Predictive testing of young children: the delusions of 'accountability'

by Terry Wrigley

Publication note:

The author's involvement in this investigation and analysis was as part of the Reclaiming Schools network, an alliance of academic researchers set up to develop and mobilise knowledge relevant to the education campaigns of the National Union of Teachers and other public campaigns (www.reclaimingschools.org).

Since this article was written, the Department for Education has withdrawn full implementation of baseline testing, at least for Autumn 2016: it had proved impossible for DfE statisticians to reconcile data from the three approved agencies and versions of assessment sufficiently for them to serve as baseline for measuring subsequent 'value added'. There is, however, no sign yet that the Department recognises the other difficulties of early assessment, including those resulting from the parameters it had set and the issue of predictive validity at the centre of this article. Beyond this, Government policy continues to depend on reductionist statistical comparisons.

The latest addition to England's heavily audited school system (Ranson 2008; Ball 2008) is the introduction of Baseline assessments as a starting point for 'holding primary schools to account' (DfE 2014a). This raises critical issues about the appropriate use of assessment data, whilst also opening up to question some common assumptions about statistics within education and perhaps other fields.

In brief, the Department for Education (DfE), though not making baseline assessment mandatory, has decreed that schools not using it will be judged on outcomes data in absolute terms, and against demanding targets. This ultimatum places particular pressure on schools serving disadvantaged neighbourhoods, including with children for whom English is not the main family language, since their outcomes are likely to be lower than the new target.

The new assessment focus is on literacy and numeracy[i], rather than the broad spectrum of development previously assessed. The assessment must be completed within the first six weeks in Reception class rather than by the end of the year.

The DfE initially licensed six different organisations to provide the new tests[ii], though this was narrowed to three (CEM, EE and NFER - see below) on the basis of the initial sales figures for the trial run. Even three very different forms of assessment could create serious headaches for the DfE in trying to make the scoring systems commensurate as a starting point for measuring subsequent progress (*value added*). However, contracting out to multiple providers has allowed the DfE to sidestep a request from the National Union of Teachers for data about predictive validity, for reasons of "commercial interests"; presumably the DfE will be able to conceal much of the data that later emerges.
The various providers are however required to follow some common rules:

1. each item must require the scorer to make a single, objective, *binary* decision 'yes' or '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)

All three conditions are problematic, and represent a desire for neatness which reveals government the Department's remoteness from the complex realities of children in early education. The achievements of four-year-olds are often not susceptible to simple yes-no confirmation – often the only honest evaluation is 'partly'[iii] or 'she didn't feel like it today' or 'he just didn't understand the question'.[iv]

The amalgamation into a single score and scale denies the unevenness of development. 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.

Ethical difficulties

The first ethical question concerns the relationship between school-level accountability and the assessment of individual children. Although the DfE's declared objective is to 'hold schools to account', this cannot be done 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 increased by the common tendency in English schools (encouraged by government policy) of arranging children in "ability groups" for literacy and numeracy teaching. *Ability* is, of course, a problematic concept, especially when applied to young children – a floating signifier which conflates the recognition that some children have had richer experiences than others with assumptions that children have different quantities of innate intelligence or potential (Hart et al 2004). Consequently, early assessment, if this entails attaching a score to a child, would be ethically questionable even if it could be done with some degree of accuracy, since positive and negative judgements could operate as self-fulfilling prophecies.

A second ethical question relates to the commercial basis on which the baseline tests are offered. This gives the provider a foot in the door, and the potential for future custom. Indeed some of the promotional materials show that the providers are well aware of this: schools are told that, for an extra fee, they could repeat the same or a similar test at the end of the Reception year to judge progress (eg CEM 2014a). This carries a strong possibility of the child's learning being distorted because teachers, operating within a system of high-stakes accountability, spend valuable time practising for the re-run of the test. (This already happens during the following school year; teachers spend time practising reading nonsense words as required for the Phonics Check.) There is also a possibility that nursery staff will be tempted to practice the test with even younger children. As one primary headteacher put 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 a number of technical issues, but ones which concern not only whether the statistical procedures adequately reflect reality, but also how statistics are read.

Of the three providers, the Centre for Evaluation and Monitoring (CEM) at the University of Durham are by far the most experienced in

predictive testing. Indeed, the test they plan to use has been over 20 years in development, having been sold on a commercial basis to numerous schools in England and internationally. There is no doubt about the expertise nor the good intentions of these academics and test developers. However, as an academic research centre which has mutated into a thriving business with over 100 employees, they operate in two distinctive discursive environments.

In such a context, it is perhaps understandable that a zealous copywriter should have advertised CEM's tests as having "excellent predictive validity" (CEM, 2014b). To give the benefit of the doubt, perhaps all that was meant was "we are better than our competitors" or even "this is as good as it gets given the difficulties of assessing four-year-olds". However, the juxtaposition between this phrase and the scientific precision of what follows within the same bullet point appears to lend it authority:

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

This correlation is typical of others to be found in CEM documents, many of 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, and to a lay reader who knows the scale runs from 0 to 1 it seems good. Does its adequacy not depend, however, on what is being correlated with what, 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 to me that, when calibrating instruments, a correlation of 0.99 was disastrous: "Bridges could fall". Pursuing that thought, should we not expect stronger correlations for predictive tests than when we are exploring the strength of various contributory factors jointly influencing an outcome? Would doctors, for example, not require a much higher correlation between a diagnostic test and the condition it purported to identify, than they would need to demonstrate that a particular treatment *might help* to alleviate pain? If scientists were to discover a 0.3 correlation between eating blueberries and living to 80, we might all be happy to eat more blueberries. However, a test which claimed to predict cancer or alzheimers two years later with a correlation of 0.7 would be unusable: there would be far too many false negatives or positives. In the first case, we are dealing with one among many contributory factors in longevity, and perhaps one that benefited some people and not others. In the second case, we are aiming for an accurate indication of an emergent illness.

We also need to consider how non-statisticians might tend to understand correlations. Many lay readers may not even realise that the scale for correlations runs from 0 to 1. Of those that do, many lay readers might assume that the movement from 0 to 1 is a uniform linear gradation, so that 0.5 represents a relationship half as strong as 1. It is not unlikely that teachers and school leaders may assume that a correlation of 0.7 means that 7 out of 10 children will hit the predicted level. Most of the potential customers for baseline tests are unlikely to understand the need to square a correlation in order to judge how much of the variance in y can be explained by the variance in x. Since $0.7^2=0.49$, a correlation of 0.7 means that only about half the variance in the later outcomes can be explained by the baseline scores.[v]

So the question remained: what does a correlation of 0.68 actually mean in terms of a test's ability to predict an individual child's future attainment? A clue came with the discovery of a short research paper by Peter Tymms (2003) of CEM. This provides not only correlations but also a Chances Table – presented as an imaginary class but based on a large dataset of the outcomes which have resulted from the baseline scores of thousands of children. (For illustration, an extract is provided below in figure 1.)
On one level the baseline scores are predictive: the table shows that children with the lowest baseline score are the most likely to score poorly in KS2 SATs, and so on. The question is rather: *how likely is it* that a child with a particular baseline score will have a particular outcome?

Towards both extremes on the scale, prediction is strong. Indeed 90% of the children with the lowest baseline score get Level 3 or below at Maths, whereas ; from 68% of those with the highest baseline score reach Level 5. Prediction is much weaker, however, nearer the centre of the baseline scores, and this is where we might expect to find the more frequently occurring scores. Of children with the midpoint score in the baseline test for Maths, 17% went on to get Level 3, 56% Level 4 and 27% Level 5 at the end of KS2.

For illustration, the following extract (Figure 1) shows KS2 levels (i.e. age 11) for "Bethany" who scored around the midpoint on baseline, "Samantha" 3th from bottom and "Rachel" 3th from top in this class of 16 children. It should be noted that in this example PIPS (Performance Indicators in Primary Schools) is used at the start of Year 5, with less than a 2 year gap before the end of KS2 outcome point: the DfE's

attempt to relate a version of PIPS at the start of Reception to the end of KS2 assessments 7 years later is likely to be more hazardous.

Name	3 or below	Level 4	5 or above
Samantha	65 %	33 %	3 %
Bethany	17 %	56 %	27 %
Rachel	8 %	49 %	53 %

Figure 1: Extract from Chances Table (Tymms 2003)

At the time 48% of children nationally attained Level 4 at the end of KS2, so this was a *very large target to hit.* A spread both sides of Level 4 suggests a very wide spread of outcomes from a single baseline score. Overall, many of the baseline scores appear to disperse across 60 or even 80 percentiles of the child population ranked by attainment. In reality, then, this predictive test appears to behave more like a sawn-off shotgun than a precision tool.
Following a Freedom of Information request, CEM provided a spreadsheet showing the percentage of children expected to secure particular outcomes from each baseline score. Despite a rider that a *model* is used with a correlation of 0.68 between PIPS as starting point and the later end-of-keystage outcomes, it is presumably rooted in their real data albeit indirectly. For each PIPS score 0-100, the spreadsheet shows the percentage of children attaining each of levels W, 1, 2C, 2B, 2A and 3 in each of Maths, Reading and Writing at the end of Key Stage 1. Thus, with the lowest possible baseline score (0), 98% are shown as reaching W (working towards Level 1). As with Peter Tymms' study, this indicates strong prediction from the extremes. However, such extreme scores could be almost nonexistent if there was a normal distribution curve.

Further enquiry revealed that the baseline scores did indeed follow a normal distribution, with approximately 68% of scores between 40 and 60, 95% between 30 and 70, and 99.8% between 20 and 80. In other words, 68% scored between 40 and 60, 27% 30-39 or 61-70, 4.8% 20-29 or 71-80. Thus the highly predictive scores at both extremes are practically irrelevant since only 0.2% of children obtain them.

A calculation was carried out to estimate the chances of a child reaching the most strongly predicted KS1 level / sub-level from each baseline score. [vi] The calculation suggests that the test, with a stated correlation of 0.68, can predict correctly for about 4 children in every 10. We should note that this is based on PIPS tests used at the end of the Reception year, and not at the start.

A further check was made on the practical utility of these baseline scores. In this case the baseline scores with the highest prediction of each outcome levels at KS1 Reading were chosen (figure 2 – most probable outcome shown in bold). Baseline scores below 20 and above 80 are omitted entirely because practically irrelevant, since only 0.2% of children altogether score in that range. Indeed it should also be noted that very few children will have scored 20, 28 or 80 because of the normal distribution referred to above. Hence the most important rows are those linked to baseline scores of 37, 44 and 53 (in italics).
The final row shows the national distribution of each outcome level or sub-level, to enable readers to judge the size of each 'target'.

Baseline score	W	1	2C	2B	2A	3
20	**59**	30	8	3	0	0
28	30	**39**	19	10	2	0
37	*9*	*28*	***26***	*24*	*11*	*2*
44	*2*	*14*	*21*	***32***	*22*	*9*
53	*0*	*3*	*9*	*24*	***34***	*30*
80	0	0	0	0	2	**98**
National distribution	2	7	8	23	27	32

Figure 2: Extract from CEM PIPS > end KS1 spreadsheet, showing baseline scores with the strongest probability of attaining each KS1 level or sub-level

In reading for example, from 44, the baseline score with the highest chance of Level 2b at KS1, only 32% actually get this level, 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 assessment next to useless, except in the case of the more extreme (but infrequent) baseline scores.

Another spreadsheet on the same design shows outcome frequencies from the start ot KS2 (i.e. Y3) to the end of KS2 (Y6). Here the

correlation is shown as 0.60, but, unlike the KS1 spreadsheet which subdivided Level 2, the most common level 4 is not subdivided into sub-levels 4a, 4b and 4c. The same method as earlier was used for reproportioning to reflect the rarity of more extreme scores. This time the initial scores made a correct prediction about 6 times out of 10. We should note, however, that we are dealing here with very large targets which are difficult to miss, since 42% of children nationally attain Level 4 and 38% attain Level 5 – 80% of children nationally are graded level 4 or 5.

This raises a further issue in interpreting claims made for predictive tests: that correlation figures might not be a sufficient guarantee of accuracy; the capacity of the test to make accurate predictions might *depend largely on the size of the outcome targets*. It is clearly far easier to predict an outcome correctly where there are fewer of them.
None of the above is to question CEM's in designing tests, choosing test items or computing the data. Rather it places a serious question mark over the viability of predictive testing with very young children and the wisdom of the DfE in pursuing this policy. (More later.)

The other providers

Another very experienced agency in assessment is the National Foundation for Educational Research (NFER). In response to emails requesting data on predictive validity, they stated that they had none because this is a new test. 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.

It is formally quite correct that the DfE refer only to school-level data. However, it seems inconceivable that teachers, schools and indeed Ofsted inspectors will not examine and track progress based on individuals. Indeed, a standard part of the Ofsted inspection process is to identify samples of low- and high-attaining children to ensure satisfactory progress, and they will expect the school to maintain an audit trail. Moreover, a response from NFER to Schools Week (27 Nov 2015)[vii] confirmed that parents and teachers would be supplied with individual profile reports for each child, 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, and the danger of using tests which were developed for children 2 or 3 years older, the response from NFER contains reference to a research paper (Muter et al 2004) in order to establish the viability of early testing. The irony is that the data in this paper actually undermines the claim that good predictions are possible in the first two years at school (see below). A particular question concerns whether test items designed for children *after* literacy or numeracy teaching, to ascertain the effect of that teaching, can legitimately be used to assess the *potential* of children to learn literacy or numeracy *before* they have been taught.

The third approved provider Early Excellence (EE) is new to the field of assessment, their core business being largely in the sale of nursery furniture and equipment. 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.

EE pride themselves on having given early years teachers the opportunity to avoid test-based assessment. There are, however, some key differences between the existing and new arrangements, and which have technical and ethical implications. It is worth reiterating that under the new arrangements:

- observations will have to take place during the first six weeks at school, rather than by the end of the year;
- the statutory 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, confirm that since the procedure is new, they have had no opportunity to assess its predictive validity by tracking pupils through from baseline to KS1. They did however conduct 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 both ends of the range, as follows.

The first example (figure 3) is of a school with high KS1 results in recent years. On the left the bars represent bands of baseline scores (each covering a fifth of the population of the 17 schools), and on the

right, KS1 outcomes (1 or below, 2c, 2b, 2a, 3 or above). The vertical axis shows the percentage of pupils in each category.

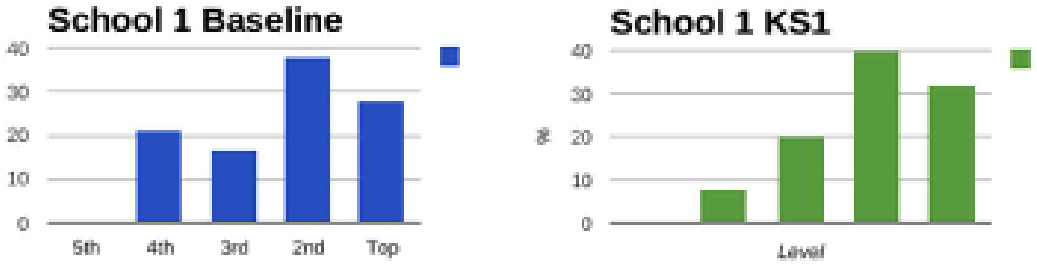

Figure 3: EE school-level data: high attaining school

Assuming that this is typical, there is clearly correspondence at school level: 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: in other words, 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, for example, or whether there was substantial movement between bands.
The situation regarding the low attaining school (figure 4) is far more problematic.

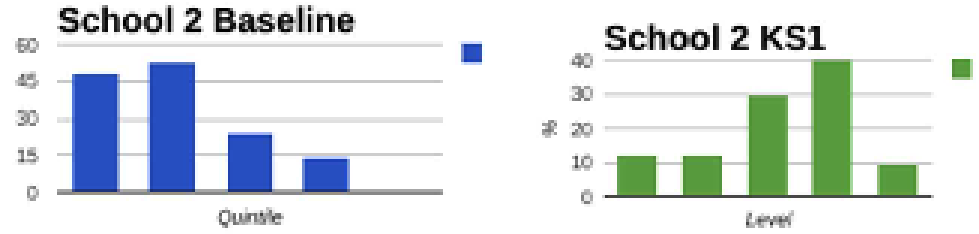

Figure 4: EE school-level data: low attaining school

In this school it is clear that most children score poorly at baseline, but a much smaller number have low attainment at the end of KS1. (The percentages on the left do not add up to 100, but even allowing for that there is a lack of correspondence.) This suggests that most of the pupils with low baseline scores will proceed to average and above average levels at KS1. As before, from this data there is no way to investigate up and down movement between bands.

The diagram highlights the serious danger that children could easily be written off as 'low potential' on the basis of their baseline scores, or that there could be such a concentration 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.

Is this surprising? Other examples of early assessment

Qualitative assessment which is flexible, provisional and sensitive to the individual child is well established in early education. What is at question here is the reliability of quantitative judgements, and particularly those used to measure the 'effectiveness' of a school or nursery.

The Early Years Foundation Stage Profile (EYFSP) is based on observations undertaken periodically in nurseries and completed by the end of Reception year. The DfE have already attempted to convert these qualitative observational assessments into numerical scores which can be matched and compared with later attainment measures. The results show 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).

As with PIPS, the DfE's own research shows that children with a higher initial score are more likely to do well later, but this is fairly meaningless since it is calculated only on the likelihood of *reaching* Level 2 as a whole, without sub-levels, i.e. a very large target, hit by around 90% of children.

Numerous correlations are provided between scores, but they are not particularly 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]. 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 assesssment (on a 9 point scale – the horizontal axis) and KS1 Reading levels / sub-levels (in percentages - the vertical axis).

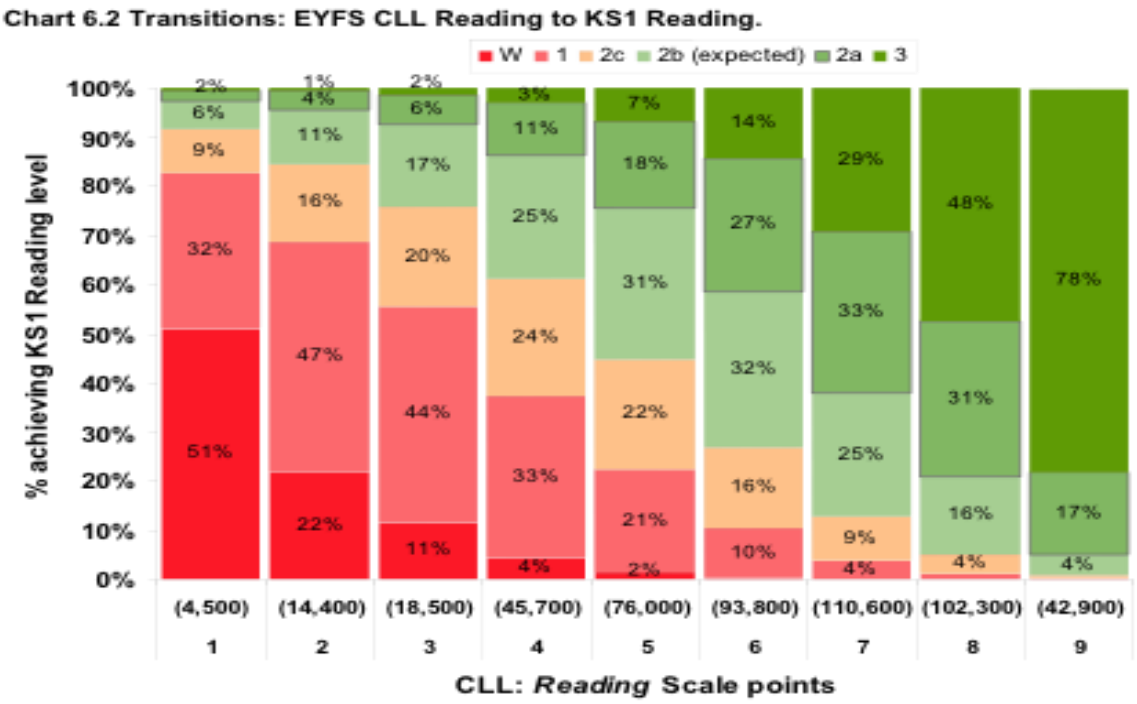

Figure 5: DfE data relating EYFSP to KS1 in Reading

We see here that children with the midpoint score (5) diverge almost equally between four bands: W or L1, 2C, 2B, and 2A or 3.

Another set of DfE data (DfE 2015) shows an interesting disjunction between the Phonics Check and Key Stage 1 Reading assessments. The phonics check is applied to all children at the end of Year 1, but repeated for those who fail 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 l3. In other words, many of the slow starters were becoming quite competent readers by the end of the following year.

This is part of a much larger problem of the accountability system when based on the notion of 'value added'. Using 'value added' data to judge school effectiveness depends on reasonably reliable norms and expectations, so that schools that deviate seriously from the norm stand out. In other words, it must be underpinned by a general assumption that progression is normally fairly smooth and linear. If progression is extremely erratic, deviation becomes meaningless.

However recent work by Education Datalab (2015) has holed the ship below the waterline. 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
- furthermore, of these children who do meet their predictions, the majority 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 reached that point via widely different levels at age 11. This is hardly the basis for systematic accountability judgements.

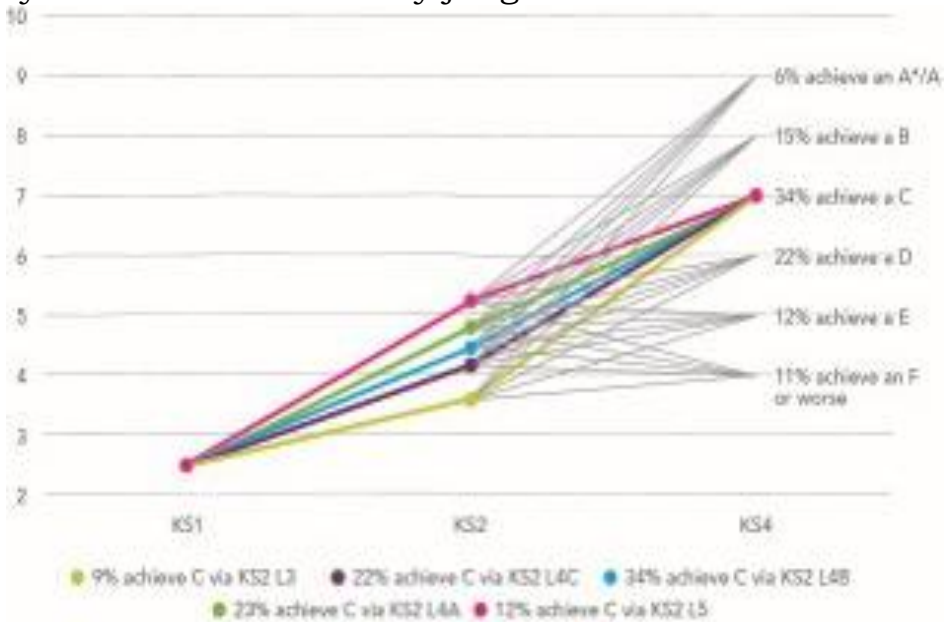

Figure 6: Education Datalab showing progression paths from KS1 to GCSE

A further finding from Education Datalab is that "children with low initial attainment have particularly unpredictable future attainment" – a conclusion which 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).

Ironically, some of the difficulties at the early stages are revealed in the 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 text of this research report is pulled apart by the contradictions between its highly positive verbal claims in favour of predictability and the statistical data which fails to back up these claims. The contradictory words and figures are frequently held together by the word 'significant', which slides between its technical / statistical use and its everyday meaning: statistical *significance* is used in ways which appear, wrongly, to suggest size or importance (i.e. the vernacular meaning), thus "Reading ability at Time 3 was predicted at significant levels by all Time 2 measures". (Time 1 = early in Reception, time 2 = early in year 1, times 3 early in year 2).

According to this research report, the correlation between scores on the same reading test on two occasions just a year apart is 0.71, with considerably lower correlations on other factors. For example:

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

One area of considerable disjunction is between reading in the sense of word recognition and reading in the sense of understanding. 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). The authors point out that "the growth of word recognition abilities is relatively uninfluenced by vocabulary and grammatical skills". The problem is that vocabulary development may go unnoticed in early testing focused on letter recognition and similar sub-skills, and might be neglected in teaching which focuses overwhelming on such sub-skills, yet it is a crucial factor in reading for comprehension.

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, and indeed this argument was used by CEM (2012) in advocating strongly for baseline assessment in response to a DfE policy consultation. [viii]
One might also argue (as CEM have done) that teachers should be less deterministic in their interpretation and use of assessment data. This may well be true, but it becomes very difficult for teachers because of

the aura of science surrounding the statistical data which conveys an impression of transparency, impartiality and certainty.
The problem we face here is a particular instance of modernity's faith in numbers, and the belief that they provide a more reliable and impartial account of reality than verbal explanation. Some insight into this can be found in Mary Poovey's fascinating historical study *A history of the modern fact* (1998) which traces back to the 16th Century not only the *prestige* attached to numbers, but the ways in which formal precision can serve to disguise a poor match with reality. This has become more intense in recent decades in the context of the 'audit society' (see Power, 1997). One of the features highlighted by Michael Power is that formal checks on systems come to appear more important than ensuring the truthfulness of the data. It is interesting, then, that the DfE's criteria for approving the various providers of baseline assessment are formalistic, relating to internal consistency rather than external applicability or truth. (See the section headed Reliability in DfE 2014c)

Power discusses two possible responses to pervasive auditing: *decoupling*, whereby service providers pay lip-service to the audit, regarding it as separate from substantive operations; and *colonization*, whereby it

> penetrates deep into the core... not just in terms of requiring energy and resources to conform to new reporting demands but in the creation over time of new mentalities, new incentives and perceptions of significance. (Power 1997:97)

When colonization occurs, auditing has corrosive side-effects and becomes a 'fatal remedy' (Sieber 1981). Power argues that, in general, both decoupling and colonization are likely to occur, creating tensions within the organisational culture and in the minds of its individual members. This is certainly apparent among teachers: accountability data is hated for creating impossible pressures but can also serve as a comfort blanket (and indeed, teachers may value competitive data because it will keep Ofsted away). For many teachers, high-stakes accountability feels alien to real educational values and relationships – their reason for being a teacher – but at the same time it has come to permeate their discourse and activity.

Both decoupling and colonization are already emerging 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. It is not valid as a basis for planning to

> support children's learning as it does not reflect the most important areas of learning and development in the early years and will not serve children's later success. [TACTYC / Early Education 2015)

However such decoupling will prove difficult if not impossible; the colonization impulse will be fuelled by professional insecurity and the anxiety to make early judgements of children's 'ability' and 'potential'. There is a substantial critical literature, in educational sociology and critical policy studies, concerning 'governance by numbers', and especially in high-stakes accountability systems such as England (eg Ozga and Lingard 2007; Ranson 2003). The deployment of assessment data to hold schools accountable has various kinds of impact. As James Scott demonstrates in his book *Seeing like a state* (1998), measurement has the power to reconstitute the world it seeks to measure. High-stakes accountability systems in education focus attention on what is most easily measurable, leading to the neglect of less quantifiable outcomes (creativity, kindness, aesthetic or ethical sensibilities, among others). Accountability pressures can impact on relationships, so that teachers begin to perceive individuals as examples of official categories (a "white British, free school meals male with SEN"), or start to view students instrumentally in terms of the benefit to the school's performance data and reputation (Fielding 2001). Thrupp and Willmott (2003:119) also comment on the triage effect whereby teachers begin to focus on helping borderline students clear a particular hurdle (turning the Ds into Cs at GCSE) while neglecting higher and lower attaining students.

The shift of high-stakes accountability downwards into Reception Year has its own risks. Firstly, 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' (Palmer 2009). Secondly, it will encourage the practice of segregating children into 'ability groups' from an early age. Thirdly, it is likely to reduce expectations and 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 (many of whom accelerate later), children with health problems, and the large numbers of children growing up in poverty.

Statistics is necessarily reductionist, in that it must lose detail in order to present a summative overview of multiple individuals and events: this is the price we pay for its analytical and representational

power. However it is important for readers and users, as well as statisticians themselves, to engage in rescuing and rebuilding the complexity of the real world from the data. When illusions are sown among educators about the accuracy and reliability of numerical judgements on young children in the interests of a draconian accountability machine, the reductionist labels can seriously distort the child's development and becomes a vicious circle of self-fulfilling prophecy.

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

CEM (2014a) Getting the measure of primary. http://www.cem.org/attachments/Getting%20the%20measure%20of%20primary%20-%20brochure%2014072014.pdf

CEM (2014b) Getting the measure of early years. http://www.cem.org/attachments/Getting%20the%20measure%20of%20early%20years%20-%20brochure%2014072014.pdf

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

DfE (2014b) Reception baseline: criteria for potential assessments. 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

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

Fielding, M (2001) Target setting, policy pathology and student perspectives: Learning to labour in new times. In M Fielding (ed) Taking education really seriously: Four years hard Labour. London: Routledge

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

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

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): 665:681

Ozga, J and Lingard, B (2007) Globalisation, education policy and politics. In B Lingard and J Ozga (eds) The RoutledgeFalmer Reader in Education Policy and Politics. London: Routledge

Palmer, S (2009) Four years bad, six years good, seven years optimal. Literacy Today, December 2009. http://www.suepalmer.co.uk/modern_childhood_articles_four_years.php

Poovey, M (1998) A history of the modern fact: Problems of knowledge in the sciences of wealth and society. Chicago: University of Chicago Press

Power, M (1997) The audit society: Rituals of verification. Oxford: Oxford University Press

Ranson, S (2003) Public accountability in the age of neo-liberal governance. Journal of Education Policy, 18(5):459:80

Scott, J (1998) Seeing like a state: how certain schemes to improve the human condition have failed. New Haven: Yale University Press

Sieber, S (1981) Fatal remedies: The ironies of social intervention. New Y0ork: Plenum Press

Stobbart, 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

Thrupp, M and Willmott, R (2003) Education management in managerialist times: beyond the textual apologists. Maidenhead: Open University Press

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

****

[i]. Some of the providers offer assessment tools for other aspects of development as extras in the same package, though this is not reflected in the scoring to be reported to the Department for Education. It is also not clear what will happen to the broadly based Early Years Foundation Stage Profile which is currently required by the end of Reception year.

[ii]. The word 'test' is used frequently in this paper, although one of the providers Early Excellence is scoring children on the basis of teachers' observations. This will be discussed on a later page.

[iii]. Examples include "Links sounds to letters, naming and sounding the letters of the alphabet". Does this mean all the sounds and letters? consistently?

[iv] For example, "What sounds are in the word 'net'?"

[v].This does not necessarily mean that only half the children receive the predicted score, but that is not far wrong, as later data reveals.

[vi].The most frequent outcome score was identified for each baseline score, the percentage obtaining that score was noted, these percentages were averaged within each of three frequency bands (40-60; 30-39 + 61-70; 20-29 + 71-80), and finally a reduced weighting was given to the second and third band to reflect the lower numbers obtaining those baseline scores. (Factor of 0.397 for band 2; 0.07 for band 3. Scores below 20 and above 80 were ignored because virtually nonexistent.)

[vii]. http://schoolsweek.co.uk/two-more-reception-baseline-tests-come-under-the-spotlight/

[viii]. Although CEM use the expression 'each child's developmental level', the DfE appear to be working to different assumptions since they insist on ignoring the child's age – a strong factor in 'developmental level' at the age of 4-5 years.