

Education Committee

Oral evidence: Primary Assessment, HC 682

Wednesday 18 January 2017

Ordered by the House of Commons to be published on 18 January 2017.

[Watch the meeting](#)

Members present: Neil Carmichael (Chair); Lucy Allan; Ian Austin; Suella Fernandes; Lucy Frazer; Lilian Greenwood; Catherine McKinnell; Ian Mearns; William Wragg.

Questions 82 - 156

Witnesses

I: **Dr Rebecca Allen**, Director, Education Datalab, **Professor Harvey Goldstein**, Professor of Social Statistics, University of Bristol, **Joanna Hall**, Deputy Director for Schools, Ofsted, and **Tim Oates CBE**, Group Director of Assessment Research and Development, Cambridge Assessment.

II: **Professor Rob Coe**, Director of the Centre for Evaluation and Monitoring, Durham University, **Dr Mary James**, Former Professor and Associate Director of Research, University of Cambridge Faculty of Education, **Catherine Kirkup**, Research Director, National Foundation for Educational Research, and **Professor Dominic Wyse**, Head of Department of Learning and Leadership, UCL Institute of Education.

Written evidence from witnesses:

- Education Datalab [[PRI0288](#)]
- National Foundation for Educational Research [[PRI0397](#)]
- Head of Department of Learning and Leadership, UCL Institute of Education [[PRI0348](#)]

Examination of Witnesses

Dr Rebecca Allen, Professor Harvey Goldstein, Joanna Hall and Tim Oates.

Q82 **Chair:** Good morning and welcome to our second session on primary assessment. Our purpose today, and it is quite explicit, is to explore the accountability system, look at the use of data and consider alternative accountability systems where appropriate. What we want to come up with is accountability and assessment that works. That is our objective and that is what will inform our report, and our questions are really geared to that end.

Without further ado, if you would like to say who you are and where you are from, for the purposes of the millions of viewers that are now tuning in.

Joanna Hall: Good morning, everyone. I am Joanna Hall, I am Deputy Director at Ofsted and my responsibility is for schools and initial teacher education.

Dr Allen: I am Rebecca Allen, I am Director of Education Datalab and I am also a reader in the economics of education at the UCL Institute of Education.

Professor Goldstein: I am Harvey Goldstein, I am Professor of Social Statistics in the Graduate School of Education at the University of Bristol.

Tim Oates: Tim Oates. I am Assessment Research and Development Group Director at Cambridge Assessment, a large non-teaching department of Cambridge University.

Q83 **Chair:** Thank you very much. I think we will get some interesting answers to our questions with that line up, so thank you all for coming today.

Assessment and accountability are closely linked in primary schools in England, as we all know. How is the accountability system at primary school impacting the quality of education?

Tim Oates: For me, the key thing is to differentiate a number of things in considering the relationship between the assessment for learning and the use of the data in examining the quality of the education that schools and the system provides overall. I think we have moved slowly towards greater clarity in each of those things.

In terms of the form of testing, we have done quite a few international comparisons in terms of what does the test look like over the total testing regime in primary, right through from the start of school to the end of year 6. We have looked at the form of testing and compared it internationally. The density of testing is not unusual internationally. Tests are used for a variety of purposes within the schools in terms of the check of reading at a very early age through to an assessment of the totality of attainment at the end of year 6. We have seen considerable



refinement in those individual instruments, their placing and their function. But I had to differentiate that from the quality of the administration and the design of the instruments so you can get the right design of the form of tests, the number of tests, what they are to be used for. The policy can be right, but you can have problems in terms of the practical administration of the instruments. That can range from the quality of the development of the instruments down to the way in which they are administered on a day-to-day basis within institutions.

The key matter is the use of the outcomes of those instruments within the institution for formative purposes, to reflect on quality, the extent to which it affects the quality of the teaching and the form of the teaching and then the use to which it is put by outside agencies such as the state. There we have seen considerable refinement over the last few years since the coalition Government. Prior to that we had much more simplistic use of accountability, much more simplistic measures that had not been refined over more than a decade, despite the accumulation of evidence on the detrimental effect, particularly in secondary schools.

So to summarise and answer your question, there are many uses to which the data are put. In some schools, this results in year 6 being very narrowly focused on drilling to the tests and sometimes inappropriate anticipation of what will occur in the tests. We have evidence that occurs in a number of institutions. There are other institutions that understand well the kind of policy statements that have been made over the last five years that the best way of preparing for the test is to teach a broad and balanced curriculum well and that will result in a high score in the tests.

Q84 Chair: We will be probing that later. Joanna, with an Ofsted perspective in mind obviously, what are your thoughts in relation to the accountability system and the quality of education?

Joanna Hall: We absolutely recognise the high stakes nature of assessment, but I think we equally recognise the importance in the common inspection framework of teaching, learning and assessment, what we have positioned as being at the heart of that framework, across the curriculum, not just focusing on English and mathematics. We also recognise that in terms of the use of the data when we are in schools and when we are inspecting those accountability national performance measures, we generate discussion with leadership teams, with teachers and sometimes with pupils and sometimes with parents. We do recognise that that is very much part of that discussion and also part of the way in which we continue to inspect schools. That is one of the key measures obviously that we do look at, but it is not the only measure.

Dr Allen: In the case of primary schools, we need to recognise there are some things that are intrinsic to primary schools that make it difficult to get reliable measures of how the children are doing and how well the school is doing. We choose to run relatively short tests, we do not take children out of school for months at a time as we do for GCSEs. So we have relatively imprecise measures of children's attainment, as we



should, I think. We have a very unreliable baseline, almost no baseline at all at the moment, and therefore we don't have a good view of what it is reasonable for schools to achieve. We can do something to fix that, but we must recognise that testing four year-olds is difficult, we are always going to have something that is quite unreliable, and we have small schools. We are making judgments on relatively few pupils.

We can talk about how to make the system better at the margins, we can talk about key stage 2 writing, we can talk about accessibility of the test for lower-attaining pupils and we can talk about getting a baseline in place. We can do things at the margin to make it better, but we have to recognise that, if we are properly thinking about the validity and the reliability of our assessment system or of our accountability system, we are making quite fragile judgments on schools. With that in mind, I think we have to lower the stakes that are associated with any single year of primary school assessment data.

Chair: We are certainly going to be talking about the baseline later in this session, but that is a very good analysis. Harvey, any comments?

Professor Goldstein: Not much, just to back up what Rebecca was saying, that our work is concerned with the inherent statistical unreliability of the data, especially when you have relatively small numbers, as in primary schools. The more you refine it—and this picks up something that Tim said—the worse the situation becomes, because then you are essentially dealing with much more narrowly-defined groups of students. That is where the refining comes in. It is drilling down to look at the actual student characteristics that make up the so-called accountability measures.

Our work is pointing out that if you are going to use these measures at all, what you shouldn't be doing is using them to make judgments about schools. You can use them as screening devices, for example, to tell you where there may be problems. I can elaborate now or later, but we conducted a very interesting experiment 20 years ago in Hampshire, in the local education authority, where we did just this. It involved the schools and the teachers. I seem to think it was extremely successful because it wasn't producing public league tables, which then became used as targets within schools, but was providing useful information that the authorities themselves, through their inspectorate, could then use to make their own judgments about schools.

Q85 **Chair:** Thank you. The key question I suppose is, do you all agree that schools should be held to account for the education of their pupils as set out in the Bew review?

Tim Oates: Again, I would appeal to the international evidence. We now have pretty robust evidence from the large international surveys, not just PISA but international surveys, alongside incisive international comparisons outside the surveys that suggest that appropriate



HOUSE OF COMMONS

accountability measures are linked to those systems that have had periods of sustained improvement.

Chair: That is as far as we need to go at the moment, I think.

Tim Oates: A small coda though, that often what is looked for in some of this research is the kind of accountability that we have in this country. So often researchers will go to another nation and say, "Do you have accountability of this form?" and that nation will say no. Beneath the surface though there are other forms of data provision from schools to local authorities or school districts that amount to the same thing, but assume a different form to the form we have in this country.

Professor Goldstein: Just to enter a strong disagreement with Tim, I do not think the international comparative surveys tell you anything about the relative quality of education going into schools. There are lots of technical reasons for that, which I am sure we don't have time to go into now. I just want to register a strong disagreement within the research community with what Tim has said.

Chair: You have registered that and it will be here forever on public record. Rebecca, anything you would like to register?

Dr Allen: I would keep performance tables, but I am concerned about a situation where head teachers lose their jobs or where schools are forcibly passed to sponsored academies on the basis of data that may be insecure and fragile. I want human beings to be the ultimate arbiter of whether everything is okay at a school.

Chair: In the world of artificial intelligence, you might be tested about that in due course. Joanna.

Joanna Hall: I would wish to add from an Ofsted perspective that we know that they are complex and demanding, the new assessment systems and performance measures. They are harder for the sector to interpret and we absolutely understand therefore at Ofsted we have to increase our scrutiny of what we do with our inspectors and their training, and also when they are out in the field working with leaders to understand the quality of education in those schools.

Chair: Thank you. We have set the scene for an intense probing of the system and how we might improve it. Catherine.

Q86

Catherine McKinnell: We know in my local area in Newcastle, in 2016—last year's assessments—22% fewer children achieved the expected standing in reading, writing and maths than in 2015. We know that was a tougher exam, but one local teacher in my area said that the goalposts were moved overnight, which created what has been testified as quite a chaotic situation for a large number of schools.

I know the Government have said they have not made a decision yet or they will not be making firm decisions on intervention based on the 2016 results. What is your view in light of the evidence you have already given



HOUSE OF COMMONS

about making decisions, but also putting schools into league tables on the basis of the 2016 data? Would you like to start, Joanna?

Joanna Hall: Yes, I would. One of the most important things is that when we set the common inspection framework, we knew these changes were coming. In terms of helping inspectors to understand what the implications of those changes are for schools and school leaders and also for the teaching community, an important part of our advice, our training and our guidance to inspectors is that this is a different bar. It is a very different set of results than inspectors may have seen in previous years. Particularly with teacher assessments, our guidance was to go with caution when you are talking to leaders about the impact of those teacher assessments and what they look like. I do believe, in terms of this first year of assessments and guiding inspectors to deeper scrutiny, that we are, when we talk to leaders, aware of that shift.

Dr Allen: There are two separate issues here. One is, were the 2016 tests reliable? I think they were. All of our judgments say that they were, in that the school year on year relationship between previous performance is stable and it is consistent, as it should be. The relationship of the pupils' outcomes compared with our prior attainment measures is exactly as it should be. The correlations are as we would expect, except for writing, which is a separate issue. But then you raised the question, should we be setting a threshold, which is ultimately arbitrary? It is set by Government in a way that has no kind of educational foundation. Should we have set it in such a way that half of schools were told that they were not over the bar? Personally, I don't like thresholds. I do not think it is meaningful, necessarily, or useful to talk about expected standards, and I don't think we had to do that. The introduction of the scaled score meant that we didn't need to do that.

Professor Goldstein: Yes, I agree with Rebecca. The notion of a threshold, especially if the threshold is kind of towards one aim, one extreme, is problematic because the uncertainty surrounding thresholds is absolutely enormous. The uncertainty surrounding just an average is large enough, especially when dealing with primary schools. If you are dealing with the one form entry schools it is absolutely enormous. The DfE recognises this in published tables. It won't publish certain results that are based on too few pupils. That again is arbitrary. If you are basing that on less than 10 pupils, why 10? If you have 11 pupils, that is just as uncertain as having nine pupils. So the notion of a threshold I don't think is a good one. If you are going to use any kind of measure do not use a threshold; it is highly variable. Again, using the national pupil database, which we have done, the year on year correlation is not very high when you have small numbers in primary schools. It is quite low. For example, if you want to say something about how schools function now, how a school is likely to function next year based on past results, you have to take into account the fact that trying to predict year on year is an uncertain basis, so last year's results do not predict results in two or three years' time. That is reflecting again the inherent uncertainty in



schools. It is bad enough in secondary school; in primary schools it is far, far worse. Anything that downplays that I think is excellent.

If I may just comment a little bit on the Ofsted model, which I think again has a rather fundamental flaw in it, in that, when inspectors go in to judge what is going in the school, they know and they utilise the statistical information that is there from the test. What you really want is an independent judgment of what is going on inside the school, which you can then put together with the statistical information. The problem about all Ofsted inspections—and it is a bit of a generalisation, but let me let that stand—is they confound two kinds of measurements. They confound the measurement the inspectors make when they go into schools, judging classrooms and teachers and so on, with the statistical evidence that is measuring something different. It would be much better and provide much more useful information if those were completely separate. That means of course changing the whole Ofsted model, which Ofsted may not like. But that would be my view about getting decent measurements.

Catherine McKinnell: Tim, do you have anything to add to that?

Tim Oates: Yes, just a few notes. I would like to pick up particularly on Rebecca's point about the underlying data. The country does need to give some attention to the technical measurement characteristics of the 2016 tests, and I concur with Rebecca. The data are looking okay in terms of measurement instruments, but we have to differentiate three things and we have to recognise one important fact. We need to be clear that there is a difference between standard-setting in respect of a test, whether it is a GCSE or a national assessment, so setting the standard, monitoring standards over time and maintaining standards over time.

The departmental statements clearly acknowledge that we had more demand in the national curriculum that was in place for a short time in schools. The cohort being tested had not had the exposure to the quality of the curriculum associated with that new demanding curriculum for the entirety of their primary education and the test had to be set for the first time in relationship to a cohort that had only taken a part of the new national curriculum. It was an exercise in standard-setting and equating was therefore very problematic in terms of the 2016 test and any prior tests. We therefore would look technically at it in terms of rank order of schools. We would look at the internal measurement characteristics that Rebecca has outlined.

I think the Department acknowledges the year zero nature of the assessments. You will find in a lot of their documents something very useful: they acknowledge the volatility of test outcomes over time. In some segments of the accountability they are now looking at windows for making judgments about schools and Cambridge has done quite a lot of work on the underlying natural volatility in GCSE examination results. It is very important to understand that this does occur. One year of depression can be accounted for by a wide variety of factors, some of



which are outside the control of the school in terms of the quality of their education.

Q87 Catherine McKinnell: Around all of this, school league tables continue to be compared regardless. I know, Harvey, you have previously written about what you see as the negative impacts of school league tables. Do you think there are any positives?

Professor Goldstein: Not many. I did allude earlier to the Hampshire experiment. They can be useful, but they could be useful as simply one bit of information in making a judgment. Most people accept the notion of accountability for schools and this is one little bit of information.

I would say two things: first, you need to recognise the inherent uncertainty, as I have already indicated, with the results and the variability, as Tim suggested, from year to year that limits their usefulness. There are three points. Secondly, they need to be refined even further. I think Tim was right in saying the first league tables that were produced were simply average scores, GSCE and then key stage 2 results and so on. Then they moved to a value-added system where they would try to take into account the prior achievement of students. That was certainly an advance. That came in, I think, in 1997. Even more recently the DfE has advanced by looking at small groups of children, for example, looking at what happens to children who have low prior attainment in different schools, so making more detailed comparisons.

The issue and the problem with that is, especially in primary schools, when you do that you are dealing with almost tiny numbers of students. If you are looking at the differential performance comparatively across schools, for those who come in with very low achievement or very high achievement or, for example, for different ethnic groups or for boys versus girls, then you begin to start dealing in comparison with very, very small numbers. The thing then peters out because you cannot do it. So there is an inherent difficulty relating to this whole uncertainty associated with small numbers.

Finally, I fall back on this notion of thinking of this as contributing evidence, whereas at the moment the problem is it is the headline evidence, that is what makes the headlines and it should not be. It should be way back in the background, of use as backing up or indicating where there may be issues, but not as the primary source for making judgments about schools.

Chair: Joanna, did you want to answer that?

Joanna Hall: Yes, I would, please. I would like to come back to what Harvey said earlier about inspectors going into schools to diagnose the problems and find the issues. I think that would be a sad place to be if we were still there. What we have done in recent years in changes to framework and changes to training is work with leaders to find absolutely, "What are the strengths of your school? Are you aware of



things that you might be working on and how are you driving improvements?" That is the first thing.

The second thing about those statistical pieces of evidence that are at the end points of key stages, they are only a component that reflects performance outcome. What about the children's performance in other year groups right across the curriculum? Yes, there are some headline measures, yes, they are discussed with leaders, but above and beyond that it is very important that in looking at teaching, learning and assessment we look at all the other year groups within the school and obviously across the curriculum.

Q88 Catherine McKinnell: Joanna, I am going to ask you, but I would like to hear from the whole panel, do you think that Ofsted currently carries out the job effectively of assessing schools and do you think there are aspects that could make the process even more effective?

Joanna Hall: Yes, we do. One of the trajectories of improvement, in terms of accountability, is the performance now of 90% of primary schools at good or better, of which 15% are outstanding. The level of challenge and debate there has been part of that improvement journey. Yes, I do.

In terms of the workforce coming in-house, we have many more serving practitioners who work with us who are practising head teachers, deputies, executive heads and that helps us bring that knowledge into the workforce. Equally, our new chief inspector has said publicly that in terms of working with the research community, like the colleague sitting to my left, that is a very important part of Ofsted's journey.

Q89 Catherine McKinnell: Rebecca, what are your thoughts in terms of primary assessment?

Dr Allen: In terms of the publication of league tables or Ofsted?

Catherine McKinnell: No, of Ofsted and its effectiveness.

Dr Allen: I personally do not think that Ofsted inspectors currently make reliable judgments. I don't think they walk out of the door of Ofsted sufficiently trained such that they are making consistent judgments when they walk into schools, but the past chief inspector did something right by bringing inspection back in-house. That sets a train in place where I think they now understand that one of their primary objectives must be to ensure consistency of inspection.

Q90 Catherine McKinnell: What do you think is the way that they can achieve that? Better training within Ofsted?

Dr Allen: Absolutely, training and testing of the inspectors. I don't want inspectors to walk out the door without passing a test that shows that they can watch videos and make consistent judgments on the videos.

Catherine McKinnell: They can have their very own Ofsted inspection.



Dr Allen: Absolutely.

Catherine McKinnell: Harvey or Tim, do you have anything to add to that about Ofsted?

Professor Goldstein: No.

Lilian Greenwood: Could I ask a quick follow-up?

Catherine McKinnell: Hang on a second, Tim is just going to answer my question.

Tim Oates: Just very briefly, I endorse absolutely Rebecca's point about the focus on the consistency of judgments. I think that is in train within Ofsted and there are processes to address that. On the reality of the curriculum, we know from other work that it is the quality of the exchange in the classroom that is absolutely fundamental. If Ofsted are unable to reach down deeply enough into schools to understand what is happening in the classroom then that is a problem. I know that the previous chief inspector and the new chief inspector are very committed to ensure that that reality is apprehended by inspectors. I do think that the use of data by Ofsted was not good 10 years ago. Again, I think it has improved dramatically. The kind of refinements we are talking about, the improvements, the interpretation of data in the research community, Sean Harford and other colleagues in Ofsted are very much on top of that.

There is an outstanding and very demanding problem though for Ofsted. That is the density and frequency of inspections. There are many schools, because they have good scores, that do not have frequent inspections and that is a problem for the system, not least in terms of the identification and dissemination of good practice.

Q91 **Lilian Greenwood:** On the use of the test around accountability, one of the comments from a number of schools in my constituency in Nottingham is that they have high levels of pupil mobility and they say, "Of course we would be able to get the pupils up to standard if they were here from reception, but a lot of them are joining part way through key stage 2". Do you think that is properly recognised, particularly in the use of league tables or do you think that is problematic?

Dr Allen: We have done quite a lot of work on this, but in the context of secondary schools we have recommended that we should weigh the contribution a pupil makes to the performance table according to the time that they have spent in the school. You could imagine a similar situation. What we do not want is to drop highly-mobile pupils from the data because it means that the day that they arrive in school the school simply has no incentives to ensure they do the best for that child that they possibly can, but we may want to lower the weight that we attach to that child's results. It is very straightforward for us to do that.

Q92 **Catherine McKinnell:** Just taking the question from the whole school



assessment and Ofsted back to the individual pupil, do you think it is right that individual pupils have been measured by their performance on the basis of a single test and how would you change that, if at all?

Dr Allen: Are you asking whether I would like more testing?

Catherine McKinnell: Do you think it is right that an individual pupil is assessed at key stage 1, key stage 2? Is it a good system?

Dr Allen: I think it is okay. As I said, I think they are reliable enough, given we do not want to spend hours and hours and hours testing children. They provide relatively imprecise measures of how children are doing but they are better than nothing. They are absolutely informative. Would I have more testing? For the sake of just providing feedback to the pupil or their parents, absolutely not. I would only be interested in extra testing in the situation where we are trying to substantially alter the timing at which we are incentivising schools to teach particular things, as we did for the phonics test.

Professor Goldstein: I think you need to distinguish the kinds of testing, particularly between what is called formative testing—which is not for accountability purposes, it is for the purposes of learning and encouraging children—and accountability testing, which is what we are talking about now. What I would do personally is reduce drastically the amount of accountability testing. It seems to me this is totally over-cured; you don't need this. It is not totally useful anyway, as I said, for making compared judgments between schools.

If you want a monitoring system of testing for the whole of education, you can do that by sampling. You can do that, as the Assessment Performance Unit did it in the 1970s and 1980s. You do not need to test every single student several times. The more you have good, formative testing that is used by the teachers to understand where pupils are and what they need to know the better. I think it is terribly important that you distinguish the term. One of the issues at the moment is that this distinction is not properly made. People just talk about testing.

Q93 **Catherine McKinnell:** Does the panel agree with that or do you agree with the single test, high stakes model for accountability?

Joanna Hall: I think the most important thing there is, whatever we are testing and however we do it, what does it lead to next? Does it give schools at points of transition, and also parents and the young people, the information they need, particularly from primary to secondary and about to transition—

Q94 **Catherine McKinnell:** Do you think it does, the current system?

Joanna Hall: I think at the moment it is very complex. We have heard from school leaders, particularly in secondary, that unless they do very focused work with their primary schools, understanding the nature of



assessment that then comes into year 7 is complex and I think that is harder.

Q95 Catherine McKinnell: Is that a no? Do you think the current system works, and if not, how do we improve it?

Tim Oates: Although Harvey disagrees with me on international comparisons, I may just mention briefly a couple of other countries. It is critical to think about purpose. What is it you are interested in and who is interested in what and why?

Catherine McKinnell: But does the current system work?

Tim Oates: The current system is producing a great deal of data that allows us to understand three important things, between school variation, within school variation and it gives us evidence about the performance of the education system as a whole. That is the key thing about purposes that I was describing. I agree with Harvey about the importance of ongoing assessment on a day by day basis—

Q96 Catherine McKinnell: But you wouldn't change the current system in terms of the single test assessment measure?

Tim Oates: There isn't a single test assessment measure.

Catherine McKinnell: In terms of proper accountability.

Tim Oates: We have a series of tests throughout schooling, all of which produce data that is used for a variety of purposes, from the phonics check through to the end of the stage 2 tests.

Q97 Ian Mearns: There is a massive importance and difference in respect of the key stage 2 tests, in terms of the way they are given to individual pupils, but also the school as a whole.

Tim Oates: Absolutely. This is why this issue of knowing what the purpose is, is important. We need to find out about between school variation, within school variation, and what is happening in the system as a whole. There are various ways of doing that. There are refinements of the way we do it now and there are alternatives. As Harvey says, you don't necessarily need to test every child every year to know what is happening in the education system as a whole.

Q98 Ian Mearns: Tim, in answer to a previous question you said that the cohort that was dealt with in the last year had not been through the curriculum for the whole time in that school and then were tested upon that. How fair is that for the school and for the pupils?

Tim Oates: It is fair if you are using it in an appropriate way. We rightly, as I mentioned—

Catherine McKinnell: But is it being used in an appropriate way, the current system?

Chair: We have to move on or we will confront a situation where some of



the pupils at primary school will be doing A levels by the time we finish. One question, Ian.

- Q99 **Ian Mearns:** This is not a question, it is an observation. I am inviting the panel: there are many unanswered elements to those questions. Would you mind writing in to us with what you think about that particular question?

Tim Oates: Yes, of course.

Chair: We are going to move on now.

Lucy Allan: I just want to quickly add my voice to what Catherine was saying—

Chair: As long as it leads quickly to your question.

- Q100 **Lucy Allan:** She was talking about local schools in her area in response to the 2016 stats, because I have similar debate in mine. A lot of schools said to me that they did not feel it fairly reflected the learning of the pupils. Moving on from that, given the importance of non-cognitive skills in children's learning in primary schools, such as motivation, perseverance, resilience, optimism, all those sort of skills, do we think that the new accountability system has sufficient emphasis on the progress as opposed to attainment? Should attainment ever really be a measure for the purposes of accountability? Could I start with Tim?

Tim Oates: It is important that even attainment measures in terms of what somebody knows and can do and can demonstrate clearly through a formal test obviously has an element of non-cognitive ability within it, because you have to be organised, you have to be reflective, you have to apply yourself to achieve those things. It is not that they are entirely independent. In terms of the emphasis on progress, yes, I think we are beginning to see a much more balanced understanding of ensuring that we collect data from institutions and on individuals in terms of the progress they have made, from distinctive backgrounds, whether those backgrounds are associated with prior attainment or from the social and economic location of the child. Again, we are seeing increasing sophistication being taken into account in terms of the development of measures and the collection of the data.

Professor Goldstein: Just on progress, children make progress at different ages, at different times, different rates and teachers of course will understand that. The problem about standardised testing is you test everybody at exactly the same time if you want some comparability, but that isn't very faithful to what goes on in terms of learning. Hence my emphasis on the formative side of assessing. If you are interested in students learning, as opposed to accountability, what you don't want is the external test, what you want to do is to emphasise and train teachers in proper formative testing.

- Q101 **Lucy Allan:** Do you think there should be more emphasis on progress then?



Professor Goldstein: Yes. All teachers will be judging progress, not on attainment just at a particular point in time. It is relative to where they were before. The judgment of progress is inherent in any kind of assessment, certainly teacher assessment.

Q102 **Lucy Allan:** Rebecca, is attainment ever the right measure and should it be more heavily skewed towards progress when looking at accountability?

Dr Allen: From the perspective of the school, I think we should be measuring progress whenever possible.

Q103 **Lucy Allan:** Is there a better way of doing it than is being done at present? Do you think we are doing the progress measurement under the new system in the right way or is there a better way of doing it?

Dr Allen: Our critical problem at the moment is that because we do not have a well-measured baseline we do not have reasonable expectations for each child of where they should be.

Q104 **Lucy Allan:** How do you think a reliable baseline could be measured?

Dr Allen: I think a reliable baseline at age four probably already exists. It already exists among one of the commercial test providers. We just need to look closely and make a good decision as to which one we are going to use, recognising its flaws, but recognising it cannot possibly be any worse than the current baseline that we are using, whether that is foundation stage profile or key stage 1. It does not have to be perfect; it cannot be perfect.

Q105 **Lucy Allan:** Is there an appetite, do you think, for that to be changed? Is there going to be a receptive response to change?

Dr Allen: Within schools? I do not know.

Joanna Hall: I think right at the heart of the common inspection framework is progress. It is that discussion with leaders about current progress of children in the school. If I am with your school for two days tell me about, from starting points, where your children began. What have they learnt? What are the concepts that they know, understand and can do and how do you plan for that in the sequences and whatever you put into your planning that then sustains that progress of that child? I think it is right at the heart of the teachers' standards, standard 6, that understanding of the craft of—as my colleagues have said—formative assessment, absolutely. Talking about that progress discussion in schools is key.

Dr Allen: Absolutely.

Q106 **Lucy Allan:** In terms of reliable baseline assessment, what would you suggest doing?

Joanna Hall: That is obviously a matter for the DfE, where they go next and decisions about what that looks like and what it tests, but I think, as Rebecca said, the reliability and validity of that are right at the heart of



HOUSE OF COMMONS

the issue. Clearly, if it were part of a new framework, we would consider that as part of inspection, but what could that look like? We would be happy to continue that dialogue with you and write to you about our further thoughts on that.

Q107 **Lucy Allan:** You share Rebecca's view that that has to change?

Joanna Hall: We have to have some kind of measure from which we can then judge progress, yes.

Professor Goldstein: Could I just make a technical point about reliability? If you are testing a four year-old, despite what Rebecca claims, you cannot get a very reliable test. Even at the age of 11 or later, you are talking about building a temporary reliability coefficient at something like 0.8. At the age of four this is going to be down to 0.7, 0.65. If you are going to use that in any kind of analysis then you have to get quite sophisticated about taking that relatively low reliability into account, otherwise you are going to come to incorrect conclusions. It is a slight technical point. It is very poorly recognised, it is not recognised in any of the analyses in national testing. It is extremely important. If you talk to measurement specialists, they will always emphasise the fact that you need to take account of that relatively low reliability. Only when you have reliabilities approaching the value of one, which is perfect reliability, can you ignore it. I just want to enter that technical caveat into this discussion.

Chair: Thank you. What this Committee would be quite interested to discover is what you think a baseline should contain, what the coverage should be. I am not inviting you to answer that question now, but I would like you to drop us a line to explain—all of you—what you think it should contain. That is at the essence of the point that Joanna has made, when she was pointing out that you have to look around the whole school and the whole coverage of what is happening in the school, so the baseline matters there, doesn't it? Obviously, given the dialogue with the Department for Education, that is a work in progress, but we want to be able to make comment on this in our report.

Ian Mearns: Once the baseline is established, in your opinion, perhaps you would also let us know what you think an intelligent accountability system looks like.

Chair: We are going to probe that in a moment. If you would all just drop us a line, that would be very helpful.

Now we are going to have assessing writing from our resident former teacher, William.

Q108 **William Wragg:** Good morning, everybody. I want to talk about writing. I probably cannot give it as much time as it requires with the time we have left this morning, but a quick question to Tim first of all, in that he chaired the expert panel for the national curriculum review that led to the spelling, punctuation and grammar test. I want to ask Tim directly, did the 2016 SPAG test fit with the panel's recommendation?



HOUSE OF COMMONS

Tim Oates: Obviously the feedback from the schools and what we have seen by way of reaction to the test suggests that the requirement is somewhat overblown at the top end of the national curriculum in terms of technical language about language. The testing appeared to be adequate in relationship to testing the content that is currently in the national curriculum in year 5 or 6. My own personal view and the recommendation that I have made to the Department is that the Department should look at the requirements at the top end in terms of language about language in years 5 and 6. The test was accurate in relation to it reflecting the curriculum content. The curriculum content needs to be examined.

Q109 **William Wragg:** Could I go along the panel, starting with Joanna, if I may, and ask what evidence there is that the spelling, punctuation and grammar test, or teaching of it discretely, improves children's writing?

Joanna Hall: We have reported very much on that as a key part of the improvements that we have seen in literacy in primary schools. The annual report last year addressed the improvements, which we have written about in things like the commentaries. That has been a key feature of some of the improvements that we have seen in children's literacy.

William Wragg: Rebecca?

Dr Allen: No comment.

William Wragg: No?

Professor Goldstein: Writing is not my expertise.

Q110 **William Wragg:** That is fine. Maybe just a further supplementary for Joanna in that case. Talk about the improvements in literacy. Do you talk in terms of the technical grasp of being able to identify complex linguistic structures or about the flair that children have for writing?

Joanna Hall: Our research obviously looks at both and on inspection we definitely would look at both. I would want to come back to you and talk to other colleagues about what we have that underpins that from a range of inspection evidence that we have.

Q111 **William Wragg:** If I may now perhaps go on to talk about the differences between the secure fit model and the best fit model. At the moment of course with the secure fit model for assessing of writing, could I ask, beginning with Tim, is it working and how would you change it, if it is not?

Tim Oates: Again, I will go back to the purpose of testing. We know that with levels they were far too coarse a means of describing children's attainment. It was only when sub-levels were introduced, just as mathematically generated subdivisions of levels, that you could begin to look at decent correlations between prior attainment and later attainment such as GCSE and so on. All of that pointed towards moving towards scores and much more finely grained evidence on attainment than levels.



HOUSE OF COMMONS

That was very, very important. There were real issues five, six years ago associated with threshold, best fit, secure fit.

The assessment of writing remains extremely problematic, of course. It has all sort of practical difficulties and there are real issues of moderation. It is right that we look at alternative approaches to assessing writing. Things like comparative judgment are being explored to see whether we can introduce new means of more consistent assessment of the artefacts that children produce in writing.

William Wragg: I will pose the same question to Joanna, if I may.

Joanna Hall: We recognise that there may be some variability in the teacher assessments this year in writing—and we have talked to inspectors a lot about that. In respect of that space, the removal of assessment about levels, and what we have reported on in terms of schools developing their own assessment system, some schools are much further on in terms of those constructs, which they are looking at within an assessment system, than others. That is something again we reported on in the annual report this year.

Q112 **William Wragg:** I don't want to encroach into the next question to do with the assessment about levels, but is that compatible with even having the best fit model? Is it compatible with the secure fit in terms of assessing writing in that way?

Joanna Hall: That is where, in terms of the workforce and what we do to look at that in a rounded view, as Tim has just said, there are challenges.

William Wragg: Doctor and Professor, do you have anything to add at all on writing?

Professor Goldstein: I don't think so.

Dr Allen: I can add something on the writing assessment and in particular the moderation of it. There were clearly very, very serious problems with the moderation of the writing, but in part they reflect the lack of clarity and guidance over the criteria by which writing should be judged. That in turn reflects the fact that it is very hard for us to write down, as a checklist, what constitutes good writing. There is such a thing as good writing and there is a shared expert understanding and the strong intuition when you see a piece of child's writing about whether it is good or not or whether one is better than the other. That is where comparative judgment works really, really well. It works well where we can conceive of a test—for example, the timed writing condition—but we deliberately want it to be open-ended and we do not want to write a mark scheme of criteria the child has to meet to do well or not. That is why in this very particular circumstance comparative judgment is such a compelling way for us to judge the standard of writing of 11 year-olds.

William Wragg: Yes, particularly given the subjectivity of what you would judge as a good piece of writing, speaking from experience.



HOUSE OF COMMONS

Dr Allen: Absolutely. You have to have the combination though of subjectivity, but also shared understanding of what constitutes good writing. If you do not have that shared understanding, comparative judgment will not work.

William Wragg: Do you have something to add on that, Tim?

Tim Oates: There is a kind of classic paradigm in assessment, that to make it consistent you have to make the task as similar as possible over time and between individuals. That is a paradigm that we need to question in relationship to writing. That is one of the assets of exploring these new models of assessment like comparative judgment. You can retain the variety and the outcomes of students, but deploy professional judgment consistently. The evidence, the background measurement characteristics, the technical characteristics, are very promising in this new approach to assessment.

Q113 **William Wragg:** There are lots of calls for urgent changes to the writing assessment framework, indeed in advance of the Government's consultation on the matter. I ask the panel: do you agree with that and what changes would you like to see, if you have not covered that already? I will start with Joanna perhaps.

Joanna Hall: We have to have that debate about, should we remove that teacher assessment entirely from that writing space.

Dr Allen: I agree. I would like to remove teacher assessment—

Tim Oates: Again, we pay attention to the data. We have been doing quite a lot of work in Cambridge comparing data and outcomes of assessment at a time when they were formal tests, external tests, well-designed, and teacher assessment. In many cases it is unfair to expect teachers to assess with the kind of precision and consistency that you can yield from a well-designed external test, although it often involves quite draconian pressures on schools to behave consistently and results in often very poorly-designed tasks in pursuit of consistency. We have to be very realistic in terms of the level of dependability that we can yield from teacher assessment and whether it is always fair to expect teachers to assess with a level of consistency that we expect when we use the data for particular purposes.

Q114 **William Wragg:** If not teacher assessment, could you just repeat what would you replace it with?

Tim Oates: In respect of things like reading, we have very, very good—

William Wragg: Particularly about writing.

Tim Oates: With writing, it is well worth exploring the use of comparative judgment. It has been used experimentally within research and yielded very, very good outcomes in terms of very high dependability. One of the key things about comparative judgment is that artefacts from children are judged on more than one occasion by more



HOUSE OF COMMONS

than one judge, so the totality of the evidence of these multiple judgments by multiple judges yields a very high level of dependability. The challenge is often logistic in terms of doing it at scale and doing it rapidly enough.

William Wragg: Indeed.

Tim Oates: We are beginning to edge towards an understanding of what we need to do and towards both technical systems and practical administration that will enable us to do it, I believe.

Q115 **William Wragg:** Just briefly, you mentioned the burden on teachers in assessing writing. Who would be the people who would do that comparative judgment?

Tim Oates: You would have judges and they would be presented probably with onscreen, scanned images of children's writing and make very rapid judgments. It is the kind of thing done by teachers in respect of GCSE at vast volume already.

Q116 **Lucy Frazer:** I used to meet with some primary school teachers who specifically raised the issue with me, as I have dyslexia, and how unfair it was on both teachers, but in particular on students, being tested on things like spelling when they were never going to be able to achieve the results that they were expected to achieve and how demoralising that was. What can we do about that in a standardised testing system?

Dr Allen: We cannot create a test where children who cannot spell are told they can spell, so the difficulty arises when we aggregate lots of different dimensions of the attainment and performance of what a child is capable of doing into one overall measure. But there are serious differences between the sub-domains and that is what creates the difficulties. In the case of these children, they will always find it difficult to achieve a good score in the spelling, punctuation and grammar test. If we were sufficiently concerned about that we could just simply separate them and have a spelling test and then have a punctuation test and then have a grammar test, but as I understand it, that is the only choice that we have.

Q117 **Lucy Frazer:** So you are not concerned about the impact that that has on the student who is almost certainly going to fail that test?

Dr Allen: I am concerned if, because they know they will fail the test, they will therefore not practise and get as good a result as they can in punctuation and grammar and indeed in spelling. Yes, I would be concerned, but then that comes back to the problem of the threshold. I do not like tests and I don't think we need them where people pass and fail. We can talk about where you are in the scale without that.

Chair: We are going to have five minutes with Ian on the subject of removal of levels.

Q118 **Ian Mearns:** The removal of assessment levels: how do you think that



HOUSE OF COMMONS

was managed? How do you think the quotas were managed? I seem to remember meeting with heads and staff in schools in my area. Some of treated it with a bit shock in terms of the way in which it was announced and then managed.

Tim Oates: I will certainly dive in. It was a principal recommendation of the expert panel and we considered a very wide body of evidence in terms of making the recommendation that levels should be lifted as a formal requirement on schools. I continue to think that we were absolutely right so to do, for a whole series of technical reasons.

Q119 **Ian Mearns:** It might be the right thing to do, but how do you think the process was managed? That is what I am asking.

Tim Oates: Absolutely, and your question is about implementation. There are two things that become a requirement when you make a major change in the system. One is to communicate the reasons for making the change and then the second is to provide alternative ways of operating within the system in order to ensure that a void doesn't open up in terms of practice.

On the first, the issue of communication of principles, there were some weaknesses in the communication strategy. It was relatively late. A number of people, including me, were involved in a lot of conferences around the country and over a period of time we began to be able to give very clear messages as to the underlying principles. There was approval of a range of schemes that went on to the DfE website. Most of those met the criteria of assessment without levels. A couple were pretty close to being reinstatement of levels and those have now been removed from the DfE website.

Schools have, in many cases, been left to divert their own arrangements and that was led principally by the notion of school autonomy that was very dominant in Government thinking at the time, a few years ago. It has led some schools to say, "We have moved away from levels" and when you look at the system they have implemented, it is levels. There has been very patchy implementation. Organisations that are providing ongoing CPD to schools have confronted this and I think that the new chief executive of the College for Teaching is aware of the need to communicate both the principles and get a good exchange about good practice at the level of process.

Q120 **Ian Mearns:** When the initial announcement was made—and I may be technically wrong on this—I think an awful lot of schools felt, "What do we do in that case? We are getting rid of levels, so what do we do?" Do you think that the process of managing that was a bit ineffective inasmuch as one system was being removed, but schools were not being told immediately what was being expected of them in terms of replacement of levels?

Tim Oates: There were problems associated with it. It was principally associated with the notion of autonomy. On a whole series of dimensions



HOUSE OF COMMONS

there was a commitment to try to increase the amount of control over process in the schools that head teachers and management groups were exercising. We needed to communicate extremely clearly the principles and evidence associated with the removal of levels and that would have helped more the development of school-based systems and the selection by schools of commercial systems that were being made available.

Q121 **Ian Mearns:** Would you accept that that wasn't done initially anywhere?

Tim Oates: I think that could have been improved.

Q122 **Ian Mearns:** From the perspective of Ofsted, has it created particular problems from Ofsted?

Joanna Hall: Yes. We asked inspectors to look very carefully at this last year and one in three schools we reported on in the annual report were still at a very early stage of developing life without levels. What does that look like? As Tim has said, that means in schools that leaders and teachers and governors have to have a secure understanding of assessment in order that they can then devise a way of assessing their children in a world that isn't about levels and telling a student, "You are a level 3B". What does that mean without that tag on that child? There are good schools that have moved on, have innovated and are doing some good work that we have reported on with assessment, but then the picture is very mixed, so absolutely, what Tim has said is right there. That is borne out in what we have seen: a very mixed picture.

Dr Allen: If you look from the perspective of Government, why on earth would they take away levels, which everyone agreed was a good thing, and put nothing in its place? Either they did it because they or experts didn't know what the right thing was to put in its place—and there is a case to say that was true at the time—or they thought that teachers had the best answers. I am not sure they ever believed that or they thought teachers didn't have the best answers. They believed in innovation and that teachers would sort it out for themselves. The problem with that is that we do not have a system of training for teachers that makes them in any way experts in assessment. You have to be knowledgeable about assessment to be able to go on and devise some sort of system or make good purchasing decisions. That is why we are in this place where even those schools that we consider are in quite a good place around new systems aren't in a great place.

Q123 **Ian Mearns:** Because of that void, was it not the case that a number of off the peg solutions were devised that were readily available online for some amounts of money that were not really solutions at all? That is part of the problem in doing things in the way that they were done.

Professor Goldstein: I am not aware of the details of how this worked, but presumably one of the key issues in abolishing levels in the three months that have banked up is the labelling one. There is the self-fulfilling part and the labelling one. It seems to me the way in which you have to judge whatever replaces it is the extent to which that has been



HOUSE OF COMMONS

removed. I am not aware of any studies that have specifically gone into schools and studied what has replaced levels in those terms. That seems to me to be the key issue here.

Ian Mearns: A quick answer from everyone: do you think that the removal of levels has brought about the desired result? Are we in a place that we would want to be with the removal of levels?

Q124 **Chair:** Yes or no is what we are looking for here.

Dr Allen: No.

Joanna Hall: Not yet.

Professor Goldstein: We do not know until we do the work.

Tim Oates: Not yet. It is a very varied picture in the UK system.

Professor Goldstein: It needs a proper evaluation.

Q125 **Chair:** Rebecca, you talked about training of teachers. Presumably logic follows that that is also very critical in terms of the head teacher?

Dr Allen: That is absolutely right. We face a challenge that we need about 17,000 primary school head teachers who are experts in judging how well their teachers are teaching and what children have learned.

Chair: Thank you very much. Last but not least, Lilian is going to be looking at alternative assessment methods.

Q126 **Lilian Greenwood:** We have been looking at the current assessment method and I think in response to earlier questions both Becky and Tim have started to talk about the use of comparative judgment for the assessment of writing, but I wanted to ask a broader question. What other methods of assessment could be used for statutory national tests in primary schools, and would these work better than the current assessment system that we have? I do not know whether you, Tim, could pick up any international experience, because you have hinted at that, but we have not heard any.

Tim Oates: I did emphasise right at the beginning that you find what you want to look for when you go to other nations, and it is critical that during the period of Finnish improvement during the 1980s and 1990s things were going on that people, when they think about Finland, have no knowledge of at all. It is because they have asked the wrong questions. In the 1980s and 1990s there was a great deal of testing going on in schools and some of those tests were devised by the teacher associations and unions, and it was all associated with making sure that learners were always identified if they were falling behind.

Q127 **Lilian Greenwood:** Were they used in the school or were they used for comparison between the schools?



Tim Oates: Interestingly, they were used in schools. There were also state grade tests introduced, which were not of every child. This is Harvey's sampling issue. The state was interested in what was happening in the system as a whole and they wanted to know that within schools children who were at risk of falling behind were identified and their problems identified accurately through the use of formal tests and then remedial action put in place. So a high density of testing is often characteristic of those systems that are improving, but with the outcomes of the tests being used on a day-to-day basis within the institution to support children.

We need to think about that when we are thinking about what kind of density of test and what sort of test do we want in our system. We do not over-test in England at the moment. It is the uses that give these low density of tests such high prominence. In reading, we need to know whether children can read, but we have plenty of standardised tests. We can test children at any age as to their reading age, so you can choose when to test them. This is often a key thing for teachers. They want to be able to choose when to test and they want to choose the test from a range that are available. Reading is internationally not difficult to test, and we ought to not really stumble over its assessment in the way in which we do.

In terms of writing we need to look at comparative judgment, we need to examine it experimentally and see what we can achieve through it, both in terms of its measurements characteristics, how it is administered and whether it is manageable.

In terms of mathematics we need to explore much more the use of bagged items. We are convinced in Cambridge that in the future we will see online adaptive tests that can be administered at a time of the school's choosing and the results will be extremely dependable. It will not be a question of issuing one test for all kids on one occasion, where some of the items are below the level of the child and some of the items are above the level of the child, because such items are a waste of time. The nice thing about adaptive testing and online adaptive testing is that very rapidly the test is only providing items that fall within the range of things the child can do, and gives you a lot of nice data on their coverage of different topics and areas of the curriculum.

All of these things are available. I have not made recommendations to the Government about exploring these as options within our existing testing regime. The key thing is don't fall into the trap of saying that what we need to do is have far fewer, smaller tests because if you have far fewer, smaller tests you just do not get enough evidence to make good judgments on that small amount of evidence that is available.

Q128 **Lilian Greenwood:** That sounds like quite a different approach though to every child sitting down and doing the maths test on one day, albeit much more reliable. What do other members of the panel think? Harvey?



Professor Goldstein: I think the key issue is training teachers in assessment. If you look at what happens in teacher training, it really is something that is tagged on somewhere, often to the end, and that has all kinds of implications for the teacher trainers, who does the training and how it is done. That is the key issue.

We can have all kinds of weird and wonderful innovations in testing, whether it is adaptive testing or not, but if you do not have teachers who are properly trained to use them they will probably be misused. That is the key question. If resources are going to be put anywhere that is where they need to be put, I think. Not so much into new forms of testing.

What I will say about adaptive testing and where you get the tests from, and Tim talked about standardised tests and off the shelf tests, is that it seems to me what we want are tests that are strongly related to the curriculum that goes on in the classroom, both in terms of the national curriculum, the common curriculum, but also the local curriculum, what is being expected locally from children because of the nature of the environment that they are in, the kinds of children they happen to be. That is very much then related to teacher capability of providing and adapting existing instruments for their own particular purposes. It is not that teachers should be there simply to use whatever tests are available. I think teachers should be able to in some sense create tests or adapt tests for their own needs.

This relates to something I said right at the very beginning. There is a need to distinguish between formative testing, which is that kind of testing where the teachers are fully in control and they have enough information and enough understanding of assessment to know whether they should be taking this particular off the shelf test, and whether they should be devising their own test, who they should be talking to, who should be helping them to do all this.

It seems to me there is this broader question of bringing teachers into assessment by training, continual professional development, to create a culture where assessment is properly understood in schools. At the moment I do not think it is. It does seem to me that it is one possible function—and I am interested to see what Joanna has to say—that Ofsted could do because of their expertise, to provide some kind of assistance in provision of training for teachers in assessment.

Q129 **Lilian Greenwood:** If teachers are taking all the responsibility in innovating, does that not make it difficult to combine the use of tests to see how pupils are progressing with the use of them for accountability measures across schools?

Professor Goldstein: Exactly, and I would emphasise very strongly the need to separate these. The problem is at the moment the accountability component dominates everything else and it distorts the curriculum, it distorts learning, it distorts children's behaviour. There is lots of evidence now about the stress that children go under. Assessments should not be



HOUSE OF COMMONS

doing that to children. Assessments should be encouraging children to learn. So a clear separation between these two. I think the use of the word “test”—and we all do it; not just standardised tests, but all tests—to me is a bad one. We should get used to talking about assessment in all its forms. Leave the word “test” to something that is more like a standardised common test.

Lilian Greenwood: Before I come to Becky can I come to Joanna, because that seems to be at the heart of what Ofsted are involved with?

Joanna Hall: I think from all the discussion we have heard today that understanding in terms of teaching, learning and assessment on a daily basis, how it informs curriculum planning, how it informs teachers’ practice, is absolutely something that for inspectors we are looking at. In terms of what Harvey has just said, we have the accountability measures and then also we have that craft of assessment. In whatever measures we use, what knowledge, understanding and skills do we want to test, in whatever assessment methodology we might use?

Also we need to come back to what Lucy mentioned earlier, which is about, for example, those children with special educational needs. Whatever we do in terms of the assessment and testing regime, the Rochford Review made 10 very clear recommendations, and that has to be considered within those kinds of discussions going forward as well.

Lilian Greenwood: Becky, is there anything you wanted to add?

Dr Allen: The exit tests at year 6 in a sense are too late for the school to be as useful as it could be, but there are lots of fantastic things that assessment can do in helping schools to get feedback and have a conversation about the nature of the curriculum and about the quality of instruction.

I want to give an example of an innovation that has taken place, because the Government has left schools to do as they please around innovating assessment, which I think is really nice. Craig Westby, who is a deputy head in Sandwell, took all the children in his school from year 1 to 6 and he set them an open-ended writing exercise in timed conditions. Every child did it. They used comparative judgment to compare every single script in the school and place them on a scale. They learned some amazing things from this about individual children; a child in year 1 who was better than the median child in year 5 at writing and the quality of their writing. They looked at the year groups that appeared to be very similar in the quality of their writing, years 3 and 4 and years 5 and 6. That allowed them to trigger a conversation about what was taking place in the curriculum in that school that meant that the classes were not making particularly stark progress over those year groups.

What he was not able to do is to talk overall about how good his school is at writing. There was no standardisation but he could innovate now and do that. He could go and find another 10 schools and say, “Hey, let us all



HOUSE OF COMMONS

set exactly the same question. Let us do it together. Let us share our scripts”.

Chair: We are right up to the borderline now.

Lilian Greenwood: One last question. At the moment the Government is committed to a period of stability and understandably that has been welcomed by teachers and schools who feel like it is permanent revolution, even though they are unhappy with the current situation. Do you think they are right to stick with stability or are changes needed more urgently? Which way is the priority?

Q130 **Chair:** Yes or no answers would go down well.

Professor Goldstein: Yes.

Lilian Greenwood: Yes to stability?

Tim Oates: I think the type of stability that we have is good, but of course within that you need to constantly monitor and fine-tune.

Chair: Yes? No? Harvey?

Professor Goldstein: Yes.

Dr Allen: I would have stability. I would remove the key stage 2 writing as it currently stands.

Chair: So a qualified yes?

Joanna Hall: I am a qualified yes as well.

Chair: Okay. I want to thank you all very much indeed. Thank you.

Examination of Witnesses

Professor Rob Coe, Dr Mary James, Catherine Kirkup, Research Director, National Foundation for Educational Research and Professor Dominic Wyse.

Q131 **Chair:** Good morning and welcome to the second panel. I am sorry that we are starting almost 20 minutes later, but the last panel was certainly thorough. I am keen to make sure that we do finish in a timely fashion so just bear that in mind, but we will try to be quick with our questions.

You know why we are here. I think all of you were listening when I set out the purpose of today's session in particular, so without further ado, starting from Catherine, would you like to say who you are and what you represent?

Catherine Kirkup: Good morning, everyone. My name is Catherine Kirkup. I am the Acting Head of Assessment at the National Foundation for Educational Research and we have been involved in developing high-quality assessments for 70 years now.



HOUSE OF COMMONS

Professor Coe: Robert Coe, Professor of Education at Durham University.

Dr James: I represent myself, because I retired from a professorship at Cambridge University three years ago, but in my career I have worked very closely with schools and teachers, so I probably give a rather different perspective from some of the people in regulatory bodies.

Chair: This is a contested area so we are expecting some debate.

Dr James: Absolutely.

Professor Wyse: I am Dominic Wyse, Professor of Early Childhood and Primary Education at University College London Institute of Education. I am also Head of Department of the department called Learning and Leadership at the IOE.

Q132 **Chair:** We have already heard that it would be fair to say that the primary accountability and assessment system has had some negative impacts on schools. Do you think the current system is working? I am really looking for a yes or a no.

Catherine Kirkup: Yes, I think that some fine-tuning needs to be done, but it is very important to hold schools to account. The national testing has a very important role to play, not only in comparing the performance of schools but also in reporting back to pupils, parents, governors and identifying not only those schools that are underperforming but those that are doing really well. That is one way in which if you find out those schools are doing well, some of them in very challenging circumstances, you can look at what works well in those schools and how we can spread that across the sector.

Professor Coe: I think there are some very significant areas that could be improved.

Dr James: I feel rather like Rob, that there are some major improvements that ought to be made.

Professor Wyse: A very strong no. I do not think the system is working at all well at the moment and drastically needs changed.

Q133 **Chair:** We have a kind of a rising graph of disapproval here, which is very good for the purpose of this inquiry.

Two things emerged from the last session that I thought were very interesting. One was the question of training of teachers, head teachers as well. Could you all comment on your understanding of why that might be a good idea? Dominic.

Professor Wyse: I would like to make a point about teacher training more generally first. That is my point. Comments have been made about the nature of teacher training and we run huge numbers of teacher training across nearly every route for primary, early years, secondary and



post-16 and so on. Although it is true that assessment does not get a huge focus in, let's say, a traditional PGCE we have to take into account the fact that students spend much longer in schools than they used to, and while they are in schools of course they are learning ideally from practising teachers, notwithstanding the comments earlier about how knowledgeable the teachers are about assessment. What trainees learn about assessment is, like many other things they learn, part of a limited amount of time.

In my view, we should look more carefully at what they do learn in teacher training and be less quick to criticise.

Q134 **Chair:** Okay, fair enough. Catherine, on the other end of the scale, what are your thoughts about this question of training of teachers of assessment systems?

Catherine Kirkup: I agree. I feel there is a need for much greater data literacy. I do think there has been an enormous number of changes and I do not think there has been sufficient support for teachers in understanding the impact of those changes and what the changes mean. I think understanding the outcomes is extremely important and assessment has to be linked to professional development.

Dr James: Can I say something about this? One of the problems is our understanding of assessment. We can see it in the debates in this room, that assessment, I think as Harvey said, is much wider than testing and there have been moves in the last 20 years or so to get more focus on not just assessment of children's learning using testing systems, but also how you use it in the classroom, because it is a classroom process. All teachers assess children all of the time. As soon as a child comes into the classroom they observe their behaviour and they make inferences of that behaviour, they make judgments about how then—

Q135 **Chair:** Do you think there is a need for more training or different forms of training of teachers with assessment in mind?

Dr James: Yes, I do, absolutely, because it is such a central part of classroom practice.

Chair: Rob, do you concur?

Professor Coe: A very strong yes. I do not think the problem is in initial teacher training, where everything is so packed full of things that teachers need to learn and the time is very short. The problem here is that we do not have a good model for teachers to continue to learn about all the complex aspects of how to be a better teacher after that initial training programme. We need to seriously reconsider what we are doing there and how that works. There may be funding implications, but mainly cultural, I think.

I am a big fan of training and assessment and I have done a number of things to try to promote that and currently am. The scale of the problem



is enormous. The amount of knowledge that teachers need to have about assessment, the gap between that and what is widely found in schools, is massive. The number of teachers in the system is colossal and I think that is also a problem for inspection, because inspectors are going into schools perhaps without a good understanding of assessment, without a good understanding of data and what it can and cannot tell you—some of the concerns that the previous panel were talking about—and therefore they are unrealistic in their interpretations and expectations of data, so we need some good training.

Q136 Chair: We are going to do some more work on this because I think it is a very interesting territory. The second thing that emerged in the first panel was the question of separating assessment from accountability. First of all, do you think that is something that should be done? Probably a more difficult one to answer is how do you think it should be done? Rob, you have nodded there. You are smiling.

Professor Coe: Bad mistake. Yes, you cannot separate these things because accountability is the key that determines how assessments are used and we cannot be in denial about that. That is the landscape. Accountability is about a combination of what kinds of measures you have in a system, in our case assessments, largely, and how those are used, what sorts of consequences, if you like, attach to them. The consequences influence the assessments. It is perfectly possible, for example, contrary to what Harvey said, to reliably assess four year-olds, but it may not be possible to do it in a way when pressure is attached to those assessments for schools to look good. Teacher assessment may also be a valuable, legitimate and worthwhile form of assessment until you say the people making those judgments, the teachers, are also the people being judged by the outcomes and then you have a clear conflict of interest. I do not think it is helpful to separate these. I think we have to see them in the round.

Chair: Is there any agreement with Rob? Catherine.

Catherine Kirkup: I agree, I do not think you can separate them. You have to think about the different purposes of assessment. You have to think about which ones are most appropriate to be used for accountability and to try to mitigate any unintended consequences of therefore using assessment. Some of those might have been mentioned already, about not focusing just on one year's results, possibly looking over a number of years.

Dr James: Can we look at it from the other end? Look not at the assessment feeding into accountability, but look at accountability of the system that we have. It seems that we have a very elaborate accountability system in England. We have the performance tables and the assessments, we have Ofsted, we have school governors, we have regional school commissioners, we have parents as well. We have reports to parents. This is overkill. At the moment we are using the assessment



HOUSE OF COMMONS

data to feed into all of those elements and I am not at all convinced that that makes it an intelligent accountability system.

We have to look at those elements and strengthen what they are. It does not seem to me to make any sense to use one measure for all the purposes that we expect of it, for formative, for summative, for accountability because the uses to which the results are put distort the whole system. In the earlier session Harvey, I think, was touching on that. In order to monitor the system as a whole, by all means use a kind of standard measure in limited areas of the curriculum. Leave it at that and you can then look at patterns. Strengthen though the Ofsted inspection process and make it much more qualitatively based, based on the observation of teaching and learning in classrooms particularly, which I think it has been moving towards. Then we can use other things internally without the performance data.

Professor Wyse: I agree largely with what Mary was saying and Harvey. We absolutely must separate out and must be clear about the purposes of assessment. I think the reason we are here is there are well-documented problems from a variety of perspectives, the confusion of accountability with, for example, assessing what children have learned is a real problem. So we absolutely must. I have recommended for a long time that we should go back to a system of national sampling. I always think back to the Assessment of Performance Unit in the 1980s, for example, which did some really good work and enabled us to understand important things about how schools were performing, how children were performing and how teachers were performing.

Q137 **Chair:** Thank you. That is another area I think we are going to return to as a Committee.

Mary, how has the removal of levels, as recommended by the national curriculum review, impacted schools?

Dr James: I do know a bit about this even in my retirement.

Chair: I know you do. That is why I am asking.

Dr James: Yes. I was partly responsible by being on the national curriculum review that recommended that, because we saw the pernicious effects of the levels system, where children were so worried about getting a 4B but had no idea what a 4B really meant. They became the label and they were not focused on the curriculum. The first decision that Michael Gove made after our report was to get rid of the levels and there were a number of reasons why it might have appealed to him at that stage.

Schools had become used to the levels system. They did not know how to think differently, because they were anxious about how they recorded. Assessment comes almost associated in their minds with recording, tracking, writing on spreadsheets, developing charts and so forth, and they did not know how else to do it. I did make a number of visits to



HOUSE OF COMMONS

schools and gave talks to schools, particularly in Newcastle for some reason, and they were asking what the alternatives were. My suggestion to them was look at the curriculum.

We have not talked much about curriculum in this discussion this morning. The curriculum should determine what is taught in the classroom and the curriculum has a progression to it. That therefore gives teachers an idea of what they might be teaching and also what they might be assessing. At the end of any particular unit of work it is quite reasonable for a teacher, they would do it as normal practice, to look over the children's performance in various activities and determine whether in fact they have met the kind of expectations in that part of the curriculum that they have been teaching. Most of them I think could assess whether the children have met their expectations.

Q138 **Chair:** Thanks very much for that. That is very helpful. Now, one of the things we seem to have picked up is the question of training of teachers, but do we think parents are familiar enough with the assessment/accountability system? Do you have any thoughts on that? Catherine.

Catherine Kirkup: Again, I think there has been a lot of confusion due to some of the changes and it is very difficult for teachers. Teachers have an extremely hard job to do because they not only have to come to terms with all these changes themselves but they have to explain those then to parents.

Just to pick up on what Mary said though, I think one of the positives of the removal of levels is that it has focused teachers' attention back on what a child understands, what they can do. If that promotes better conversations with parents then it is a very good thing, rather than saying to a parent, "Oh, your child is a 3B" and the parent has no idea what that is. That has been one of the positives. Having been into a lot of schools, the removal of levels has promoted a lot more meaningful conversations with parents.

Chair: I want to move on to the next question. Does anyone have anything to add to what Catherine has just said? Okay. Catherine, over to you.

Q139 **Catherine McKinnell:** How do you think that the current assessment could be improved for children with special educational needs? Rob, you are nodding again.

Professor Coe: That is a good question and an important question. It very often tends to be a bit of an afterthought that we think about getting it right for about 80% or something and then we won't worry too much. That is definitely one of the things that will be on my list of those significant improvements that could be made to what we have in the current system.



HOUSE OF COMMONS

I do think there are some very specific problems about the assessment and the accountability system in relation to children with special educational needs, but I do not think they are insuperable. In general, those issues, if you like, magnify some of the problematic areas that affect everybody and therefore they are a good place to start, rather than be an afterthought. We should think about, as your question in the previous session, children with dyslexia. How are we assessing them? How are we talking about the progress they have made or the impact the school has had on them if we think that is what the accountability system should tell us? What kind of assessment would we need to do that? What kinds of accountability structures around that would we need? If we started from that place I think we would come to a very different model of what the whole system would do. It is not a precise, clear answer to that question, but it is a hard question that needs a lot more thinking about.

Dr James: When we are thinking about children with special educational needs there is a whole variety of those. We ought to think in terms of special educational needs about those children who have multiple and profound difficulties. Also those children who are probably achieving above the expectation for their year, that poses an issue for the system we have, as well as those with specific learning difficulties. I happen to have a son who is dyslexic and was diagnosed at the age of eight. He has a Master's degree now, but I had to spend an awful lot of time telling him he wasn't stupid.

It worries me that putting all children through a common system may challenge them—I think there is room for challenge, because some children with special educational needs are not challenged sufficiently because they are being labelled as not able. That is a real problem, but on the other hand we have to avoid labelling so that the reverse does not happen, so that they then become passive learners as well. There are real issues with this and in some ways you could say all children have special educational needs, because most of them have strengths and weaknesses. That is why I think we have real problems when we have very narrow definitions in high stakes tests of what is worthwhile learning. If it is focused on one aspect of reading, and I have looked at the test and it is about word recognition and comprehension, is that reading? I am not sure. If it focused on that, if it focused on certain aspects of mathematics, punctuation and spelling, the curriculum gets collapsed to that. Children then label themselves.

I would much prefer a system that is much broader and is much more a profile that you can develop for children and it would then have to depend not just on testing systems but on the strengthening of teachers' own assessment in classrooms. I think it can be done. We go back to the teacher training and the teacher education. It can be done if we put in the professional development for them. There may be objections about workload. In order to make space for that, something else has to go.



Catherine McKinnell: That is very helpful.

Catherine Kirkup: I was just going to say very quickly—and it was picked up in the earlier session—at the moment the floor standards that are there are an attainment threshold and progress measure. If the focus was on progress then all children, including children with special educational needs, are able to make progress.

Catherine McKinnell: Yes, that is a very good point. Thank you.

Professor Wyse: I wholly agree that a high stakes system is particularly problematic for these children we are talking about. In terms of reading, to give a good example of something that is very rigorously researched and shown to be effective, reading recovery is a very well-researched system that has 10 years of practice located in my department, as it happens—the International Literacy Centre was originally funded by Government—so we have strong answers to help with these problems. Of course it is built on a diagnostic system of assessment that is personal to the child and carried out by an expert teacher, but we know from research that is the best way those children are helped.

Q140 **Catherine McKinnell:** I think, Catherine, you may have already answered this question. Being able to differentiate between the attainment levels of a wide range of pupils, do you think that the current assessment system is able to do that? One of the concerns is that it has raised the bar to a significantly high level, so that even the highest attaining pupils are not reaching the highest level and therefore it is not giving sufficient credit or attention to those who are at the lower attainment levels. Comment would be helpful.

Catherine Kirkup: It was a very difficult ask to develop tests that covered such a wide range of ability, yet at the same time there was a change in the model, because if the model is being used as a baseline for later progress, for the Progress 8 then you also have to have a test that has no ceiling effect, because you need to be able to challenge the most able so that their progress can be measured as well.

2016 is the first year. There needs to be a full evaluation of that data. I do not have access to all the technical data but the Standards and Testing Agency have. I think it needs to look at whether there was sufficient differentiation at each sector of the ability range, and if there was to be an external evaluation of that data then we would be very happy to participate.

Q141 **Catherine McKinnell:** Presumably the issue of threshold also applies in the sense that no matter what your attainment level, you are measuring progress rather than pass or fail, effectively, which is the effect for some children.

Catherine Kirkup: You could still look at the attainment of the national sample, but you could make the floor standard basically on progress rather than on both.



Dr James: I think the threshold issue is quite crucial. In the report, and I have looked back at the report of the national curriculum review, we made a fairly subtle distinction. We were reporting on the curriculum, not the assessment system particularly, although we could not help but say something about that because Lord Bew was doing the other at the same time, but we were talking about mastery as a really key concept, that we would want children to master the content of the curriculum and that they should be secure in that before they went on to the next aspect of the curriculum.

It was a kind of input expectation. It was a high expectation of all children. It was not a threshold assessment level and I think we said in that that we could see the reason for scaled scores as an alternative at national level to the levels, but we were not advocating a threshold that, "This is what you should attain". As soon as you do that, as we know, as soon as we say they have to get 100 then that is what teachers will drill to and they become quite adept at doing that kind of thing. You therefore will not get the spread and the push forward for all children.

Q142 **William Wragg:** Just very briefly on that, Mary, if I may, you mentioned there the mastery of things, but I want to talk just briefly about writing, because you mentioned children with dyslexia particularly. Would you say that the move from the best fit model to the secure model was disadvantageous for those children in terms of the assessment of writing?

Dr James: I was an English teacher once upon a time in secondary schools and I have worked with this. It is a really big issue with the writing. There were problems with the best fit, as we know, because first of all you have to determine the criteria and you had a collection gathered together and you want to say, "Well, do children fit this mostly?" but those became generalised criteria. There were extraordinary interpretations of what that meant. For instance, you have to use certain numbers of sentences and children making full stops the size of golf balls to make sure that the assessor knew that they were making a full stop. So there were very strange effects of this and it did not cover the scope, but when you got the secure fit it means there is a kind of one mark target.

I am not sure I am up-to-date entirely on the comparative judgment, but I think that looks more promising as long as you build it into a system where teachers not only look at the work in their classroom and their school, but in other schools as well. There is a huge benefit to do that, and I think you can therefore, by looking at children's writing, say, "What is the quality in this?" and you can determine the criteria from the writing. That is a useful way of doing it, but Dominic is a literacy person so he has probably got—

Professor Wyse: Yes, if I may. Are we coming back to writing?

Chair: Not really, no. That was a supplementary from William put as a specific question to Mary.



Professor Wyse: Could I just say something on writing before we finish?

Chair: Yes, if you are quick.

Professor Wyse: The problem we have, and I am glad Mary mentioned the curriculum, is that we have these pages and pages of grammatical terms that children should learn and therefore tests and teacher assessment seeking to assess those, but there is no rationale in research or even good scholarship that says we should have those appendices. They should have never been there, and in my view they are there for ideological reasons, not for any other.

Professor Coe: This question about should we have outcomes reported as thresholds, pass, fail, school ready, secondary ready, expected standard, however you want to label that, this might be a rare occasion where all the experts, unless anybody else is going to say no, agree that we should not do that. It is damaging for all sorts of reasons. It drives bad behaviour in the system and we got rid of it for secondary, with Progress 8 becoming the main measure. I know we still have A to C or whatever that becomes, but in primary it is still there and really big. We have this expected standard and there is a narrative about the curriculum being demanding and the tests have to be hard because that shows that we have a proper education system, that we can hold our heads up against Singapore or whatever, but that is inconsistent with this view about we want children to make progress and we want their performance to be positive and affirmed, wherever it is on the scale. We just need to reconcile that.

Chair: Ian Austin, we are looking at alternative assessment methods.

Q143 **Ian Austin:** Yes. I would like to ask what changes you think the Government should make to the assessment system. How do you think it could be reformed, what they would look like and how quickly do you think any changes you think should be introduced could be introduced? I do not know who would like to kick off with that.

Dr James: I think you ought to get rid of the performance tables. I think the performance tables are the cause of a lot of the difficulty, if not almost all of it. I do not believe that any other country, to my knowledge, publishes performance tables that create such a high level of anxiety in schools and so much hangs on it: teachers' jobs, head teachers' jobs, the closure of schools. Everything hangs on these performance tables.

Another thing that we have not discussed is we have talked a lot about reliability in the previous session but not about the validity of the tests. Are we basing so much on so little?

Q144 **Ian Austin:** What would you replace performance tables with then? How would you enable parents and everybody else to make judgments about—



Dr James: The parents have access to the curriculum if they wish to, and they have annual reports, probably more frequently than that to their children. The reporting to parents is crucially important. It is the performance that children make on the curriculum, the progress they make on the curriculum that is important. You do not have to make reference to tests there. I do not know if you have talked to Dame Alison Peacock, but in her school her reports to year 6 parents, as well as the others, are just done in narrative form. It is hugely detailed, very rich. I did ask her when I looked at these, "Is no parent going to ask you what level their child has attained in the SATs?" She said, "I wondered if they would do that". Only one had ever done it, and the reason was that the reports were so rich that they knew what their child was able to do or could not do and they did not need a score on a particular national test.

You can argue if you have a good school system and that you are really engaging with your parents and you get them to understand what their children are attempting to do at every stage in their learning, then I think you can have alternatives to the testing system. I do not think parents are particularly interested in the scores of their children at key stage 2 or key stage 1; they are much more interested in what the detail of their learning is and whether they are happy at school as well. Some of the parents are anxious because their children are not happy; some of them go to bed crying before the key stage 2 test.

Ian Austin: What do other people on the panel think about this?

Professor Wyse: I repeat that I think we need to move to a system of sampling in terms of one way of assessing the effectiveness of the system. Also, if what Ofsted did was made entirely separate from the league tables, as was argued earlier— Ofsted reports are of course available publicly, and should be. When done well, they are a terrific source of evidence about how effective a school is in all ways, the teaching, the management and so on. That is sufficient, in my view, for a parent wanting to know how good a school is. Given the problems we have with the accuracy of the test scores anyway, as other people have said, parents should not be making judgments on those bases.

Q145 **Ian Mearns:** Dominic, is that entirely true? If a school was judged good say four years ago, it might now get a one-day drive-by, one inspector, and if on a cursory one day drive-by analysis by one inspector it still appears to be good, it will carry on being good? The validity of those inspections is called into question to a certain extent.

Professor Wyse: I accept the problem with the frequency of inspections and draw a parallel with the health service, where they have national standards monitoring, which is not only UK-based, which I think is annual, but also a European mechanism that is also annual. There are ways we could change the ways we inspect, although I do think inspectors observing teachers actually teaching is still a massively important part of that process and it is time-consuming.



HOUSE OF COMMONS

Dr James: At one time we had local authorities also, and local authorities had their own inspectors and their advisers.

Ian Mearns: I have a vague memory of local authority inspections.

Catherine Kirkup: I think sampling is good for monitoring standards, but it does not give you information about all schools. I would retain the national testing but I think, as others have suggested, there is too much focus on one year's results and I would like a move to rolling averages and trends so that you can look at how a school is performing over time but still look at the overall attainment of all schools.

Professor Coe: Prefacing this to say there are all sorts of problems with interpreting test data, assessment data, judging schools, but this would be going from a frying pan to a fire to say, "Let's put our faith in Ofsted judgments instead". I am not sure that that would be progress.

I am going to interpret your question as being about assessment and accountability together because I have already said I think they are inseparable. I will give you six headlines and I will flesh these out in a written note perhaps, because I know we do not have much time.

Number one is not thresholds: talk about averages as a main outcome. Number two is address the issues about conflicts of interest where teachers' judgments are used to judge teachers; we need to think more carefully about that. Number three is a focus on progress. If you want to judge the impact that schools have fairly, we need to have progress measures. There are two important implications of that. One is that you need some kind of baseline and that is a thorny one. The other is that you need strong caveats around the precision of progress measures, which I think Harvey was hinting at and he is right about that.

Number four is about design, where the design of accountability systems needs to recognise that people will cheat. Let's not design systems that work on paper if nobody cheats; let's design them with that in mind, so that people cannot cheat. Maybe that is an unacceptable message; I don't know.

Number five is about the quality of the assessment, which maybe is the question. There are major things we can do to improve the actual technical quality of the assessment: not too narrow, just focused on little bits of the curriculum, they must be broad; not too predictable, so we know exactly what we have to teach to get kids to do well in the tests, even if there are lots of other things that we ought to be teaching that we know are not going to be tested. We can use more independent sources of information and integrate them better into a judgment about what children have learnt and what progress they have made. Bring it back to validity. We need to see evidence of the validity of assessments, that the kinds of uses and interpretations that we want to make of them can be supported.



HOUSE OF COMMONS

The last is about expertise and capacity, which is training.

Q146 **Chair:** We would like you to flesh that out in a note.

Professor Coe: Yes, I will. It will be interesting.

Chair: Thank you for all those points.

Q147 **Ian Austin:** Do you think the teacher assessments should be included? If so, how could that—

Professor Coe: There is scope for doing that. There was a lot of talk about comparative judgment previously and that is a nice model that does use teacher assessment. Your question was about who should do it. Tim, I don't know if you quite answered it, but the answer is that teachers should do it and teachers can do because those judgments are effectively moderated through the system. The big question about comparative judgment is how that operates in a high stakes environment. We have seen it work in experimental situations and in schools; that was the one Becky was talking about. It is quite an exciting prospect. My worry is that we would introduce it as a solve-all, solve all our problems in a single stroke, and find that some of those same problems are there because it is the high stakes rather than the assessment that drives the problems.

Catherine Kirkup: If I can just add this about comparative judgment, one caveat that has not been mentioned is that it is mainly focused on making a holistic judgment and sometimes that overlooks the detailed information that teachers need to then feed back to learners. Therefore, we have to ask, if teachers are going to be involved in that, is that the best use of their time? I do not know whether you want to have writing as part of a summative assessment, but if you are, perhaps you need to think about this: comparative judgment is a very effective and efficient way, but it is making holistic judgments. Perhaps that is fine for summative assessment, but it is not improving teachers' professional understanding of what makes a good piece of writing.

Dr James: Very quickly, I think the teacher's judgment is very valuable, in fact crucial, for formative purposes, and we are quite able to develop systems for them to do summative assessments of children's outcomes. It is the use that is made of those data that is the crucial thing. If those summative assessments are for the children themselves, for receiving teachers, for parents, I do not think there is a problem because it is focused on the child and their learning and where they are going. It is when you take those data and then feed them in to an assessment system that looks at the school as a whole and judges the quality of the school, that is where you have a problem. It comes back again to the publication of those results as performance tables of the school that is the crucial problem.

Chair: Thank you very much. We are now going to go over to the baseline assessment with Ian Mearns.



Q148 Ian Mearns: In your six-point plan, Rob, you mentioned the thorny issue of baseline assessment and the progress measures. Do you think the Government should continue to develop a viable baseline assessment measure?

Professor Coe: Yes, I do, although there are all sorts of problems with it. One should say that I have a conflict of interest here as one of the providers for the baseline—

Catherine Kirkup: Ditto.

Professor Coe: As does Catherine. However, we have been doing baseline assessment of four year-olds for more than 25 years. Can you do it reliably? Yes, you can. Does it give meaningful information to teachers that is really helpful to plan their teaching? Yes. To parents? Absolutely, yes.

When you introduce high stakes into that and compulsion, the whole thing changes and that is something that we need to think quite carefully about. As a baseline, obviously the pressures are different. The incentive is to want to do badly on a baseline if it is progress you are going to be measured on. There is a whole lot of thinking that needs to be done there, but I think if it is a requirement that we can judge schools—and that seems to be what drives all this. Some people might say, “No, we do not want to do that; let’s not have accountability at all” but my view is we should have that. I think the evidence does support overall small positive effects of accountability.

There is lots of devil in the detail on that, of course. If you want to judge the performance of schools and the effective schools, you must have some kind of baseline measure, so we need to solve that problem. I think we came reasonably close to doing it and then the landscape seemed to change under us, but there is some proof of concept work already in place.

Catherine Kirkup: I agree. I think it is possible to develop a child-appropriate assessment for that age group. It does need to be kept very simple if it is being used as a baseline for accountability. There are lots of things that practitioners need to know about those children, but those do not need necessarily to be part of the baseline. We need to look at what the relationship is between the baseline and later attainment and then measure as little as possible that gives us the most information.

Q149 Ian Mearns: Mary, are you sceptical?

Dr James: I am sceptical, yes. We submitted to the previous consultation a paper from the Assessment Reform Group here. You could understand the reasons for it and the desire to measure progress in school, but I think the difficulties outweigh it. If you are thinking of small schools and that you possibly will not have the same cohort going through that school, you have a limited number of children doing it, so the reliability is going to be problematic. The perverse incentive for



HOUSE OF COMMONS

teachers—which is a very real one—of increasing their value-added, if you like, their value, by depressing the results of the early stages is quite marked because anything that you introduce is liable to have perverse consequences. It is always important to say, “What is the worst that can happen with this?” because to be sure it is going to happen. There are all sorts of reasons why it will not work.

Q150 Ian Mearns: But if you can put accountability to one side, isn’t measuring progress a good diagnostic tool for looking at the needs of the individual child?

Dr James: Absolutely. That is what I think I am saying, certainly. There is nothing wrong with tests if they are good tests, if they are reliable and valid. Reliability, we must remember, is the reliability of the group, not the reliability for an individual child. The chances of that test being reliable for the child, an individual, are limited.

Professor Coe: It is not; it is perfectly reliable. Harvey was wrong about that.

Dr James: I think there are others. This is another debate.

Professor Coe: There is clear evidence. It is black and white. It could not be simpler.

Dr James: It depends on the instrument anyway.

Professor Coe: Yes, it does.

Dr James: I think there are real problems with it.

Chair: Dominic, quickly.

Professor Wyse: There are significant risks with having a baseline that is high stakes, so in my view the attention should be on high-quality teacher assessment and the points made about training teachers are very well made. That is where we should focus. There is clearly also a high political risk, it seems to me, to be seen as a country that formally tests the youngest children in the world, if that indeed is the case. Anyway, timescales for these things are a real problem. It seems to me you need at least a good two years to look seriously at some of these implications that people are talking about before you implemented something, if you were going to.

Q151 Ian Mearns: Does anybody see a scenario where baseline tests for four year-olds are published in terms of a results measure?

Professor Wyse: They are bound to be, aren’t they? If freedom of information requests are made, newspapers will do it anyway.

Q152 Ian Mearns: What would a league table show? A child on entry is on entry; the school would have had no responsibility for that child until the date of entry, so what we are doing there is measuring what parents



HOUSE OF COMMONS

have done in terms of educating their own children prior to them going to school.

Professor Coe: I think the answer, Ian, is that we don't really know. There is good evidence about testing and the benefits of testing for learning—of assessment, if you want to prefer that word—the kinds of information it can give you. There is a strong case for wanting to solve this problem: how can we use assessment well to support learning? We also know there are downsides, the downsides people have talked about, narrowing the curriculum, labelling, stress, for example. The challenge is an optimisation one. It is about trading off; it is about finding ways of doing assessment, designing the assessments themselves, thinking about the kinds of consequences that attach to them, thinking about issues like publication and how it can be done. Those are really complex problems and there is no good research about, "Here is the formula, this is the way to make it work". We need to do more work on that.

Q153 **Ian Mearns:** I know that at least two of the panel are entirely sceptical about this. I am therefore asking Rob and Catherine, are there any particular risks that exist over the introduction of a baseline measure?

Professor Coe: Yes, there are massive risks. It is deeply unpopular with the vast majority of early years teachers, so that is a big problem.

Catherine Kirkup: Again, it is the use to which it is put. Baseline assessments for young children are not new. If we go back to before the introduction of the foundation stage profile, which I was involved in, before then there were lots of different baseline schemes. They were not statutory. There were about 98 different ones—some produced by local authorities, there was the old QCA baseline scheme—and some of them more formal than the ones we are talking about now, that we used to see what children could and could not do on entry. The problem now is that if it is going to be used for accountability, that introduces all these different perceptions of the baseline assessment and it is a big issue that we have to look at. First, I don't think it should even be looked at at individual level. You are looking at a cohort and you are looking at progress so you are only measuring it in order to measure progress later so I do not think they should be published at all.

Q154 **Ian Mearns:** Putting to one side whether it is desirable or not, could I ask you all to pen a few thoughts as to what should be included in a baseline measure, if that is the direction in which we are going?

Dr James: Yes. Can I say one thing though about the baseline? We have to ask the question about what early years education is about and is it just preparation for secondary schools at the age of four? This is where the early years specialists will come down and say, "We are about children's development, socially, physically, as well as cognitively" and so forth. To narrow it down to preparation for spelling, punctuation and grammar is completely distorting, bearing in mind the fact that our



HOUSE OF COMMONS

children go into formal education two and three years earlier than they do in other countries and we do not do noticeably better as a result of that.

Professor Wyse: Can I add that one important aspect of this is the totality of the curriculum? We must assess properly the whole curriculum within the constraints of the time and pragmatics and that is a massive problem for primary as well. This emphasis on maths and English typically, even science has been lost, is damaging and needs separate work anyway.

Chair: Last but not least, Lucy Frazer on design and implementation.

Q155 **Lucy Frazer:** The new assessment system has been labelled as chaotic by many teachers and the Standards and Testing Agency, which has been responsible for the implementation of primary school tests. When there was a review conducted by the Department for Education, even they said that the STA is “broadly fit for purpose” but there are still some issues.

Dr James: Broadly fit for purpose?

Lucy Frazer: Yes. Do you think the STA is the right organisation to oversee the development of the primary school tests?

Professor Coe: It depends a bit on what the alternative is, I think.

Lucy Frazer: Anyone else?

Dr James: I think there is an issue, but I did look at the previous evidence session in December and there were questions about the quality of the tests. Some of the teachers were criticising them for not probably even presenting them to children first to see how they reacted. I don’t know what the foundation for that is.

Catherine Kirkup: I can say there that they are presented to children. There is a very rigorous development process with lots of trialling in schools, so I am not sure where that sort of comment came from.

Lucy Frazer: Do you think it is the right agency, Catherine?

Catherine Kirkup: There has to be a very rigorous process of development. There have been changes over a number of years, not just for the 2016 review. Prior to that there had been changes in the way the tests are constructed. We are one of the test development agencies that work with the Standards and Testing Agency, so I have to put that out there. The ways in which the test development agencies work with the Standards and Testing Agency have changed and the way in which the tests are constructed has changed, so whether it is time to consider whether all the changes have had positive impact, I don’t know. That is for others to say. However, I do know that the tests go through a very rigorous process of development. They go out to schools in two very large trials and they do take input from teachers, markers and expert review groups. I know a lot about how the process works, but whether there should be fine-tuning of that, that is for somebody external to say.



Professor Wyse: From my point of view, the input of expert teachers, academics, people with other expertise early on in the development is the important thing for me, not once a prototype has already been developed, and taken it into schools, "What do you think of this?" That is not so good.

Q156 **Lucy Frazer:** Should Ofqual have a role in it?

Professor Wyse: I don't think I have the expertise to answer that one.

Dr James: It seems to me there is a lot of expertise from outside, not having a vested interest in it and in test development agencies that have been represented to you. They are very sophisticated in that and as long as that kind of level of attention to detail is used within the STA, then I think there is reason for saying, yes, perhaps they should retain that.

I know that certainly when there was discussion over the baseline, something like 16 potential providers came forward with all sorts of different kinds of offers. Whether that was a good thing or not, because it was putting it in a competitive marketplace, I am not too sure, but I do think there are lots of vested interests in the system staying as it is and that is a problem and needs to be looked at.

Chair: Thank you very much, Lucy, and thank you all very much indeed for coming today. Ian asks you, as I asked the previous panel, to say something by letter about baseline, which would be really helpful. I think Rob is going to give us his six-point plan. We have a 12-point plan to consider from yesterday; your six-point plan will be really helpful as well.