Standards & Testing Agency

Phonics screening check

2013 post-administration technical report

Contents

Table of figures

About this document

What is this document about?

This document provides further evidence of the validity and reliability of the phonics screening check via a quantitative item analysis of the data from the live 2013 administration. This document should be considered alongside the previous technical reports that were published in February 2012[1] and December 2012[2] and the Statistical First Release[3] published in September 2013.

The Department for Education (DfE) has commissioned an independent evaluation of the phonics screening check culminating in a final report in spring 2015. This evaluation will provide valuable information about the impact of the check on phonics teaching. The first interim report[4] was published on 20 May 2013.

Who is this document for?

This document is primarily aimed at a technical audience, but contains information that will be of interest to all stakeholders involved in the phonics screening check, including schools.

[1] http://www.education.gov.uk/schools/teachingandlearning/assessment/keystage1/a00200415/phonics
[2] http://www.education.gov.uk/schools/teachingandlearning/assessment/keystage1/a00200415/phonics
[3] https://www.gov.uk/government/publications/phonics-screening-check-and-national-curriculum-assessments-at-key-stage-1-in-england-2013
[4] https://www.gov.uk/government/publications/evaluation-of-the-phonics-screening-check-first-interim-report

Introduction

The government has established a check of phonic decoding at the end of year 1 with the results of this check to be made available to parents.

The phonics screening check was piloted in June 2011, and rolled out nationally in 2012. The check focuses solely on decoding using phonics and confirms whether children have reached the expected standard by the end of year 1, identifying children who need additional support from their school to catch up.

In 2013 children in year 2 who previously did not meet the expected year 1 standard or were not tested in 2012 were required to re-take the check. The purpose of the check remains the same for the year 2 children.

The phonics screening check consists of 20 real words and 20 pseudo-words. The pseudo-words provide the purest assessment of phonic decoding because they are new to all children, so there is no unintended bias based on visual memory of words or vocabulary knowledge. The pseudo-words are presented alongside a picture prompt (a picture of an imaginary creature) and children are asked to name the type of creature. This approach makes it clear to children that they are reading a pseudo-word, which they should not expect to be able to match to their existing vocabulary. The real words include between 40 per cent and 60 per cent less common words, which children are less likely to have read previously. Less common words are included so that the majority of children will need to decode using phonics rather than rely on sight memory of words they have seen before.

The phonics screening check is made up of two sections with items in each section relating to specified elements of the content domain. Items within each section are ordered according to orthographical representation with real and pseudo-words grouped together. Each section contains 20 items.

It is necessary to start with easier words in section 1 to make the phonics screening check accessible and to provide some information to teachers if their children are unable to decode relatively simple words. However, the words at the end of the phonics screening check are around the level of difficulty we expect children to reach by the end of year 1.

The technical report, published in December 2012, concluded that:

> Having examined all of the evidence gathered so far through the pilot and the live sample, the Department is satisfied that the phonics screening check is sufficiently valid for its defined purpose and has acceptable levels of reliability.

Year 2 children

The policy decision to require year 2 children who previously did not meet the expected year 1 standard or were not tested in 2012 to re-take the check means that the DfE must provide assurances that the phonics screening check is sufficiently valid and reliable for these children as an extension of the original purpose.

STA collected a sample of item level data from schools taking the phonics screening check in June 2013 in order to examine the performance of items when taken by year 2 children.

This technical report presents the quantitative item analysis from the sample item level data collection, in order to provide further evidence of validity and reliability of the phonics screening check, as set out in Ofqual's Regulatory framework for national assessment arrangements (Ofqual, 2011[5]).

[5] www2.ofqual.gov.uk/files/2011-regulatory-framework-for-national-assessments.pdf

Sample selection

Two samples of maintained schools, with year 1 children and/or year 2 children, were drawn using data from the autumn 2012 school census and Edubase. The samples were stratified by region and key stage 1 attainment in reading (based on data from 2012). The achieved year 1 sample contained 10,416 children from 270 schools. The achieved year 2 sample contained 8,500 children from 533 schools.

Figures 1 and 2 show the representativeness of the samples compared to the population across key stage 1 attainment, type of establishment, and region. The samples are representative of the population of schools in 2013.

		Population		Year 1 Phonics sample	
		Count	%	Count	%
Average 2012 key stage 1 reading point score	Lowest 20%	3370	21.1	53	19.6
	2nd lowest 20%	3277	20.5	58	21.5
	Middle 20%	2965	18.5	51	18.9
	2nd highest 20%	3244	20.3	55	20.4
	Highest 20%	2947	18.4	51	18.9
	Missing data	204	1.3	2	0.7
Type of establishment	Academy converters	763	4.8	9	3.3
	Academy sponsor led	148	0.9	0	0.0
	Community school	8376	52.3	152	56.3
	Community special school	460	2.9	3	1.1
	Foundation school	495	3.1	4	1.5
	Foundation special school	28	0.2	0	0.0
	Free Schools	33	0.2	0	0.0
	Voluntary aided school	3412	21.3	58	21.5
	Voluntary controlled School	2292	14.3	42	15.6
Region	East Midlands	1521	9.5	25	9.3
	East of England	1867	11.7	31	11.5
	London	1707	10.7	29	10.7
	North East	863	5.4	15	5.6
	North West	2445	15.3	40	14.8
	South East	2346	14.7	40	14.8
	South West	1802	11.3	30	11.1
	West Midlands	1725	10.8	30	11.1
	Yorkshire and the Humber	1731	10.8	30	11.1
Total		16007	100.0	270	100.0

Figure 1 2013 Year 1 Phonics sample representation

		Population		Year 2 Phonics sample	
		Count	%	Count	%
Average 2012 key stage 1 reading point score	Lowest 20%	3385	21.2	109	19.7
	2nd lowest 20%	3280	20.5	116	21.0
	Middle 20%	2965	18.5	106	19.2
	2nd highest 20%	3247	20.3	118	21.3
	Highest 20%	2945	18.4	96	17.4
	Missing data	182	1.1	8	1.5
Type of establishment	Academy converters	762	4.8	27	4.9
	Academy sponsor led	149	0.9	5	0.9
	Community school	8375	52.3	292	52.8
	Community special school	483	3.0	7	1.3
	Foundation school	491	3.1	18	3.3
	Foundation special school	25	0.2	0	0.0
	Free Schools	19	0.1	0	0.0
	Voluntary aided school	3410	21.3	123	22.2
	Voluntary controlled School	2290	14.3	81	14.6
Region	East Midlands	1520	9.5	56	10.1
	East of England	1870	11.7	64	11.6
	London	1696	10.6	57	10.3
	North East	866	5.4	32	5.8
	North West	2448	15.3	83	15.0
	South East	2348	14.7	81	14.6
	South West	1794	11.2	61	11.0
	West Midlands	1729	10.8	61	11.0
	Yorkshire and the Humber	1733	10.8	58	10.5
Total		16004	100.0	313	100.0

Figure 2 2013 Year 2 Phonics sample representation

Summary statistics

Whole check statistics

Figure 3 shows the summary check performance from the children in the sample. The average score is just over three quarters of the total marks. The average score for each section is over half marks.

	Year 1 Sample			Year 2 Sample		
	Whole check	Section 1	Section 2	Whole check	Section 1	Section 2
Number of children	10416	10416	10416	8500	8500	8500
Mean score	31.40	17.48	13.92	31.72	17.44	14.28
Standard deviation	9.11	3.78	5.72	8.02	3.37	5.05
Cronbach's alpha	0.952	0.901	0.927	0.934	0.863	0.900
Standard error of measurement	2.0	1.3	1.6	2.1	1.3	1.7

Figure 3 Whole check statistics

Cronbach's alpha is a measure of the internal consistency of a test or assessment, with a maximum value of 1. The high value of Cronbach's alpha indicates that, in general, performance on individual items correlates positively and highly with the scores on the other items within the check. This is consistent with the Cronbach's alpha identified during the pilot and in the 2012 live sample. Values of Cronbach's alpha of more than 0.9 are generally considered excellent. However, due to the nature of items in the phonics screening check, single words to be read by a child are likely to lead to high values of alpha because of their similarity. The Cronbach's alpha values found in the year 2 sample are slightly lower than those in the year 1 sample. This is likely because the cohort taking the check in the second year is only a portion of the overall year 2 population.

Another indication of the reliability of the phonics screening check is the standard error of measurement. The standard error of measurement is an estimate that allows the user to determine a confidence interval around an observed score. In the case of the 2013 phonics screening check the standard error of measurement is 2.0 in the year 1 sample and 2.1 in the year 2 sample. This means that we can be 95 per cent confident that a child's 'true score' is within five marks of their observed score. This is an indicator of the quality of the assessment and consistent with the standard error of measurement calculated from the 2012 assessment.

As is to be expected, given the specification, children performed better on section 1 than on section 2 of the phonics screening check. The Cronbach's alpha for section 1 is lower than section 2. However, since large numbers of children are scoring high marks on

section 1, there is less opportunity for the section to discriminate between higher and lower performers; hence a slightly lower value of Cronbach's alpha is to be expected.

Figures 4 and 5 show the distribution of total score for the year 1 sample and year 2 sample. The year 1 distribution is similar to that seen in the national data, published in the Statistical First Release in September 2013[6].

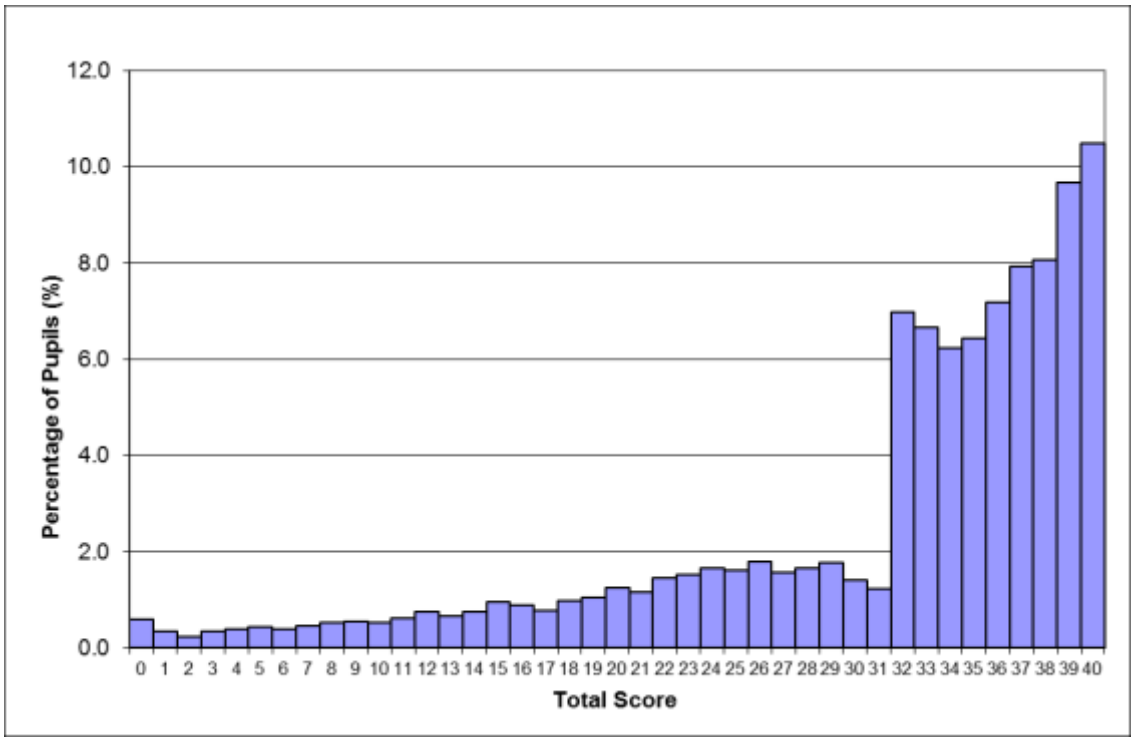

Figure 4 Total score distribution for Year 1 sample

[6] https://www.gov.uk/government/publications/phonics-screening-check-and-national-curriculum-assessments-at-key-stage-1-in-england-2013

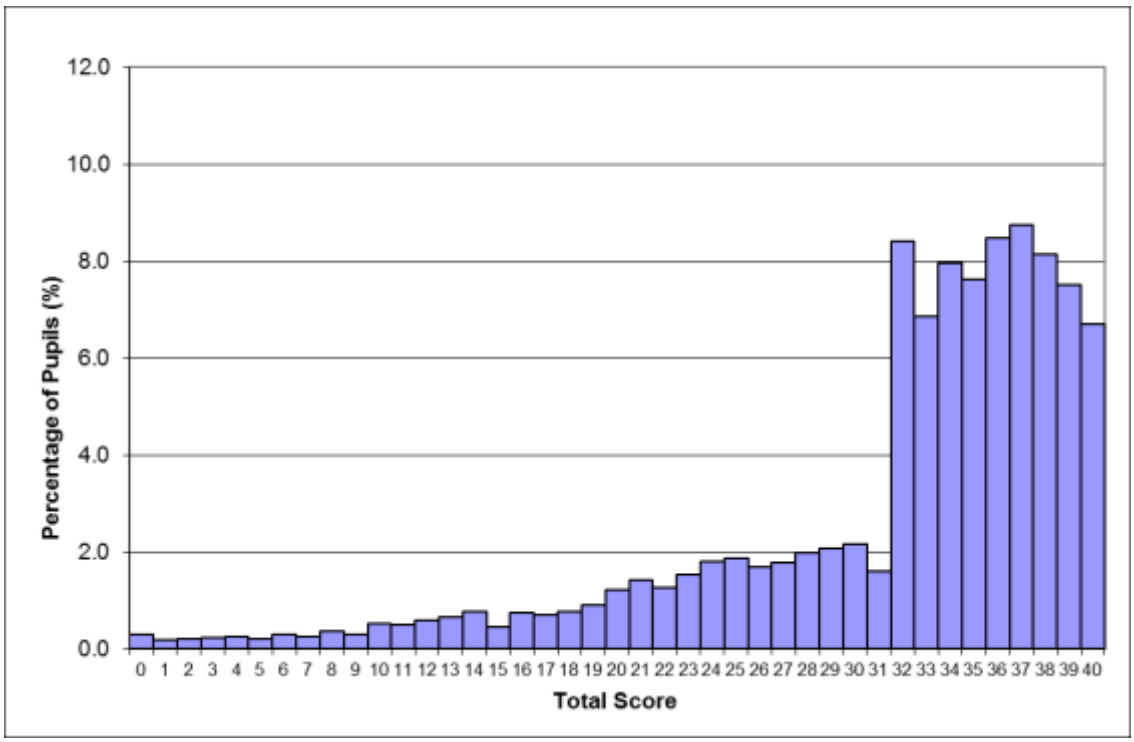

Figure 5 Total score distribution for Year 2 sample

There are some small differences in the distributions between the year 1 sample and the year 2 sample. For year 1 children the most common mark scored was 40 and for year 2 children it was 37, with peaks also at 32, which was the threshold on the 2013 phonics screening check. Seventy per cent of both year 1 and year 2 children taking the check achieved the expected standard or higher. In both the trial and the live samples, there is a steady negative skew which means that most of the children in both samples were in the upper end of the distribution.

As seen in 2012, there is a difference between the score distribution seen in the live administration of the phonics screening check and that seen in the pilot in that a spike in the distribution was not observed in the trial. This is due to the fact that an expected standard was not available at the time of trialling, while teachers were provided with the threshold mark in the scoring guidance for the live check. A Topic note[7] on the 2012 phonics screening check results, published in May 2013, provides a full discussion of the peak at the threshold.

The purpose of the phonics screening check is to identify children who might need further support in order to catch up. The results of the check should be used in line with the purpose of the check which means that if a child has not met the expected year 1 standard, then the school should consider what extra support the child needs to improve

[7] https://www.gov.uk/government/publications/evaluation-of-the-phonics-screening-check-first-interim-report

their decoding knowledge. The level of support should be decided by the school, taking into account the child's precise score on the check, and other information about the child's reading. The unusual shape of the distribution may lead some to conclude that there is an element of misclassification. From the data available, it is not possible to determine whether there has been deliberate misclassification and if so, the cause. An interpretation of the area around the threshold peak is consistent with teachers accounting for potential misclassification in the check results, and using their teacher judgment to determine if children are indeed working at the expected standard. Misclassification around the threshold, for any test, threatens the interpretation of the outcome for those children near the threshold. Classification accuracy will be examined later in the report.

Figure 6 shows the facilities and discriminations, calculated using classical test theory methods, for items in the check.

Item	Year 1 sample		Year 2 sample	
	Facility	Discrimination	Facility	Discrimination
fot	97.4	0.289	97.2	0.275
keb	91.1	0.413	89.7	0.397
gan	96.7	0.327	96.6	0.332
ulp	90.1	0.532	89.6	0.449
poth	90.6	0.554	91.2	0.454
shan	94.8	0.489	95.5	0.387
veen	94.0	0.506	94.1	0.419
quorg	77.1	0.530	72.0	0.441
drap	89.3	0.513	90.0	0.440
flarm	79.7	0.686	81.0	0.623
lect	91.0	0.528	91.4	0.487
voisk	68.2	0.621	64.5	0.549
thin	91.2	0.533	93.2	0.474
peck	91.6	0.449	91.0	0.391
torn	83.2	0.611	85.3	0.562
cheek	90.5	0.526	90.2	0.446
trap	91.7	0.517	93.5	0.470
snarl	73.5	0.688	73.9	0.604
milk	93.7	0.570	96.1	0.474
moist	72.2	0.619	68.0	0.559
quigh	66.2	0.512	59.6	0.407
herks	76.3	0.633	76.9	0.575
jorb	81.2	0.633	82.9	0.578
zale	61.5	0.612	62.2	0.536
bluns	83.2	0.615	84.8	0.548
skarld	68.5	0.638	68.2	0.556
splot	87.0	0.643	89.0	0.568
strabe	48.2	0.536	44.5	0.462
toy	90.3	0.656	94.5	0.558
spike	64.7	0.666	68.0	0.612
fuel	36.5	0.452	37.5	0.397
name	81.8	0.733	88.3	0.658
props	76.5	0.617	81.7	0.563
spoilt	71.5	0.656	70.9	0.601
scram	77.0	0.496	77.6	0.419

Item	Year 1 sample		Year 2 sample	
	Facility	Discrimination	Facility	Discrimination
strike	61.6	0.648	62.5	0.595
panic	71.1	0.529	71.5	0.486
second	68.4	0.641	73.7	0.595
tantrum	55.4	0.556	60.3	0.524
reaching	65.5	0.682	73.2	0.634

Figure 6 Classical item statistics

For one-mark items, such as those in the check, facilities are equivalent to the percentage of children who answered each item correctly. Discrimination relates to the ability of an item to differentiate between high and low performers, specifically, the relationship between child performance on an item and their total score. Items with high discrimination will help ensure that children are appropriately classified as having met or not met the expected year 1 standard. Items with low discrimination will tend to lead to increased misclassification. It should be noted that the calculated discriminations are corrected point biserial correlations, as such values greater than 0.30 are acceptable.

In the year 1 sample the facilities range from 36.5 (fuel) to 97.4 (fot). In the year 2 sample the facilities range from 37.5 (fuel) to 97.2 (fot). As expected, the facilities are generally higher for words in section 1 than for words in section 2. Comparing real and pseudo-words of similar structure (that is excluding the first page of three letter pseudo-words and the last page of two syllable real words), in the year 1 sample the average facility for pseudo-words was 81.6 and the average facility for real words was 75.4. In the year 2 sample the figures were 81.1 and 77.5. This is different from the pattern observed in the pilot and 2012 live sample, where the facilities for pseudo-words were lower than those for real words. This could be an indication that there is now more emphasis in the classroom on decoding pseudo-words.

The discriminations are generally good or very good. The discriminations for the first few items are lower, although still acceptable. This is to be expected given that the facilities for these items are so high, leaving little opportunity to discriminate between high and low performers. The discriminations are generally higher in the year 1 sample than in the year 2 sample. This is also to be expected given that only those who did not pass the check in year 1 were included in the year 2 sample.

Results by subpopulation

The Statistical First Release provides detail on the outcomes of the check by the subpopulations of gender, children for whom English is an additional language (EAL) and children with special educational needs (SEN). This does not conflict with the conclusions regarding subpopulations and minimising bias from the previous technical reports.

Item response theory

A two-parameter item response theory (IRT) model was estimated using the software package Mplus v7.11[8].

Other IRT models are available, however, the two-parameter model is considered to be the most suitable in this context as estimating both difficulty and discrimination is meaningful and it is clear that estimating discrimination is the most appropriate route because of the range of values obtained. This makes the one-parameter model less appropriate because only difficulty is estimated. Estimating a lower asymptote parameter in a three-parameter model is possible but meaningful interpretation of this parameter in this context is unclear.

IRT assumption checking

There are two main assumptions in item response theory: unidimensionality and local independence. The assumption of unidimensionality suggests a single underlying construct in the data, commonly referred to as ability. In the case of the phonics check it would be the ability to decode using phonics. The assumption of local independence assumes that the items are not related to each other except through child ability. It is well established that IRT is robust to minor violations of these assumptions; and that it is important to evaluate these assumptions.

The assumption of local independence was tested using Yen's Q3 statistic. For any pair of items the Q3 statistic is calculated as the correlation between the extent to which children achieve above or below their expected score given their ability on one item and the extent to which they achieve above or below their expected score on the other item. The estimates of ability for each child and the item parameters derived from the IRT model were used to calculate the expected score on each item for each child. From this, the difference between the expected score and actual score and the correlations between these differences were calculated. For the assumption of local independence to be upheld these correlations should be close to zero. The average Q3 statistic for all 40 items in the check was -0.02 both samples, indicating that the degree of violation of local independence is relatively small.

Unidimensionality was tested with factor analysis and was found to be well within expectations of good model fit for a unitary construct. Hu and Bentler (1999)[9] recommend that model fit be considered good if the Tucker Lewis Index (TLI) is not less than 0.95 and the root mean square error of approximation is not more than 0.05. The TLI and RMSEA values were largely within these recommendations - for the year 1 sample the

[8] Muthén, L. K., & Muthén, B. O. (1998-2014). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
[9] Hu, L-t and Bentler, P. M. (1999). Cut-off criteria for fix indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modelling, 6(1), 1-55.

TLI was 0.95 and the RMSEA was 0.05. The figures for the year 2 sample were 0.93 and 0.045.

The evidence presented on the IRT assumptions supports the use of IRT to analyse the phonics data. With respect to item fit, Yen (2006) advises that 'definitive conclusions about the best way to measure item fit cannot yet be drawn' and that large sample sizes increase the number of items misfitting. Examining item fit graphically shows that the vast majority of items fit the model. This provides further evidence of the appropriateness of the methods used.

The Department is therefore confident that the IRT model chosen fits the data and continues to be appropriate for the analysis of the phonics screening check data.

Results from IRT analysis

The scale on which the IRT operates is different from classical test theory and generally revolves around a mean ability of zero and standard deviation of one. The scale of item difficulty ranges from -3.37 to 0.45 for the year 1 sample and from -3.59 to 0.48 for the year 2 sample. This means that items with a value less than zero are less difficult than items with a difficulty greater than zero. The discrimination scale is a bit more difficult to interpret, but the general principle is, as with classical test theory, the larger the value the better. The scale of discriminations on the 2013 phonics screening check ranges from 0.77 to 2.21 for the year 1 sample and from 0.62 to 1.77 for the year 2 sample. Figure 7 shows the difficulty and discrimination from the IRT models for each item on the 2013 check.

Item	Year 1 sample		Year 2 Sample	
	Difficulty	Discrimination	Difficulty	Discrimination
fot	-3.37	0.81	-3.59	0.70
keb	-2.27	0.77	-2.24	0.70
gan	-3.09	0.84	-3.16	0.80
ulp	-1.80	1.06	-2.03	0.81
poth	-1.77	1.17	-2.15	0.85
shan	-2.24	1.21	-2.76	0.87
veen	-2.14	1.20	-2.50	0.87
quorg	-1.07	0.89	-1.01	0.67
drap	-1.81	0.97	-2.10	0.80
flarm	-0.89	1.65	-1.05	1.32
lect	-1.86	1.09	-2.03	0.95
voisk	-0.51	1.36	-0.44	1.14
thin	-1.85	1.13	-2.20	1.00
peck	-2.15	0.88	-2.38	0.71

Item	Year 1 sample		Year 2 Sample	
	Difficulty	Discrimination	Difficulty	Discrimination
torn	-1.20	1.23	-1.40	1.09
cheek	-1.83	1.07	-2.08	0.82
trap	-1.94	1.09	-2.24	1.00
snarl	-0.64	1.73	-0.75	1.26
milk	-1.92	1.53	-2.41	1.27
moist	-0.67	1.31	-0.56	1.14
quigh	-0.59	0.87	-0.42	0.62
herks	-0.83	1.30	-0.94	1.10
jorb	-1.06	1.30	-1.24	1.12
zale	-0.26	1.64	-0.35	1.16
bluns	-1.20	1.23	-1.43	1.00
skarld	-0.51	1.42	-0.59	1.06
splot	-1.35	1.44	-1.63	1.16
strabe	0.10	1.47	0.21	1.09
toy	-1.50	1.79	-2.02	1.55
spike	-0.32	2.11	-0.48	1.55
fuel	0.45	1.24	0.48	0.92
name	-0.91	2.21	-1.35	1.77
props	-0.86	1.24	-1.21	1.05
spoilt	-0.60	1.49	-0.63	1.27
scram	-1.15	0.80	-1.38	0.64
strike	-0.24	2.06	-0.31	1.56
panic	-0.76	0.93	-0.86	0.81
second	-0.50	1.49	-0.76	1.22
tantrum	-0.12	1.26	-0.31	1.03
reaching	-0.34	2.11	-0.66	1.53

Figure 7 IRT item statistics

Differential item functioning

Differential item functioning (DIF) was examined using a sub-sample of the data. Group differences in item difficulty were calculated for gender (boy / girl), as this was the only background characteristic that was collected.

Seven items exhibited negligible DIF in the year 1 sample and six exhibited negligible DIF in the year 2 sample. These are shown in Figure 8. There are no clear explanations for the differential item functioning of these items. Only two items exhibited DIF in both samples. The existence of DIF only indicates that sub-groups appear to respond differently from each other relative to what would be expected, it does not necessarily mean that the items are biased.

Item	Favours	Significance	Sample
drap	girls	5%	Year 1
thin	girls	5%	Year 1
milk	girls	5%	Year 1
fuel	boys	1%	Year 1
scram	boys	5%	Year 1
strike	boys	5%	Year 1
reaching	girls	5%	Year 1
veen	boys	5%	Year 2
herks	boys	5%	Year 2
props	girls	5%	Year 2
spoilt	girls	5%	Year 2
scram	boys	5%	Year 2
reaching	girls	5%	Year 2

Figure 8 Items displaying DIF

Classification accuracy

Classification accuracy refers to how precisely children have been classified. Various methods of estimating classification accuracy have been developed, both under classical test theory and item response theory. Two procedures appropriate for the 2013 phonics check (a single administration of dichotomously scored items) have been used to estimate the classification accuracy on the probability scale from 0 to 1.

The software BB-CLASS (Brennan, 2004)[10] was used to implement the Hanson and Brennan (HB, 1990)[11] procedure. This is a procedure based on classical test theory. The classification accuracy index obtained from the HB procedure in BB-CLASS was 0.947 for the year 1 sample and 0.938 for the year 2 sample.

The software IRT-CLASS (Lee and Kolen, 2008)[12] was used to implement the Lee (2008)[13] method. This is based on item response theory. The classification accuracy index obtained from IRT-CLASS was 0.934 for the year 1 sample and 0.921 for the year 2 sample.

The two values from the two procedures are very similar, and suggest that the probability that a child is misclassified is less than eight per cent. This is in line with the findings in the 2012 technical report. There also does not appear to be much difference between classification accuracy for the year 1 and year 2 samples. This provides evidence to support the statement that the test is sufficiently reliable for year 2 as well as year 1.

[10] Brennan, R. L. (2004). 'BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy' (Version 1.0) (CASMA Research Report No. 9). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from www.education.uiowa.edu/casma).
[11] Hanson, B.A., & Brennan, R. L. (1990). 'An investigation of classification consistency indexes estimated under alternative strong true score models'. Journal of Educational Measurement, 27, 345-359.
[12] Lee, W., & Kolen, M. J. (2008). *IRT-CLASS:* A computer program for item response theory classification consistency and accuracy (Version 2.0) Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from www.education.uiowa.edu/casma).
[13] Lee, W. (2008). Classification consistency and accuracy for complex assessments using item response theory (CASMA Research Report No. 27). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from www.education.uiowa.edu/casma).

Conclusion

This section of the report will focus on synthesising the analysis presented above to provide evidence of validity and reliability as set out in Ofqual's Regulatory framework for national assessment arrangements (Ofqual, 2011[14]).

The DfE has stated that the purpose of the phonics screening check is to confirm whether or not children have learned phonic decoding to an expected year 1 standard such that those children who have not met that standard are provided with additional support to catch-up. The level of support should be decided by the school, taking into account the child's precise score on the check, and other information about the child's reading. Children who did not meet the expected standard in year 1 were required to re-take the check in year 2.

The Ofqual regulatory framework for national assessments (2011[15]) states that an assessment should 'generate outcomes that provide a valid measure of the knowledge, skills and understanding that the learner is required to demonstrate as specified by the assessment objectives'. The DfE believes that the evidence from the pilot and subsequent analysis of 2012 and 2013 item level data analysis provides sufficient evidence that the check is a valid assessment of phonic decoding.

The one outstanding question relating to validity identified previously: Are children who have not met the expected standard on the phonics screening check in need of additional support? Will be addressed in the independent evaluation mentioned in the Introduction.

The conclusions here will focus on the comparisons of item level data from two nationally representative samples, one for year 1 and one for year 2, of children who took the phonics screening check in June 2013. The similarity between the item level results for year 1 and year 2 as well as comparing back to 2012 provides the DfE with the evidence that the check performed as it was designed, for both year 1 and year 2 children.

The total score distribution for the year 1 sample shows the most common score was 40 marks, as it was in 2012. The total score distribution for the year 2 samples shows the most common score was 37. In both distributions there was a rise in the number of children gaining marks from the point of the expected standard mark.

The item statistics across analysis methods show that the first few items do not discriminate as well as later items. This is likely to be because the items are designed to be easier than later items in order to ease the children into the check. This is a pattern identified in 2012 and has maintained in 2013. The facilities were of the expected range, section 1 items were slightly easier than those in section 2. Interestingly this year, in both year 1 and year 2 samples, pseudo-words had a higher average facility than real words.

[14] www2.ofqual.gov.uk/files/2011-regulatory-framework-for-national-assessments.pdf
[15] Ibid.

This could be an indication that there is now more emphasis in the classroom on decoding pseudo-words.

There were seven items that exhibited negligible DIF in the year 1 sample and six items that exhibited negligible DIF in the year 2 sample. However, there did not appear to be any substantive explanation for the difference and therefore no evidence of bias was found in these items.

The Ofqual Regulatory framework for national assessments (2011[16]) states that an assessment should 'generate outcomes that provide a reliable measure of a learner's performance' and that:

> Reliability is about consistency and so concerns the extent to which the various stages in the assessment process generate outcomes which would be replicated were the assessment repeated. Reliability is a necessary condition of validity, as it is not possible to demonstrate the validity of an assessment process which is not reliable. The reliability of an assessment is affected by a range of factors such as the sampling of assessment tasks and inconsistency in marking by human markers.

To demonstrate sufficient reliability for the phonics screening check, the following aspects have been considered:

- internal consistency;
- classification consistency;
- classification accuracy; and
- consistency of scoring.

The internal consistency reliability in the form of Cronbach's alpha was high, indicating strong interrelationships between the items. While this is good news, it is important to examine other measures of reliability, for example, the standard error of measurement. The standard error of measurement is an estimate that allows the user to determine a confidence interval around an observed score. In the case of the two samples collected of the 2013 phonics screening check the standard error of measurement is 2.0 for the year 1 sample and 2.1 for the year 2 sample. This means that we can be 95 per cent confident that a child's 'true score' is within five marks of their observed score. This is likely to have a greater impact for children close to the threshold. The similarity of the values across the two samples suggests that the assessment is providing reliable results in those two year groups.

Classification accuracy was calculated using two different methods and provided similar results, which indicated that less than eight per cent of children would have been misclassified in the phonics screening check in either sample. This is the same level of classification accuracy as was found in the 2012 live data.

[16] Ibid.

Having examined all of the evidence gathered so far through the pilot and the live samples over two years, the DfE is satisfied that the phonics screening check is sufficiently valid for its defined purpose and has acceptable levels of reliability. Given the evidence presented in this report regarding the performance of year 2 children, the DfE is satisfied that the check is sufficiently valid and reliable for these children as an extension of the original purpose.

Standards & Testing Agency

Any enquiries regarding this publication should be sent to us at assessments@education.gov.uk.

This document is also available from our website at: www.education.gov.uk.