Standards & Testing Agency

Phonics screening check

2012 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 Year 1 phonics screening check via a quantitative item analysis of the data from the live 2012 administration. This document should be considered alongside the first technical report on the pilot that was published in February 2012[1] and the Statistical First Release[2] published in September 2012. Further information will be provided in the Topic note on the 2012 phonics screening check results, due to be published in Spring 2013.

The Department has commissioned an independent evaluation of the phonics screening check over the next three years. This will provide valuable information about the impact of the check on phonics teaching. The first interim report will be available in spring 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 Year 1 phonics screening check, including schools.

[1] http://www.education.gov.uk/schools/teachingandlearning/assessment/keystage1/a00200415/phonics
[2] www.education.gov.uk/researchandstatistics/datasets/a00213773/phonics-screening-ks1-england-2012

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 (referred to as 'the check' or 'phonics 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.

The phonics 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 February 2012 concluded that 'Having examined all of the evidence gathered through the pilot, the Department is satisfied that the Year 1 phonics screening check is sufficiently valid for the defined purpose, with acceptable levels of reliability, which is fair for children and manageable for schools. However, as stated previously, additional analysis will be carried out to ensure that the Department can be more confident in this assertion.'

To conduct this analysis and provide further evidence, STA collected a sample of item level data from schools taking the phonics screening check in June 2012.

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

check, as set out in Ofqual's Regulatory framework for national assessment arrangements (Ofqual, 2011[3]).

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

Sample selection

A sample of maintained schools with Year 1 children was drawn using data from the autumn 2011 school census and Edubase. The sample was stratified by region and Key Stage 1 attainment in reading (based on data from 2011). The achieved sample contained 12,190 children from 313 schools.

Figure 1 shows the representativeness of the sample compared to the population across Key Stage 1 attainment, type of establishment, and region. The sample is representative of the population of schools taking the phonics check in 2012.

		Population		Phonics sample	
		Count	%	Count	%
Average 2011 Key Stage 1 reading point score	Lowest 20%	3249	20.2	58	18.5
	2nd lowest 20%	3331	20.7	69	22.0
	Middle 20%	3199	19.9	63	20.1
	2nd highest 20%	3135	19.5	61	19.5
	Highest 20%	3010	18.7	60	19.2
	Missing data	157	1.0	2	0.6
Type of establishment	Academy converters	321	2.0	5	1.6
	Academy Free Schools	11	0.1	0	0.0
	Academy sponsor led	36	0.2	1	0.3
	Community school	8896	55.3	188	60.1
	Community special school	491	3.1	4	1.3
	Foundation school	436	2.7	11	3.5
	Foundation special school	16	0.1	0	0.0
	Non-maintained special school	16	0.1	0	0.0
	Voluntary aided school	3513	21.8	68	21.7
	Voluntary controlled School	2342	14.6	36	11.5
	LA nursery school	3	0.0	0	0.0
Government office region	East Midlands	1535	9.5	31	9.9
	East of England	1868	11.6	35	11.2
	London	1699	10.6	33	10.5
	North East	875	5.4	17	5.4
	North West	2448	15.2	47	15.0
	South East	2355	14.6	45	14.4
	South West	1805	11.2	38	12.1
	West Midlands	1749	10.9	34	10.9
	Yorkshire and the Humber	1747	10.9	33	10.5
Total		16081	100.0	313	100.0

Figure 1 2012 Year 1 Phonics sample representation

Summary statistics

Whole check statistics

Figure 2 shows the summary check performance from the children in the sample. The average score is nearly three quarters of the total marks. The average score for each section is over half marks.

	Whole check	Section 1	Section 2
Number of children	12190	12190	12190
Mean score	29.25	16.87	12.38
Standard deviation	10.30	4.48	6.30
Cronbach's alpha	0.959	0.923	0.935
Standard error of measurement	2.1	1.2	1.6

Figure 2 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. 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.

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 2012 phonics screening check the standard error of measurement is 2.1. 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 consistent with the standard errors of measurement identified during the pilot and suggests that the pilot seems to have been a good indicator of the quality of the live assessment.

As is to be expected, given the specification, children performed better on section 1 than on section 2 of the check. 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.

Figure 3 shows the distribution of total score. This distribution is the similar to that seen in the national data, published in the Statistical First Release in September 2012[4].

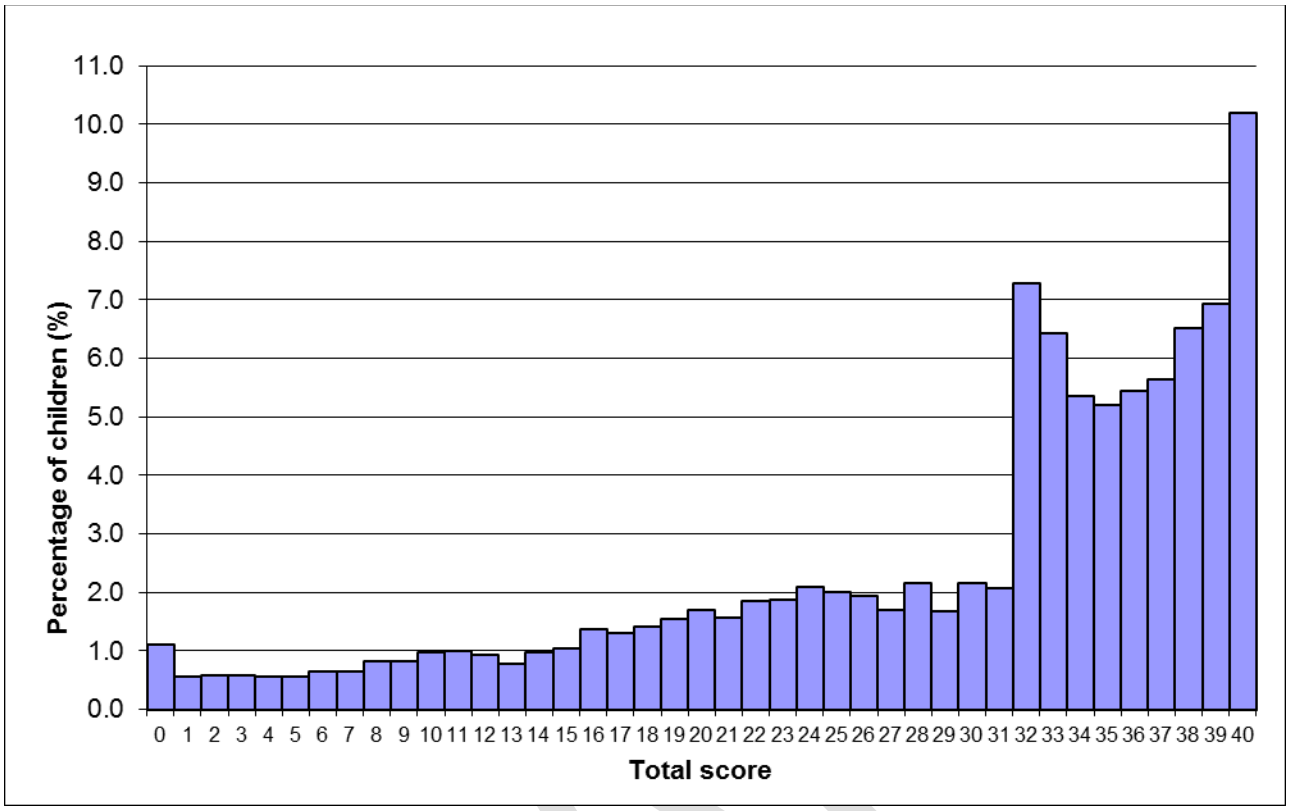

Figure 3 Total score distribution

The most common mark scored was 40, with another peak at 32, which was the threshold on the 2012 phonics screening check. Fifty eight per cent of children taking the check achieved the expected standard or higher. There is a difference between the score distribution seen in the live administration of the check and those seen in the pilot in that a spike in the middle of 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. In both the trial and the live sample, there is a steady negative skew which means that most of the children in both samples were in the upper end of the distribution.

The purpose of the 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 standard, then the school should consider what extra support the child needs to improve 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. 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

[4] www.education.gov.uk/researchandstatistics/datasets/a00213773/phonics-screening-ks1-england-2012

children are indeed working at the expected standard. Classification accuracy will be examined later in the report.

An analysis of the peak at 32 has shown that compared to a smoothed distribution there may have been misclassification in approximately four per cent of children near the threshold. A full discussion of the peak at the threshold will be provided in in the Topic note on the 2012 phonics screening check results, due to be published in Spring 2013.

Figure 4 shows the facilities and discriminations, calculated using classical test theory methods, for items in the check.

Item	Facility	Discrimination	Item	Facility	Discrimination
pib	87.5	0.488	kigh	65.2	0.621
vus	93.1	0.404	girst	57.4	0.605
yop	93.8	0.376	baim	65.2	0.673
elt	87.0	0.582	yune	43.7	0.549
desh	88.0	0.583	flods	74.1	0.561
chab	86.5	0.526	groiks	47.4	0.569
poil	63.2	0.649	strom	72.8	0.508
queep	85.4	0.537	splaw	45.2	0.560
stin	84.3	0.658	fair	62.6	0.663
proom	86.1	0.600	flute	47.5	0.619
sarps	60.5	0.616	goat	76.2	0.702
thend	84.7	0.659	shine	58.2	0.697
chip	94.0	0.486	crept	76.9	0.584
jazz	91.4	0.522	shrubs	69.4	0.630
farm	80.7	0.695	scrap	77.3	0.605
thorn	76.0	0.675	stroke	48.2	0.621
stop	92.8	0.572	index	81.5	0.711
truck	84.0	0.616	turnip	59.4	0.671
jump	94.5	0.522	waiting	65.1	0.729
lords	74.1	0.738	portrait	44.5	0.604

Figure 4 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 standard. Item 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.

The facilities range from 43.7 (yune) to 94.5 (jump). 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), the average facility for pseudo-words was 69.4 and the average facility for real words was 75.2). This difference is similar to that found in the analysis of the pilot data.

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.

Results by subpopulation

The Statistical First Release provides detail on the outcomes of the check by the subpopulations of gender, EAL and SEN. This does not conflict with the conclusions regarding subpopulations and minimising bias from the previous technical report.

Further analysis on subpopulations will be reported in the Topic note, due to be published in spring 2013.

Item response theory

A two-parameter item response theory (IRT) model was estimated using the software package Mplus v5.2[5].

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 that we call 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 was calculated and the correlations between these differences. 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, indicating that the degree of violation of local independence is relatively small.

Unidimensionality was tested with confirmatory factor analysis and was found to be well within expectations of good model fit for a unitary construct. Bentler and Hu (1999) 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 within these recommendations - the TLI was 0.97 and the RMSEA was 0.048.

[5] www.statmodel.com

The evidence presented on the IRT assumptions clearly 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 is appropriate for the analysis of the Year 1 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 -2.47 to 0.23. This means that items with a difficulty 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 2012 phonics screening check ranges from 0.82 to 2.36. Figure 5 shows the difficulty and discrimination from the IRT model for each item on the 2012 check.

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
pib	-1.72	0.92	kigh	-0.43	1.30
vus	-2.31	0.90	girst	-0.18	1.34
yop	-2.47	0.86	baim	-0.38	1.67
elt	-1.44	1.25	yune	0.23	1.48
desh	-1.49	1.31	flods	-0.89	0.97
chab	-1.56	1.01	groiks	0.12	1.32
poil	-0.33	1.51	strom	-0.92	0.82
queep	-1.46	1.04	splaw	0.18	1.42
stin	-1.16	1.58	fair	-0.30	1.66
proom	-1.35	1.32	flute	0.12	2.07
sarps	-0.28	1.29	goat	-0.74	1.82
thend	-1.17	1.61	shine	-0.14	2.36
chip	-2.06	1.33	crept	-0.96	1.07
jazz	-1.81	1.24	shrubs	-0.59	1.24
farm	-0.94	1.78	scrap	-0.95	1.15
thorn	-0.77	1.55	stroke	0.10	1.97
stop	-1.75	1.78	index	-0.95	1.91
truck	-1.22	1.32	turnip	-0.20	1.81
jump	-1.96	1.76	waiting	-0.33	2.30
lords	-0.63	2.10	portrait	0.19	1.98

Figure 5 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.

Five items exhibited negligible DIF. These are shown in Figure 6. There are no clear explanations for the differential item functioning of these items. 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
sarps	girls	0.01%
farm	girls	5%
yune	girls	5%
index	boys	1%
turnip	boys	5%

Figure 6 Items displaying DIF

Classification accuracy

Classification accuracy refers to how precisely children have been classified. Reasonable estimates of classification accuracy are only valid now that the phonics screening check has been administered in all schools. Various methods of estimating classification accuracy have been developed, both under classical test theory and item response theory. Two procedures appropriate for the 2012 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)[6] was used to implement the Hanson and Brennan (1990)[7] procedure. This is a procedure based on classical test theory. The classification accuracy index obtained from the HB procedure in BB-CLASS is 0.940.

The software IRT-CLASS (Lee and Kolen, 2008)[8] was used to implement the Lee (2008)[9] method. This is based on item response theory. The classification accuracy index obtained from IRT-CLASS is 0.927.

The two values are very similar, and suggest that the probability that a child is misclassified is less than eight per cent. The predicted misclassification from the trial of the check in 2011 indicated around a ten per-cent misclassification rate, and the analysis from the live 2012 check suggests it is better than ten per cent.

[6] 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).

[7] 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.

[8] 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).

[9] 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[10]).

The Ofqual regulatory framework for national assessments (2011[11]) 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 Department believes that the evidence from the pilot provided sufficient evidence that the check was a valid assessment of phonic decoding.

There was one outstanding question relating to validity in the previous technical report: Are children who have not met the expected standard on the phonics screening check in need of additional support? This question was not able to be addressed in this analysis and will be discussed in the independent evaluation.

The Department has stated that the purpose of the phonics check is to confirm whether or not children have learned phonic decoding to an expected 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.

The analysis of the item level data from a representative sample of 12,000 children who took the phonics screening check in June 2012 provides the Department with the evidence that the check performed as it was designed.

The total score distribution shows the most common score was 40 marks, as it was in the 2011 trial. There is a rise in the number of children gaining marks from the point of the expected standard mark. There is a difference between the score distribution seen in the live administration of the check and those seen in the pilot, however, it should be noted, that most of the children in both samples were in the upper end of the distribution.

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. The facilities were of the expected range, section 1 items were slightly easier than those in section 2. As in the pilot, there was a small difference in facility between the real and pseudo-words with similar structures.

There were five items that functioned differently between boys and girls, but there did not appear to be any substantive explanation for the difference and therefore no evidence of bias was found in these items.

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

The Ofqual Regulatory framework for national assessments (2011[12]) 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 should be 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 2012 phonics screening check the standard error of measurement is 2.1. 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 therefore likely to have a greater impact for pupils close to the threshold.

Reasonable estimates of classification accuracy are only valid now that the phonics screening check has been administered in all schools. 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. It appears that teachers may have accounted for misclassification in approximately four per cent of children near the threshold and this is within the limits of classification accuracy.

Classification consistency and consistency of scoring were examined in the first technical report. Classification consistency refers to the extent to which children are classified in the same way in repeated applications of a procedure. Evidence from the check-re-check study conducted in the pilot indicated approximately 90 per cent of children had been consistently classified. Consistency of scoring relates to the extent to which children are classified in the same way when scored by different teachers. Evidence from the inter-rater reliability study conducted in the pilot suggested 92 per cent of children have been consistently classified.

[12] Ibid.

This leaves the question of potential misclassification of pupils near the threshold of 32 marks. As stated previously, compared to a smoothed distribution, approximately 4 per cent of pupils may have been misclassified. However, this figure is low and it is likely that the pupils concerned are working very close to the expected standard. As a result, although there may be a small amount of misclassification, we do not believe that this materially impacts the validity of the assessment.

Having examined all of the evidence gathered so far through the pilot and the live sample, the Department is satisfied that the Year 1 phonics screening check is sufficiently valid for its defined purpose and has acceptable levels of reliability.

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.