%
	3	1	10%	9 to 10%
	4	0	0%	9 to 10%
	5	2	20%	9 to 10%
	6	0	0%	9 to 10%
	7	1	10%	9 to 10%
	8	2	20%	9 to 10%
	9	1	10%	9 to 10%
	10	0	0%	9 to 10%
EE(1)_NFER(2) Total		**10**		
EE(2)_CEM(1)	0	0	0%	9 to 10%
	1	1	11%	9 to 10%
	2	1	11%	9 to 10%
	3	1	11%	9 to 10%
	4	1	11%	9 to 10%
	5	0	0%	9 to 10%
	6	1	11%	9 to 10%
	7	1	11%	9 to 10%
	8	1	11%	9 to 10%
	9	1	11%	9 to 10%
	10	1	11%	9 to 10%
EE(2)_CEM(1) Total		**9**		
EE(2)_NFER(1)	0	0	0%	9 to 10%
	1	2	20%	9 to 10%
	2	0	0%	9 to 10%
	3	1	10%	9 to 10%
	4	0	0%	9 to 10%
	5	2	20%	9 to 10%
	6	0	0%	9 to 10%
	7	2	20%	9 to 10%
	8	1	10%	9 to 10%
	9	1	10%	9 to 10%
	10	1	10%	9 to 10%
EE(2)_NFER(1) Total		**10**		

The distribution of schools by decile for the test combinations where schools had chosen NFER as their test provider.

Test Combination	Decile	Count of Schools	Actual %	Target %
NFER(1)_CEM(2)	0	1	9%	9 to 10%
	1	1	9%	9 to 10%
	2	1	9%	9 to 10%
	3	1	9%	9 to 10%
	4	2	18%	9 to 10%
	5	1	9%	9 to 10%
	6	1	9%	9 to 10%
	7	0	0%	9 to 10%
	8	2	18%	9 to 10%
	9	0	0%	9 to 10%
	10	1	9%	9 to 10%
NFER(1)_CEM(2) Total		**11**		
NFER(1)_EE(2)	0	1	13%	9 to 10%
	1	0	0%	9 to 10%
	2	1	13%	9 to 10%
	3	1	13%	9 to 10%
	4	1	13%	9 to 10%
	5	0	0%	9 to 10%
	6	1	13%	9 to 10%
	7	1	13%	9 to 10%
	8	1	13%	9 to 10%
	9	1	13%	9 to 10%
	10	0	0%	9 to 10%
NFER(1)_EE(2) Total		**8**		
NFER(2)_CEM(1)	0	2	18%	9 to 10%
	1	1	9%	9 to 10%
	2	2	18%	9 to 10%
	3	1	9%	9 to 10%
	4	1	9%	9 to 10%
	5	1	9%	9 to 10%
	6	0	0%	9 to 10%
	7	0	0%	9 to 10%
	8	2	18%	9 to 10%
	9	0	0%	9 to 10%
	10	1	9%	9 to 10%
NFER(2)_CEM(1) Total		**11**		
NFER(2)_EE(1)	0	0	0%	9 to 10%
	1	0	0%	9 to 10%
	2	2	22%	9 to 10%
	3	1	11%	9 to 10%
	4	1	11%	9 to 10%
	5	1	11%	9 to 10%
	6	1	11%	9 to 10%
	7	0	0%	9 to 10%
	8	2	22%	9 to 10%
	9	0	0%	9 to 10%
	10	1	11%	9 to 10%
NFER(2)_EE(1) Total		**9**		

Distribution of schools over region also deviated from the target. For example there were a disproportionately high number of schools from the North West and lower number of schools from London.



Geographical distribution of allocated schools for Baseline Reception study 2015.

The distribution of schools by region for the test combinations where schools had chosen CEM as their test provider.

Test Combination	Region	Count of Schools	Actual %	Target %
CEM(1)_EE(2)	East Midlands	1	9.1%	9.7%
	East of England	0	0.0%	11.8%
	London	2	18.2%	10.8%
	North East	0	0.0%	5.2%
	North West	4	36.4%	14.6%
	South East	0	0.0%	15.5%
	South West	1	9.1%	11.1%
	West Midlands	1	9.1%	10.6%
	Yorkshire and the Humber	2	18.2%	10.7%
CEM(1)_EE(2) Total		**11**	**100.0%**	**100.0%**
CEM(1)_NFER(2)	East Midlands	1	9.1%	9.7%
	East of England	1	9.1%	11.8%
	London	1	9.1%	10.8%
	North East	1	9.1%	5.2%
	North West	3	27.3%	14.6%
	South East	1	9.1%	15.5%
	South West	1	9.1%	11.1%
	West Midlands	1	9.1%	10.6%
	Yorkshire and the Humber	1	9.1%	10.7%
CEM(1)_NFER(2) Total		**11**	**100.0%**	**100.0%**
CEM(2)_EE(1)	East Midlands	0	0.0%	9.7%
	East of England	3	30.0%	11.8%
	London	0	0.0%	10.8%
	North East	0	0.0%	5.2%
	North West	3	30.0%	14.6%
	South East	1	10.0%	15.5%
	South West	1	10.0%	11.1%
	West Midlands	0	0.0%	10.6%
	Yorkshire and the Humber	2	20.0%	10.7%
CEM(2)_EE(1) Total		**10**	**100.0%**	**100.0%**
CEM(2)_NFER(1)	East Midlands	1	7.7%	9.7%
	East of England	0	0.0%	11.8%
	London	0	0.0%	10.8%
	North East	2	15.4%	5.2%
	North West	5	38.5%	14.6%
	South East	1	7.7%	15.5%
	South West	1	7.7%	11.1%
	West Midlands	1	7.7%	10.6%
	Yorkshire and the Humber	2	15.4%	10.7%
CEM(2)_NFER(1) Total		**13**	**100.0%**	**100.0%**

The distribution of schools by region for the test combinations where schools had chosen EE as their test provider.

Test Combination	Region	Count of Schools	Actual %	Target %
EE(1)_CEM(2)	East Midlands	1	11.1%	9.7%
	East of England	1	11.1%	11.8%
	London	0	0.0%	10.8%
	North East	0	0.0%	5.2%
	North West	2	22.2%	14.6%
	South East	1	11.1%	15.5%
	South West	1	11.1%	11.1%
	West Midlands	2	22.2%	10.6%
	Yorkshire and the Humber	1	11.1%	10.7%
EE(1)_CEM(2) Total		**9**	**100.0%**	**100.0%**
EE(1)_NFER(2)	East Midlands	1	10.0%	9.7%
	East of England	0	0.0%	11.8%
	London	2	20.0%	10.8%
	North East	0	0.0%	5.2%
	North West	3	30.0%	14.6%
	South East	1	10.0%	15.5%
	South West	2	20.0%	11.1%
	West Midlands	0	0.0%	10.6%
	Yorkshire and the Humber	1	10.0%	10.7%
EE(1)_NFER(2) Total		**10**	**100.0%**	**100.0%**
EE(2)_CEM(1)	East Midlands	0	0.0%	9.7%
	East of England	2	22.2%	11.8%
	London	0	0.0%	10.8%
	North East	0	0.0%	5.2%
	North West	3	33.3%	14.6%
	South East	1	11.1%	15.5%
	South West	1	11.1%	11.1%
	West Midlands	2	22.2%	10.6%
	Yorkshire and the Humber	0	0.0%	10.7%
EE(2)_CEM(1) Total		**9**	**100.0%**	**100.0%**
EE(2)_NFER(1)	East Midlands	1	10.0%	9.7%
	East of England	0	0.0%	11.8%
	London	2	20.0%	10.8%
	North East	0	0.0%	5.2%
	North West	4	40.0%	14.6%
	South East	1	10.0%	15.5%
	South West	0	0.0%	11.1%
	West Midlands	1	10.0%	10.6%
	Yorkshire and the Humber	1	10.0%	10.7%
EE(2)_NFER(1) Total		**10**	**100.0%**	**100.0%**

The distribution of schools by region for the test combinations where schools had chosen NFER as their test provider.

Test Combination	Region	Count of Schools	Actual %	Target %
NFER(1)_CEM(2)	East Midlands	1	9.1%	9.7%
	East of England	0	0.0%	11.8%
	London	1	9.1%	10.8%
	North East	0	0.0%	5.2%
	North West	3	27.3%	14.6%
	South East	1	9.1%	15.5%
	South West	1	9.1%	11.1%
	West Midlands	3	27.3%	10.6%
	Yorkshire and the Humber	1	9.1%	10.7%
NFER(1)_CEM(2) Total		**11**	**100.0%**	**100.0%**
NFER(1)_EE(2)	East Midlands	1	12.5%	9.7%
	East of England	1	12.5%	11.8%
	London	1	12.5%	10.8%
	North East	2	25.0%	5.2%
	North West	2	25.0%	14.6%
	South East	0	0.0%	15.5%
	South West	0	0.0%	11.1%
	West Midlands	0	0.0%	10.6%
	Yorkshire and the Humber	1	12.5%	10.7%
NFER(1)_EE(2) Total		**8**	**100.0%**	**100.0%**
NFER(2)_CEM(1)	East Midlands	0	0.0%	9.7%
	East of England	0	0.0%	11.8%
	London	0	0.0%	10.8%
	North East	2	18.2%	5.2%
	North West	3	27.3%	14.6%
	South East	0	0.0%	15.5%
	South West	2	18.2%	11.1%
	West Midlands	4	36.4%	10.6%
	Yorkshire and the Humber	0	0.0%	10.7%
NFER(2)_CEM(1) Total		**11**	**100.0%**	**100.0%**
NFER(2)_EE(1)	East Midlands	1	11.1%	9.7%
	East of England	1	11.1%	11.8%
	London	0	0.0%	10.8%
	North East	0	0.0%	5.2%
	North West	1	11.1%	14.6%
	South East	1	11.1%	15.5%
	South West	1	11.1%	11.1%
	West Midlands	1	11.1%	10.6%
	Yorkshire and the Humber	3	33.3%	10.7%
NFER(2)_EE(1) Total		**9**	**100.0%**	**100.0%**

The comparatively poor distribution of schools across stratification characteristics was a direct result of the recruitment context where recruitment period was short, rejection rates to calling were high and most schools came from the letter invitation to register via the web. Despite this outcome compromising the stratification characteristics was a necessary action to achieve the primary objective of recruiting sufficient pupils to conduct the study.

Appendix 3: Detailed description of assessment content

CEM

The assessment measures levels of literacy and mathematics development with an optional measure of personal, social and emotional development.

The literacy section assesses the following skills and knowledge which research has shown to be important in the development of literacy:

- pattern matching
- concepts about print
- repeating words
- vocabulary
- letter recognition
- word recognition
- reading and comprehension

The mathematics section assesses the following skills and knowledge that have been found to be important for development:

- counting and numerosity
- shape identification
- number identification
- quantity and spatial position
- addition and subtraction problems
- mathematics problems

Items are drawn from an item bank consisting of 247 items covering the areas of literacy and mathematics as shown in the table below. Items are drawn from the item bank using a rule-based algorithm. Children will receive different numbers of questions in each section and an average of 105 questions in total. The proportion of items they see are broadly similar to the proportions represented below.

Main area of content	*Proportion of assessment*
Literacy	76% (approx.)
Mathematics	24% (approx.)
Total	100%

EE

The Early Excellence Baseline Assessment (EExBA) consists of assessment data items derived from national guidance designed to support the current statutory EYFS Framework.

The main areas of content within EExBA consist of:

- Characteristics of Effective Learning
- Personal and Social and Emotional Development
- Communication and Language
- Physical Development
- Literacy
- Mathematics

Each of these main areas of content are subdivided into the following specific components which detail the skills, knowledge and behaviours assessed in each of the areas:

Characteristics of Effective Learning

- Engagement
- Motivation
- Creativity and Critical Thinking

Areas of Learning and Development

- Personal Social and Emotional Development
    - Self Confidence and Self Awareness
    - Managing Feelings and Behaviour
    - Making Relationships
- Communication and Language
    - Listening and Attention
    - Understanding
    - Speaking
- Physical Development
    - Moving and Handling
    - Health and Self-care
- Literacy
    - Reading
    - Writing
- Mathematics
    - Numbers
    - Shape Space and Measures

The scoring system is weighted and distinguishes between the Areas of Learning and Development (AoL&D) which consist of Personal Social and Emotional Development, Communication and Language, Physical Development, Literacy and Mathematics and the Characteristics of Effective Learning (CoEL). Each statement attained for the AoL&D is worth 1 mark, each statement attained for the CoEL is worth 2 marks; therefore 18 marks are available for the 9 CoEL statements. The rationale for this is based upon the importance of CoEL as indicators of possible future attainment and 'School Readiness'

Main area of content	Proportion of assessment
EYFS Area of Learning and Development – Communication and Language	25.5%
EYFS Area of Learning and Development – Literacy	17%
EYFS Area of Learning and Development – Mathematics	17%
EYFS Area of Learning and Development – Personal Social and Emotional Development	13%
EYFS Area of Learning and Development – Physical Development	8.5%
EYFS Characteristics of Effective Learning	19%
Total	**100%**

NFER

The core areas of content in the NFER Reception Baseline are communication, language, literacy and numeracy. Schools will also be offered a further optional module assessing learning dispositions and personal, social and emotional development.

Communication, Language and Literacy Tasks

Task	Description	Abilities assessed
1	Vocabulary – simple receptive	Receptive language
2	Vocabulary – simple expressive	Expressive language
3	Vocabulary – complex receptive	Receptive language Grammatical skills
4	Phonics - segmenting	Phonological awareness
5	Phonics - blending	Phonological awareness
6	Picture sequencing & story prediction	Comprehension, understanding of story structure, drawing inferences
7	Reading simple words	Phonological awareness, letter knowledge
8	Reading simple sentences	Phonological awareness, letter knowledge, grammatical skills, comprehension
9	Writing their name	Phonological awareness, letter knowledge
10	Listening comprehension	Comprehension, understanding of story structure, drawing inferences

The language and communication checklist measures listening and attention, receptive and expressive language and communication skills.

Numeracy tasks

Task	Description	Abilities assessed
1	Counting	Arithmetic
2	Identification of numerals	Numeral identification
3	Sequencing numerals	Relations between numbers
4	Finding more	Mathematical reasoning
5	Finding less	Mathematical reasoning
6	Practical addition	Arithmetic
7	Practical subtraction	Arithmetic
8	Written addition	Arithmetic
9	Halving	Mathematical reasoning
10	Shapes	Awareness of mathematical structure
11	Pattern recognition	Mathematical reasoning

Overall assessment

Main area of content	Proportion of assessment
Communication and language	14%
Literacy (including vocabulary)	48%
Numeracy	38%
Total	100%

Standards & Testing Agency

About this publication:

enquiries www.education.gov.uk/contactus

download www.gov.uk/government/publications

Reference: 978-1-78644-177-5 STA/16/7595/e