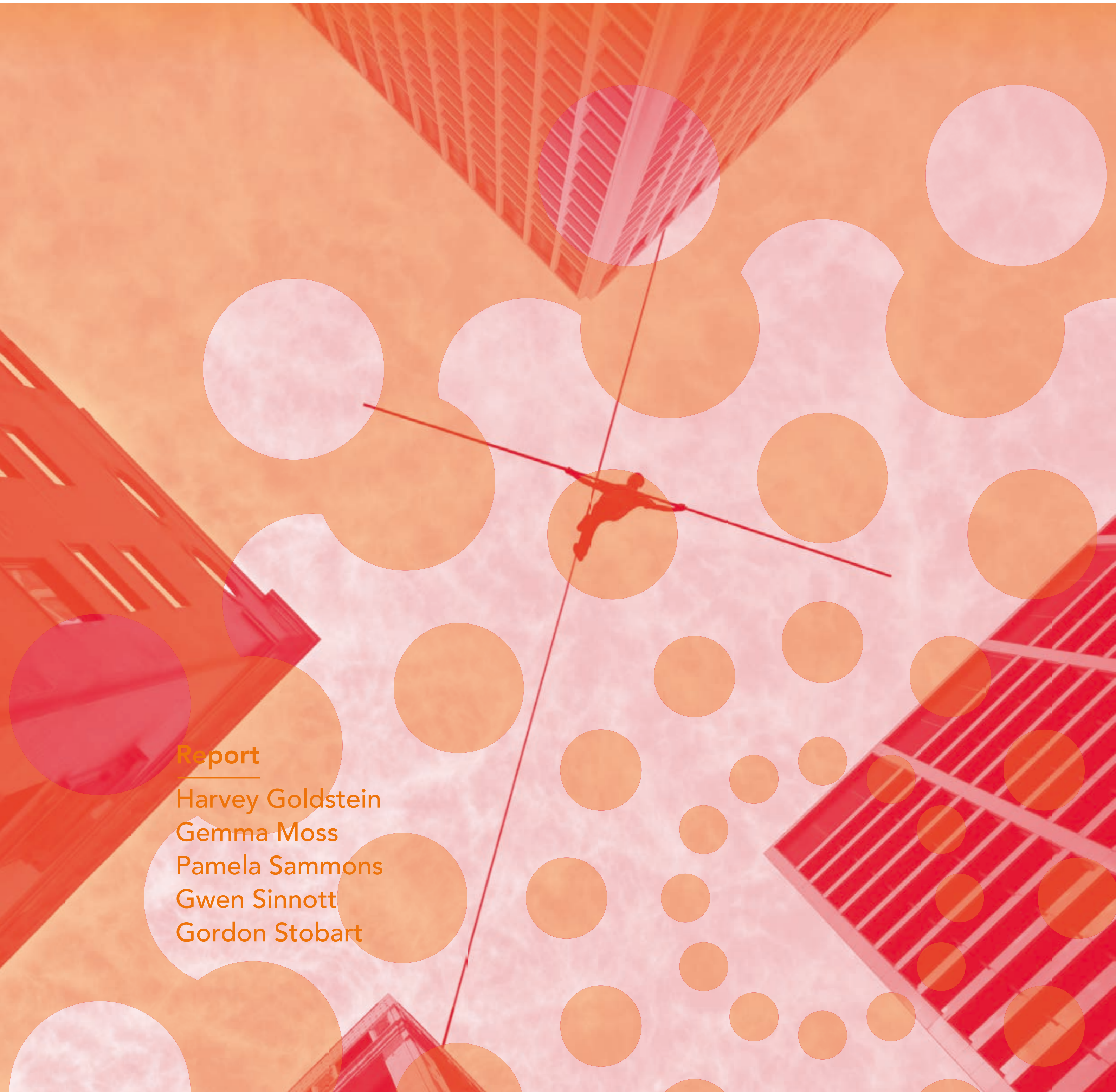
BERA
BRITISH EDUCATIONAL RESEARCH ASSOCIATION

A BASELINE WITHOUT BASIS

THE VALIDITY AND UTILITY OF THE PROPOSED RECEPTION BASELINE ASSESSMENT IN ENGLAND



Report

Harvey Goldstein
Gemma Moss
Pamela Sammons
Gwen Sinnott
Gordon Stobart

ABOUT BERA

The British Educational Research Association (BERA) is the home of educational research in the United Kingdom. We are a membership association committed to advancing knowledge of education by sustaining a strong and high quality educational research community.

Together with our members, BERA is working to:

- advance research quality
- build research capacity
- foster research engagement.

Since its inception in 1974, BERA has expanded into an internationally renowned association with both UK and non-UK based members. It strives to be inclusive of the diversity of educational research and scholarship, and welcomes members from a wide range of disciplinary backgrounds, theoretical orientations, methodological approaches, sectoral interests and institutional affiliations. It also encourages the development of productive relationships with other associations within and beyond the UK.

Aspiring to be the home of all educational researchers in the UK, BERA provides opportunities for everyone active in this field to contribute through its portfolio of distinguished publications, its world-class conference and other events, and its active peer community, organised around 30 special interest groups. We also recognise excellence in educational research through our range of awards. In addition to our member-focussed activity, we aim to inform the development of policy and practice by promoting the best quality evidence produced by educational research.

***

This report presents the views of an expert panel established by BERA in early 2018 to consider the evidence behind the government's proposal to use a baseline assessment test of pupils in reception to hold schools in England to account for the progress that pupils have made at the end of key stage 2.

Contents

Summary 1

Introduction 6

The validity of the proposed test 6

1. Is it legitimate to use baseline assessment for school accountability purposes seven years hence? 8

1.1 The ethics of testing young children for accountability purposes 8

1.2 Is it feasible to test four-year-olds for accountability purposes? 9

1.3 Can baseline tests in reception be used to calculate school value-added at key stage 2? 10

1.4 Is the development brief for the test appropriate? 10

2. Will the proposed tests be accurate or fair? 13

2.1 How reliable will the baseline tests be? 13

2.2 What impact will the tests have on learners? 13

2.3 Can fair judgements be made using the baseline data? 14

3. What recognition is being given to contextual factors in the interpretation of the data? 16

3.1 The impact of pupil and teacher mobility 16

3.2 The impact of socioeconomic and family factors 17

4. Will this form of accountability lead to useful comparisons of schools? 19

4.1 The utility of school performance data for parental school-choice 19

5. What is the likely impact of these accountability measures on pupils and schools? 22

5.1 Delaying feedback 22

5.2 Gaming baseline scores 22

5.3 Distorting younger children's provision 23

5.4 Adverse impacts on children in reception 24

5.5 Unintended consequences: 'labelling' versus individualised support 24

5.6 The narrowing of the curriculum throughout primary schooling 25

6. Are there better alternatives to baseline testing? 26

6.1 Examples of intelligent accountability in action 27

Conclusion 29

References 31

About the panel 36

Published in July 2018 by the British Educational Research Association

British Educational Research Association (BERA)
9–11 Endsleigh Gardens
London WC1H 0EH

www.bera.ac.uk
enquiries@bera.ac.uk
020 7612 6987

Charity Number: 1150237

Download

This document is available to download from:
https://www.bera.ac.uk/researchers-resources/publications/a-baseline-without-basis

Citation

If referring to or quoting from this document in your own writing, our preferred citation is as follows.

> Goldstein H, Moss G, Sammons P, Sinnott G and Stobart G (2018) *A baseline without basis: The validity and utility of the proposed reception baseline assessment in England*, London: British Educational Research Association. https://www.bera.ac.uk/researchers-resources/publications/a-baseline-without-basis

Permission to share

SUMMARY

This report sets out the case against the government's proposal to use a baseline assessment test of pupils in reception to hold schools in England to account for the progress that pupils have made at the end of key stage 2.

When the government's plans were published in 2018, BERA convened an expert panel to consider whether the evidence from the assessment literature could justify such a test being used for this purpose. The conclusion of the expert panel is that it cannot. This report is intended to inform public debate by providing an accessible account of the reasons why the proposals are flawed.

In the panel's view the proposed baseline assessment will not lead to accurate comparisons being made between schools, as policymakers assume. Perhaps most importantly, they will not work in the best interests of children and their parents.

The panel has drawn the following conclusions on the basis of the evidence set out in the body of this report.

1. Under these proposals, children will be exposed to tests that will offer no formative help in establishing their needs and/or in developing teaching strategies capable of meeting them.
2. Any value-added calculations that will be used to hold schools to account will be highly unreliable.
3. This is an untried experiment that cannot be properly evaluated until at least 2027, when the first cohort tested at reception has taken key stage 2 tests.

The panel arrived at their view by seeking answers to the following six questions, summarised below.

1. Is it legitimate to use baseline assessment for school accountability purposes seven years hence?

It is both ethically and methodologically questionable to use reception baseline assessment (RBA) for such a purpose. As currently proposed, RBA is likely to produce results with little predictive power and dubious validity.

The assessment of very young children may be ethically justifiable when used to support a child's learning, in which case they stand to benefit directly. However, the government's RBA will be used solely for school accountability, a purpose for which the test is not fit.

The research evidence demonstrates that any early-years assessment system will have little predictive power. Aggregating scores in a proposed 20-minute test, covering the three domains of literacy, numeracy and self-regulation to produce a single number, is misguided. Besides its inherent unreliability, it would ignore the fact that children may perform differently in each domain, and that some domains are better predictors of progress in different areas of the curriculum than others. Generalising from a cohort to the school would be unwise given the limited sample size in each primary school. Furthermore, no proposals have been made regarding how predictive validity will be investigated and reported across different years.

For accountability purposes, it remains unclear whether the reception baseline tests are intended to align, in terms of method and content, with the relatively narrow formal testing at key stage 2 against which pupils' progress in the intervening years will be measured. Insisting on a close alignment may result in a narrowing of the early-years and primary curricula.

2. Will the proposed tests be accurate or fair?

There is good reason to question the reliability of the data that the test will produce, and the ways in which that data will be interpreted and acted upon.

The panel expects the baseline tests to show low levels of reliability because, firstly, no indications have been given that age effects will be controlled for at both the initial baseline test and the outcome tests at key stage 2 – yet this is essential if the data is to be used for school accountability purposes, for two reasons.

- Just a few month's difference in age has been shown to produce pronounced developmental differences at reception age. Autumn-born children have demonstrated a strong advantage in attainment over their younger, summer-born peers in assessments similar to the one proposed.
- Pupil cohorts within primary schools are statistically small, and often have uneven distributions of younger and older children. Schools serving more children who are young for their year of entry may appear to have less favourable effects on children's later attainment than those that serve children who are old for their year, unless age and season of birth are accounted for with sufficient precision.

Secondly, pupil mobility poses a problem if the RBA is intended as a measure of pupil progress seven years hence: either mobile pupils will have to be taken out of the progress measure in all schools, resulting in varying numbers of children being 'missing' from the accountability measure; or baseline assessment results will 'follow' pupils between schools, resulting in schools being held accountable for pupils' progress despite being unaware of their starting points, and having been responsible for only part of pupils' school lives. Teacher turnover, and the likelihood of a change in head teacher over a seven-year period, will also muddy the issue of accountability.

3. What recognition is being given to contextual factors in the interpretation of the data?

It is generally recognised that the only proper way to make comparisons between schools is to make adjustments for the prior attainments of their pupils when they enter those schools, and to control for other relevant characteristics of pupil intakes such as parents' educational levels, family income and having English as an additional language. Such adjustments lead to what are known as 'value added' comparisons. There is strong evidence that these characteristics affect both attainment and relative attainment in value-added measures. However, under the government's current proposals, school-level attainment at year 6 will be adjusted for using the reception baseline assessments alone, and *without* controlling for any contextual factors. This approach cannot lead to fair comparisons.

4. Will this form of accountability lead to useful comparisons of schools?

The available evidence suggests not.

Little research has been carried out on the efficacy of using pupil progress measures to hold schools to account at the primary level. However, research at the *secondary* level has found that, when ranking schools in this way:

- value-added scores suffer from considerable statistical uncertainty due to low sample sizes
- data used to inform parents' choice of school is extrapolated from the results of students who entered school several years earlier, and is thus significantly 'out of date'
- the fact that specific sets of value-added school effects will prevail at any given time means that predictions for new cohorts based on this data will be very weak.

The analogous uncertainties will almost certainly be greater for the reception baseline tests, because:

- the time-gap between reception and year 6 is greater than between year 6 and year 11
- baseline tests of very young children will be prone to greater inherent measurement error.

5. What is the likely impact of these accountability measures on pupils and schools?

The results themselves will do little to help secure positive outcomes for pupils, teachers or parents in either the short or longer term.

The government intends to hold the baseline test data until the cohort reaches key stage 2. It is not yet known whether it will release a limited set of data to schools during the test year. Certainly, publishing the data at the point at which it is collected in reception could encourage the production of statistically worthless ranked league tables of school performance. Conversely, while non-disclosure of the data may prevent the over-interpretation of individuals' and schools' results from a potentially unreliable test, it is likely to frustrate teachers and parents, who may well ask, 'Why administer a test that doesn't help teaching and learning?'

While the assessment is not intended to have any diagnostic value for schools and individual children, teachers administering the test will see children's scores. This could mean that some children – particularly the summer-born, those with English as an additional language and those with special educational needs – could be unnecessarily labelled as low-ability at the very beginning of their education, with the risk that premature judgements about their abilities may then become 'self-fulfilling'.

6. Are there better alternatives to baseline testing?

Baseline testing reflects a more general trend in public services towards using 'performance indicators chosen for ease of measurement and control rather than because they measure quality of performance accurately'.[1] However, there are alternatives. An 'intelligent accountability' approach would allow practitioners to use their professional judgement more fully in the assessment process – gathering deeper and more meaningful data that can take account of contextual factors, help to support individual pupils, and inform improvement planning both within and between schools.

[1] O'Neil O (2002) *A Question of Trust: The BBC Reith Lectures 2002*, Cambridge: Cambridge University Press: 54

Both the Surrey value-added initiative and the London Education Research Network, among other examples, have demonstrated that principles of intelligent accountability can readily be adopted and put into practice to support school improvement and spread good practice in the use of data.

***

The panel believe that the government's proposals for the reception baseline assessment are flawed, unjustified and wholly unfit for purpose. They would be detrimental to children, parents, teachers and the wider education system in England. We publish this report in the hope of informing public debate by offering an accessible and thorough account of why these proposals must be comprehensively rethought.

INTRODUCTION
THE VALIDITY OF THE PROPOSED TEST

This report sets out the case against the government's proposal to use a baseline assessment test of pupils in reception to hold schools in England to account for the progress that pupils have made at the end of key stage 2.

When the government's plans were published in 2018, BERA convened an expert panel to consider whether the evidence from the assessment literature could justify such a test being used for this purpose. The conclusion of the expert panel is that it cannot. This report is intended to inform public debate by providing an accessible account of the reasons why the proposals are flawed.

In the panel's view the proposed baseline assessment will not lead to accurate comparisons being made between schools, as policymakers assume. Perhaps most importantly, they will not work in the best interests of children and their parents.

In considering the evidence and arriving at their conclusions, the panel has paid particular attention to the validity of the proposed test.

> *'**Validity** refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.'*
>
> (AERA et al 2014: 11)

The first step in evaluating the validity of any test is to establish its purpose. Only then can we judge whether it achieves this purpose, and is thus a valid test.

Unusually for a national test, the new baseline test in reception has only one overriding purpose.[2] This is to provide data on the achievement levels of pupils on entry into their reception class in primary schools. This data will then be used seven years later to make 'value-added' comparisons among schools, comparing the baseline assessment to the same pupils' key stage 2 scores in year 6.

In its evaluation the panel considered whether this is a legitimate purpose, and whether the proposed test will be able to achieve the desired outcome 'fairly and accurately', as Justine Greening argued it would (DfE 2017a: 3). The panel

[2] Paul Newton (2007) has identified at least 18 purposes for which assessments can be used, ranging from monitoring national standards to providing teachers with information about individual pupils.

considered the six specific questions set out below – each of which will be explored in detail in the six chapters that make up the remainder of this report.

1. Is it legitimate to use baseline assessment for school accountability purposes seven years hence?
2. Will the proposed tests be accurate or fair?
3. What recognition is being given to contextual factors in the interpretation of the data?
4. Will this form of accountability lead to useful comparisons of schools?
5. What is the likely impact of these accountability measures on pupils and schools?
6. Are there better alternatives to baseline testing?

1. IS IT LEGITIMATE TO USE BASELINE ASSESSMENT FOR SCHOOL ACCOUNTABILITY PURPOSES SEVEN YEARS HENCE?

1.1 The ethics of testing young children for accountability purposes

The panel asked, **Is it ethical to ask very young children to sit a baseline attainment test, on entrance to school, from which neither they nor their teachers will receive any direct benefit?**

Responses to the Department for Education (DfE's) consultation on the baseline tests suggest that many early years practitioners believe it is unethical to test children who have just arrived in school, often from very diverse backgrounds, and who may be settling in to an unfamiliar environment (Bradbury and Robert-Holmes 2016). The decision to conduct the test as soon as possible in the autumn term heightens these concerns.

Assessment of young children may be justified if the purpose is diagnostic or formative, and used to support a child's learning. This was the rationale of the early years foundation stage assessments, and other diagnostic tools such as Performance Indicators in Primary Schools (PIPS) (see Tymms et al 2014).

However, the results of the reception baseline assessment will be used for school accountability purposes, rather than to support individual children's learning. It is noteworthy that the Centre for Educational Measurement, which ran one of the three baseline schemes piloted in 2015, withdrew from the test development bidding process on the grounds that it was 'verging on the immoral' to use the test for accountability purposes alone (quoted in Bradbury et al 2018: 16). Another provider, Early Excellence – whose scheme was the only one among the three to be solely observation-based – withdrew in November 2017, calling the government's proposals 'self-contradictory, incoherent, unworkable and ultimately inaccurate, invalid and unusable' (Ward 2017).

These ethical concerns are supported by findings from the US, where there has been extensive research into early years assessments, particularly in relation to the concept of school-readiness (Shepard et al 1998; LaParo and

Pianta 2000; Kim and Suen 2003). The general finding of this research is that '[i]nstability is more the case than not in early childhood development, and tests of accountability that overlook the implications of this variability will mislead policymakers, the public and children's teachers' (Meisels and Atkins-Burnett 2008: 543).

1.2 Is it feasible to test four-year-olds for accountability purposes?

The panel asked, **Will the testing process meet the requirements that fair testing involves?**

The research evidence suggests that, in any early-years assessment system, multiple assessment measures are required and results should always be interpreted with caution (ibid). A survey of 44 studies by Kim and Suen concluded that 'the predictive power of any early assessment from any single study is not generalizable, regardless of the design or quality of the research' (2003: 561).

This variability was evidenced in the initial pilot of baseline testing, in which three different assessment methodologies were used: observational (the Early Excellence baseline assessment); a computer-based test (the Centre for Evaluation and Monitoring's BASE assessment); and a resource-based assessment using a mixture of tasks and observational checklists (the National Foundation for Educational Research [NFER's] reception baseline assessment). An evaluation study concluded that the results from the different tests lacked sufficient comparability to create a fair starting point from which to measure pupils' progress (STA 2016: 20).

Since then, the government has announced NFER as its preferred supplier for the assessment (Ward 2018a). However, **switching to a single provider with a single assessment methodology will only mask the problem of how a particular test format determines the results of that test, particularly among young children with little or no experience of test-taking**. If a different test were used, the results for many children would differ – as the government's own study illustrated (STA 2016). Indeed, similar conclusions were reached in previous research by the Qualifications and Curriculum Authority, which confirmed that baseline tests vary in the extent to which they can accurately measure differences in performance between groups of pupils (Sammons et al 2000).

If the results for individual children would differ on different tests, so too will the predictive ability of any single test in terms of the government's stated intention of comparing schools' effectiveness with 'value added' calculations seven years later.

1.3 Can baseline tests in reception be used to calculate school value-added at key stage 2?

The panel asked, **Will the test be fit for purpose?**

Modern validity theorising centres on 'construct validity'. Only once we know what the construct, domain or skill is that is being tested can we decide whether a particular assessment is fit for purpose. Two major threats to validity are not adequately sampling the domain, and assessing elements that are not part of the construct.[3] This raises particular issues in the case of the reception baseline assessment – a test designed to be used for accountability purposes.

In explaining why the government had ruled out observational tests, early years minister Nadhim Zahawi said, 'The data from the baseline needed to correlate with key stage 2 assessment so that "like for like" comparisons could be made' (Ward 2018b). Such a statement is inherently misleading, as correlations will exist even for unlike tests, and could have been explored using data from observational tests.

In fact, it is not yet clear how close an alignment is really intended between a reception baseline and the tests to be conducted at the end of key stage 2, nor what the consequences would be of making this the goal. In line with their pilot test, the developer, NFER, proposes 'child-friendly and accessible' assessment tasks, using physical resources such as 'counting teddy bears, plastic shapes and picture sequencing cards' (NFER, no date). However, this assessment approach was not designed principally to correlate with the far narrower formal testing that takes place at key stage 2. If policymakers insist on close alignment with concepts tested at key stage 2, then the baseline tests may well be narrowed. (This has happened before, albeit in the key stage 3 context – see Whetton 2009: 143–145). As things stand, it will be seven years before we can know how valid any proposed alignment is.

Whatever is tested at the baseline stage must be, first and foremost, aspects of cognitive development that are appropriate for that age. Certainly, the content of the baseline test should not be based on or treated as preparation for the content of the key stage 2 tests.

1.4 Is the development brief for the test appropriate?

The development brief for the tests (DfE 2017a) stipulates that the test will be 20 minutes long, will be accessible to 99 per cent of the cohort, and will offer a wide spread of marks with no more than 2.5 per cent of

[3] Accounts of validity theory often adopt Samuel Messick's (1989) classic 'threats to validity': *construct under-representation* and *construct-irrelevant variance*. An example of the former would be focussing a language test on writing and ignoring speaking; an example of the latter would be a maths test that requires such a high level of reading skill that it rewards good readers rather than good mathematicians.

takers receiving full marks.

The panel asked, **Are these proposals appropriate?**

One of the threats to a test's validity is the way in which it is scored and marks are aggregated (Crooks et al 1996). Crooks et al describe these threats in further detail in the following terms.

- Scoring fails to capture some important qualities of task performance.
- Undue emphasis on some criteria, forms or styles of response.
- Lack of intra-rater or inter-rater consistency.
- Scoring is too analytic or holistic.
- Aggregated tasks are too diverse.
- Inappropriate weights are given to different aspects of performance (ibid: 270).

NFER proposes that the 20-minute test will cover the three areas of literacy, numeracy and self-regulation. The scoring of these different areas is likely to fall foul of this checklist by being too diverse to meaningfully aggregate.

An individual pupil's score will vary by whatever weightings are given to the three constructs included in the assessment. Adding up children's scores in each area in order to produce a simple overall score, as is currently proposed, would ignore the fact that children may perform differently in different domains, and that some domains may be better predictors of later key stage 2 results than others. The major longitudinal, DfE-funded 'Effective Pre-school, Primary and Secondary Education' research (EPPSE 3–16+) found that pre-reading was a better predictor of later reading at the ages of six and seven, while early number concepts were a better predictor of later maths, and the baseline measure of children's 'independence and concentration' was the stronger predictor of later self-regulation. In each case, they found significant background effects related to child age, gender, ethnic group, family socioeconomic status, parents' education and home learning environment on baseline attainment (at school entry) and in attainment in years 1, 2, 5 and 6 (Sammons et al 2002, 2003, 2004, 2008a, 2008b; Sylva et al 2004, 2006).

This evidence bolsters the argument that the three domains of early literacy, numeracy and self-regulation are indeed distinctive, and are better treated separately when predicting children's later attainment. However, although the test could offer a profile across the three constructs, the sampling of each domain would be so limited in a 20-minute test that it would tend to have both poor validity and lower reliability for the different constructs covered. It is not clear whether and how the test developer will address these scoring issues through the trial and pilot phases.

The government's intention to produce one overall score is misguided. The unreliability inherent in administering a 20-minute test of a range of skills to young children has not been estimated and reported by NFER; nor have any proposals been forthcoming regarding how predictive validity will be investigated and reported across different years. Nothing has been said on the importance of contextualising results to obtain 'fair' measures that can account for differences in school intake – even though the evidence is that tests may indeed have different predictive validity for different groups of pupils (Tymms et al 2014: 21, 31).

2. WILL THE PROPOSED TESTS BE ACCURATE OR FAIR?

2.1 How reliable will the baseline tests be?

The panel expects the baseline tests to show low levels of reliability as a consequence of both their format and the inevitable variations in administration as teachers seek to explain to young children – just settling into school and with no prior experience of test-taking – what is required.

Reliability refers to 'the consistency of outcomes that would be observed from an assessment process were it to be repeated… [It] is about quantifying the luck of the draw' (Newton 2009: 51). In the case of the reception baseline assessment, its level of reliability will be affected by whether pupils would have received different results had they taken it on a different day, taken a different version of the test, had a different assessor, or been introduced to the test in a different way.

The fact that test-takers would be young and inexperienced also poses considerable reliability problems. Many will not have taken such a test before, and may still be anxious about starting school. Even if teachers are trained to administer these tests in a comparable way, there will be inconsistencies between teachers and between schools in terms of the levels of support they offer to different children. Indeed, given the age and inexperience of the children, test administration will need to be individualised, and this could itself prove a source of considerable inconsistency both between children and between different classes and schools. This additional unreliability will further lower the predictability of the key stage 2 test scores.

2.2 What impact will the tests have on learners?

It cannot be assumed, as the policy currently appears to, that a short baseline test will have no impact on the test-taker: such tests always have an impact, whether directly or indirectly (Stiggins 2000). The intention of the baseline test is to sum up, in 20 minutes, what a child is bringing into school (NFER, no date). Background factors such as deprivation, home language and age may mean that some children will have limited success on the tests, and/or find them disproportionally stressful. The children who struggled most with the test may also be labelled as a consequence – something that can become 'self-fulfilling' (Hart et al 2004; Boaler 2009). This is a particular risk if teachers are encouraged to make premature judgments about children's abilities and/or their family context from the test-taking process. Yet NFER and the DfE have set out no plans to systematically evaluate the impact of the test on pupils, teachers, parents and schools to ensure that its introduction has no negative consequences.

2.3 Can fair judgements be made using the baseline data?

The panel asked, **How will the results be interpreted and used?**

The widely accepted definition of validity as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (AERA et al 2014: 11) emphasises the need for careful interpretation of the results. A test can be well-constructed and scored, yet the results can then be misunderstood or misused. For example, some media outlets have reported that year 6 children scoring at level 3 in reading are 'unable to read' or 'illiterate' – an inaccurate and misleading interpretation (see for example Clark 2010; see also HoC-ESC 2005: 6, 34).

This is a particular risk in the case of test data collected from the very young. The scoring of the test must adjust properly for the age of the child, for the simple reason that at such an early stage of life a few months' difference in age may lead to pronounced developmental differences – six months is a considerable developmental period at the age of four. We discuss this in further detail below.

2.3.1 The effects of child age by month

The proposed baseline reception test will be taken by children who vary in age by up to 12 months. This is a very important consideration, as it is widely recognised that there are important age-related developmental effects that are especially striking among young children. Well-constructed tests for the very young are therefore frequently age-standardised in order to explicitly control for every month in age difference.

Typically, autumn-born children show a strong advantage in attainment over their younger, summer-born peers (differences being most marked between the oldest autumn-born and youngest summer-born children, who have nearly a year of age between them). An early value-added study of pupils' progress from reception baseline to the end of key stage 1, conducted for formative school improvement purposes (rather than for accountability), revealed large age effects for all key stage 1 areas covered (reading, writing, maths and science) (Sammons and Smees 1998). Importantly, the authors noted the following finding.

> *'Older pupils did better in all areas than younger members of the year group. Because prior attainment [at reception entry] is controlled, this means that older pupils made more progress over the infant years, as well as having higher initial attainments at entry.'*
>
> (Ibid: 398)

This problem remains intractable, as has been revealed by more recent research comparing children's performance upon entry to school and the progress made in the first year of school. Tymms et al (2014) found that in their data the correlations between test scores and child age range between 0.21 for phonological awareness and 0.30 for early mathematics. For the total score the age correlation is 0.31, which represents a substantial age effect (ibid: 32). Indeed, Tymms et al also note, in relation to the age effect for personal social and emotional development,

> *'The older the child on starting school, the higher the ratings tend to be on each item. The effect sizes are modest but clear; the older children were seen to concentrate more, to feel more comfortable, to communicate better, to have better relationships and so on. There was a fairly constant effect across all items.'*
>
> (Ibid: 36)

Thus, statistical control for age effects is highly desirable, while to ignore age effects on attainment is likely to prove misleading.

2.3.2 Age cohorts in primary schools

For the government's stated purpose for the reception baseline assessment – school accountability – the age effect is particularly problematic, since pupil cohorts within primary schools are, statistically speaking, very small (often only one or two classes of children), and the distribution of younger and older children can be quite uneven. Schools that serve more children who are *young* for their year at entry may appear to have less favourable effects on children's later attainment in year 6 than schools serving more children who are *old* for their school year. Unless age effects are controlled for at both the initial baseline test and for outcome tests at key stage 2, it will not be possible to establish how schools' value-added results will be affected by the proportion of younger or older children in their cohorts in any given year.

3. WHAT RECOGNITION IS BEING GIVEN TO CONTEXTUAL FACTORS IN THE INTERPRETATION OF THE DATA?

3.1 The impact of pupil and teacher mobility

For accountability purposes, the baseline data will first be interpreted at the school level at the end of key stage 2. Yet there are no proposals to take into account, when measuring progress over the seven years from reception to the end of primary school, the length of time for which a pupil has been in the same school or that a headteacher has been in post (and accountable for pupils' progress), or the rate of teacher mobility.

Mobility in school, in relation to both teachers and pupils, is a much-discussed issue. Already, the government has decided that infant, first, middle and junior schools will be exempted from the baseline-to-end-of-key-stage-2 accountability measures because the continuity for pupils and accountability for headteachers is disjointed when pupils move from one such setting to another (DfE 2018). (In 2017, the proportion of pupils in England who would be thus affected was roughly 13 per cent.[4])

Yet research also tells us that about 20 per cent of pupils move school in England at non-standard times of year (Sharma 2016). This also varies by region: in London, non-standard admissions are 20 per cent higher than in other regions in England. Not only is there an inconsistent rate of mobility across areas of England and between schools within those areas, but there are also differences in the characteristics of pupils who are more mobile (ibid). Schools with very low percentages of free school-meal pupils generally have very low levels of pupil mobility; by contrast, schools in disadvantaged and diverse areas tend to experience greater pupil mobility (Rodda 2013). Socioeconomically disadvantaged pupils are over-represented among mobile pupils, as are pupils with special educational needs (SEN) and/or disability, while the mobile pupil population is more ethnically diverse than the overall pupil population (ibid). The likely difference in contextual factors for these children and, consequently, their learning needs should be recognised if baseline scores are to be treated appropriately when children move school.

[4] Author calculations based on DfE 2017b.

Considerable work remains to be done before we properly understand the impact of using a measure of pupil progress in areas of the country where there is disproportionate pupil mobility across schools. If mobile pupils were to be *taken out* of the progress measure in all schools, then on average only a couple of pupils per class would then be 'missing' from the school accountability measure; however, in some schools there could be many more. Alternatively, if the reception baseline assessments results were to *follow* a pupil who moves school during the primary phase, as is current practice between key stages 1 and 2, then a school will be held accountable for a pupil's progress from a starting point that they do not know, and that school may only have a relatively short part of the pupil's school life in which to make up progress.

For all these reasons, pupil progress scores that span from the reception baseline to key stage 2 are unlikely to easily translate into simple judgements about what a school has added. Similar issues are raised by the movement of headteachers. A recent study reported that while 84 per cent of primary school headteachers remained head of the same school from year to year, retention rates are falling (Lynch et al 2017). These rates are lower among schools that have recently been deemed inadequate by the Office for Standards in Education, Children's Services and Skills (Ofsted) or become academies or multi-academy trusts, and among those that have a higher percentage of low-attaining pupils. Against this backdrop, there is clearly a need for strategies to 'retain effective head teachers within the profession and to build a stronger pipeline of new head teachers' (ibid: 4).

While retention remains difficult, across seven years the chances of a change of headteacher in any primary school will be quite high; some schools may experience several such changes. Similarly, many teachers may move schools across a seven-year time span. As with pupil mobility, more work needs to be done before we understand the usefulness of an accountability measure that has such a far-reaching end-point. What will it really measure, if many heads and school leaders, as well as teachers and pupils, do not stay with a school for the full seven-year period between the proposed reception baseline assessment and key stage 2?

3.2 The impact of socioeconomic and family factors

There is strong evidence that other child characteristics also affect both attainment and relative attainment in value-added measures. Different effects have been found not only for age but for the early-years home-learning environment, parents' educational levels, family socioeconomic status, family income and neighbourhood disadvantage, as well as English as an additional language (EAL) (see research on the millennium cohort study, as well as other longitudinal research funded by the DfE such as the EPPSE research – Melhuish et al 2008; Sammons et al 2002, 2003, 2004, 2008a, 2008b, 2008c, 2015).

Since 2010, successive governments have decided to ignore such effects by dropping the contextualisation of schools' results in accountability comparisons. However, this does not make these effects disappear: they remain, but become a source of unmeasured bias. In their study, Tymms et al (2014) found significant effects for disadvantage, related to the disadvantage of the neighbourhood in which a child lived (from the neighbourhood's score on the Income Deprivation Affecting Children Index), but they did not study the effects of family income using a child's free school meal (FSM) status. Sammons and Smees' (1998) study on baseline assessments revealed significant effects on both attainment at baseline and value-added attainment related to a child's FSM status over and above the effects of age. **Ignoring such effects makes the proposed use of the reception baseline assessment for school accountability particularly inappropriate, since comparisons will not take proper account of differences between schools in terms of the characteristics of their pupil intakes.** This will systematically favour schools serving fewer disadvantaged pupils, and penalise schools serving higher numbers of disadvantaged children.

4. WILL THIS FORM OF ACCOUNTABILITY LEAD TO USEFUL COMPARISONS OF SCHOOLS?

England, relative to most other countries, already employs unusually high-stakes accountability procedures in its education system. Schools are judged both by the results of national tests and examinations – against which school performance targets are set – and by the outcomes of systematic school inspections by Ofsted.[5] Both systems carry serious consequences for schools, particularly if they perform below target thresholds, or are judged by Ofsted to require improvement.

4.1 The utility of school performance data for parental school-choice

The government justifies the use of accountability data to rank schools and produce league tables as being in the interests of parental choice. While little research on the use of school comparisons ('league tables') to inform parental choice of schools at the primary-school level has been published, there is extensive literature on such choices at the secondary level. This evidence helps to draw out the implications of using progress measures from reception to key stage 2 for such a purpose – which is crucial, given that it is quite clear from the DfE's response to the consultation on baseline testing that this is the main purpose of the proposals (DfE 2017a: 15).

It is generally recognised that the only proper way to make comparisons between schools is to make adjustments for the prior attainments of pupils when they enter those schools, and to control for other relevant characteristics of pupil intakes. Such adjustments lead to what are known as 'value added' comparisons. In the case of the baseline tests, school-level attainment at year 6 will be adjusted for using the reception baseline assessments but *without* controlling for any contextual factors such as age, FSM, EAL or SEN status. This approach cannot lead to fair or useful comparisons.

[5] West et al (2011) observe that, while there are various types of accountability, in English education it is the managerial and market forms that dominate. This stems from policies based on choice and competition, which necessitate standardised public data to aid comparison and choice. Other countries have been less keen to adopt such policies (Mattei 2012).

Leckie and Goldstein (2011) used data from the national pupil database to conduct an in-depth analysis of the adequacy of secondary school value-added rankings as a means of informing parental school choice. Most significantly, they demonstrated that secondary school comparisons based on pupil examination results at year 11, adjusted for key stage 2 attainment at year 6, led to value-added league tables that were of very limited use when making choices between schools. They concluded that attempting to rank schools in this way is unsatisfactory for several reasons.

The value-added score itself is subject to considerable statistical uncertainty, resulting from the limited number of students who make up the school population and upon which this score can be based.

- School comparisons will become even less reliable if they attempt to ascertain effectiveness for sub groups of children, such as low-achievers, since they will be based on smaller samples
- Any data used to inform a parent's choice of school must be extrapolated from the results of a cohort of students who entered the schools six or so years *prior to* the current year of entry. In this respect, the data is always at least six years 'out of date' (Leckie and Goldstein 2011).[6]

Since a specific set of value-added school effects will prevail at any given time, all of this raises the question, What is the prediction for children starting school *next* year? To make such a prediction, essentially you need to add to the current 'error' (as expressed in the usual confidence intervals), the uncertainty of prediction from the past cohort to the one starting now. Leckie and Goldstein (ibid) arrived at an estimate of the school-level correlation across two cohorts five years apart of just 0.64 – so combining these uncertainties will give you a very weak prediction. Likewise, for pure accountability purposes it is not useful to refer to events that occurred seven years in the past on the basis of simplistic measures.

As far as we are aware, comparable analyses do not yet exist for reception baseline tests – yet the analogous uncertainties are almost certainly greater, and much wider confidence intervals would have to be placed around any school comparisons made on the basis of them. First, the time gap between reception and year 6 is already greater than between year 6 and year 11. Second, the baseline tests will be less reliable than the key stage 2 tests used by Leckie and Goldstein (ibid) because of the inherent measurement error associated with the proposals as outlined above. This will be compounded by the fact that age cohorts are much smaller in primary than in secondary schools. Given all these difficulties, it would seem that the final outcome of the reception baseline assessment will be of very little practical use.

[6] Leckie and Goldstein (2011) presented a simple way of visualising these uncertainties – as graphs that give clear pictorial comparisons by allowing the user to vary the factors affecting the uncertainty of the actual value-added scores, and by looking at comparisons of different schools (216, 219).

For all these reasons, it would be irrational to pursue baseline assessment for the stated purpose of school value-added comparisons without first commissioning a comprehensive study, by independent researchers, of the likely utility of the effort. The current plans (DfE 2017a) do not take note of past DfE work on, or academic studies of, school effectiveness. Pressing ahead with baseline testing as a means of making comparisons between schools is likely to prove a waste of time and money, given the many problems that we have described above.

5. WHAT IS THE LIKELY IMPACT OF THESE ACCOUNTABILITY MEASURES ON PUPILS AND SCHOOLS?

5.1 Delaying feedback

The government intends to hold the baseline test data until the cohort reaches key stage 2. It is unclear whether there will be a limited release of data, perhaps only of school aggregate scores, to schools during the test year. If limited data were to be released during the test year, it would encourage the production of completely inadmissible ranked league tables of school performance. Conversely, non-disclosure of the data will frustrate teachers and parents who will rightly ask, 'Why administer a test that doesn't help teaching and learning?' – even if this approach may prevent overinterpretation of individuals' and schools' results from a potentially unreliable test.

In practice, after the previous round of baseline testing many schools did not actually use the resulting data for any purpose: it was seen as just another externally imposed task they had to complete. As many schools continued to use their own assessments anyway, this had negative consequences for teachers' workload.

All of this is likely to stoke resentment at having to put children through a 'useless' test, and at the costs of development and administration – estimated to be £10 million pounds at a time of considerable budget reductions in most schools.

5.2 Gaming baseline scores

The well-known Campbell's law states:

> *'The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.'*
>
> (Campbell 1976: 85)

There is now extensive evidence that when tests results are used in high-stakes accountability systems in education – such as NAPLAN in Australia and key stage 2, GCSE and GCE performance tables in England – they distort those systems, as schools react to them by 'gaming' in order to obtain better results (Koretz 2002; Hursh 2005; Boyle and Bragg 2006; Stobart and Eggen 2012; Klenowski and Wyatt-Smith 2012).

Baseline tests in reception are different in the sense that good scores may not necessarily be seen as beneficial to schools. However, this may not inhibit game-playing; rather, it simply poses new questions to schools.

- Would it be better, for instance, to start from a low baseline so that the school's later value-added scores seem more impressive, as Professor Robert Coe suggested in evidence to the Education Select Committee (Education Committee 2017: Q148)? The temptation could be to administer the testing as early as possible, in order to obtain the lowest possible measure and capture the progress that settling in constitutes. However, this could interfere with the process of building positive relationships, and could be stressful for a child who is still orienting herself in an unfamiliar environment (see section 5.4 below).
- Conversely, would it be better to demonstrate to parents that this school has a 'good' intake as measured by the overall baseline test result?

Whatever the ways in which schools 'play' the tests, it will lead to a great deal of variability and further reduce the reliability, and therefore validity, of the test in terms of its stated purpose of school accountability.

5.3 Distorting younger children's provision

With baseline tests there is always debate about when the baseline should be measured in order to capture the greatest possible amount of progress. Many primary schools with reception classes will also have nursery classes, a growing number of which will be catering for two-year-olds – generally the more vulnerable two-year-olds who are receiving free early education. What will the impact of baseline be on the nursery curriculum?

The authors' consultations with heads and teachers on the new proposals have confirmed that many schools see the early years as part of one key stage, and already monitor progress in the foundation stage from the beginning of nursery using the early years foundation stage profile.[7] Implementing the reception baseline could split this existing work within the early years into two distinct phases (before and after the baseline assessment). This could mean a reduction in the time and resources spent on the more rounded development work appropriate for all pre-

[7] See https://www.gov.uk/government/publications/early-years-foundation-stage-profile-2018-handbook

schoolers in order to prepare older nursery children for the assessments at reception. If the test itself is narrow, then pressure to prepare pupils for it may narrow the pre-school curriculum in harmful ways (and may also have wider effects on the primary curriculum – see section 5.6 below). Moreover, parents may feel under pressure to secure coaching for their pre-school-age children in order to ensure that they are 'test-ready', in the belief that this would secure the best results for their children.

5.4 Adverse impacts on children in reception

The impact of the baseline assessment on the children sitting it has been a key concern for early years practitioners. Children start school with a wide range of experiences, and 'settling in' is a very important part of reception schooling. The baseline assessment has the potential to disrupt that process. Concerns have been raised about children who might not be able to complete a task, and the risk that they will feel they have failed, or that school is too hard. The time it takes to settle in to school varies between children; age as well as context is a consideration (see section 2.3.1); and the longer-term impact of a difficult start to school needs to be understood. Each of these factors will reduce the chance that the test will produce reliable data.

This is why the majority of early years practitioners favoured the Early Excellence baseline programme[8] when they had the opportunity to choose (Ward 2015). It fitted better with early-years good-practice, which is based on professional judgement and the understanding of the developing child as a learner (TACTYC and Early Education 2015). It also allowed practitioners to ensure that settling-in time was as stress-free as possible, particularly for children who had not spent time away from home or caregivers before, who were new to the English language, or who were unfamiliar with the sorts of activities or resources found in a school environment.

5.5 Unintended consequences: 'labelling' versus individualised support

The reception baseline assessment is not intended to have any diagnostic value for schools or individual children. However, since teachers administering the test will see the scores that children produce, the baseline test could result in some children being unnecessarily labelled as low-ability at the very start of their formal education. This is likely to be of particular concern for summer-

[8] See http://earlyexcellence.com/eexba/. The Early Excellence baseline scheme was chosen by over 11,000 out of 17,000 primary schools, making it the most popular choice from a practitioner's perspective for baseline testing in 2015/16 (Ward 2015).

born pupils, EAL children and those with SEN. Unless contextual information is collected, the data would not indicate the potential reasons for why these children achieve low scores.

Practitioners value the more observational style of assessment precisely because it encourages them to use their professional judgment in a more fluid way in order to support the development of young children, regardless of their starting point. It has an immediate positive purpose. This was reflected in the fact that observational tests were the most popular choice for baseline assessments in 2015/16 (Ward 2015). As things stand, the time spent on the baseline tests may result in time lost collecting more useful information about young learners that would enable early years staff to plan and discuss the best support for individuals during the foundation stage.

5.6 The narrowing of the curriculum throughout primary schooling

It is widely recognised that high-stakes assessments for accountability purposes can have unintentional impacts on teaching and the way in which a curriculum is delivered. One of the concerns about using baseline tests to measure progress according to indicators in the key stage 2 tests is the narrowing of the curriculum that may follow from it. A relentless emphasis on literacy and numeracy all the way through primary school is already resulting in less time and focus being given to foundation subjects such as science, and limiting access to the arts. Some schools have increasingly been using intensive data collection and monitoring in order to make decisions on lesson planning, with a great deal of teaching-time being devoted to pre-planned drills (Bradbury and Robert-Holmes 2016). Introducing baseline tests with a narrow focus on literacy and numeracy may well further entrench such practices.

6. ARE THERE BETTER ALTERNATIVES TO BASELINE TESTING?

The panel considered alternatives to baseline testing that could answer questions about school accountability in a much more productive way.

In her 2002 Reith lectures, the philosopher Onora O'Neill critiqued then-current accountability measures in the public services for their emphasis on 'performance indicators chosen for ease of measurement and control rather than because they measure quality of performance accurately' (2002: 54). She called for intelligent accountability, which would place more trust in professionals and pay more attention to self-governance, because 'since much that has to be accounted for is not easily measured it cannot be boiled down to a set of stock performance indicators' (ibid: 58). O'Neill's vision was of accountability that 'provides substantive and knowledgeable independent judgement of an institution's or professional's work' (ibid: 58).

Terry Crooks (2007) provides six principles for intelligent accountability in education.

1. It preserves and enhances trust among the key participants in the accountability processes.
2. It involves participants in the process, offering them a strong sense of professional responsibility and initiative.
3. It encourages deep, worthwhile responses rather than surface window-dressing.
4. It should recognise and attempt to compensate for the severe limitations of performance indicators in capturing educational quality.
5. It provides well-founded and effective feedback to support good decision-making.
6. It leaves the majority of participants more enthusiastic and motivated in their work (adapted from Crooks 2007).

This is a far more defensible approach than using a single measure, based on aggregated results from a 20-minute test, seven years later as an indicator of a primary school's contribution to pupil progress.

6.1 Examples of intelligent accountability in action

Examples of intelligent accountability in the early years include collaborations between academics, local authorities and schools that have encouraged reflection on the value of the data collected and the purposes to which it can best be put (Yang 1999).

6.1.1 The Surrey value-added initiative: supporting school improvement

The Surrey value-added initiative collected reception baseline data that could be used by practitioners to support individual pupils, while also informing school improvement planning more broadly (Sammons and Smees 1998). It took into account and made explicit the important role of child age, and other background effects. It also provided separate measures of school performance in different areas (rather than using one total measure), while taking into account the statistical uncertainty associated with calculating value-added measures of school effects. Schools received their own results alongside (in anonymised form) those of other schools in their local authority area. Participation was voluntary, and schools agreed not to use their contextualised value-added results for marketing purposes. The intention was to support schools in the formative use of data, to ask 'intelligent' questions, and to focus on improvement. Importantly, to this end the local authority provided teachers and schools with guidance and resources on individual education plans to support children whose baseline scores suggested that they might need extra support (ibid).

6.1.2 The London Education Research Network: good practice in data use

The London Education Research Network has championed the notion of using performance data effectively in order to aid school improvement for many years, and the network has shared good practice in data-use across London boroughs. Their approach relies on good partnership-working that can foster open and productive conversations between local authorities' school improvement staff, their education data teams, and school leadership teams. The aim of the partnership is to underpin robust self-evaluation at school level by providing good comparative performance information and comprehensive training on using national and local data tools. All parties are encouraged to ask intelligent questions of the appropriate data, and to reflect upon how to feed the answers most productively back into practice.

To do this well requires adequate finance and the appropriate discharge of responsibilities across different service levels. The reduction in the size of local authority school improvement services, along with local authorities' diminished roles in and responsibilities for their schools, is making this kind of conversation harder to maintain. However, in some instances – such as in Wandsworth, Hounslow, Southwark and Lambeth borough councils in London – school improvement services have continued as a 'traded service'. These local authorities have continued to provide effective training at the local level in the use of data and research to support school improvement. Lambeth has summarised and documented some of the research that it has conducted to explore this practice through a case study of their own journey (Demie 2013). Working with the updated Ofsted framework, the local authorities participating in the LERN network have placed a greater emphasis on the quality of the curriculum and real-time observation of children's work and progress. Lambeth council has also continued to provide effective training at the national level in order to share good practice in using data and research to support school improvement (ibid). This approach could be championed further as an alternative means of making effective use of pre-existing local and national data.

***

Each of these examples demonstrate that principles of intelligent accountability can readily be adopted and put into practice.

CONCLUSION

We have raised a number of serious concerns about the government's plans to introduce a baseline assessment in reception in order to create a measure of pupil progress at key stage 2 for accountability purposes. We consider the proposals as they stand to be flawed.

In the published plans, no account has been taken of the need to:

- identify and control for age and other contextual factors (such as socioeconomic status, gender, ethnicity, EAL and SEN), using suitably fine measures when comparing pupil performance
- evaluate the impact of pupil mobility at the school level, particularly in more disadvantaged schools and areas, when comparing baseline with key stage 2 data.
- take care that schools serving more disadvantaged pupils are not themselves disadvantaged by the way in which value-added measures are calculated, drawing on past DfE and other effectiveness research
- make available, for public discussion and scrutiny, full reliability data on the administration and scoring of the tests.

No assurances have been given that there will be an equalities impact assessment during the testing of the new baseline materials to examine the impact on different groups of children – even though this is standard practice in test development (AERA et al 2014). It is also completely unclear how government and the test developer, NFER, intend to use testing and piloting phases to adjust and adapt the test design – or, indeed, to assess its fitness for purpose.

In light of these unanswered questions, and the evidence set out in the body of this report, the panel draws the following conclusions.

1. Under these proposals, children will be exposed to tests that will offer no formative help in establishing their needs and/or in developing teaching strategies capable of meeting them. The morality of this is questionable, as is the use of time, money and resources for an assessment that is of little or no direct value to those involved in it. **We know of no other assessment systems internationally that offer so little formative feedback, and in which the data lie dormant for such a substantial period (seven years in this case).** The ethical case for this practice has not been made.

2. **Any value-added calculations that will be used to hold school to account will be highly unreliable.** Any 20-minute test of four-year-olds will be strongly affected by age, first language and home background. To make no adjustment for these factors, and to use the combined scores to determine a baseline from which to judge the school seven years later, violates recognised international testing standards (AERA et al 2014; ETS 2004). It also ignores decades of school effectiveness research and evidence from the DfE's own past work on contextual value-added indicators (Sammons et al 2000, 2008b; Leckie and Goldstein 2017; Reynolds et al 2014).

   Any presentation of school value-added scores for accountability purposes should recognise their inherent statistical unreliability by indicating the confidence intervals around them (Foley and Goldstein 2012). These confidence intervals would reveal that making fine (rank order) distinctions between schools in the form of ranked league tables would be invalid. As we argue in more detail above (see section 4.1), analysis of secondary-school data suggests that the data produced under the proposed policy will be of extremely limited utility for making school choices and holding schools to account (Leckie and Goldstein 2011). This problem arises because the initial tests themselves will have low reliability, and the long time-gap between baseline and key stage 2 further reduces predictability from the former to the latter. This problem is exacerbated by the small size of pupil cohorts in primary schools, and the extent and variation in rates of pupil mobility between primary schools. Given the high-stakes nature of this assessment as it is currently proposed, there may also be the added complication of schools 'playing the system': would a lower baseline score advantage the school in any future value-added calculation, for instance?

3. **This is an untried experiment.** To properly evaluate the proposed new baseline system one would have to wait until at least 2027, when the first cohort tested at reception has taken key stage 2 tests. Without this evidence, we argue that it would be unethical to impose such a system on pupils and schools. At best, the current proposals could be considered a pilot, one that should be subject to independent evaluation in seven or eight years' time. Even so, because of the seven-year time-lag, any value-added data on schools will be of very little use, and will not provide a secure basis on which to judge a school's current offer to four-year-olds (there will be significant turnover among teachers and headteachers within a typical primary school over a seven-year period).

In conclusion, the panel believe that the government's proposals for the reception baseline assessment are flawed, unjustified, and wholly unfit for purpose. They would be detrimental to children, parents, teachers, and the wider education system in England. We publish this report in the hope of informing public debate by making an accessible and thorough account of why these proposals must be comprehensively rethought.

REFERENCES

American Educational Research Association [AERA], American Psychological Association, National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing (US) (2014) *Standards for Educational and Psychological Testing*, Washington, DC

Boaler J (2009) *The Elephant in the Classroom: Helping Children Learn and Love Maths*, London: Souvenir Press

Boyle B and Bragg J (2006) 'A curriculum without foundation', *British Educational Research Journal* 32(4): 569–582.

Bradbury A and Robert-Holmes G (2016) *"They are children… not robots, not machines": The Introduction of Reception Baseline Assessment*, London: National Union of Teachers and Association of Teachers and Lecturers. http://www.teachers.org.uk/files/baseline-assessment--final-10404.pdf

Bradbury A, Jarvis P, Nutbrown C, Roberts-Holmes G, Stewart N and Whitebread D (2018) 'Baseline assessment: Why it doesn't add up', London: More Than a Score. https://morethanascorecampaign.files.wordpress.com/2018/02/neu352-baseline-a4-16pp-crop.pdf

Campbell D T (1976) 'Assessing the impact of planned social change', *Evaluation and Program Planning* 2(1): 67–90

Clark L (2010) 'One teenager in five leaving school unable to read or do maths', *Daily Mail*, 7 May 2010. http://www.dailymail.co.uk/news/article-1274947/One-teenager-leaving-school-unable-read.html

Crooks T J (2007) 'IPL: Principles for Intelligent Accountability, with Illustrations from Education', inaugural professorial lecture, University of Otago, 4 October 2017. https://www.listennotes.com/e/e039390bebe3436b9f0f510375a14bce/ipl-principles-for-intelligent-accountability-with-illustrations-from-education/

Crooks T J, Kane M T and Cohen A S (1996) 'Threats to the valid use of assessment', *Assessment in Education* 3(3): 265–285

Demie F (2013) *Using Data to Raise Achievement: Good practice in schools*, London: Lambeth Research and Statistics Unit, London Borough of Lambeth. https://www.lambeth.gov.uk/rsu/sites/lambeth.gov.uk.rsu/files/Using_Data_to_Raise_Achievement-Good_Practice_in_Schools_2013.pdf

Department for Education [DfE] (2017a) *Primary assessment in England: Government consultation response*, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644871/Primary_assessment_consultation_response.pdf

Department for Education [DfE] (2017b) 'Schools, pupils and their characteristics: January 2017', statistical first release. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

Department for Education [DfE] (2018) 'Measuring Progress in Primary Schools', London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/698912/Measuring_progress_in_primary_schools.pdf

Education Committee (House of Commons select committee) (2017) 'Education Committee: Oral evidence: Primary Assessment, HC 682: Wednesday 18 January 2017', London. https://www.parliament.uk/business/committees/committees-a-z/commons-select/education-committee/inquiries/parliament-2015/primary-assessment-16-17/publications/

Educational Testing Service (ETS) (2004) *ETS international principles for fairness review of assessment*, Princeton, NJ

Foley B and Goldstein H (2012*) Measuring success: League tables in the public sector*, London: British Academy Policy Centre. https://www.britac.ac.uk/publications/measuring-success-league-tables-public-sector

Hart S, Dixon A, Drummond M J and McIntyre D (2004) *Learning without Limits*, Maidenhead: Open University Press

House of Commons Education and Skills Committee [HoC-ESC] (2005) *Teaching Children to Read: Eighth Report of Session 2004–05: Report, together with formal minutes, oral and written evidence*, London: Stationery Office. https://publications.parliament.uk/pa/cm200405/cmselect/cmeduski/121/121.pdf

Hursh D (2005) 'The growth of high-stakes testing in the USA: Accountability, markets, and the decline in educational quality', *British Educational Research Journal* 31(5): 605–622

Kim J and Suen H K (2003) 'Predicting children's academic achievement from early assessment scores: A validity generalization study', *Early Childhood Research Quarterly* 18(4): 547–566

Klenowski V and Wyatt-Smith C (2012) 'The impact of high stakes testing: the Australian story', *Assessment in Education: Principles, Policy & Practice* 19(1): 65–79

Koretz D (2002) 'Limitations in the Use of Achievement Tests as Measures of Educators' Productivity', *Journal of Human Resources* 37(4): 752–778

LaParo K M and Pianta R C (2000) 'Predicting children's competence in the early school years. A meta-analytic review', *Review of Educational Research* 70(4): 443–484

Leckie G and Goldstein H (2011) 'Understanding uncertainty in school league tables', *Fiscal studies* 32(2): 207–224

Leckie G and Goldstein H (2017) 'The evolution of school league tables in England 1992–2016: "Contextual value-added", "expected progress" and "progress 8"', *British Educational Research Journal* 43(2): 193–212

Lynch S, Mills B, Theobald K and Worth J (2017) *Keeping Your Head: NFER Analysis of Headteacher Retention*, Slough: National Foundation for Educational Research. https://www.nfer.ac.uk/keeping-your-head-nfer-analysis-of-headteacher-retention/

Mattei P (2012) 'Market accountability in schools: policy reforms in England, Germany, France and Italy', *Oxford Review of Education* 38(3): 247–266

Meisels S and Atkins-Burnett S (2008) '26: Evaluating Early Childhood Assessments: A Differential Analysis' in McCartney K and Phillips D (eds) *Blackwell Handbook of Early Childhood Development*, Hoboken, NJ: Blackwell-Wiley: 533–549

Melhuish E, Sylva K, Sammons P, Siraj-Blatchford I, Taggart B and Phan M (2008) 'Effects of the Home Learning Environment and preschool center experience upon literacy and numeracy development in early primary school', *Journal of Social Issues* 64(1)*:* 95–114

Messick S (1989) 'Meaning and Values in Test Validation: The Science and Ethics of Assessment', *Educational Researcher* 18(2): 5–11

Newton P E (2007) 'Clarifying the purposes of educational assessment', *Assessment in Education: Principles, Policy & Practice* 14(2): 149–170

Newton P E (2009) 'The reliability of results from national curriculum testing in England', *Educational Research* 51(2): 181–212

National Foundation for Educational Research [NFER] (no date) 'Information About the 2018 Reception Baseline Assessment Trial', webpage. https://www.nfer.ac.uk/for-schools/participate-in-research/information-about-the-2018-reception-baseline-assessment-trial/

O'Neil O (2002) *A Question of Trust: The BBC Reith Lectures 2002*, Cambridge: Cambridge University Press

Reynolds D, Sammons P, De Fraine B, Van Damme J, Townsend T, Teddlie C and Stringfield S (2014) 'Educational effectiveness research (EER): A state-of-the-art review', *School Effectiveness and School Improvement* 25(2): 197–230

Rodda M, with Hallgarten J and Freeman J (2013) *Between the cracks: Exploring in-year admissions in schools in England*, London: RSA Action and Research Centre. https://www.thersa.org/discover/publications-and-articles/reports/between-the-cracks

Sammons P and Smees R (1998) 'Measuring Pupil Progress at Key Stage 1: Using baseline assessment to investigate value added', *School Leadership and Management* 18(1): 389–407

Sammons P, Sylva K and Mujtaba T (2000) *What Do Baseline Assessment Schemes Measure? A Comparison of the QCA and Signposts Schemes: Report prepared for the Qualifications and Curriculum Authority*, London: Institute of Education, University of London

Sammons P, Sylva K, Melhuish E C, Siraj-Blatchford I, Taggart B and Elliot K (2002) *The Effective Provision of Pre-school Education Project: Technical Paper 8a: Measuring the impact on children's cognitive development over the pre-school years,* London: Institute of Education, University of London and Department for Education and Skills. http://dera.ioe.ac.uk/18189/11/EPPE_TechnicalPaper_08a_2002.pdf

Sammons P, Sylva K, Melhuish E C, Siraj-Blatchford I, Taggart B and Elliot K (2003) *The Effective Provision of Pre-school Education Project: Technical Paper 8b: Measuring the impact on children's social behavioural development over the pre-school years,* London: Institute of Education, University of London and Department for Education and Skills. http://dera.ioe.ac.uk/18189/12/EPPE_TechnicalPaper_08b_2003.pdf

Sammons P, Sylva K, Melhuish E, Siraj-Blatchford I, Taggart B Elliott K and Marsh A (2004) *The Effective Provision of Pre-school Education (EPPE) Project: Technical Paper 11: The continuing effect of pre-school education at age 7 years,* London: Institute of Education, University of London and Department for Education and Skills. http://dera.ioe.ac.uk/18189/15/EPPE_TechnicalPaper_11_2004.pdf

Sammons P, Sylva K, Melhuish E, Siraj-Blatchford I, Taggart B and Jelicic H (2008a) *Influences on Children's Development and Progress in Key Stage 2: Social/behavioural outcomes in Year 6*, London: Department for Children, Schools and Families. http://dera.ioe.ac.uk/18192/1/DCSF-RR049.pdf

Sammons P, Sylva K, Melhuish E, Siraj-Blatchford I, Taggart B and Hunt S (2008b) *Influences on Children's Attainment and Progress in Key Stage 2: Cognitive outcomes in Year 6*, London: Department for Children, Schools and Families. http://dera.ioe.ac.uk/18190/1/DCSF-RR048.pdf

Sammons P, Anders Y, Sylva K, Melhuish E, Siraj-Blatchford I, Taggart B and Barreau S (2008c), 'Children's Cognitive Attainment and Progress in English Primary Schools During Key Stage 2: Investigating the potential continuing influences of pre-school education', in Roßbach H G and Blossfeld H P (eds) *Frühpädagogische Förderung in Institutionen*, Wiesbaden: VS Verlag für Sozialwissenschaften

Sammons P, Toth K, Sylva K, Melhuish E, Siraj I and Taggart B (2015) 'The long-term role of the home learning environment in shaping students' academic attainment in secondary school', *Journal of Children's Services* 10(3): 189–201

Sharma N (2016) 'Pupil Mobility: What does it cost London?', London: London Councils

Shepard L A, Kagan S L and Wurtz E (eds) (1998*) Principles and recommendations for early childhood assessments,* Washington, DC: National Education Goals Panel. http://govinfo.library.unt.edu/negp/reports/prinrec.pdf

Standards and Testing Agency (STA) (2016) *Reception baseline comparability study: Results of the 2015 study*, London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/514581/Reception_baseline_comparability_study.pdf

Stiggins R J (2000) *Student-Involved Classroom Assessment,* London: Pearson

Stobart G and Eggen T (2012) 'High-stakes testing – Value, fairness and consequences', *Assessment in Education: Principles, Policy & Practice* 19(1): 1–6

Sylva K, Melhuish E, Sammons P, Siraj-Blatchford I and Taggart B (2004) *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 12: The Final Report: Effective Pre-School Education*, London: Institute of Education, University of London and Department for Education and Skills. http://dera.ioe.ac.uk/18189/16/EPPE_TechnicalPaper_12_2004.pdf

Sylva K, Melhuish E, Sammons P, Siraj-Blatchford I, Taggart B and Sammons P (2006) 'PART 1: Influences on children's attainment, progress and social/behavioural development in primary school', in *Promoting Equality in the Early Years: Report to The Equalities Review*, Wetherby: Department for Communities and Local Government: 22–62. http://ro.uow.edu.au/cgi/viewcontent.cgi?article=2176&context=sspapers

TACTYC and Early Education (2015) 'Guidance on Baseline Assessment in England: Summary', Watford. http://tactyc.org.uk/wp-content/uploads/2015/02/Summary-Baseline-Assessment-Guidance.pdf

Tymms P, Merrell C, Hawker D and Nicholson F (2014) *Performance Indicators in Primary Schools: A comparison of performance on entry to school and the progress made in the first year in England and four other jurisdictions*, London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/318052/RR344_-_Performance_Indicators_in_Primary_Schools.pdf

Ward H (2015) 'Thousands of primaries opt for test-free baseline assessments', *Times Educational Supplement*, 22 May 2015. https://www.tes.com/news/thousands-primaries-opt-test-free-baseline-assessments

Ward H (2017) 'Early Excellence pulls out of bid for "unworkable" Reception baseline test', *Times Educational Supplement*, 29 November 2017. https://www.tes.com/news/early-excellence-pulls-out-bid-unworkable-reception-baseline-test

Ward H (2018a) 'Reception baseline assessment to be developed by NFER', *Times Educational Supplement*, 11 April 2018. https://www.tes.com/news/reception-baseline-assessment-be-developed-nfer

Ward H (2018b) 'Banning teachers' preferred method of baseline assessment will relieve "burden", says minister', *Times Educational Supplement*, 6 March 2018. https://www.tes.com/news/banning-teachers-preferred-method-baseline-assessment-will-relieve-burden-says-minister

West A, Mattei P and Roberts J (2011) 'Accountability and sanctions in English schools', *British Journal of Educational Studies* 59(1): 41–62

Whetton C (2009) 'A brief history of a testing time: national curriculum assessment in England 1989–2008', *Educational Research* 51(2): 137–159

Yang M, Goldstein H, Rath T and Hill N (1999) 'The Use of Assessment Data for School Improvement Purposes', *Oxford Review of Education* 25(4): 469–483

ABOUT THE PANEL

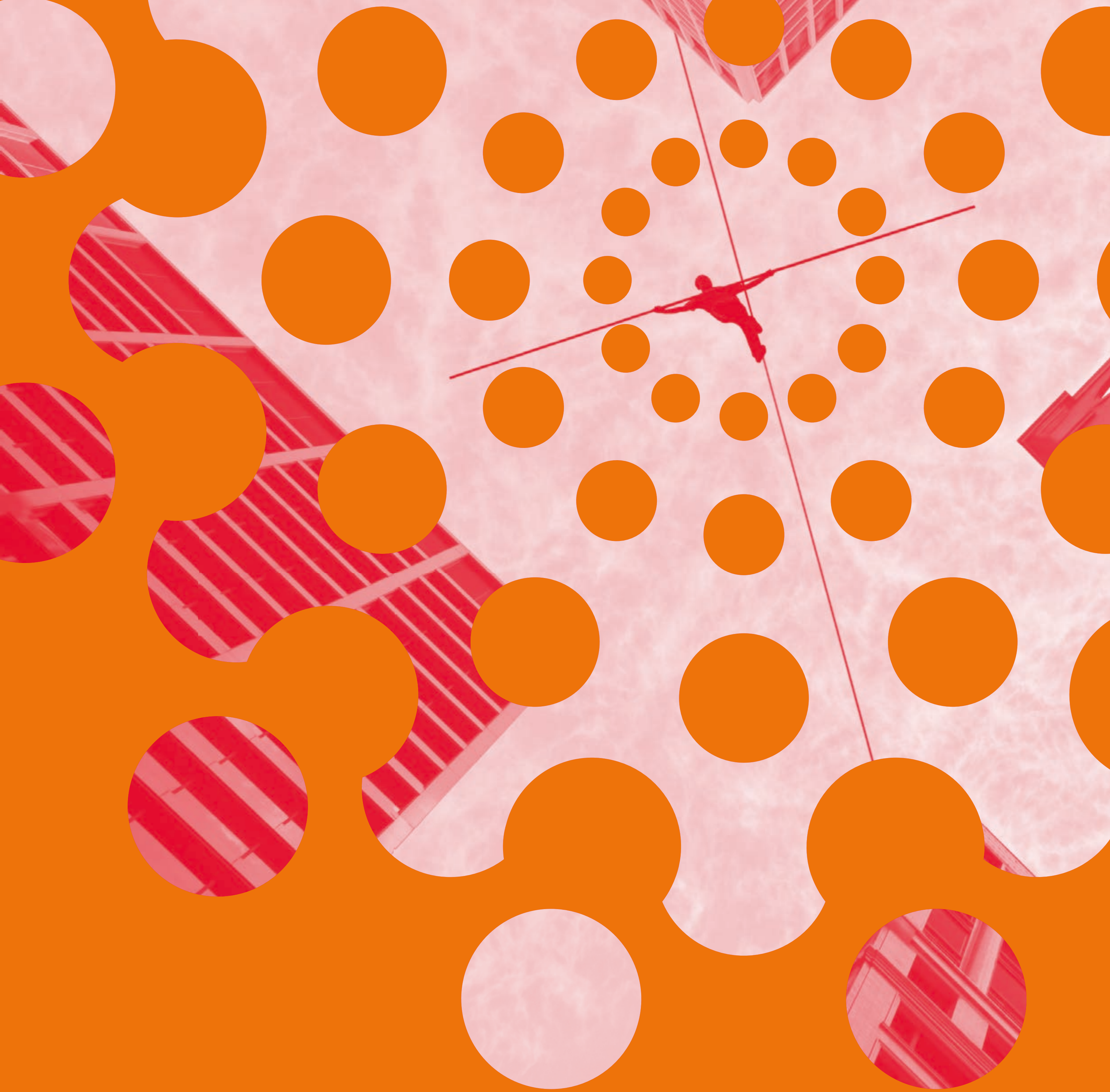
Harvey Goldstein is Professor of Social Statistics at Bristol University, and a fellow of the British Academy. He has extensive experience of analysing educational data.

Gemma Moss is the former President of the British Educational Research Association (BERA), and is Director of the International Literacy Centre and Professor of Literacy at the UCL Institute of Education.

Pamela Sammons is Professor of Education at the Department of Education, University of Oxford, and a Senior Research Fellow at Jesus College, Oxford. She is internationally known for her research on equity, educational effectiveness and improvement.

Gwen Sinnott is an independent education performance consultant. She owns and runs SINNOTT Learning Solutions, and is the former head of management information and analysis at Southwark children's services and past president of London Education Research Network. She has extensive experience of working with pupil data to support school improvement at local authority level, and with school leaders directly.

Gordon Stobart is Emeritus Professor of Education at the UCL Institute of Education and Honorary Research Fellow at the Oxford University Centre for Educational Assessment (OUCEA). He is well known as a policy researcher with internationally recognised expertise in assessment.



BERA
BRITISH EDUCATIONAL RESEARCH ASSOCIATION