

# Northumbria Research Link

Citation: Wrigley, Terry and Wormwell, Louise (2016) Infantile accountability : When big data meet small children. *Improving Schools*, 19 (2). pp. 105-118. ISSN 1365-4802

Published by: UNSPECIFIED

URL:

This version was downloaded from Northumbria Research Link: <http://northumbria-test.eprints-hosting.org/id/eprint/47972/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)



**Northumbria  
University**  
NEWCASTLE



**UniversityLibrary**

## **Infantile accountability: when big data meets small children**

**Terry Wrigley, Northumbria University and Louise Wormwell, Newman University, England**

*corresponding author terrywrigley@gmail.com*

### **Abstract**

This article examines a government attempt to impose testing of four-year-olds as a baseline against which to 'hold primary schools accountable' for children's subsequent progress. It examines the various forms of baseline testing in this experiment, and analyses the misleading claims made for the 'predictive validity' of baseline scores. The article also takes a broader look at standardised ways of tracking children's attainment and progress to the end of primary school, and tacit assumptions of linear progress underpinning large-scale data-based accountability processes.

**Keywords:** high-stakes testing, baseline, assessment, accountability, ability, potential

A tiny bundle's thrown into the world  
its nappy not yet pinned,  
the vicar takes his fee before he christens it,  
but its dreams have already been dreamt away  
it is betrayed and sold.

It's red and tender still, and yet they know  
how big the margin, what's the target rate,  
what to teach it and what to hide.  
Its life is stolen, future pathways planned,  
already forfeited and thrown away.

These verses from Hans Magnus Enzenberger's poem *Geburtsanzeige* (Announcement of birth) provide a powerful reminder of the ways in which even small children are impacted by macroeconomic structures and the tangled interrelationship between knowledge and power (*power/knowledge*, eg Foucault 1980). This article focuses on a particular manifestation of the neoliberal regime of truth, namely the attempted imposition of testing on 4 year olds in England.

As part of a raft of tests at different points in schooling, a determined attempt has been made by Government to extend national standardised testing for accountability reasons down to the start of Reception class, where the youngest children are just turned 4 and the oldest barely 5. This was intended to replace broader and more integrated forms of development assessment up to the end of Reception. Other steps have been instituted around the same time, including a phonics test at the end of Year 1 (see the article by Margaret Clark in this special issue) and the restoration of written tests in English, maths and science at the end of Year 2 in place of summative assessment by teachers. Tests at the end of primary, i.e. Year 6, are also being modified to include addition, discrete tests of grammar, spelling and punctuation. Teachers, heads and inspectors are given

guidance on the expected progress in between, generally measured in steps which can be achieved in about 3-4 months (although the situation has been confused, inexplicably, by the abandonment of a standard measure of ‘levels’!) The illusion behind all of these is that smooth, evenly calibrated progress is normal and measurable, and that teachers and schools can be judged successful or failing according to whether pupils make the expected amount of progress.

### **The background to baseline**

The Department for Education’s Policy Paper 2010-2015 *Government policy: school and college funding-accountability* (2015) states that the baseline test is being introduced with the intention to give schools greater accountability for their education provision. The purpose, according to the Standards and Testing Agency

“... is to support the accountability framework and help assess school effectiveness by providing a score for each child at the start of reception which reflects their attainment against a pre-determined content domain and which will be used as the basis for an accountability measure of the relative progress of a cohort of children through primary school.” (Standards and Testing Agency, May 2014)

The DfE invited test providers to bid for licenses against detailed criteria. Six of the submissions met approval, coming from a range of known and unknown educational publishers. The chosen six providers were announced and schools were encouraged to sign up to one of them. Advice from various quarters informed practitioners what to look for when choosing.

However the competition was not over: to be successful and remain an ‘approved’ provider, the publishers had to obtain at least 10% of the market share. It was then reported that Early Excellence had been selected by 12,000 of the possible 17,000 providers, making it impossible for all of the other five to reach the required market share (Gaunt, 2015), which resulted in three of the six being eliminated and some schools having to switch. The popularity of Early Excellence may have come as a surprise to Government ministers and officials, who generally mistrusted assessment by teachers.

The vast majority of children in Reception class (age 4) in autumn 2015 were duly assessed, using one of the three providers. Though strictly speaking non-statutory, schools choosing not to implement it were threatened with draconian accountability sanctions based on extremely high absolute attainment benchmarks at age 11 without regard for the pupils’ socio-economic characteristics. A vigorous campaign was mounted by Better without Baseline, a broad coalition of early education groups, teacher trade unions and researchers. This coalition sought to persuade

schools against using any of the providers, since even Early Excellence, though observation based, were working towards problematic criteria set by government.

At the time of writing, the tests have just been withdrawn by Government, whose evaluators proved unable to reconcile scores from different versions of the tests, although the Government has not abandoned the principle of baseline testings for accountability purposes and is considering a replacement. What this will be remains to be seen, but following the announcement that the baseline assessment had been abandoned Early Years forums on social media were speculating about a crude ‘school readiness’ check.

The extension of accountability measurements to younger children, or *datafication* (Roberts-Holmes 2014), is accompanied by *schoolification*, or the spread of traditional formal schooling into the pre-school stage (Moss 2013). Both of these are nested within a wider neoliberal educational politics from which even young children are not exempt. (See Moss 2014; also the *Editorial* of this special issue).

The particular focus of this article, however, is the notion that baseline tests have ‘predictive validity’, in other words, accuracy in ascertaining not only the ‘ability’ of the child (whatever that means!) but also its potential attainment in later years of schooling. This is, of course, well aligned with the wider neoliberal policy framework, as they both see educational processes in terms of ‘readiness’ for future functioning (Evans 2013).

Framing early childhood care and education in this way sits in serious contradiction to the progressive tradition which has grown up, and been struggled for, over 200 years (Pestalozzi, Owen, Froebel, Dewey and many more), and which is based on respect for the child’s present and emergent feelings, interests and characteristics. This is not to accept the view that children are isolated individuals whose development is somehow spontaneous: we must work with the Vygotskian model of an encounter between the growing child and the culture it inhabits (eg Kozulin et al, 2003). For example:

The growth of the normal child into civilization usually involves a fusion with the processes of organic maturation. Both planes of development – the natural and the cultural – coincide and mingle with one another. The two lines interpenetrate one another and essentially form a single line of sociobiological formation of the child’s personality. (Vygotsky 1960:47, in Wertsch and Tulviste 2005:72)

There is a different sense of time and directionality, experience and autonomy between this quotation and the neoliberal emphasis on measurable linear progression which, in the worst case, treats children like caged hens.

## The regulatory framework

The latest attempted addition to England's heavily audited school system (Ranson 2003; Ball 2008) is the introduction of Baseline assessments as a starting line for 'holding primary schools to account' (DfE 2014a). Although not legally mandatory, schools are being persuaded that this would allow them to be judged fairly on progress rather than against an absolute and extremely high attainment target. This is particularly threatening to schools in poorer areas.

The new assessment focus is on literacy and numeracy, rather than the broad spectrum of development previously assessed through observations. Although some of the providers offer other aspects of development within the same package, this is not reflected in the scoring.

The assessment was to be completed within the first six weeks in Reception class rather than by the end of the year. This was a complete contrast to established Early Years practice where observations across seven areas of development and Characteristics of Effective Learning were carried out across the year, culminating in the completion of the Early Years Foundation Stage Profile.

Schools were allowed to choose between three government-approved providers (originally six, but the others had few takers). It was predictable that this would create problems of commensurability, as each will use a different assessment process. However, the various providers were required to follow some common rules:

1. each item must require the scorer to make a 'single, objective, *binary* decision', in other words yes / no
2. the assessments must culminate in 'a score for each child on a *single scale*'
3. the scores 'must *not* be age-standardised'. (DfE 2014b).
4. The assessment had to be carried out in English
5. The scale on which scores are reported must ensure the full range of attainment is appropriately distributed across the range with fewer than 2.5% of children achieving full marks

All of these conditions are problematic, and represent a desire for neatness which shows the Department for Education's (DfE) remoteness from the complex realities of children in early education:

- i) The achievements of four-year-olds are often not susceptible to simple yes-no confirmation – often the only honest evaluation is 'partly' or 'she didn't feel like it today' or 'he just didn't understand the question'. As an example, the question "Does the child link sounds to

letters, naming and sounding the letters of the alphabet?” begs the question: how many sounds and letters? consistently?’

- ii) The amalgamation into a single score and scale denies the unevenness of development. It also fails to take into account the enormous and unpredictable leap from using language to subjecting it to analysis, as required by many test items (eg “What sounds are in the word ‘net’?”)
- iii) The refusal to consider the child’s age is extraordinary, given the large developmental differences to be found during this year of life: the youngest children are just turned 4 and the oldest around 5. Although one provider Early Excellence is using observation- rather than test-based assessment, these conditions remain problematic.
- iv) The condition stating that the assessment had to be carried out in English contradicts guidance for assessment in the Early Years. The Early Years Foundation Stage Profile Handbook (2016) states that with the exception of the areas of Communication and Language and Literacy, practitioners should assess the development of children who speak English as an additional language in their home language as well as English. This has been the guidance for several years.
- v) The reporting scale restrictions appear to limit attainment.

### **Ethical difficulties**

The first question concerns the relationship between school-level accountability and the assessment of individual children. Schools cannot be “held to account” without assessing individual children. In the world of social action, such data is never simply descriptive, it is *performative* or *productive* (Ball 2008; Hursh 2008; Lingard 2009; Ranson 2003; Stobbart 2008): the data from baseline tests can affect the way a teacher regards and teaches that child, and even the way the child is perceived by its own parents. These dangers are exacerbated by the prevalence of “ability grouping” for literacy and numeracy teaching. *Ability* is, of course, a problematic concept, especially when applied to young children: it conflates the fact that some children have had richer experiences than others with assumptions that children have different quantities of innate intelligence (Hart et al 2004). Consequently, early assessment which involves attaching a score to a child would be ethically questionable even if it could be done accurately, as both positive and negative judgements can become self-fulfilling prophecies.

Secondly, since the assessment packages are commercially provided, giving the provider opportunity for future custom, there is a strong possibility of the child’s learning being distorted by

teachers devoting time to practising for the re-run. Nurseries will also be tempted to practice the test with even younger children. As one primary headteacher expressed it, it creates a:

downward pressure that will inevitably lead to three and four year old boys in nursery spending more and more time at writing tables orientating letters, writing their name and improving their pencil grip. (Crilly, 2016)

As with the first ethical question, this would still be a problem *even if* baseline assessment made accurate forecasts; *it does not*.

### **Statistical claims for predictive validity**

This section deals with various technical issues, which concern not only whether the procedures adequately reflect reality, but also how data is read. Particular attention is paid here to the Centre for Evaluation and Monitoring (CEM) at the University of Durham which is by far the most experienced in predictive testing. Indeed, their test has been developed and sold for over 20 years on a commercial basis to numerous schools in England and internationally. This is not to question the expertise or good intentions of staff at CEM, but rather the viability of the overall concept in government policy.

CEM's advertisement made the following claim:

- Excellent predictive validity – *correlates at 0.68 level with age 11 assessments*.

To give the benefit of the doubt, perhaps the advertising copywriter became over enthusiastic; or maybe all that was meant by *excellent* was “We are better than our competitors” or simply “This is as good as it gets when assessing four-year-olds”. However, the correlation does seem to lend authority to the promotional claim.

The correlation of 0.68 is typical of others to be found in CEM documents, which seem to hover around 0.7. The question is: *what does this mean in reality?*

0.7 is widely regarded as a strong correlation, since correlation scales run from 0 to 1. Doesn't it matter, though, *what is being correlated with what, in which circumstances and for what purpose?* In other words, is there such a thing as a ‘good correlation’ in the abstract?

A former civil engineer pointed out that 0.99 is disastrous when calibrating instruments: “Bridges could fall.” Pursuing that thought, a correlation of 0.3 between eating bananas and living to the age of 80 might be persuasive: bananas might be just one of several contributory factors to longevity, but significant enough to be worth the trouble of eating them. On the other hand, a test which claimed to predict cancer or alzheimers two years later with a correlation of 0.7 would be unusable:

there would be too many false negatives or positives. In the second case, i.e. a predictive test, much higher levels of accuracy would be necessary.

Furthermore, most non-statisticians would be unaware of the need to square a correlation in order to ascertain just how much of the variance in y can be explained by variance in x. Thus, a correlation of 0.7 squared gives 0.49. In other words, only half the variance in y can be explained by the variance in x.

A clue emerged as to the real meaning of a 0.68 correlation in a research paper by Peter Tymms (2003) of CEM. Alongside correlations, it provided a ‘chances table’ showing just how likely it is that a particular baseline score leads to particular outcomes. Following a successful Freedom of Information request, a more extended dataset was provided by CEM. Each row of the spreadsheet showed a baseline score, ranging from 0 to 100, while each column gave the percentage of children with that score who would achieve a particular KS1 level or sub-level.

CEM clarified that this followed a normal distribution curve, with over two-thirds of pupils obtaining between 40 and 60, but hardly any scoring below 20 or above 80. In fact very low or very high baseline scores proved highly predictive, but, because of the distribution curve, hardly any pupils actually gain such scores. On the other hand, the mid-range scores – those obtained by most pupils – made poor predictions. This is shown in Figure 1 below. The sample scores in the left-hand column are those with the greatest likelihood of each particular Key Stage 1 outcome. For example, the greatest likelihood of a level 1 is from a baseline score of 28, and 39 percent of pupils (in bold) with that baseline score attain level 1.

Readers will see immediately the poor quality of predictions from the mid-range baseline scores. Thus, the pupils with the greatest chance of obtaining level 2c are those with a baseline of 37, but only a quarter of this subset do end up with level 2c. (Levels have now been replaced, but this historical data is still a valid way of judging the predictive validity of the assessment tool.)



Baseline score	<1	1	2c	2b	2a	3
20	<b>59</b>	30	8	3	0	0
28	30	<b>39</b>	19	10	2	0
37	9	28	<b>26</b>	24	11	2
44	2	14	21	<b>32</b>	22	9
53	0	3	9	24	<b>34</b>	30
80	0	0	0	0	2	<b>98</b>
National distribution	2	7	8	23	27	32

*Figure 1: Extract from CEM PIPS > end KS1 spreadsheet, showing baseline scores with the strongest probability of attaining each KS1 level or sub-level*

Thus, in reading, from 44 - the baseline score with the highest chance of Level 2b at KS1 - only 32% actually get 2b, whereas 16% of children with this same baseline score get 1 or below, 21% 2c, 22% 2a, and 9% level 3. This enormous divergence makes even CEM's highly developed version of baseline testing next to useless, except from the more extreme (but rare) baseline scores.

Based on the indications of a normal distribution of scores, with most pupils scoring mid-range, calculations were made of the overall likelihood of a pupil reaching the most probable level for their baseline score. This showed, on average, correct predictions for only 4 children in every 10. Judging by the most experienced provider then, baseline assessment is more like a *sawn-off shotgun* than a precision tool. The consequences of inappropriate labelling for 6 out of 10 children are obviously very serious.

It should be noted that this derives from using CEM's PIPS test at the end of Reception. Using it at the start of Reception is likely to be even less predictive.

Another spreadsheet showed rather more success in using the PIPS test at the start of Year 3 to predict end of Key Stage 2 outcomes (i.e. nearly four years later), with accurate predictions for 2 children in every 3. That may be because testing is more reliable with older children; it may also be because the outcome levels in this case were not sub-divided: level 4 (undivided) and level 5 are very large buckets into which to throw a ball. (Nationally 42% attain level 4 and 38% attain level 5.)

## The other providers

The National Foundation for Educational Research (NFER) also has substantial experience in assessment and research, but stated, quite correctly, that they had no longitudinal data on which to stake a claim for predictive validity. In email correspondence, they asserted that

There is no intention on our part to use baseline assessment outcomes to make predictions about individual children. It is also my understanding that the progress between school entry and the end of key stage 2 will be measured / reported by the DfE at the *cohort* level.

This is formally correct: the DfE do refer only to school-level data. However it is inconceivable that teachers, schools and government inspectors would not examine and track individual progress. Indeed, a subsequent press statement from NFER confirmed that parents and teachers would be supplied with individual profile reports, and that these would form a basis for teachers to “identify the next steps for children”.

When questioned about the viability of testing very young children, NFER’s response made reference to a research paper (Muter et al 2004) which supposedly supported their case. Ironically much of its data actually undermines the claim that good predictions are possible in the first two years at school. The question remains on the table of whether a test which was designed to monitor how well children *had* learned to read, i.e. *following* literacy instruction, can be used appropriately on children *before* such instruction.

The third approved provider Early Excellence (EE) is new to the field of assessment; their core business appears to be largely in the sale of nursery furniture and equipment and in staff development. The EE baseline assessment is based not on a test but on observations which are similar, in many respects, to those which schools already carry out during the Reception year. That is probably the reason why, at this stage, Early Excellence are the most popular of the three with schools, and indeed EE pride themselves on having kept open this option of an observation-based assessment.

There are, however, important differences between the existing and new arrangements, as a consequence of the rules set by the government department. It is worth reiterating that under these regulations:

- observations have to take place during the first six weeks at school, rather than by the end of the year;
- the data requirement is for literacy and numeracy, marginalising other aspects of the child’s development;

- only a simple yes-no answer is permitted to each question or criterion;
- the observations must lead to a single composite score for each child.

Early Excellence, like NFER, confirmed that since the procedure was new, they had no longitudinal data for individual pupils and therefore were unable to assess its predictive validity by tracking pupils through from baseline to KS1. They had however conducted a pilot with 17 schools in order to establish the likelihood of some similarity between a school’s baseline scores and its recent KS1 outcomes. Sample data was shared for two of these schools at opposite ends of the range, as follows.

The first example (figure 2) was a school with high KS1 (end of Year 2) results in recent years. On the left the bars represent bands of baseline scores (each covering a fifth of the population of the 17 sample schools), and on the right, KS1 outcome levels (1 or below, 2c, 2b, 2a, 3 or above). The vertical axis shows the percentage of pupils in each category.

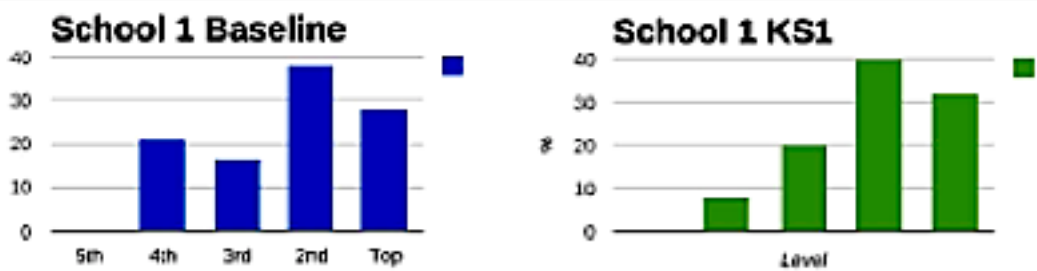


Figure 2: EE school-level data: high attaining school

There is clearly correspondence at school level, since the diagram to the right superimposes reasonably well onto the diagram on the left. There is however no evidence here of correspondence at an *individual* level, since these are two different sets of children and no longitudinal data is available: for example, we simply cannot tell, from this data, how many pupils scoring in the top band at baseline went on to the highest level at KS1 or whether there was substantial cross-over between bands.

The situation regarding the low attaining school (figure 3) is far more problematic.



Figure 3: EE school-level data: low attaining school

Most children in this school score poorly at baseline, but far fewer have low attainment at the end of KS1. (The percentages on the left don't add up to 100, but even allowing for that error, there is a clear lack of correspondence.) This suggests that most of the pupils with low baseline scores proceed to average and above average levels at KS1. As with the high-attaining school, on the basis of this data there is no way to investigate up and down movement between bands.

The diagram highlights the danger of children being written off as 'low potential' on the basis of low baseline scores. It is also possible that teachers might concentrate so much on improving the assessed skills for a re-run of this assessment at the end of the year that longer-term development could be jeopardized.

### **Standardised tests and diverse children**

One of the specifications of the Baseline Assessment was that it had to be delivered in English (STA, 2014) - a procedure which TACTYC (2014), one of the early years organisations opposing baseline, believed made the assessments potentially discriminatory. In a recent pilot study undertaken by a team led by Professor Margaret Clark at Newman University, in an urban area of central England, it was found that, in just three schools, the children in Reception spoke at least 16 languages other than English. Of these 117 pupils, 52 spoke a language other than English as their first language (Clark, 2016). Many of these children may have entered reception and endured the baseline without an understanding of the language they were being assessed in. It was inevitable that the baseline score they received would not be an accurate reflection of their true capabilities.

This was reinforced by research carried out by a team at the Institute of Education of University College London for two major teacher trade unions the NUT and ATL. It revealed that 68% of staff and parents surveyed did not believe the baseline helped identify the needs of children with English as an Additional Language (EAL).

Given the changing population in our schools and the proportion of EAL children currently in schools this was neither fair nor representative.

### **Wider doubts on predictability**

Qualitative assessment which is provisional and sensitive to the individual child is well established in early education. What is at question here is the reliability of quantitative judgements which on the surface appear more definitive.

The Early Years Foundation Stage Profile (EYFSP) is based on observations undertaken periodically in nurseries and completed by the end of Reception year, and has been used for a

number of years. The DfE had attempted to convert these qualitative observational assessments into numerical scores which could be matched and compared with later attainment measures. The results of this matching showed a very limited continuity in the progression of individual children. There is only space here to highlight some points, but extensive details and explanation can be found in chapter 6 of DfE (2010).

Numerous correlations are calculated between scores, but they are not strong. The best predictor for KS1 Reading is the average for EYFSP Communication Language and Literacy, with a correlation of 0.68 (p62) [see earlier explanation about the need to square correlations, section *Statistical claims for predictive validity*]. Indeed the report admits that only 55% of the variation in KS1 average points scores (Reading, Writing and Maths) can be explained by the Early Years profile (p57).

The following table (figure 5) shows in more detail the relationship between the Foundation Stage Reading assessment (on a 9 point scale – the horizontal axis) and KS1 Reading levels / sub-levels (the percentages for each baseline score hitting each level or sub-level – the vertical axis).

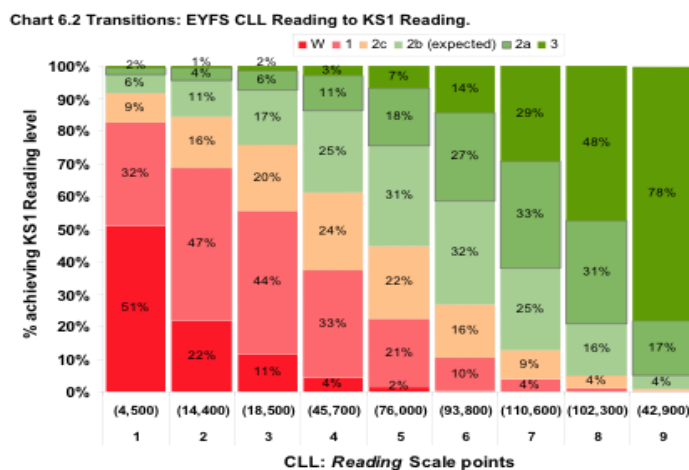


Figure 4: DfE data relating EYFSP to KS1 in Reading

We see here that children with the midpoint score (5) diverge almost equally into four broad bands: W or L1; 2c; 2b; and 2a or 3. Although the precise scoring methods have since changed, this is the most complete published version of data and is included here to demonstrate the poor predictive value of early assessment.

A more recent set of attainment data (DfE 2015) shows a further disjunction, this time between the Phonics Check and Key Stage 1 Reading assessments. The phonics check is applied to all children at the end of Year 1, and those who fail have to repeat it at the end of Year 2. Key Stage 1 Assessments are applied to all children at the end of Year 2. Of pupils who failed the phonics check in Year 1 but passed it on the retake (i.e. at the same time as the KS1 assessments), 13% were

awarded Level 1, 25% 2c, 41% 2b, 17% 2a and 4% level 3. In other words, many of the slow starters were quite competent readers by the end of the following year.

This is part of a much larger problem of the accountability system. Using ‘value added’ data to judge school effectiveness depends on reasonably reliable norms and expectations for progression, such that schools that deviate seriously from the norm will stand out. In other words, comparisons between schools in terms of their relative effectiveness must be underpinned by a general assumption that progression is normally smooth and linear. If progression is erratic, deviation becomes meaningless. However recent work by Education Datalab (2015) has holed the ship below the waterline. Tracing individual pupils between statutory End of Key Stage levels, its researchers have revealed that:

- only 55% of children get the KS2 level (age 11) which matches their KS1 levels (age 7)
- only a third of children getting the average level (2B) at age 7 get the average grade (C) at 16
- even these children (i.e. the third who start average and do meet their predicted average outcomes) generally do so via a route that includes period of slow and more rapid progress.

As the researchers express it, “More children get to the ‘right’ place in the ‘wrong’ way than get to the ‘right’ place in the ‘right’ way!” The following graph (figure 6) shows the divergence from an initial Level 2b at age 7 to age 11 and age 16: children with the average level at age 7 who reach the (expected) average level at age 16 have travelled via widely different levels at age 11. This is hardly a sound basis for systematic accountability judgements. (Again, the grade boundaries and names have just been changed, but the conclusions remain valid.)

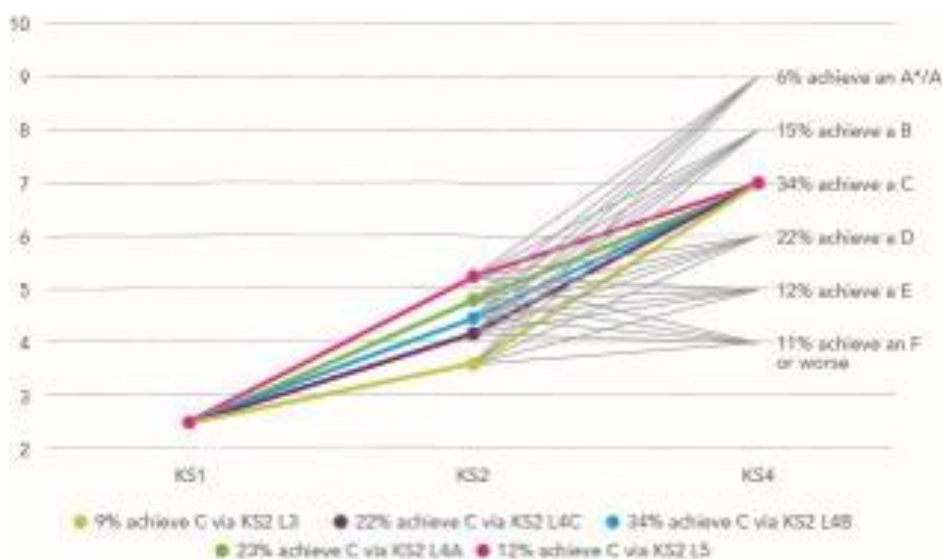


Figure 5: Education Datalab showing progression paths from KS1 to GCSE

A further finding from Education Datalab is that “children with low initial attainment have particularly unpredictable future attainment” – a conclusion which, as argued earlier, makes it very dangerous to label these children with early quantitative judgements.

Part of the explanation is provided by the cumulative impact of poverty. A small proportion can be explained by differences between schools. Some of it is simply human unpredictability (character differences, biographical accidents, and so on).

### **Exploring reasons for the instability of early assessment**

Reference was made earlier to a research paper (Muter et al 2004) which NFER cited in support of their claim that assessment from a very early stage is viable (see above –*The Other Providers*). Rather than establishing the reliability of such assessment, as asserted, its data actually reveals how tenuous that can be. In fact, there are frequent contradictions in the report between verbal claims that predictability is strong and statistical data undermining these assertions.

This research is based largely on a study which assessed children at three intervals a year apart, referred to as Times 1, 2 and 3. The data is revealing for a number of reasons. Firstly, the correlation between scores on the *same* reading test on two occasions just a year apart is only 0.71, with considerably lower correlations on other factors. There are also considerable disjunctions between tests focusing on different aspects of reading, which demonstrate that there are considerable problems in predicting future reading performance from tests of sub-skills.

For example:

- various phonemic tests at time 2 have correlations of .42, .55 and .40 to Early Reading at time 3 [i.e. a year later]
- the relationship between phonemic tests at time 1 and Early Reading at time 3 are weaker still, at .34, .24 and .13 [i.e. two years later]
- letter knowledge at time 1 has only a .56 correlation with Early Reading at time 3.

The text reveals considerable disjunction between reading in the sense of pronouncing letters or words and reading in the sense of understanding. It explains that vocabulary knowledge and grammatical skills (i.e. tacit syntactic awareness) are as important as phonetic skills and early word recognition in explaining success in reading for comprehension even by Time 3 (early in Year 2); and also that “the growth of word recognition abilities is relatively uninfluenced by vocabulary and grammatical skills”. This adds to concerns that early testing based on sub-skills such as letter recognition might divert from the wider development which is also necessary for children to become truly literate.

## **Some reflections**

One of the arguments that could be used in favour of early testing is that it mitigates against possible bias on the teacher's part. Indeed this argument was used by CEM (2012) when advocating strongly for baseline assessment in response to a DfE policy consultation.

One might also argue that teachers should be less deterministic in their interpretation and use of assessment data. This may well be true, but it becomes very difficult given the aura of science around the statistical data which suggests transparency, impartiality and certainty.

Teachers suffer considerable psychological tension in all this. They sense the distance between the world of numbers and real children. The numerical data is felt as alien to their reason for becoming a teacher, and yet the discourse of accountability data has come to permeate their professional activity and sometimes even seems like a comfort blanket - after all, good data does help to keep Ofsted's inspectors away.

There are particular tensions with regard to baseline testing. Two professional associations TACTYC and Early Education have advised school leaders that it would be best not to adopt baseline testing, but if schools are compelled to, they should 'put away the resulting data and forget about it until children reach the end of KS2' [TACTYC / Early Education 2015) because it will not help the children's learning. This is bound to entail anxiety, however, and there is systemic pressure on teachers to make early judgements of children's 'ability' and 'potential' (whatever these words mean!) and to teach accordingly.

The shift of high-stakes accountability downwards into Reception Year presents multiple risks.

- 1) It threatens to undermine age-appropriate practices of early years education, whose roots go back to 19th Century reformers such as Froebel and Pestalozzi, and replace these practices with formal patterns of teaching and learning – a process which has been called 'schoolification' (OECD 2006).
- 2) It tends to reinforce the practice of segregating children into 'ability groups' from an early age.
- 3) It can reduce expectations and therefore place a ceiling on the development of children it labels as having low ability or potential, with particular risks for boys (many of whom are slower to develop), children for whom English is an additional language (who tend to accelerate later), children with health problems, children in care, and the vast numbers of children growing up in poverty.



- 4) It is interesting that in the transient nature of a school's population there would be no guarantee that the cohort of pupils that started the school in reception would be the same cohort that completed the year 6 assessments. Without tracking each individual, there would be no fair way of measuring a school's progress accurately. In any future accountability measure this would need to be considered and measures taken to factor in the issue of pupil mobility.
- 5) The focus on measuring literacy and numeracy comes from a misguided belief that 'earlier is better', but also negates the fact that there are other crucial skills and activities which are more fundamental to successful learning, including self-regulation, co-operation, spoken language and engagement in play (Whitebread and Bingham 2011).

Although a majority of schools showed a preference for the provider offering observation-based assessment rather than tests, the constraints set by the Department for Education meant that children would still be labelled with a single score. In other words, observation-based assessment under these rules is, in finality, equally reductionist and stigmatising. Baseline assessment which sows illusions of predictability can seriously distort the child's development and becomes a vicious circle of self-fulfilling prophecy.

It is a relief that the Government has abandoned this first attempt to assess children as young as 4 years 0 months as a baseline against which to 'hold schools to account'. In many respects, this U-turn is a tribute to the energetic campaign by the Better Without Baseline coalition of teacher unions, early years organisations and academic researchers. Without this activity it could so easily have slid into 'normal practice'. There is however no room for complacency, as the government has not abandoned the principle and is seeking other means as part of a growing raft of high-stakes accountability procedures at every stage of education.

## References

- Ball, S (2008) *The education debate*, 2nd edition. Bristol: Policy Press
- CEM (2012) *Primary assessment and accountability under the New National Curriculum – Consultation October 2012*.  
<http://www.cem.org/attachments/CEM%20Response%20to%20Consultation%20on%20Assessment%20in%20Primary%20Schools%208th%20October%202013.pdf>
- Clark, M.M (2016) *Baseline Assessments their value and validity in assessing young children on entry to school*. Baseline Assessment: What research is telling us. Newman University  
<http://www.newman.ac.uk/24feb/4560/> (accessed April 2016)
- Crilly, L (2016) *Understanding the diversity of children's needs* (MA assignment, Leeds Beckett University)
- DfE (2010) *Achievement of Children in the Early Years Foundation Stage Profile*. Research Report DFE-RR034
- DfE (2014a) *Reforming assessment and accountability for primary schools: Government responses to consultation on primary school assessment and accountability*. Published March 2014

- [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/297595/Primary\\_Accountability\\_and\\_Assessment\\_Consultation\\_Response.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297595/Primary_Accountability_and_Assessment_Consultation_Response.pdf)
- DfE (2014b) *Reception baseline: criteria for potential assessments*.  
[https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/415142/Baseline\\_criteria.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/415142/Baseline_criteria.pdf)
- DfE (2014c) *Reception baseline: criteria for potential assessments*.  
[https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/415142/Baseline\\_criteria.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/415142/Baseline_criteria.pdf)
- DFE (2015) 2010 to 2015 government policy: school and college funding and accountability. London Crown .
- DfE (2015) *Phonics screening check and national curriculum assessments at key stage 1 in England*, SFR 32/2015, 24 September 2015. Table 14: Key Stage 1 reading level by phonics prior attainment.
- Education Datalab (2015) *Seven things you might not know about our schools*.  
<http://www.educationdatalab.org.uk/getattachment/Blog/March-2015/Seven-things-you-might-not-know-about-our-schools/EduDataLab-7things.pdf.aspx>
- Enzensberger, H (1986) *Gedichte 1950-1985*. Frankfurt a.M.: Suhrkamp
- Evans, K (2013) “School readiness”: The struggle for complexity. *Learning Landscapes* 7(1): 171-186
- Foucault, M (1980) *Power/Knowledge: Selected interviews and other writings, 1972-1977*, edited by Colin Gordon. Brighton: Harvester
- Gaunt, C. (2015) <http://www.nurseryworld.co.uk/nursery-world/news/1151726/baseline-approval-process-falters> (accessed Sept 2015)
- Hansard <http://www.theyworkforyou.com/wrans/?id=2015-11-17.16662.h&s=Education+Assessments> (Accessed November 2015)
- Hart, S, Dixon, A, Drummon, M and McIntyre, D (2004) *Learning without limits*. Maidenhead: Open University Press
- Hursh, D (2008) *High-stakes testing and the decline of teaching and learning*. New York: Rowman and Littlefield
- Kozulin, A, Gindis, B, Ageyev, V and Miller, S eds (2003) *Vygotsky’s educational theory in cultural context*. Cambridge: Cambridge University Press
- Lingard, B (2009) Testing times: The need for new intelligent accountabilities for schooling. (*QTU Professional Magazine*) [http://www.qtu.asn.au/files/1313/2268/2362/vo24\\_lingard.pdf](http://www.qtu.asn.au/files/1313/2268/2362/vo24_lingard.pdf)
- Moss, P (2013) The relationship between early childhood and compulsory education: A properly political question. In P Moss (ed) *Early childhood and compulsory education: Reconceptualising the relationship*. Abingdon: Routledge
- Moss, P (2014) *Transformative change and real utopias in early childhood education: A story of democracy, experimentation and potentiality*. Abingdon: Routledge
- Muter, V, Hulme, C, Snowling, M and Stevenson J (2004) Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental Psychology* 40(5): 665-81
- OECD (2006) *Starting strong II: Early childhood education and care*.  
[http://www.unicef.org/lac/spbarbados/Implementation/ECD/StartingStrongII\\_OECD\\_2006.pdf](http://www.unicef.org/lac/spbarbados/Implementation/ECD/StartingStrongII_OECD_2006.pdf)
- Ranson, S (2003) Public accountability in the age of neo-liberal governance. *Journal of Education Policy*, 18(5):459-80
- Roberts-Holmes, G (2014) The ‘datafication’ of early years pedagogy: ‘if the teaching is good, the data should be good and if there’s bad teaching, there is bad data’. *Journal of Education Policy* 30(3):302-315
- Standards and Testing Agency (2014) *Reception baseline: criteria for potential assessments*. London: Crown
- Stobart, G (2008) *Testing times: The uses and abuses of assessment*. London: Routledge

TACTYC / Early Education (2015) *Guidance on baseline assessment in England*. (28 February)  
<https://www.early-education.org.uk/sites/default/files/Baseline%20Assessment%20Guidance.pdf>

Tymms, P (2003) *Performance indicators in primary schools: Feedback report Key Stages 1 and 2*

Wertsch, J and Tulviste, P (1992) L.S.Vygotsky and contemporary developmental psychology.  
*Developmental Psychology* 28(4):548-57

Whitebread, D & Bingham, S (2011) *School readiness: a critical review of perspectives and evidence*.  
TACTYC Occasional Paper No. 2. TACTYC Retrieved at <http://tactyc.org.uk/occasional-paper/occasional-paper2.pdf>