



On the sensitivity of the phonics screening check (Duff, Mengoni, Bailey & Snowling, 2015): erratum and further analysis.

Journal:	<i>Journal of Research in Reading</i>
Manuscript ID	Draft
Manuscript Type:	Brief Report
Keywords:	Phonics, Screening, Sensitivity, Specificity, Predictive Value

On the sensitivity of the phonics screening check (Duff, Mengoni, Bailey & Snowling, 2015): erratum and further analysis.

Abstract

Background: Duff et al. (2015) evaluated the sensitivity and specificity of the phonics screening check against two reference standards. This report aims to correct a minor data error in the original article and to present further analysis of the data.

Methods: Calculation of predictive values of the phonics screening check in addition to sensitivity and specificity, and evaluation of agreement between the reference tests.

Conclusions: 1) Predictive values are important indicators of screening test quality. 2) The positive predictive value of the phonics check is low (0.31) when compared to a standardised reading test but high (0.84) when compared to teachers' phonics phases judgements, reflecting poor agreement ($\kappa=0.27$) between reference tests. 3) Results have implications for practice in terms of choice of reference standard, and choice of threshold criterion for children to pass the screening check. Longitudinal data are needed to assess the predictive validity and utility of the check.

Highlights

1. What is already known about this topic:

- The importance of phonics in learning to read is widely acknowledged.
- The phonics screening check was introduced into UK schools in 2012 to ensure that all children develop phonic decoding skills.
- Estimates of the sensitivity and specificity of the phonics screening check, compared with two established 'reference' measures, were reported by Duff et al. (2015).

2. What this paper adds:

- We correct a minor error in the report of the original data by Duff et al. (2015).
- We draw attention to the importance of including predictive values, alongside sensitivity and specificity, in the evaluation of screening test validity. We also propose an alternative statistic for comparing the two reference measures.
- We show that applying this further analysis to the data in Duff et al. (2015) reveals: i) the numbers of incorrect (false-positive and false-negative) outcomes in the phonics check, and ii) the marked difference in these numbers depending on the choice of reference measure.

3. Implications for theory, policy or practice:

- Reports of screening test validity should include positive and negative predictive values.
- A fundamental consideration for evaluating the validity of the phonics screening check is the choice of reference measure.
- Longitudinal data are needed to assess the predictive validity and utility of the phonics check.

Children learning to read must develop the ability to recognise words by figuring out how the letter combinations (graphemes) they see in print are related to the sounds (phonemes) they hear when the word is spoken. This process of 'decoding' is widely regarded as a fundamental requirement in reading, to such an extent that a review by Rose (2006), undertaken for the UK government, said *"... words must be decoded if readers are to make sense of the text. Phonic work is therefore a necessary but not sufficient part of the wider knowledge, skills and understanding which children need to become skilled readers and writers ..."* In order to effect change in teaching practice, to ensure emphasis on the development of phonic skills, the phonics screening check for children in Year 1 was introduced in UK primary schools in June 2012 *"to confirm that all children have learned phonic decoding to an age-appropriate standard"* (Department for Education, 2012; p4).

The phonics screening check comprises 40 items, each of which is either a real word (e.g. day, shin, grit, best) or a pseudo-word (geck, blan, terg, fape). As can be seen from the examples here, pseudo-words (also called non-words) are letter strings that are not typically recognised as valid English words, but they can be pronounced. Non-words test the ability to decode from print-to-sound with relatively little benefit of real word knowledge. Children are tested individually towards the end of (UK) Year 1- the child is asked to say each word or non-word in turn, and every item is marked correct or incorrect. Thus, each child receives a score between 0 and 40 indicating the number of items named correctly. Next, a threshold is applied to determine a meet/miss outcome. Each year since the introduction of the test in 2012, the threshold value (set by Government) has been 32, meaning that a child who gets 32 or more items correct 'meets the standard', while a child who gets less than 32 items correct 'misses the standard'. The intention is that children who miss the standard at this stage should be given further support by the school to improve their phonic decoding skills.

It is the application of the meet/miss (pass/fail) criterion that makes this a screening test. Screening tests aim to determine the presence or absence of a condition of interest. A 'fail' (positive test) indicates that the condition is present and requires attention, whereas a 'pass' (negative test) indicates that it is absent. In health care, for example, a screening test for diabetes would use a blood glucose threshold. A person whose blood glucose level exceeds the threshold has a positive test and is considered to be diabetic, one whose blood glucose is below threshold has a negative test, and is considered not to be diabetic. In the case of the phonics screening check, a child who achieves a score of 32 or more has a negative test, suggesting that phonic decoding ability is good and does not

1
2
3 require further support. On the other hand, a score of <32 is a positive test, suggesting presence of a
4 decoding difficulty that requires further attention.

5
6 The analogy with health screening is useful because it highlights the importance of thinking about the
7 implications of a screening test result being wrong. In the diabetes example, a false-positive outcome
8 would be a person judged by screening to be diabetic but who, when tested further, turns out not to
9 be diabetic. This error may be undesirable, but it is at least safe. On the other hand, a false-negative
10 outcome, representing a person with diabetes who is not identified by screening, is clearly unsafe.
11 Both types of screening error have costs, so measures of screening test quality are generally
12 concerned with evaluating error rates in an effort to minimize one or both types of false result. Such
13 measures are most commonly expressed in terms of test *sensitivity* and *specificity*. This is the
14 approach taken in the study by Duff, Mengoni, Bailey & Snowling. (2015) on which this report is
15 focused. In health care however it is common practice, and beneficial, to report the diagnostic
16 accuracy of screening tests using other statistics in addition to sensitivity and specificity (Daly &
17 Bourke, 2000). The aims of this report are to extend the analysis of Duff et al. (2015) with these
18 additional, simple statistics and to discuss the implications of the results. Before introducing this
19 analysis, we must draw attention to a minor error in the reporting of the original data.

20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 **Duff, Mengoni, Bailey & Snowling (2015) - Erratum**

35 In order to evaluate the sensitivity and specificity of the phonics screening test, Duff et al. (2015)
36 compared it to two different reference standards: *"As with any new screening test, it is important to*
37 *compare its ability to identify risk for difficulties with that of more established approaches. Thus, the*
38 *classification function of the phonics screening check was compared with that of standardised*
39 *measures of reading, and a routine teacher assessment (phonic phase judgements)."* Data for these
40 comparisons are presented conventionally (Table 3, p117) as 2x2 tables representing the numbers of
41 children who meet and miss the threshold criterion on the phonics screening check versus the
42 standardised reading test in a sub-sample of 160 children (from full sample of 291), and on the
43 phonics screening check versus teachers' phonic phases judgements in the same sub-sample of 160
44 children (and also on the full sample of 291). The data given for the first of these are in Table 1.

Table 1. Incorrect data for Phonics Screening vs Standardised Test (Duff et al., 2015, Table 3)

Phonics Screening	sub-sample	Standardised Test		
		miss (+)	meet (–)	
	miss (+)	14	31	45
	meet (–)	2	133	135
		16	164	180

Note in Table 1 that the row and column sums in the table add to a total of 180, rather than the stated sub-sample size of 160. This is not immediately apparent in the published article, as these sums are not shown. The correct values for this table may in fact be deduced from details given in the text by Duff et al. (2015, p15) and these are shown in Table 2.

Table 2. Correct data for Phonics Screening vs Standardised Test (Duff et al., 2015)

Phonics Screening	sub-sample	Standardised Test		
		miss (+)	meet (–)	
	miss (+)	14	31	45
	meet (–)	2	113	115
		16	144	160

Fortunately, this minor error does not change the conclusions made by the authors. However, below we discuss an alternative way of looking at the data.

Sensitivity, specificity and predictive values

The principle underlying analysis of the sensitivity of phonics screening is that the quality of a screening test should be judged by comparing the miss (+) and meet (–) classifications obtained using the screening test with those obtained using some other ‘gold standard’ reference test. The reference test is assumed to represent the best possible measure of the true status of each individual. Presentation of the data in a 2x2 table enables us to visualize how well the classifications from the screening test compare with those from the reference test. Classifications that are the same on the two tests (miss/miss & meet/meet) indicate correct or ‘true’ screening test results, while those that are different (miss/meet & meet/miss) indicate incorrect or ‘false’ screening test results. All measures of screening test quality are therefore concerned with evaluating the numbers of true and false classifications, and the improvement of screening test quality is aimed at minimizing the numbers of

false classifications.

Sensitivity (Se) and specificity (Sp) are the most commonly reported measures of screening test quality. As noted above, Duff et al. (2015) give sensitivity and specificity values for comparisons of the phonics screening check with two different reference tests: a standardised reading test and teachers' assessments of phonic phases. Values reported, using the sub-sample of 160 children, are: 1) Phonics Screening Check vs Standardised Test: Se = 0.88 (14/16), Sp = 0.82 (133/164) – Table 1. Note: due to the Erratum corrected in Table 2 above, the correct specificity value should be Sp = 0.78 (113/144), but sensitivity is unchanged. 2) Phonics Screening Check vs Phonic Phases: Se = 0.60 (37/62), Sp = 0.92 (90/98) – Table 3.

Table 3. Data table (sub-sample) for Phonics Screening vs Phonic Phases (Duff et al., 2015)

		sub-sample		Phonic Phases		
				miss (+)	meet (–)	
Phonics Screening	miss (+)			37	8	45
	meet (–)			25	90	115
				62	98	160

Sensitivity and specificity values may be thought of as probabilities and, crucially for this discussion, the way in which we read each probability is determined by the denominator in the calculation. For example, in comparison with the standardised test, we see that Se = 0.88 (14/16). Here the denominator 16 is the total number of children who miss the criterion on the standardised test. Thus we read sensitivity as the probability that a child will miss the standard on the phonics screening *given* that they miss on the standardised test – the denominator indicates what we are given. Likewise, the (corrected) specificity Sp = 0.78 (113/144) is read as the probability that a child will meet the standard on the phonics screening *given* that they meet it on the standardised test.

The detail in these readings of sensitivity and specificity is important because, in practice, we are not given the numbers of children missing and meeting the reference standard. When the screening check is administered, teachers do not know which children will achieve the reference standard. Indeed, the point of the phonics screening is that it should be able to predict the results that would be obtained using the reference test. Therefore, we need to look at other probabilities. Our question is: "What is the probability that a child will miss on the standardised test *given* that they miss on the phonics screening?" That is, the denominator of the statistic needs to be the total number missing the

phonics screening standard. Here we see (Figure 2) that this number is 45 and that 14 of these also miss on the standardised test, thus the probability of interest is $PPV = 0.31$ ($14/45$). This probability (PPV) is called the positive predictive value. There is a corresponding negative predictive value (NPV), which is the probability that a child will meet the standardised test level *given* that they meet the standard of the phonics screening. From Figure 2 we calculate this to be $NPV = 0.98$ ($113/115$). Corresponding values of PPV and NPV may be calculated from Table 3 for comparison of the phonics screening with phonic phases judgements. For the purposes of further discussion, all the statistics presented above are summarized in Table 4.

Table 4. Statistics of phonics screening test quality compared with two reference tests.

Phonics Screening Check compared to:	Se	Sp	PPV	NPV
Standardised Reading Test	0.88	0.78	0.31	0.98
Phonic Phases (Teacher Assessments)	0.60	0.92	0.82	0.78

The merits of reporting predictive values alongside sensitivity and specificity are recognised by medical statisticians. Daly and Bourke (2000), for example, say: *“Although sensitivity and specificity may seem adequate in determining the validity of a ... test ... For the practising clinician or organizer of a screening programme, the predictive value of a test is a most important parameter”*. In Table 4 we see a characteristic that is not uncommon in practice where, with reference to the standardised reading test, the phonics screening test has high sensitivity and specificity yet its positive predictive value is low (see Daly & Bourke, 2000 for discussion of a similar example). Reporting of sensitivity and specificity alone may therefore give a misleading impression of how the test will perform in practice.

Implications for practice

1. Meeting the Standard of the Phonics Screening Check

If, for the moment, we consider the standardised reading test to be the reference/benchmark procedure for determining children at risk of reading difficulties, then Table 4 shows that the positive predictive value of the phonics screening check is low (0.31). This indicates that only 31% of the children who failed the phonics screening also failed the standardised test or, to put it another way, 69% of those who missed the standard on the phonics check met it on the standardised test (i.e. they were ‘false-positives’). So, rather than the phonics screening check slightly overestimating the number

of children at risk, as is implied by its relatively high (0.88) sensitivity, in fact it grossly overestimates the number at risk. On the other hand, the phonics check detects almost all those who are at risk (NPV of .98 indicates only 2% false-negatives). Arguably, this outcome is acceptable because the screening process is simple to administer and the chances of missing a child with genuine difficulty are low. However, there are other considerations to take into account. Let us begin by considering the distribution of reading skills in the population. It is reasonable to assume that reading, like other human traits, is normally distributed – 50% of children can be expected to score above and 50% below the mean, which some 68% falling within one standard deviation either side of average. Duff et al., (2015) found that the standard set for the Phonics Screening Check (32/40) was roughly equivalent to a standard score of 108 on the single word reading test they administered – i.e. about half a standard deviation *above* the mean for the UK population on which the test was standardized. One might ask – is it reasonable to expect *all* children to reach this standard? Might there be a limit to the capacity of some children to do so, at least by this age? Are we expecting too much of some children and indeed of their schools? This is not to argue against the need to monitor children's progress through the early phases of reading development – it is certainly important to do so. However, a single check is unlikely to be sufficient to ensure that all children are on track. Rather, we can predict, based on our knowledge of the predictors of individual differences in reading, that some children (e.g. those with oral language difficulties) will have difficulties meeting the standard – yet the screening check is silent with regard to fundamental weaknesses in language and cognitive skills which may be preventing such children from thriving in the phonics regime.

2. Adjusting the Phonics Screening Check Threshold

Ideally we desire a situation in which numbers of false-positives and of false-negatives following screening are minimal. However, in general, the pass/fail characteristic of screening tests means that we face a trade-off in these misclassifications. As false-positives increase (i.e. PPV decreases) so false-negatives decrease (NPV increases) and vice-versa. The balance may be adjusted by changing the pass/fail threshold. As noted earlier, the threshold for the phonics screening check has been fixed at 32/40 since its introduction. The implication of the low PPV value is that this threshold is too high; as suggested above, too many children are failing the test who are found not to be at risk when assessed by a standardised test. The effect of lowering the threshold would be to reduce the number of false-positives, but the penalty would be more false-negatives – children who pass the phonics

check but then perform poorly on the standardised test. In the light of this inevitable trade-off, the decision to be made in practice is whether it is preferable to have false-positives or false-negatives in phonics screening. Supporters of phonics screening would surely argue the former; that is, the test threshold should be high to identify all children whose phonics skill level may put them at risk, even if some of those who ‘fail’ the test are actually not at risk – the principle of ‘better safe than sorry’. But, while the cost of a simple screen seems low, we might question that assumption. If preparing pupils for the check involves a significant focus on phonics then there may be costs elsewhere in the curriculum and not least for the primary goal of reading, which is understanding. Furthermore, the costs of intervention are high if there are many children failing to meet the standard but who are on course to develop along normal lines. As the Rose report (2006) recommended, phonics should be taught systematically *within a language-rich curriculum*.

3. What is the appropriate reference test?

The above discussion begs the question – what is the appropriate reference standard against which to compare children’s reading progress? The statistics in Table 4 show a marked difference in how the phonics screening check performs against the two different reference tests. In relation to the standardised test, the phonics screening check has a high false-positive rate of 69% (1-PPV) but a very low false-negative rate of only 2% (1-NPV). In relation to the phonic phases assessment, however, the phonics screening check has a much lower false-positive rate of 18%, but now the false-negative rate has increased to 22%, the latter representing the proportion of children who meet the phonics screening standard yet miss the teachers’ phonic phases standard. It is apparent that the two established approaches, the standardised reading test and teachers’ phonic phases judgments, are setting quite different standards. This becomes evident when their classifications are compared directly (Table 5):

Table 5. Data table (sub-sample) for Phonic Phases vs Standardised Test (Duff et al., 2015)

sub-sample		Standardised Test		
		miss (+)	meet (–)	
Phonic Phases	miss (+)	15	47	62
	meet (–)	1	97	98
		16	144	160

From analysis of Table 5, Duff et al. (2015) report the sensitivity and specificity of phonics phases as

.94 (15/16) and .67 (97/144) respectively. We could add values of PPV and NPV as previously, but the assumption behind using all these measures is that we are comparing a screening test with an established reference 'gold' standard. In this context, however, both phonics phases and the standardised test have been used separately as reference standards for evaluation of the phonics check. Therefore, rather than consider that the standardised test is the gold-standard, which sensitivity and specificity analysis implies, we can take a neutral approach and just evaluate how well the two reference standards agree. This involves using a different statistic called a kappa coefficient, which can also be calculated easily from the data in the table above. Daly and Bourke (2000) say "*Kappa can be employed any time agreement between two qualitative tests is being sought ...*". Kappa values can be thought of as correlation coefficients, going from 0 (agreement no better than would occur by chance) to 1 (perfect agreement). In this case the value of kappa is .27, which would be interpreted as rather poor agreement.

So, the two reference standards used here are classifying children quite differently and a fundamental issue is which of the two is more valid. A clear position on this should precede discussion on the validity of the screening check threshold. One way of possibly resolving this issue can be found in Duff et al., (2015) who report that, the Phonics Screen correlates more strongly with a measure of nonword reading than does the Phonics Phases (not surprisingly). In contrast, the Phonics Phases correlate more strongly with a measure of reading comprehension. This may at least in part be because teachers' judgements about their pupils' phonics progression include assessments of knowledge not only of regular, consistent letter-sound correspondences, but also of inconsistent relationships which pertain to English reading and spelling. Their judgements are also based on observations made over a number of weeks and are hence more likely to reflect the totality of each child's competence. More worryingly, the finding of the strong correlation between the Phonics Screen and nonword reading might be taken to suggest that children are developing phonics as a 'splinter skill' divorced from real word reading whereas teachers may be more sensitive in their judgements to the range of skills which define a 'good reader'. Neither of these hypotheses receives direct support from the available data and hence they are speculative. To properly assess the utility of the Phonics Screening Check and the validity of the threshold set requires longitudinal data. We need to know of the longer term literacy outcomes of those who meet and fail to meet that standard. It is also important to ascertain how many children who failed to meet the standard received intervention, what

was the nature of that intervention, and what was their response to it.

Conclusions

1. When evaluating classifications of children on the phonics screening check against those on a reference test, we recommend that positive and negative predictive values should be reported alongside sensitivity and specificity.
2. The complements of the positive and negative predictive values are the probabilities of false-positive and false-negative errors respectively. An important decision for the teaching & research communities is which type of error should be minimized. The 'safer' approach is to minimize false-negatives, which is what the current phonics screening threshold tends to do, but the penalty for this is the need to accept the occurrence of false-positives.
3. The apparent validity of the phonics screening check in identifying children at risk of reading difficulties depends crucially on the standard against which it is compared. The data from Duff et al. (2015) show that teachers' classifications of children based on judgments of phonic phases agree poorly with classifications on a standardised reading test. Perhaps the most important question then, before asking 'How good is the phonics screening check?' is to ask 'What should be the reference standard?'
4. To fully assess whether the screening check is fit for purpose we need to know if it has achieved the stated aim of raising reading levels, how many children who failed to meet the standard received appropriate evidence-based intervention and what was their response to the intervention.

References

Daly, L.E., & Bourke, G.J. (2000). *Interpretation and uses of medical statistics*. 5th edition. Wiley-Blackwell.

Department for Education (2012). Assessment framework for the development of the Year 1 phonics screening check. Retrieved 18 July 2016, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230810/Phonics_assessment_framework.PDF

Duff, F.J., Mengoni, S.E., Bailey, A.M., & Snowling, M.J. (2015). Validity and sensitivity of the phonics

screening check: implications for practice. *Journal of Research in Reading*, 38(2), 109-123.

doi: 10.1111/1467-9817.12029

Rose, J. (2006). *Independent review of the teaching of early reading: Final Report*. London, Department for Education and Skills Publications.