Department for Education

Reception baseline research: results of a randomised controlled trial

Research Brief

July 2015

Deaprtment for Education

Contents

Introduction

In March 2014, the Department for Education published its response to a consultation on reforming assessment and accountability in primary schools. The response set out the Department's intention to change the way it hold primary schools to account, by introducing a Reception Baseline assessment, which will be the only measure used to assess the progress of children from entry (at age 4-5) to the end of key stage 2 (age 10-11). From 2016 onwards, all schools that wish to demonstrate progress for accountability purposes will have to adopt an approved Reception Baseline scheme. In 2023, when this cohort of pupils reaches the end of key stage 2, the Reception Baseline will be the starting point used to measure pupil progress for all-through primary schools. Schools can opt to use an approved baseline assessment from September 2015 if they wish to do so.

DfE commissioned research to gain a greater understanding about how the proposed reception baseline could be implemented and to identify effective ways of communicating the results to parents[1]. The research comprised two strands. The first was a randomised controlled trial carried out in the autumn term 2014 by DfE in partnership with the Centre for Evaluation and Monitoring (CEM) at Durham University, which aimed to investigate schools' behaviour changes in response to the accountability reforms. This is reported here. The second strand was a qualitative study undertaken by the National Foundation for Educational Research (NFER)[2].

The randomised controlled trial aimed to explore whether schools' perceptions of the purpose of the reception baseline test led to differences in pupils' early attainment, and in particular if there was any evidence of 'gaming'. This summary sets out the key findings drawn from this study.

Methodology

Durham University's Centre for Evaluation and Monitoring (CEM) administered the research and collected the data. Schools which had previously used CEM's Performance Indicators in Primary Schools (PIPS) Baseline Assessment were invited to take part in the research. To encourage participation, schools were given access to new reporting materials for parents.

A sample of 153 schools with 5,368 eligible pupils was split into two groups[3], with one group told that the reception baseline test would be used for accountability purposes (the

[1] 'Parents' includes primary carers throughout.

[2] https://www.gov.uk/government/publications?keywords=&publication_filter_option=research-and-analysis&topics%5B%5D=all&departments%5B%5D=department-for-education&official_document_status=all&world_locations%5B%5D=all&from_date=&to_date=

[3] The sample size was in line with agreed expectations.

Accountability group) while the other was told that the test would only be used for teaching and learning (the **Teaching and Learning** group). The schools were divided into the two groups at random and tests confirm that the two groups of schools were similar in terms of school type, size, location, pupil gender, Year 1 Phonics pass rate, Key Stage 1 Level 2 pass rate and the proportion of pupils eligible for Free School Meals (FSM), who have English as an Additional Language (EAL) and who have a Special Educational Need (SEN).[4]

This research design increases the likelihood that any differences in pupil attainment observed can be attributed to the difference in how the baseline test was framed rather than differences in the characteristics of the schools.

The data were analysed internally by the Department for Education. Pupil level linear regression models were built to identify the independent effect of the baseline test 'framing' on the average pupil scores on the test, controlling for other factors. As the pupils were clustered within schools, they did not form a truly independent random sample. A design effect was calculated to take account of this clustering and determine if there was a genuine difference between the two groups.

CEM also undertook independent retests on 250 pupils at the sampled schools to investigate whether there was a long-term impact of the difference in 'framing' the purpose of the test.

Results

- When results are adjusted to take account of the clustering in the sample, there is no strong evidence that framing the reception baseline test as an accountability measure as opposed to a teaching and learning aid resulted in a reduction in test results. This may be due to the small sample size rather than the absence of an effect.

More Detailed Findings

- There was a small difference between the mean total scores of the Accountability and Teaching and Learning groups. Pupils based in schools told that the reception baseline was an accountability measure had scores which were on average 2.7 marks (or 4.2%) lower than those who were informed that it was a teaching and learning aid.[5]

[4] See Annex A for further information.

[5] The mean score for the Accountability group was 61.8 compared to 64.5 for the Teaching and Learning group.

- Scores for the Accountability group were generally lower across the distribution of attainment compared to the Teaching and Learning Group.

- Regression analysis controlling for pupil age and gender but not taking the clustering into account suggested that there was a statistically significant difference between the baseline test scores for the Accountability and Teaching and Learning groups.[6]

- Further regressions were run on the total scores for the maths, reading and phonics subject areas to determine whether this impact was observed for all subject areas. The mean marks in the teaching and learning group were 25.9 for maths, 30.1 for reading and 8.6 for phonics. Those told that the baseline test is an accountability measure saw reduced scores of 4.2% in maths, 4.8% in reading and 1.2% in phonics. The differences between the groups in maths and reading scores were statistically significant. The difference between the phonics scores of pupils in the Accountability and Teaching and Learning groups was not statistically significant.

- When the impact of the design effect is considered, the difference in means of 2.7 becomes insignificant at the 95% level.[7] In other words, when findings are adjusted to take account of the clustering in the sample, there is no strong evidence that framing the reception baseline test as an accountability measure as opposed to a teaching and learning aid resulted in a reduction in test results. However, it is possible that there was an effect which was not detected due to the small sample size.

- CEM undertook independent retests on 250 students. Since these were carried out at a later date, both groups saw increases in the mean score. The mean increase in score was 13.0 marks for the Accountability group and 12.4 marks for the Teaching and Learning Group (a difference of 0.6 marks). Given the small sample size of retests, it is not possible to demonstrate any statistical significance of the effect seen using linear regression.

[6] See Annex B for further information.

[7] See Annex C for further information.

Conclusion

In the two treatment groups sampled, the mean score within schools told that the baseline test was an accountability measure (Accountability group) was 2.7 marks (4.2%) less than those told it is a teaching and learning resource only.

This reduction was also seen across two subject areas making up the test – maths and reading – with the largest effect seen in the reading subject area with a 4.8% decrease for the Accountability group compared to the Teaching and Learning group. However, the difference for phonics scores was not statistically significant after controlling for pupil gender and age.

The overall result would be statistically significant at the 95% level if the data were from an independent random sample. However once the correlation between pupils within schools is taken into account, the result is no longer statistically significant.

Independent retests undertaken by CEM were not carried out on a large enough sample to evidence any statistically significant difference but do show a slightly larger mean increase in the score for pupils in the accountability group.

Therefore, while it does appear that the way the test was framed to schools may be influenced pupil scores on the test, it has not been possible to confirm this given the sample data available.

Annex A: School characteristics within the treatment groups

School characteristics[8]

	Accountability	Teaching and Learning
Mean school size	241 pupils	242 pupils
School type		
Academy - Converter Mainstream	4	6
Academy Sponsor Led	0	1
Community School	27	29
Foundation School	2	3
Free School - Mainstream	1	0
Voluntary Aided School	31	30
Voluntary Controlled School	13	6
School location		
North East	2	6
North West	40	38
Yorkshire and the Humber	2	6
East Midlands	4	2
West Midlands	3	5
East	4	2
Inner London	1	1

[8] Information about Y1 Phonics pass rate, KS1 Level 2 pass rate, mean FSM rate, mean EAL rate and mean SEN rate all taken from the School Census 2014. Other data collected by CEM.

Outer London	2	0
South East	6	5
South West	14	10
Pupil Gender		
Proportion Female	48.0%	48.5%
Proportion Male	52.0%	51.5%
Year 1 Phonics Pass Rate	79.6%	77.7%
Key Stage 1 Level 2 Pass Rate	90.6%	91.0%
Mean FSM Rate	14.5%	15.7%
Mean EAL Rate	8.8%	12.9%
Mean SEN Rate	6.8%	7.5%
Total number of schools	78	75
Total number of pupils	2,844	2,524

Pupils eligible for Free School Meals (FSM), who speak English as an Additional Language (EAL) or who have Special Educational Needs (SEN) have lower attainment on average. It might be expected therefore that Teaching and Learning schools would perform less well as they have more representation from these groups. This suggests that the reduction of marks observed in the Accountability group cannot be attributed to differences in school-level characteristics.

Annex B: Table of Coefficients for Pupil Level Linear Regressions

	Coefficient	Standard Error of Coefficient	t-statistic
Dependent Variable: Reception Baseline Test Score Model 1			
Accountability* (reference Teaching and Learning)	-2.672	0.778	-3.434
R-squared = 0.002			
Dependent Variable: Reception Baseline Test Score Model 2			
Accountability* (reference Teaching and Learning)	-2.641	0.743	-3.555
Male* (reference Female)	-5.332	0.743	-7.176
Age in days*	0.075	0.003	25.000
R-squared = 0.091			
Dependent Variable: Maths Score			
Accountability* (reference Teaching and Learning)	-1.079	0.248	4.351
Male* (reference Female)	-0.930	0.248	3.750
Age in days*	0.026	0.001	26.000
R-squared = 0.092			
Dependent Variable: Reading Score			
Accountability* (reference Teaching and Learning)	-1.458	0.481	3.031
Male* (reference Female)	-3.472	0.480	7.233
Age in days*	0.039	0.002	19.500
R-squared = 0.064			
Dependent Variable: Phonics Score			

Accountability (reference Teaching and Learning)	-0.103	0.124	0.831
Male* (reference Female)	-0.930	0.124	7.500
Age in days*	0.010	0.001	10.000
R-squared = 0.063			

* p<0.01 (significant at the 99% level)

Annex C: Calculation of the Design Effect

$\mathrm{DEFF} = 1 + \delta\,(\mathrm{n} - 1)$, where

DEFF is the design effect

δ is the intraclass correlation coefficient ('ICC'), and

n is the average size of the cluster (or average number of pupils per school)

$$\mathrm{ICC} = \frac{\text{Variance between school means}}{\text{Total variance in the population}} = \frac{97.7}{806.8} = 12.1\%$$

Therefore;

$$\mathrm{DEFF} = 1 + 0.121(35.1 - 1)$$
$$= 5.13$$

This can then be used to calculate 95% confidence intervals around the difference in means of 2.7 as follows;

$$2.7 \pm 1.96 \times \sqrt{\mathrm{DEFF}} \times \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

Where;
sd_i is the standard deviation of each treatment group, and
n_i is the size of each treatment group

So

$$= 2.7 \pm 1.96 \times \sqrt{5.13} \times \sqrt{\frac{28.587^2}{2844} + \frac{28.288^2}{2524}}$$

$$= 2.7 \pm 3.5$$

As this confidence interval spans zero the difference in means is not statistically significant.

Department for Education

Reference: DFE-RB476

ISBN: 978-1-78105-508-3

This research was commissioned under the under the 2010 to 2015 Conservative and Liberal Democrat coalition government. As a result the content may not reflect current Government policy. This research was commissioned under the under the 2010 to 2015 Conservative and Liberal Democrat coalition government. As a result the content may not reflect current Government policy.

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

Any enquiries regarding this publication should be sent to us at Konstantina.DIMOU@education.gsi.gov.uk or www.education.gov.uk/contactus

This document is available for download at www.gov.uk/government/publications