SCREENING PHONICS IN ENGLAND: A CAUSE FOR CONCERN?

James Law and Thomas King

School of Education, Communication and Language Sciences

Newcastle University, UK

Abstract

Objective

Phonological awareness is an essential component of language development and in 2012 the English government introduced a new national assessment, the Phonics Screening Check, for all children in their second year of primary school. The Department of Education publishes annual returns (Statistical First Releases) of the results of the checks across the country. In this paper we reflect on these findings and compare the performance of different groups within the population.

Design

The first release of data was published by the Department for Education in September 2012 and this process was repeated in September 2013. In both cases these report presented secondary analyses of the distribution of scores from the Phonics Screening Check for the population aged 5-6 years of England. Here we take these analyses further by examining implications of graphs of the test distribution are discussed for those children in receipt of free school meals, those with pertinent identified needs (moderate learning difficulties, speech language and communication needs and autism spectrum disorders) at statement and School Action Plus levels and, finally, we ask to what extent the results varied for children born in different months of the year.

Results

We observe a consistent tetramodal distribution with both floor and ceiling effects and a dip and spike at the screen pass mark. The pattern across the two years is almost identical. The patterns vary somewhat for the different subgroups but it is very clear that those carrying out the screen are conscious of the screen pass mark and what it means and the curve suggests that children are "being given the benefit of the doubt" and pushed over the threshold of the test.

Conclusions

Notwithstanding teething troubles of the first national administration, it is clear that the guidance on the use of this test does not adequately counter the pressure to pass children. This leaves teachers in an unfortunate situation but as it is, it tells us that such early testing is not well integrated into teaching practice and is unlikely to be very accurate.

BACKGROUND

Controversy has surrounded the systems and structures for promoting phonics in English schools, and specifically systematic synthetic phonics, as the approved approach for introducing reading to all children as they enter primary school.. It is uncontentious that the development of early phonic skills is directly related to oral language and reading skills and that instruction to promote phonemic awareness helps children learn to read [1 2] and thus engage effectively in schooling. It seems logical therefore to try to identify children who are likely to have difficulties acquiring these skills as early as possible in a child's development with a view to introducing effective intervention. In 2012, the Department for Education (DfE) in England introduced a new national assessment for schools, the Phonics Screening Check (PSC) [3]. This is a universal screening test of word and non-word reading applied to children aged 5-6 years (Year 1) in maintained schools throughout England. The PSC is novel in two specific ways: it is the first national implementation of a screening test on a measure of educational development in English schools; leaving aside various attempts at introducing developmental screening within the health system in the preschool years[1], it is the youngest age at which a nationally standard, formal educational assessment has been made of children in England.

In developing and piloting the PSC, the DfE produced an extensive technical report [4] which evidences adherence to Ofqual [5] guidance on national tests. OFQUAL[2] has five specific criteria: validity, reliability, minimising bias, comparability and manageability which the report considers. In order to evaluate the assessment once it had been implemented, as opposed to in its pilot stage, a further technical report was commissioned which would investigate item response, including differences observed in success rates for words and non-words. A formal evaluation of the phonics teaching and screening programme was also commissioned from the National Foundation for Educational Research (NFER) for the period 2012-5 to report at stages, following through to outcomes at older ages; the first report of this commission appeared in May 2013 alongside a DfE publication on the topic. To support the introduction of the test extensive advice was independently produced by The Communication Trust [6] with particular relation to those with Speech, Language and Communication Needs (SLCN). Standard inter-rater and test-retest checks were carried out, leading to confidence (at 95% level) that any individual child assessment would be within 5 marks of a true value [4]. A feature of the check at its introduction was the fact that schools needed to report the results to parents, indicating whether the child's performance had fallen below the test threshold and indicating that additional support was warranted.

Screening for any condition is traditionally required to meet a set of criteria well recognised in healthcare but equally relevant in education [7]. The condition should be an important health (or educational) problem. There should have an (untreated) natural history that is adequately understood, there should be a recognizable latent or early symptomatic stage and a defined target population. Facilities for diagnosis and treatment should be available, there should be an agreed policy on whom to treat, and there should be an accepted treatment for those identified. The test used in screening should be suitable (simple, sensitive, specific, reproducible, validated, safe, and with a known distribution and cut-off points) and should be acceptable to the population. The cost of case finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on

[1] Developmental screening is based on the chronological age of the child whereas educational testing is based on the stage the child is at in the educational system.

[2] OFQUAL is the regulator for qualifications and examination standards in England.

care as a whole. Any screening program should be a continuing process and not a 'once and for all' project.

The technique for evaluating the PSC was rather different from the approach conventionally adopted in health context where performance on a screen is compared to performance on a "gold standard". This gold standard is conventionally a diagnostic test or an agreed combination of clinical judgement and measurement with the results reported in terms of a set of key indicators (specificity, sensitivity, positive and negative predictive ability and likelihood ratios). A screen threshold is determined by comparing the relative specificity[3] and sensitivity[4] at any given point on the measure. These tend to be reciprocally related with one rising as the other falls [8]. The threshold is then set following the application of Receiver Operating Characteristics or a comparable technique. There has been a long track record of developing such measures at least with regard to early language delays [9 10] but, to date, they have not been recommended for adoption because, while appealing at face value, their productivity figures are often considered problematic (sensitivity, on the whole, is too low) and too little is known about the effectiveness of intervention or indeed the natural history of the condition, traditionally understood as criteria for adoption of screening programmes.

To establish the pass mark for the PSC, a bookmarking approach was used [11]. Rather than comparing to another test as a gold standard, this uses an ordered list of items, ordered by difficulty to establish a suitable benchmark for achievement. Several iterations are followed of a process in which a suitable level is decided individually and then discussed with the group in order to establish a consensus of an appropriate standard. This was followed in two separate gatherings of teaching professionals and the two (similar) levels averaged to reach a suitable standard. This was translated across to the number of correct answers which a child would be giving if they were able to work at the agreed level, and a score of 32 out of 40 was agreed for the national implementation, from a range of 31-34 which were used for the pilots. Thus the bookmarking process relied on the professional knowledge of the teachers to agree what was a suitable level for children at that stage in school.

As is standard with Statistical First Releases, the data supporting the release was published alongside the release, giving figures for pass rates in LAs, different school types and for children with different characteristics such as having special educational needs (SEN) or in receipt of free school meals (FSM). However, the release also presented the distribution of test marks, as recorded by schools and transmitted by LAs, to allow understanding of the effect of choosing a certain pass/fail boundary on the pass rate. While the statistical release provides some 'commentary' on the data reported, there is little to comment on as the data are from the first year of the test with no past data to compare with the results. However, there is an expectation that LAs and schools will act on the results to investigate variation and to improve future performance so in future comparisons will be important.

Given the concerns about sensitivity mentioned above, it is important to examine the performance of the PSC for subgroups in the population who are more likely to be at risk of the type of difficulties that the screening check was intended to identify. One group which is clearly of significance is children who come from socially disadvantaged backgrounds, namely those in receipt of Free School Meals. We would also expect children with identified

[3] Specificity is the likelihood of an identified problem being a real problem. Low specificity results from a high number of false positives or over referrals.

[4] Sensitivity is the likelihood of detecting a problem when it does exist. Low sensitivity results missing cases identified by the gold standard leading to under referral.

special educational needs to perform very differently but we could not necessarily predict which groups (categories of primary need) would fare worse than others. One might assume that the pattern would be a function of the level of learning difficulty and those with more specific difficulties might do better. Finally, it is important to consider the age of the children. Whereas a health screen would traditionally be tied to the child's specific age, the PSC was administered at one point in the school year, raising concerns that the attainment of younger children would differ from that of older children, a phenomenon that has been well described elsewhere [12].

METHODOLOGY

Research questions

We address the following research questions:-

- To what extent does the PSC differentiate between those with and without difficulties in the use of phonics?
- To what extent is this pattern of performance replicated for children from socially disadvantaged backgrounds (i.e. in receipt of free school meals)?
- To what extent is the pattern different for key groups of children with special educational needs?
- How is the test result related to the age of the children?

Data and analyses

The data are drawn from the Statistical First Release (SFR) which reports data on 612,660 children in England in the period 2012-2013; the children are in their second year of compulsory schooling (Year 1) and will be generally six years of age (5;09-6;09).[13] Of this group 69% met the expected standard of 32/40 (compared with 58% in the first year); 73% of girls passed the PSC as compared to 65% of boys (a similar difference to the first year). Some eligible pupils (less than 1%) were absent throughout the test period and so did not take it. A relatively small number (around 2%) of children were disapplied from the test, i.e. did not start it. These proportions are both similar to the first year of the PSC.

When the statistical first release [14] reported the results of the national rollout of the screen, it contained a plot of the marks achieved by the number of children achieving that score. While the statistics originally provided by the DfE in the SFR provide a lot of information on outcomes for groups, we requested and obtained more detailed data on the distribution of marks for specific groups. This gave us data sufficient to examine the effects of free school meals status, SEN groupings and age (within school year) which showed some interesting patterns. To establish whether these would persist we requested data for 2012-2013, which is what we discuss here. Comparable data exist for the first year of the screen (2011-2012) and we include comparable graphs in an appendix to the main text.

RESULTS

A replication of original plot from the statistical release is provided as Figure 1. Since we were interested in the effect of SEN we have removed all children identified with a primary SEN designation[5] at this age (44,095).

[5] Children identified at School Action Plus (SA+) are required to have a primary need identified; SA+ corresponds to support for the individual child being sought from outside the school.

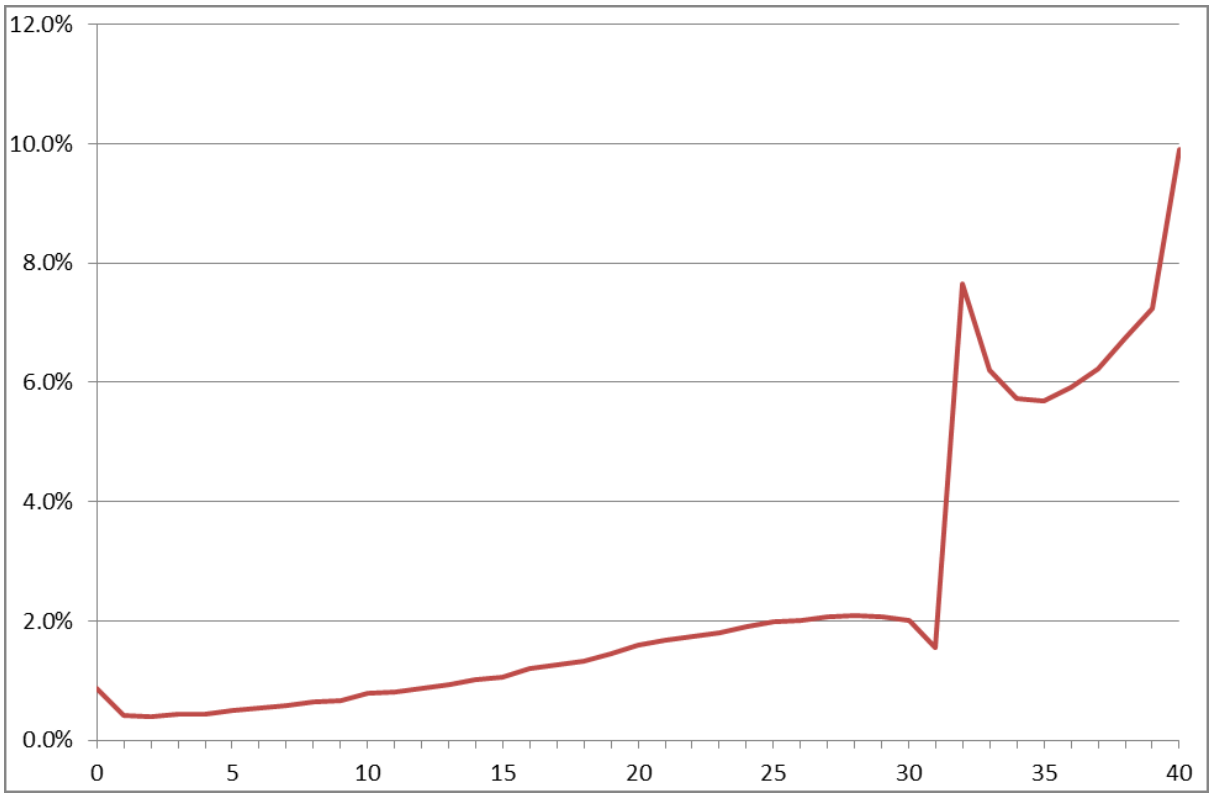

Figure 1 The Phonics Screening check results for the children without SEN (2013)

While we might predict a positively skewed distribution with greater differentiation at the lower end we see that the graph has a tetramodal distribution with three distinct patterns within the scores, namely a floor effect[6], a distinct spike at the pass mark before falling away and rising to give a ceiling effect. While even the release [14] remarked on the 'spike' at the pass mark, there are several other points to note about the plot. A steady increase with a mode somewhere towards the right of the distribution would be expected from the pilot experience [4] and the floor and ceiling are also expected. However, although the distribution follows this pattern to a score of around 28, it then peaks and even drops substantially to the score of 31 before spiking and falling away, the subsequent U-shape shows a larger than expected number of children scoring 39, if there was a simple ceiling effect.

We then looked at the comparable graph for those receiving Free School Meals (Figure 2). The pattern is identical at the spike although the proportion is higher at the lower stages and correspondingly lower at the higher end. Nonetheless the proportion of the children at the cut point is very similar at 7.5%. This is contrary to what might be expected given the understanding that difficulties in acquiring phonic skills are more common in more disadvantaged children. [15]

[6] A floor effect occurs when there are a group taking the test for which the level is too high and they all receive the same score, in this case zero, although they may not have equal ability. A ceiling is the corresponding effect at the top of the scale. Both are common in developmental tests.

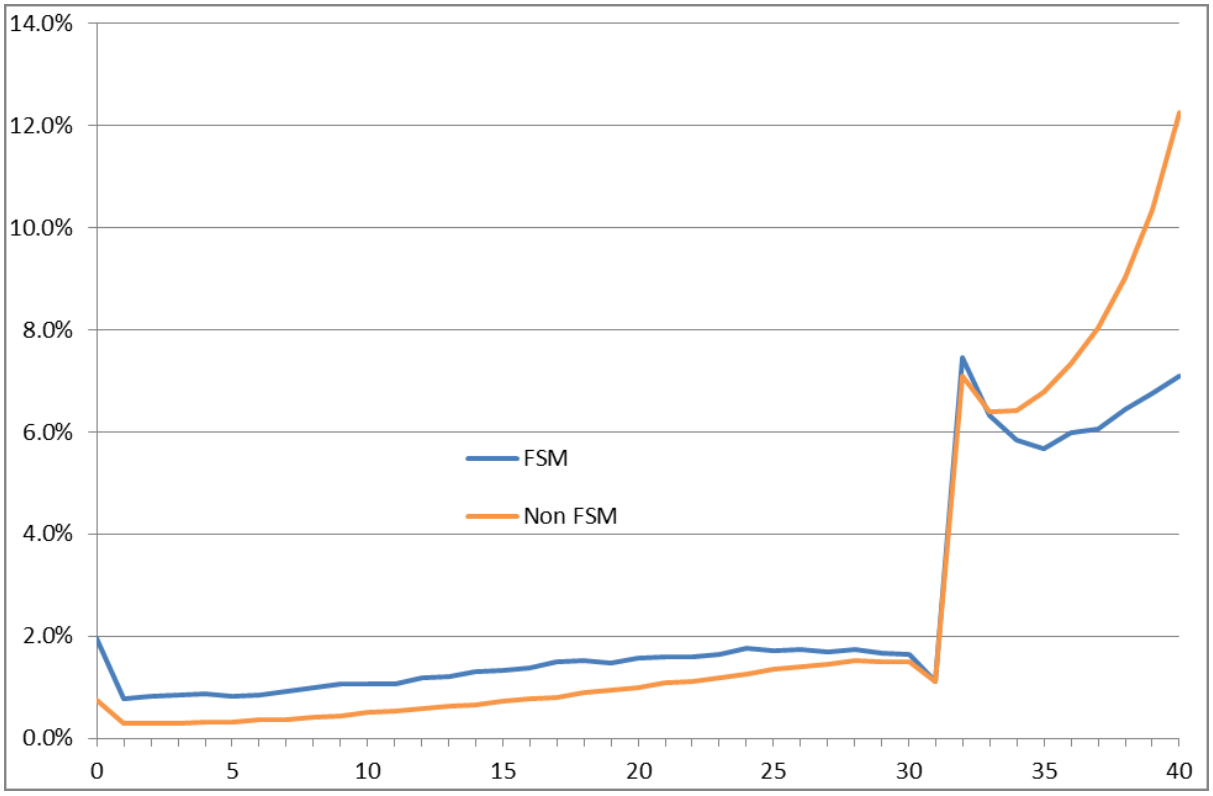

Figure 2 The Phonics Screening Check results by receipt of Free School Meals (FSM) for 2013

We then turn to the children identified as having SEN and have separated out three groups where we might predict there would be differences in their performance on the PSC (Figure 3) i.e. those with Moderate Learning Difficulty (MLD), Speech Language and Communication Need (SLCN), and Autism Spectrum Disorder (ASD). Firstly, note that the unevenness of the line reflects the lower sample size. Then it is important to observe that the floor effect for these groups is so substantial that it would dominate the graph and so the children scoring zero are omitted from Figure 3 but detailed in Table 1. Thus although children labelled with ASD score well if they are able to perform at all, a very large proportion register a score of zero, while children labelled with MLD follow a shallower profile which sees a large number of children scoring zero, one and two. The pattern is again broadly the same but the proportion at the peak is lower with distinct differences between the groups, ASD child fare better than SLCN and MLD respectively. The scores of the MLD group drop more dramatically than those of the other groups such that they appear at one point to be faring worse than the children with SLCN.

Floors, ceilings and disapplication are more important in considering the results for children with SEN. Of particular significance are the results for the children with ASD. Not only do they have few individuals with this label in the lower scoring groups but they obtain correspondingly higher scores above the threshold and are the only group to exhibit a ceiling effect. Given that many children with ASD also have learning difficulties and it is perhaps those that are scoring at the floor of the test this finding is noteworthy but that they appear to be doing better than other children with special needs at the higher end is of considerable interest. Those with ASD are also much more likely to have been disapplied at 33% of the cohort, compared to 9% with SLCN and 13% for MLD. Thus the category ASD is one of extreme variation, with 28% passing the PSC, but 42% either not being entered or scoring zero on the test.

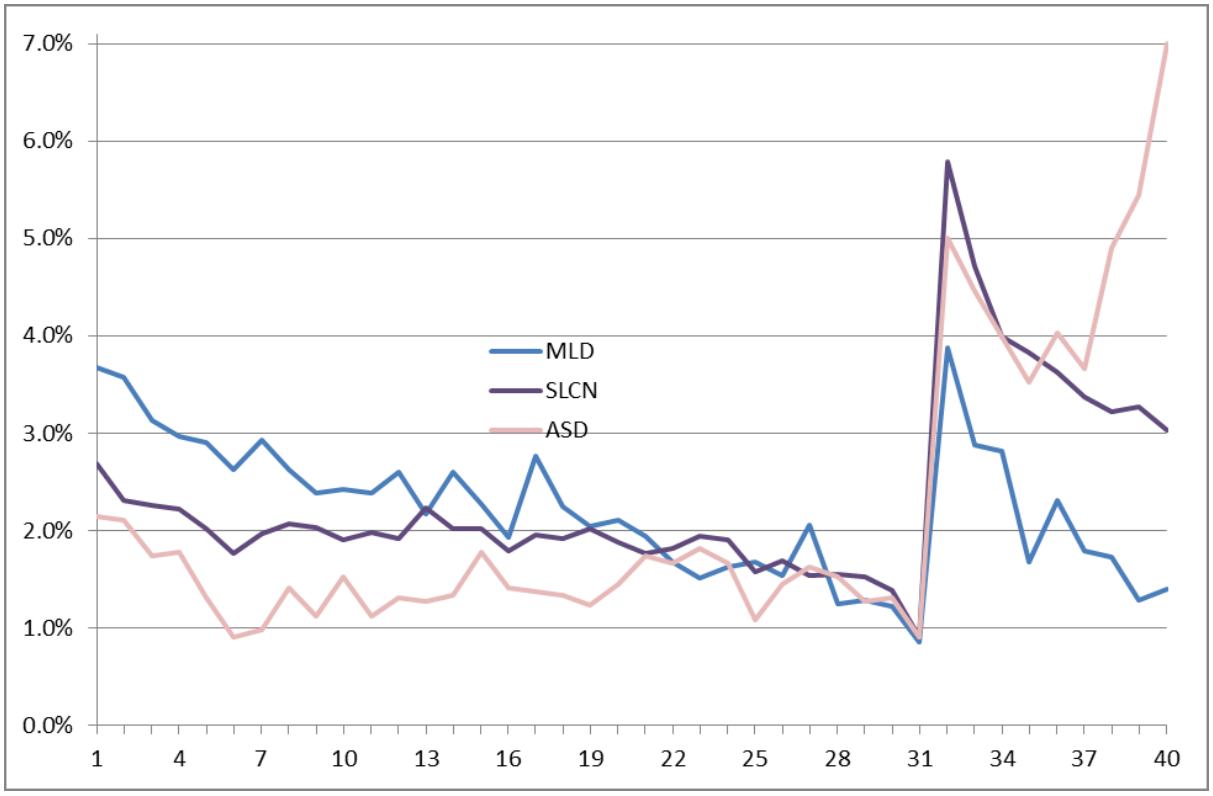

Figure 3 Phonics Screening Check results for children with certain designated disabilities (2013)

Table 1 SEN proportions of disapplication, zero scores and pass rates (2013)

Primary Need	MLD	SLCN	ASD	All SEN
Entered Zero[7] (%)	11	7	13	8
Disapplied (%)	13	9	33	16
Cohort Zero[8] (%)	10	6	9	7
Passed (%)	17	32	28	29
Total	5100	19013	4117	44095

Finally, we compare the performance of the children depending on their month of birth (Table 2). While we hope that there is little difference between development when children sit their final examinations and leave school 16-18, at the age of six, when the true age of the child will vary from 5;09 to 6;09 this is less plausible. To illustrate this we show the pass rate in age groups by month of birth which can be seen to decrease smoothly from 79% in the September born, through to 61% in the August born. Given that the threshold on the screen is absolute and given the exigencies of child development it is hardly surprising that younger children on the whole perform less well and this obviously has implications for those deemed to be in need to intervention.

Table 2 Trend in PSC pass rate with child's month of birth (2013)

Month of Birth	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Pass rate (%)	79	78	77	75	73	72	70	68	67	65	63	61

[7] The proportion of the children who were entered in the test who received a score of zero.
[8] The proportion of children from the whole cohort who received a score of zero as their result.

DISCUSSION

It would be very difficult to claim from Figures 1-3 that the PSC effectively differentiates between those with and without difficulties in phonics. The DfE topic report on the experience in 2012 [3] casts doubt on the reliability of scoring, reporting that the displaced scores for the whole group are around 12%, meaning that the pass rate should have been 46% (instead of 58%). The technical report of the rollout suggests that the 'spike' represents a displacement of around 4% (i.e. one per class of children) of scores in a subsample [16]. Further it offers: "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." [16]. A different way of expressing this would be the term 'confirmation bias' whereby scores recorded are influenced by the tester's expectations of the result which is well understood in medical trials and is one of the reasons that assessors are blinded to the treatment status of those that they assess.

An alternative interpretation, and, this is more worrying, is that teachers are forcing children into the pass category because they are concerned about class and potentially school performance in relation to external drivers – such as national targets. While this issue of bias is clearly a problem, it would be much more of a concern if it was organised to target certain groups or planned systematically as was the case in the reporting of ambulance response times [17] when recorded times were adjusted to meet the target time of 8 minutes, by a combination of managerial pressure and deliberate manipulation. While we are unable to comment on whether Local Authorities differed extensively in their reporting it is possible that marked disparities would occur.

At this point it is worth returning to the type of screening model advocated in healthcare contexts and specifically developmental screening the process by which children are identified as needing to be in receipt of additional services because of delays in their general development (gross and fine motor movement) or specific aspects of development such as speech and language. Such procedures tend to take place at very specific time points in the child's development obviating the need to account for the child's chronological age. There are alternatives to this such as the Ages and Stages Questionnaire (Squires, Twombly, Bricker & Potter 2009) which covers all of early childhood but this is banded in three month periods to avoid the type of problem thrown up by the PSC. In addition because we have neither a gold standard with which to draw comparison nor any sense of what a score either side of the threshold means across time it is very difficult to make a judgement about whether the finding that a child's score falls either side of the threshold makes any difference at all. Indeed one could argue that any potential benefit from identification has to be contrasted with the negative implications of telling a parent that their child has "failed" the screening test which is a requirement of the administration of the PSC. Poor specificity corresponds to over identification means that children who do not need extra support will receive it unnecessarily. Poor sensitivity means that children are missed. What it means to be overidentified or missed can have implications for resources and parental, and potentially child, anxiety. The findings may well have implications for the schools themselves.

One implicit assumption any test of child development is that as the child improves their score on the test improves i.e. the screen score increases monotonically with development. This allows the consistent setting of a benchmark as all the children who have advanced beyond just being able to read the words at that level will do better. The very gradual effect at the ceiling in the distribution of scores suggests that the best readers were making more

mistakes than might be expected. This might be explained by children progressing with their reading try to read non-words as real words (e.g. reading 'thend' as 'the end') when weaker children simply decode and pass the test. This is related to some qualitative evidence that fluent readers were not scoring as highly as those who are more functional decoders [18]. Children who have good decoding skills but also have other strategies they use to read may well make mistakes and need no remedial action, but others have extensive ability to read familiar words and use context but limited ability to decode which may need intervention.

If we compare the two years of the test we do observe some differences (see the appendix for comparable Figures 4 & 5). For the general distribution, while the pattern is very similar there are more children reaching ceiling and the dip after the pass mark is less pronounced. Similarly the overall proportion of children with lower scores is smaller. The number of children in receipt of free school meals who score zero on the test has dropped. There are more children in both FSM and non-FSM groups that obtain the top score but the effect is more marked for the non-FSM group. Trying to compare whether the distortion at the pass mark has receded in the second year of the test, we can say that this is not supported by the evidence. Indeed, in the FSM group it seems the displacement is slightly larger than in the first year of the test, a contrast which is not apparent in the remainder of the cohort. In general the distribution is the same which has prompted the DfE to withhold the pass mark until after the tests have been taken for 2014.

Finally we turn to compare the results of the specific disability groups between the two years. In all the SEN groups, the number disapplied has increased slightly causing the proportion of children scoring zero to reduce which seems a pragmatic response to the experience of the first year. The ASD group has performed similarly across the two years although more of them appear to have been "given the benefit of the doubt" at the pass mark in the first year. In the second year it the SLCN group has the higher peak that the pass mark. The pattern at the ceiling of the check is similar, with autistic children outperforming the two other groups. It is salient to observe that the proportion of ASD children scoring the top mark rose from 5 to 7%. The SLCN rose slightly while the MLD group remained the same across the two years but it is difficult to interpret patterns in these groups without more detail. Indeed the data highlight the sheer number of children with a primary need of SLCN and the fact that they consistently score across the range of the test.

Limitations of the study

There is a risk that the unusual nature of the distribution of these data is, in part, a function of the way that teachers were introduced to the PSC. It is possible that new guidance could address these issues although the fact that the pattern is so similar in year 2 suggests that this issue has yet to be addressed. These data are, of course, aggregate data from the whole of England. They do not allow us to tease out the performance of individual local authorities and indeed individual schools to check whether the patterns described are being driven by the way that the process has been introduced in some areas rather than others. They do not allow us to explain what appears to be happening from the teacher's perspective. We found anecdotal evidence about this but it would be very helpful to carry out systematic qualitative analyses of their experience of administering the test and their expectations of the PSC relative to other sources of information that they may have about the children's early reading abilities.

CONCLUSIONS

It is clear that trying to identify young children with poor phonic skills and offering them appropriate support makes the search for a valid and useful screen warranted. Unfortunately the PSC is not it – at least not in its present form. While it might reasonably be suggested that this test is experiencing teething troubles associated with its introduction, it may be that the curious shapes of the distributions in these graphs could reflect a fundamental problem with the measure itself and how it is being used. It is clear from these graphs that no further extrapolations should be made about the phonic skills of children in the first two years of its application in primary school unless the measure and its application are properly examined. We understand that for the third iteration of the "screen", teachers will not be given the pass mark until after they have completed the test. But this still raises the issue of whether the screen is being appropriately used with this age group or indeed whether adequate support is likely to be put in place for the third of pupils who fail the test and finally what parents think about the check's utility.

ACKNOWLEDGEMENTS

We thank Sally Marshall, Data and Statistics Division, Department for Education, for advice on the availability of data and assistance in providing the ad hoc release used in the research in 2012. We thank Anneka Nelson-Girtchen, Education Standards Evidence and Dissemination Division, Department for Education, for providing the data from 2013. Maria Mroz, Department of Education, Newcastle University advised on the experience of the test in schools.

Thomas King is funded through the Centre for Research Excellence in Child Language by the Australian NHMRC. James Law is a Principal Investigator on this programme grant.

REFERENCES

1. Melby-Lervag M. The relative predictive contribution and causal role of phoneme awareness, rhyme awareness and verbal short-form memory in reading skills: A review. Scandinavian Journal of Educational Research 2012;56:101-18

2. Ehri L, Nunes S, Willows D, et al. Phonemic awareness instruction helps children learn to read: Evidence form the National reading panel's meta-analysis. Reading Research Quarterly 2001;36:250-87

3. Townley L, Gotts D. Topic Note: 2012 Phonics Screening Check. London: Department for Education, 2013.

4. Standards and Testing Agency. Year 1 phonics screening check Pilot 2011: Technical report. London: Department for Education (DfE), 2012.

5. Ofqual. Regulatory Framework for National Assessments: National Curriculum and Early Years Foundation Stage: Ofqual, 2011.

6. Symbol UK. Communicating Phonics: A guide to support teachers delivering and interpreting the phonics screening check for children with speech, language and communication needs. London: The Communication Trust, 2012.

7. Wilson J, Jungner G. Principles and practice of screening for disease. WHO Public Health Papers 1968;34:1-163

8. Sackett D, Haynes R, Guyatt G, et al. Clinical Epidemiology: A Basic Science for Clinical Medicine. London: Little Brown, 1991.

9. Law J, Boyle J, Harris F, et al. Screening for speech and language delay: a systematic review of the literature. Health Technology Assessment 1998;2(9):1-184

10. Nelson H, Nygren P, Walker M, et al. Screening for speech and language delay in preschool children: systematic evidence review for the US Preventive Services Task Force. Pediatrics 2006;117(2):e298-e319

11. Cizek G, Bunch M. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests Thousand Oaks, CA: Sage, 2007.

12. Crawford C, Dearden L, Greaves E. When you are born matters: evidence for England. London: Institute for Fiscal Studies, 2013.

13. Marshall S. DfE: Phonics Screening Check and National Curriculum Assessments at Key Stage 1 in England, 2012/2013 (SFR37/13): Department for Education, 2013.

14. Marshall S. DfE: Phonics Screening Check and National Curriculum Assessments at Key Stage 1 in England, 2011/2012 (SFR21/12): Department for Education, 2012.

15. McDowell KD, Lonigan CJ, Goldstein H. Relations among socioeconomic status, age, and predictors of phonological awareness. Journal of speech, language, and hearing research: JSLHR 2007;50(4):1079-92.

16. Standards and Testing Agency. Phonics screening check 2012 Technical Report. London: Department for Education, 2012.

17. Bevan G, Hamblin R. Hitting and missing targets by ambulance services for emergency calls: effects of different systems of performance measurement within the UK. Journal of the Royal Statistical Society Series A (Statistics in Society) 2009;172(1):161-90

18. Walker M, Bartlett S, Betts H, et al. Evaluation of the Phonics Screening Check: First Interim Report London: National Foundation for Educational Research, 2013.

APPENDIX

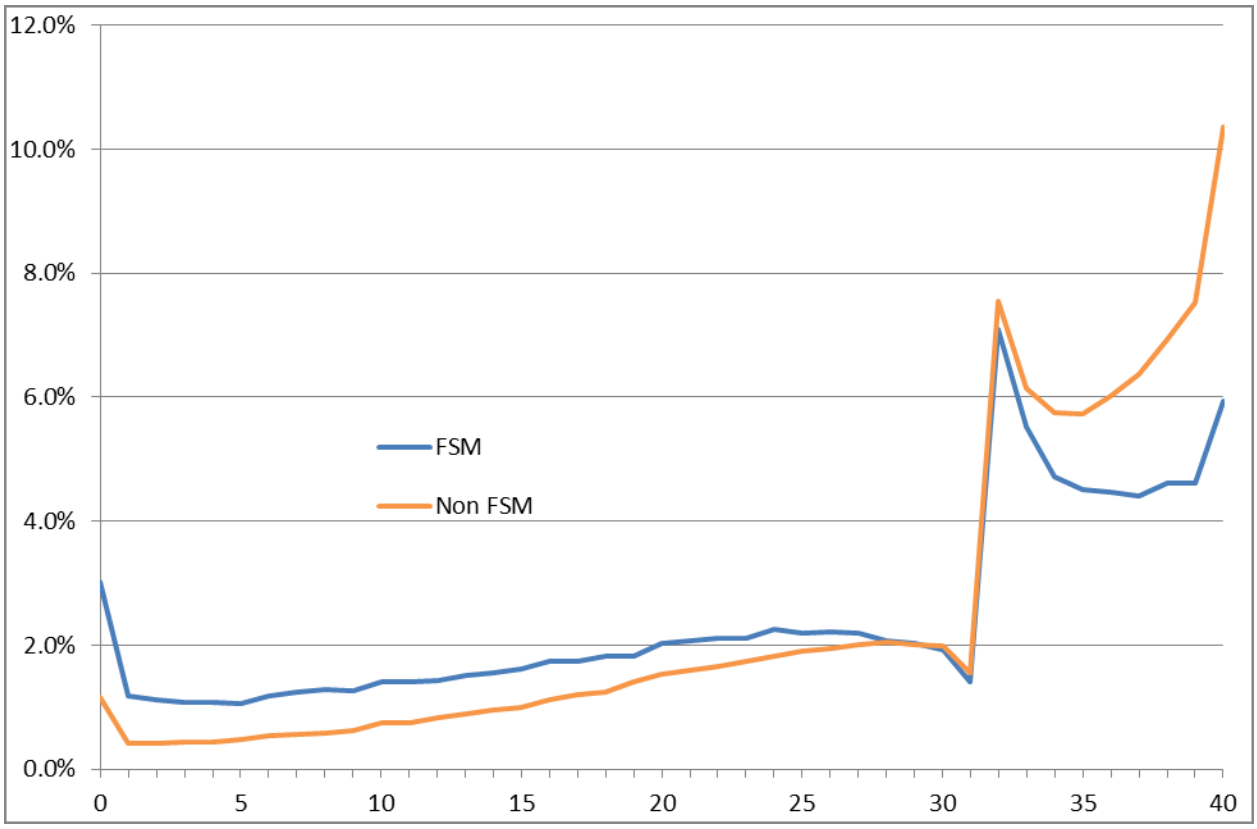

Figure 4 The Phonics Screening Check results by receipt of Free School Meals (FSM) for 2012

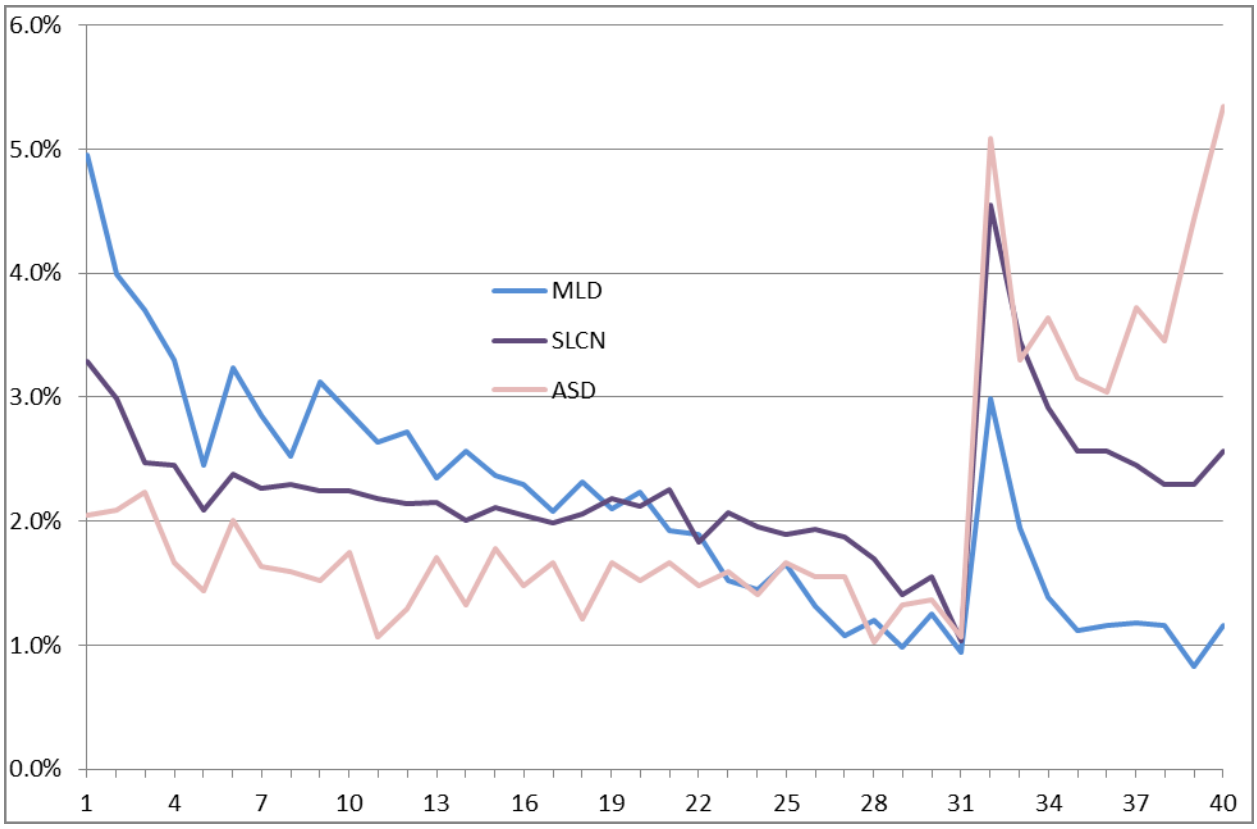

Figure 5 Phonics Screening Check results for children with certain designated disabilities 2012