The British
Psychological Society
Psychology of Education Section

The Psychology of Education Review

Volume 41 Number 1, Spring 2017

Accountability measures: The factory farm version of education

ISSN: 2396-8788 (Online)

Printed and published by the British Psychological Society.

First published April 2017.

The British Psychological Society
St Andrews House, 48 Princess Road East, Leicester LE1 7DR, UK
Telephone 0116 254 9568 Facsimile 0116 227 1314
Email mail@bps.org.uk Website www.bps.org.uk

Incorporated by Royal Charter Registered Charity No 229642

Editorial

Katy Smart

WELCOME to the 2017 special Open Dialogue edition of *The Psychology of Education Review (PER)*. This year's Open Dialogue has been initiated by Professor Merryn Hutchings and is an extension of her recent report *Exam Factories? The impact of accountability measures on children and young people*. Her thesis is that there are significant negative impacts associated with the use of pupil testing for the purposes of school accountability, in terms of the curriculum and pedagogy within schools as well as to the detriment of the emotional health and wellbeing of the pupils.

Professor Hutchings' paper *Accountability measures: The factory farm version of education* discusses the findings of this report and considers suggestions for changes to the prevailing system. It is then followed by nine invited responses from experts in the field, three providing an international perspective. I would like to thank: Dr Selma Babayiğit, Professors John Hattie and Janet Clinton, Dr Pam Jarvis, Professor Christine Merrell, Professor Dave Putwain, Dr Guy Roberts-Holmes, Wendy Symes, Assistant Professor Nate von der Embse and Associate Professor Christina Wikstrom for their considered responses. The dialogue concludes with the author being given the opportunity to reply to these comments. I am sure you will find this a very interesting and informative debate.

Never having visited Edinburgh, which by all accounts is a beautiful city, I am particularly excited that this year the Psychology of Education Section will be holding our Annual Conference at the Holyrood Hotel in Edinburgh on the 27th and 28th October. The theme of this year's conference is Learning, Teaching and Assessment. The guest speakers include Professor Christine Merrell from Durham University and Professor Lisa Woolfson from the University of Strathclyde. Professor Jo-Anne Baird from the University of Oxford will be delivering this year's Vernon-Wall Lecture on the Friday. We invite submissions for oral and poster presentations as well as workshops and symposia; the deadline for submission is Sunday 2nd July. For further details of the conference programme, submissions and registration visit https://www.kc-jones.co.uk/pes2017. Registered undergraduate and postgraduate students can apply for a bursary of 50 per cent of the conference fee – please refer to the conference website for further details. Prior to the official start of the conference join us on Thursday 26th October in the evening for our 'Psychology in the Pub' event; this is a free event open to both BPS members and non-members.

Our autumn edition of *PER* will feature a selection of papers from presentations at our 2016 conference in Birmingham, in addition to individual contributions that I am sure our members will find of interest. We welcome individual papers on any aspect of psychology and education; if you would like to submit a contribution to an issue of *PER* then please get in touch with me, the Editor Katy Smart, katy.smart@bristol.ac.uk. *PER* publishes a variety of articles including full research papers, book reviews and short reports on research in progress.

And don't forget to keep up-to-date with what is happening in the Psychology of Education Section by following us on Twitter https://twitter.com/BPS_PES.

Katy Smart
Email: katy.smart@bristol.ac.uk

The British Psychological Society
Psychology of Education Section

Annual Conference 2017

27th October – 28th October

Macdonald Holyrood Hotel
Edinburgh

Learning, Teaching and Assessment

Keynote Addresses:

PROFESSOR CHRISTINE MERRELL
Department of Education, Durham University

PROFESSOR LISA WOOLFSON
School of Psychological Sciences and Health, University of Strathclyde

Vernon Wall Lecture
PROFESSOR JO-ANNE BAIRD
Department of Education, Oxford University

Details for Abstract Submission and Registration available at

https://www.kc-jones.co.uk/pes2017

Accountability measures: The factory farm version of education

Merryn Hutchings

Introduction

THE REPORT *Exam Factories? The impact of accountability measures on children and young people* (Hutchings, 2015) is the outcome of an investigation commissioned by the National Union of Teachers. The report draws on previous research findings together with new data from a survey of teachers in England (achieving 7922 responses) and interviews with headteachers, teachers and pupils.[1] It argues that, not only are current accountability measures in England having a profoundly negative impact on teachers and students, but also that they are resulting in an impoverished form of education which does not meet the needs of universities and employers. This paper highlights the main findings of the report, and considers suggestions for change.

Accountability measures

The notion that schools should be accountable has long been a feature of education. In England and Wales, school inspection and national testing to ensure that schools were providing value for money were established in the mid-19th century. Inspectors were appointed soon after the government first provided any funding for schools. In 1839, the Committee of the Council for Education recommended that 'no further grant be made, now or hereafter, for the establishment or support of … schools, unless the right of inspection be retained' (cited in Dunford, 1976, p.4).

However, concerns about the quality of education and effective use of government funds continued, and in 1861, the Newcastle Commission recommended a further measure to ensure accountability. They proposed:

> …to institute a searching examination by competent authority of every child in every school to which grants are to be paid, with the view of ascertaining whether these indispensable elements of knowledge [reading, writing and arithmetic] are thoroughly acquired, and to make the prospects and position of the teacher dependent to a considerable extent on the results of this examination. (Newcastle Commission, cited by Rapple, 1992, p.304)

It was assumed that such testing would raise standards; the report argued, 'If teachers had a motive of this kind to see that all the children in their charge really learned to read, write and cypher thoroughly well, there can be little doubt that they would generally find means to secure that result' (Newcastle Commission, cited by Rapple, 1992, p.304). Following this, annual examination of every pupil was conducted by school inspectors, and the school forfeited a proportion of its grant for each child who failed to reach the required standard.

Despite a catalogue of negative impacts (narrowing of the curriculum, teacher and pupil anxiety, teaching to the test, learning by rote at the expense of understanding, various forms of cheating, decline in number

[1] I acknowledge with gratitude the role of Dr Naveed Kazmi in conducting some of these interviews.

and quality of teachers), payment by results continued for 35 years before it was finally abandoned (Rapple 1992). National testing ceased, and inspectors took on a more advisory role.

But in the last 30 or so years, accountability has assumed a new centrality in the policies of governments across the world (O'Neill, 2002). In England, the drive to increase accountability in education has been consistent over that period, regardless of which political party has been in government. The Education White Paper *Choice and Diversity* (DfE, 1992, p.2) identified greater accountability as one of 'five great themes [that] run through the story of educational change in England and Wales since 1979'.[2] A number of measures were designed to achieve greater accountability: a national curriculum which set out standards to be achieved; national testing and publication of school league tables; a considerable increase in the number of school inspections, which were designed to judge schools rather than to advise; and greater school choice for parents.

At that time, a market-driven approach to accountability (Leithwood, 2001) was dominant in England and Wales. It was expected that if parents were offered a choice of schools, and were provided with clear information through inspection reports and school league tables, they would send their children to more successful schools, while those that were less successful would fail to attract pupils and would therefore close. However, this assumes that parents choose rationally on the basis of information provided. But research has shown that only about half of those choosing schools consult inspection reports and league tables, or regard them as key factors in their choices. They are often more concerned with location, and their impressions of whether the school would suit their child (Francis & Hutchings, 2013; Wespieser et al., 2015). Moreover, the notion of letting some schools fail and close, while others expand, is simply not practical.

Consequently, while the rhetoric of choice and markets remains, the current form of accountability is largely management accountability (Leithwood, 2001). This involves inspection; student testing; targets for schools, teachers and students; and performance-related pay for teachers; together with sanctions for those that 'fail'. Of these, testing has the most profound impacts on the experiences of pupils. However, in the *Exam Factories (EF)* research, headteachers reported that their main concern was inspection. A key factor in this concerned the role of pupil attainment and progress data. Headteachers argued that inspectors arrive in school having formed a judgement on the basis of data, and that other things in the school are overlooked. In this way, the inspection is not distinct from the testing regime; rather, Ofsted, the inspection agency, issues the judgement and imposes sanctions, which could involve change of governance, job loss for the headteacher and possibly other teachers, loss of performance-related pay, and a degree of public humiliation.

The national tests carried out in primary schools aim 'to provide standard information to parents and to give a picture of school performance' (DfE, 2014, p.6). Seven-year-olds and 11-year-olds take tests in maths and English, and a phonics check for six-year-olds was introduced in 2012. In secondary schools 16-year-olds take GCSE examinations which act both as qualifications for the individual and as the key measure used for accountability. When testing was first introduced, the main measure was the percentage of pupils achieving the expected standard. Now pupils' progress since the previous tests is also considered, and in secondary schools this has become the main measure for accountability purposes (DfE, 2013a).

Impacts of accountability measures

The impacts of accountability measures on

[2] The others were quality, diversity, increasing parental choice, and greater autonomy for schools.

children and young people are discussed in the following sections: the curriculum; pedagogy; emotional wellbeing and mental health; and groups that are most affected. Accountability measures also impact on teachers. This is not the specific focus of this paper, but it is important to recognise that accountability measures have increased teachers' stress levels and workload, and that these in turn contribute to the overall impact on pupils.

The curriculum

When the stakes attached to test results are high, teachers focus their teaching narrowly on what is being tested; Berliner (2011, p.299) argues that this is a 'rational response'. Koretz (2015) distinguishes three strategies: reallocation of time and resources between subjects (e.g. spending more time on the subjects tested); reallocation within subjects (e.g. focusing on specific aspects of that are tested at the expense of those that are not); and coaching, which he uses to mean focusing on the format of the test (the way questions are asked, instructions to candidates, etc.).

All of these strategies were reported in the *EF* research. While teachers across all age groups reported curriculum narrowing, it was noted that this was most extreme in year groups doing tests or exams. One teacher noted that 'in Year 6 [11-year-olds]…the curriculum's narrowed to reading, writing and maths because that's what we're held accountable for and we've got to get those children to a certain level.' Teachers also noted that low-achieving and disadvantaged[3] pupils experienced a particularly narrow curriculum because they are often removed from other lessons to do extra maths and English. While middle-class families may be able to compensate for the limited school curriculum, this is not generally the case for those from disadvantaged backgrounds, who are less likely to have access to wider learning and cultural opportunities outside school (Berliner, 2011, Francis & Hutchings, 2013; Neelands et al., 2015). In this way, curriculum narrowing exacerbates the gap between disadvantaged and more affluent pupils.

Reallocation of time between subjects has clearly been a government aim; successive governments have asserted the importance of maths and English (e.g. DES, 1988; DfES, 2005; DfE, 2012b, 2013a), and in the last five years have emphasised academic subjects in secondary schools (DfE, 2012a, 2013a). While there is agreement that that literacy and numeracy are key skills for life, the increased emphasis on maths and English inevitably means that other aspects of the curriculum are allocated less time. This was a concern for teachers in the *EF* research: a secondary teacher argued that it was wrong to reduce the time allocated to drama because it is 'a subject which is invaluable in gaining life skills, (teamwork/cooperation, presentation, speaking and listening)…and which really helps build confidence and self-esteem.' Neelands et al. (2015) similarly argue that creative and cultural education is essential 'for the creativity and wellbeing of the population' (p.42) and for the cultural and creative industries. They argue that is crucial that such opportunities are not limited to those who are socially advantaged. Similarly, the Confederation of British Industry (CBI, 2015) deplored the reduction in time dedicated to science teaching in primary schools, and in particular, the limited opportunities in primary and secondary schools to carry out practical experiments. They conclude that too few young people are studying science and technology subjects at school and beyond, and that shortages of scientists, engineers and technologists will limit Britain's effectiveness in certain industries and in an increasingly competitive global market.

While some of the pupils interviewed in the *EF* research accepted the dominance of English and maths, others argued that what they learned in these subjects would not all be of use in their future lives. Some

[3] The DfE definition of pupil from disadvantaged backgrounds includes those who have been eligible for Free School meals at any time in the last six years, and looked after children.

argued strongly that they should be learning more things that were practically useful, and primary pupils argued for more science.

Reallocation of time within subjects has also been a government aim. The Coalition government introduced a phonics test for all six-year-olds in 2012, and a spelling, punctuation and grammar test for 11-year-olds in 2013. While there is agreement that it is crucial that children should learn to read and write effectively, educationists have expressed serious concerns about teaching and testing these elements in abstraction. Dombey (2010) reports that research does not support the value of focusing so strongly on synthetic phonics, but rather shows that the most effective strategy for teaching reading is a balanced approach with a strong message that literacy is about communication. Similarly, the UK Literacy Association (UKLA, 2013, p.2) asserts that 'grammatical knowledge should be neither taught nor tested outside the context of purposeful writing.' Teachers in the *EF* research argued that the use of meaningless syllables in the phonics test contradicts the key notion that reading is about meaning, and phonics has taken time away from other aspects of the literacy curriculum such as reading or listening to stories; less than half the early years teachers said this regularly happened.

The *EF* research showed that many teachers focused on the format on the test or exam in their efforts to enable children to pass. One wrote: 'In year 6 from Christmas onwards, we will be training them to pass SATs tests – with test after test after test. No fun at all for the children.'

One consequence of the strong curriculum focus on tests and exams was that many children (particularly in secondary schools) only considered it to be worth learning things that are tested. Moreover, all the pupils interviewed (including those in primary schools) asserted that their test or exam results would influence and potentially limit their future options. While this is probably true for 16-year-olds, the tests taken in primary school have no ongoing impact on the child; their sole purpose is accountability. Even secondary schools generally ignore the results of national tests taken in primary schools because they know that children are coached to succeed.

Pedagogy

There is extensive international evidence, reviewed by Lobascher (2011), that high stakes testing and accountability measures discourage creative teaching. This happens for several reasons.

First, when pupils have to take high-stakes tests, there is a strong motivation for teachers to do everything possible to enable them to pass. In Victorian England, Rapple (1992) shows that this led to a focus on memorisation rather than understanding: many children were taught to recite the reading book by heart. Berliner (2011, p.299) argues that testing has a similar impact in modern US classrooms: 'many of the instructional activities in the curriculum areas tested are of a low level in terms of the cognitive processes that are called for by students. Drill in memorisation and practice of algorithms predominates.' In the *EF* research, one teacher argued 'national expectations cannot be met without lots of repetition and rote learning.'

This of course reflects both the nature of the tests (done individually, and often involving multiple choice or very brief written answers) and of the knowledge-based curriculum. Nicky Morgan (2015a), former Secretary of State for Education, argued explicitly that children should acquire knowledge rather than learning critical thinking, because critical thinking, in her view, is not possible without knowing the facts. The counter-argument is that we do not know what facts will be important in the future so children need to learn how to access factual knowledge, and to evaluate and use it in a range of contexts. Moreover, universities are looking for young people who can engage in critical thinking. A Vice-Chancellor argued:

> The problem we have with A-levels is that students … tend not to take

> a critical stance; and they tend not to take responsibility for their own learning. But the crucial point is the independent thinking. It is common in our institution that students go to the lecture tutor and say, 'What is the right answer?' (Smith, quoted in Children, Schools and Families Committee, 2008, para 129)

A second factor discouraging creative teaching is that current accountability structures in England involve age-based expectations of what pupils should achieve. Thus teachers feel that they must cover the entire syllabus for that age group regardless of whether or not it is appropriate in terms of each child's previous learning and development, and feel unable to allow time for reflection and consolidation of learning. This pressure affects all age groups.

It was a particular concern for early years teachers, who described having to make children sit down and tackle academic work in a way that the teachers considered to be inappropriate to their level of emotional maturity. This was leading to 'silly' behaviour and lack of motivation, particularly among summer-born boys. A teacher of five-year-olds reported that they were set maths and spelling homework, but added, 'Some year 1 children are not ready for a formal style of learning, but teachers feel under pressure to make progress despite knowing that socially and physically the children need more time to learn through play.' Guided play has been shown to be the most effective strategy for supporting young children's learning (e.g. Weisberg et al., 2013). Roberts-Holmes (2015, p.303) draws on evidence to show that early years teachers' pedagogy 'has increasingly narrowed to ensuring that children succeed within specific testing regimes which interpret literacy and numeracy in very particular ways.'

A further pressure is exerted by perceptions of what inspectors require, based on previous experience or report from colleagues. In the *EF* research, many teachers reported that their schools required standardisation in lesson structure. For example a secondary teacher said that her school had 'specific start of lesson procedures' and a 'policy of 10-minute silent working periods during every lesson'. A primary teacher reported that use of PowerPoint was 'mandatory' in her school and noted, 'the PowerPoints will be uniform for each class. Only adaptations allowed would be adjusting certain frames to suit the lesson.' Such requirements, teachers argued, made their lessons 'dull' or 'boring', and limited the possibilities of spontaneity. A secondary teacher claimed:

> You are not allowed to teach any more. You have to deliver a subject in a generic way, just the same way as every other teacher…. This does not allow for any creativity or originality that pupils thrive on. As a result pupils are bored. They know the format of the lesson before you start and rather than see this routine as helpful and logical they see it as dull and boring.

Many teachers said their schools emphasised the importance of having written evidence of learning that can be shown to inspectors; this militated against learning through activity and practical experiments. Requirements for uniformity were reported by a significantly higher percentage of those working in vulnerable and challenging schools (those with low attainment or negative Ofsted judgements or with a higher number of disadvantaged pupils). Teachers in a school judged by inspectors to require improvement commented that they had previously prided themselves on their imaginative and creative lessons, but that in preparation for their next inspection they had moved to more uniform (and dull) lesson structures.

Thus, as a result of accountability pressures, teachers at all levels were teaching in ways that they felt were not in the best interests of the children they taught, or in the ways that substantial amounts of research have shown are most effective.

Pupils in the *EF* case study schools said the most memorable lessons were those that were 'different'; they talked positively about sessions where they made models, engaged in role play, and so on.

A linked finding was that many teachers said they did not know their pupils as well as they had in the past. This resulted from their workload, pressure to cover the curriculum and standardised lesson structures. They had little time to talk with pupils, and said that they tended to think of them in terms of categories (special needs, disadvantaged, gifted, English as an additional language etc.) rather than as individuals.

Children's emotional health and wellbeing

Perhaps the most obvious impact of the pressure has been in emotional responses. The World Health Organisation (2012) found that 11-year-old and 16-year-old pupils in England feel more pressured by their school work than is the case in the vast majority of other European countries. ChildLine (2014, 2015), the counselling service for children and young people, reports that school and exam pressures are one of the biggest causes of feelings of stress and anxiety among children and young people. Teachers in the *EF* research said anxiety about tests affected a wide range of pupils, including those who are high-attaining and conscientious (particularly girls), as well as those with low attainment or special needs. Primary teachers talked about pupils' anxiety about national tests; one said, 'You just see them sat there, a 10- or 11-year-old kid in complete meltdown.'

A further concern is the impact of failure. There are two issues here: the impact of failure on motivation, and the impact of overall failure in schooling. Accountability measures are designed so that substantial numbers of pupils fail to reach the expected level; this is the only way they can differentiate between schools. In 2016, one in five six-year-olds failed to reach the expected standard in the phonics check, and almost half of 11-year-olds failed to meet the expected standard in their national tests. Thus large numbers experience failure. The government view has been that children need to know when they are failing:

> It is a misguided notion on the part of some educational theorists that if work is graded some children and their parents will think of themselves as failures. Pupils need to be told when they are doing badly just as much as they need to be told when they are doing well. (DfE, 1992 para 1.41)

Research suggests that responses to failure vary with the stakes attached to the test. While many do at times encounter failure in low stakes assessments, and cope with it, both Amrein and Berliner (2003) and Mac Iver (2016) have both showed that high-stakes failure often leads to discouragement and a downward academic spiral. Harlen and Deakin Crick (2002, p.5), in a systematic review of research, reported that it is particularly low achievers who become 'demotivated by constant evidence of their low achievement.' This then further increases the gap between low- and high-achieving students. Pupils interviewed in the *EF* research talked about the negative impact of poor marks; an 11-year-old explained that poor marks 'make people that aren't as good and don't have enough confidence in themselves, have, like, less confidence,' and a secondary special needs coordinator said that there had been an increase in the number of pupils entering secondary school with poor self-esteem who did not believe they could succeed in academic work.

The logic of accountability is that some pupils will fail at the end of their school careers. In 2015, 30 per cent of 19-year-olds had not achieved Level 2 (GCSE) qualifications in English and maths (DfE, 2016a), and this number is likely to increase as GCSEs become more rigorous and challenging from 2017. Accountability measures have defined a single academic path which all pupils are expected to follow and have made maths and English central to this; they are therefore

qualifications which many employers require (Wolf, 2011). Therefore, even if young people have other skills and attributes that could be useful in a job, they are likely to be rejected. The impact on this group should be a key concern for policy makers.

The large increase in numbers of children and young people with mental health problems is well documented, and it is clear that there are many reasons for this (including for example, bullying and social media) but there is now evidence that the intense pressure related to school work is one of the causes (e.g. ChildLine, 2014; Sharp, 2013; *The Times,* 2015). In the *EF* research, teachers reported tests or exams had been the immediate trigger for mental health problems such as self-harming or anorexia.

Recent governments have been aware of the extent of mental health problems in school. They have made a number of suggestions for improving this (including character education to develop resilience, and more competitive sport). But their main proposal is that schools should provide better mental health support, and they have allocated funding for a pilot scheme (DfE, 2015). However, a more appropriate response would be to tackle underlying causes of the problem. As one special needs coordinator in the *EF* research argued, 'You can't be counselling them for what you are putting them through at school.' Public figures have made this argument. Natasha Devon, who was appointed to be the DfE's mental health champion in 2015, argued that the pressure to achieve that children and young people are under at school is a key cause of mental health problems. Her position was subsequently discontinued, as a result of reorganisation (Aitkenhead, 2016). Similarly Tanya Byron, clinical psychologist, has argued that the target-driven school system is a factor in young people's mental health problems (Scott, 2015). There is no evidence that either woman's analysis has been accepted.

The DfE's response to reports of anxiety, stress and mental health problems related to testing is to state that tests should not be stressful and that good schools manage them appropriately (BBC, 2016), thus blaming teachers for what government policy has brought about.

Which children and young people are most affected?

The previous sections have shown that all pupils are affected by accountability measures to some extent. Overall, in the *EF* research, the teachers who reported most pressure on pupils, and uniform (boring) lessons, came from schools that had low attainment and poor inspection grades (characteristics which are closely related because of inspectors' focus on data). There is a strong correlation between these two factors and proportion of disadvantaged pupils in the school. Thus disadvantaged pupils are the most likely to be in schools where accountability measures are affecting the curriculum and pedagogy, and teachers are stressed.

A second group for concern are those with special educational needs or disabilities (SEND), whether they attended mainstream or special school. Teachers in special schools reported that they are required to teach maths and English to all their pupils, rather than being able to focus on the skills that they need (which may be, for example, life skills to enable them to live independently).

A major concern is that both disadvantaged children and those with SEND are now being perceived by some schools as a liability that may jeopardise the school's ability to achieve high test results (Galton & MacBeath, 2015). A headteacher recently spoke about the covert selection strategies that some heads use to ensure an intake of high attaining, or affluent, pupils, and thus avoid the potential negative impact on attainment of disadvantaged pupils (Garner, 2015).

Do current accountability measures achieve government aims

Successive governments have claimed that accountability measures serve various purposes (see, for example, DfE, 2010, 2014, 2016b; DfEE, 1997):

1. They provide an accurate picture of school standards to inform parents, governors and the community.
2. They inform teachers about how well their students are doing, and thus inform future teaching.
3. They enable government to identify 'failing' schools and to take measures to change them.
4. They contribute to closing attainment gaps between disadvantaged pupils and their peers, and thus reduce inequality in society.
5. They will help raise school standards, and thus will ensure that the UK can compete in the global economy.

The question is no longer whether school standards are higher than they were in the past, but whether they are higher than those of international competitors (DfE, 2010). Thus the PISA international tests have assumed an increased importance.

It is important to consider, then, whether each of these claims is valid. This depends partly on what is understood by 'school standards'. Recent governments use this term to mean test/exam scores. But, as earlier sections have shown, test scores only represent performance under specific conditions in relation to some aspects of the content of particular subjects. They cannot represent overall learning and skills. Moreover, scores are limited by the quality of the test itself. For example, it was argued that the 2016 primary reading test materials discriminated in favour of middle-class children (Bloom, 2016).

But even when test scores rise, does this mean standards have risen? Koretz (2008, 2015) demonstrates that higher scores do not necessarily indicate that children have greater understanding or knowledge, but simply that they have been more effectively prepared for that particular test. Some American researchers (e.g. Carnoy & Loeb, 2002; Hanushek & Raymond, 2005) have reported that accountability measures do affect standards beyond those in the specific test used for accountability; in states with 'strong' accountability (in the form of sanctions) scores on another test also rose. But other research, using a wider range of alternative tests, found that results on these did not improve (Amrein & Berliner, 2002). Presumably one factor in this must be the degree to which the alternative tests cover identical topics to the accountability test, and use the question formats that children have been coached in. In this country, the number of secondary pupils achieving the expected level in GCSE exams increased by 14 percentage points between 2006 and 2012 (DfE, 2013b), but in the same time period scores in the PISA international tests (taken in a sample of schools) did not increase (Wheater et al., 2014). One explanation of this difference could be that teachers prepare their students to take GCSEs, which are high-stakes, but do not undertake specific preparation for PISA, which has no such impacts. In this light, it will be interesting to see the PISA results to be published in December 2016. However, these are complex arguments and it is not possible to go into them fully here.

The second claim above – that tests are essential for teachers to plan effectively – does not stand examination. Assessment is certainly essential, but does not have to involve national tests (Harlen, 2014; Wiliam, 2010). Any good teacher assesses children all the time in both formal and informal ways, and uses their assessments to inform their teaching. Moreover, teachers recognise that the results of national tests reflect intensive test preparation, and so pay little attention to the results.

The third claim relates to identification of 'failing' schools and taking measures to change them. This argument is clearly flawed in that tests do not accurately represent children's learning. Furthermore, a major problem with this argument is that research has shown that schools are responsible for only a small proportion of children's learning. Gorard (2010, p.59) reported that 'to a very large extent schools simply reflect the local population of their intakes.' Wiliam

(2010) cited OECD analysis which showed that in the US, only eight per cent of the variability in maths scores related to the quality of education provided by the school, and his analysis of data in England showed that the school contributes only seven per cent of the variability between pupils.

Currently schools at the bottom of the league tables, and with poor Ofsted grades, tend to be those in disadvantaged areas. Recognition of this has resulted in a move from using attainment figures for accountability purposes to focusing on pupil progress. However, progress is also strongly affected by factors outside the school (Allen, 2015; McEachin & Atteberry, 2016). Rasbash et al. (2010) examined students' progress and concluded that only 20 per cent of the differences between individuals related to their experience in school; the rest related to home background, the neighbourhood and so on. It is clearly not acceptable to assume that disadvantaged children will have poor attainment and progress – but is also not acceptable to punish teachers who choose to work in areas of deprivation. While governments have claimed that accountability measures can 'close the gap' between disadvantaged children and their peers, a range of evidence in this paper suggests this is not the case, and even that accountability measures can have the opposite effect. The way forward here is surely to tackle wider economic inequality in society, rather than to blame schools and teachers.

The final purpose of accountability measures listed above is to raise school standards so that the UK can compete in the global economy. The arguments above suggest that testing does not necessarily raise standards (depending on what is meant by standards). But if the UK is to compete in the global economy, young people presumably need to have the characteristics employers are looking for. But this is not currently the case. The CBI argued that 'the current exam system risks turning schools into exam factories that are churning out people who are not sufficiently prepared for life outside the school gates.' They warned that businesses were 'concerned that the examination system in place in recent years has placed young people on a continuous treadmill of assessment,' and that while young people were academically stretched, they failed to show the 'series of attitudes and behaviours that are vital for success – including determination, optimism and emotional success' (quoted by Garner, 2014). In relation to this last point, the damage to children's self-esteem and mental health reported in the *EF* research must inevitably impact on their ability to cope as adults and contribute to the economy. The Institute of Directors (2016, p.14) similarly refers to schools as 'exam factories' which are 'squeezing out creativity and the joy of learning.' They argue that 'an over emphasis on testing comes at the expense of teaching children to employ the creativity and entrepreneurial talents they will need to insulate them against the unpredictability of the future.' Despite these arguments, as mentioned earlier, employers do tend to use English and maths GCSEs as signifiers of general ability – though Neelands et al. (2015, p.45) cite research showing 'a negative relationship between high test scores in reading, maths and science and the development of entrepreneurial mindsets and skills.'

Even Nicky Morgan, the former Education Secretary, was aware of these problems. She said:

> … we run the risk of creating a generation who excel at passing exams, writing essays, absorbing information, but children without the skills they need to tackle the challenges that lie ahead and participate in society as active citizens, to make the right decisions and build their own moral framework. (2015b)

Nevertheless, she still argued that national testing was the essential. Yet the evidence above suggest that current accountability measures are not an effective strategy to achieve government aims.

What are the alternatives?

It is generally agreed that it is unrealistic to hope for a retreat from use of accountability measures. But there are possibilities for changing them. Some American psychologists whose area of expertise is in testing have suggested that in the light of the use of tests for accountability purposes, new approaches are needed to test design and validation. Haertel (2013) argues that, to maximise the benefits of testing and minimise negative outcomes, test validation should attend to both intended and unintended testing consequences, and to indirect and direct test effects. Similarly Koretz (2015, p.1) argued that 'the current uses of tests for accountability require major changes to several aspects of educational measurement,' and in particular, that test preparation by teachers has implications for validity. However, neither researcher goes so far as to argue that testing all children for accountability purposes will *inevitably* have negative impacts, regardless of the test used.

Other suggestions are more radical. Harlen (2014, p.35) argues that in primary schools 'monitoring national standards should be based on sample surveys using a large number of items within a rolling programme which extends beyond the core subjects.' She describes the way this is already done in some countries. Because only a sample of pupils take the tests, results are not reported at school or pupil level, though comparisons can be made between groups (by gender, disadvantage, etc.) and strengths and weaknesses across the curriculum can be identified. Some systems use a range of assessment methods including, for example, group tasks, which could encourage schools to focus on collaboration as well as individual work. Monitoring of this type would avoid many of the problems in current accountability systems, but would not meet the government aim of identifying 'failing' schools. While this strategy could work at primary level, the issues in secondary schools are different because GCSEs are used both as individual qualifications and for accountability.

Harlen (2014) also proposes a greater role for teacher assessment across all subject areas. This would be used to advance pupils' learning. She suggests teachers should also undertake summative assessment, which should be moderated, and reported to parents as part of accountability to them. However, it is not clear whether she then envisages the information being collected nationally, a strategy that would have obvious dangers of a return to the current system.

Exam factories?

I did not aim to write a report with the title *Exam Factories* but I was quite taken aback by the number of teachers who used this and similar metaphors to describe their own schools. A primary teacher said:

> Everything is about test results; if it isn't relevant to a test then it is not seen as a priority. This puts too much pressure on pupils, puts too much emphasis on academic subjects and creates a dull, repetitive curriculum that has no creativity. It is like a factory production line chugging out identical little robots with no imagination, already labelled as failures if they haven't achieved the right level on a test. (Primary teacher)

The factory metaphor encapsulates the intense pressure that schools and teachers are under to deliver exam results, a loss of creativity, an emphasis on uniformity (all pupils should cover the *same* curriculum at the same rate and pass the *same* exams), a decline in the quality of personal relationships, and a target driven management style.

A similar metaphor was used by Edward Holmes, writing about payment by results – he said teachers were 'treated as machines' (1911, p.111). And in the US in 1888, Emerson White discussed 'the propriety of making the results of examinations the basis

for…determining the comparative standing or success of schools.' His conclusions are still relevant:

> They have perverted the best efforts of teachers, and narrowed and grooved their instruction; they have occasioned and made well-nigh imperative the use of mechanical and rote methods of teaching; they have occasioned cramming and the most vicious habits of study; they have caused much of the overpressure charged upon the schools, some of which is real; they have tempted both teachers and pupils to dishonesty; and, last but not least, they have permitted a mechanical method of school supervision. (White 1886, pp.199–200)

This paper has shown that there is substantial evidence that accountability measures have a great many negative impacts on teachers, children and young people, and on the quality of education. And there is evidence that the same negative impacts have been experienced whenever high stakes testing has been imposed – both in the 19th century and in other countries today. It has not been possible in this paper to do justice to the range of evidence available; I have simply aimed to highlight some of the concerns that have been raised by my research and reading of the literature, much of which focuses on accountability structures in the US. It would be useful to have more research in this country to inform a comprehensive review of accountability strategies with a view to radically changing them.

The author

Merryn Hutchings
Emeritus Professor of Education,
Institute for Policy Studies in Education,
London Metropolitan University,
166–220 Holloway Road,
London N7 8DB.
Email: m.hutchings@londonmet.ac.uk

References

Aitkenhead, D. (2016). Sacked children's mental health tsar Natasha Devon: 'I was proper angry'. *The Guardian,* 13 May.

Allen, R. (2015). We cannot compare the effectiveness of schools with different types of intakes. Education datalab. http://educationdatalab.org.uk

Amrein, A. & Berliner, D. (2002). High-stakes testing, uncertainty, and student learning, *Education Policy Analysis Archives, 10*(18).

Amrein, A. & Berliner, D. (2003). The effects of high-stakes testing on student motivation and learning. *Educational Leadership,* February 2003.

BBC (2016). Primary pupils 'feel test pressure' – survey, 9 May 2016. http://www.bbc.co.uk/news/education-36229995

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education 41*(3) 287–302.

Bloom, A. (2016). 'Sats reading tests too middle class, and would have no relevance to inner-city children,' teachers say. *TES,* 10 May.

Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

CBI (2015). *Tomorrow's world: Inspiring primary scientists.* CBI.

ChildLine (2014). *Can I tell you something? ChildLine Review* 2012–2013. http://www.nspcc.org.uk/globalassets/documents/research-reports/childline-review-2012-2013.pdf

ChildLine (2015). *Under pressure ChildLine annual review 2013–2014.* http://www.nspcc.org.uk/globalassets/documents/annual-reports/childline-review-under-pressure.pdf

Children, Schools and Families Committee (2008). *Testing and Assessment, HC 169-I.*

DES (1988) *Education Reform Act.* London: HMSO.

DfE (1992). *Choice and Diversity: A new framework for schools.* London: HMSO.

DfE (2010). *The importance of teaching: The schools White Paper 2010.* Cm 7980. London: The Stationery Office.

DfE (2012a, 20 November). Only highest-quality vocational qualifications to count post-16 [Press release].

DfE (2012b, 11 June). New primary curriculum to bring higher standards in English, maths and science [Press release].

DfE (2013a). Reforming the accountability system for secondary schools: Government response to the February to May 2013 consultation on secondary school accountability

DfE (2013b). GCSE and equivalent results 2011 to 2012, revised. https://www.gov.uk/government/statistics/

DfE (2014). *Reforming assessment and accountability for primary schools.* Government response to consultation on primary school assessment and accountability. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297595/Primary_Accountability_and_Assessment_Consultation_Response.pdf

DfE (2015, 3 December). Hundreds of schools benefit from £3m mental health investment [Press release].

DfE (2016a). Level 2 and 3 attainment in England: Attainment by age 19 in 2015. https://www.gov.uk/government/statistics/level-2-and-3-attainment-by-young-people-aged-19-in-2015

DfE (2016b). *Educational excellence everywhere.* White Paper Cm 9230. www.gov.uk/government/publications

DfEE (1997), *White Paper: Excellence in Schools.* London: HMSO.

DfES (2005). *14–19 Education and Skills.* White Paper Cm6476. London: HMSO.

Dombey, H. (2010). *Teaching Reading: What the evidence says.* Leicester: UKLA.

Dunford, J.E. (1976). *Her Majesty's inspectorate of schools in England and Wales 1860–1870,* Durham theses, Durham University. Available from: Durham E-Theses Online: http://etheses.dur.ac.uk/9794/

Francis, B. & Hutchings, M. (2013). *Parent Power: Using money and information to boost children's chances of educational success.* London: Sutton Trust.

Galton, M. & McBeath, J. (2015). *Inclusion: Statements of intent.* Cambridge University and NUT.

Garner, R. (2014). Schools are becoming 'exam factories' which don't equip students for the world of work, claims CBI. *The Independent,* 17 January.

Garner, R. (2015). Leading headteacher exposes 'underhand' tactics used by schools to get round curbs on selection, *The Independent,* 24 March.

Gorard, S. (2010). Education can compensate for society – a bit. *British Journal of Educational Studies, 58*(1), 47–65.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement, 11,* 1–18.

Hanushek, E. & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297–327.

Harlen, W. (2014). *Assessment, standards, and quality of learning in primary education.* Cambridge Primary Review Trust.

Harlen, W. & Deakin Crick, R. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning,* EPPI-Centre Review. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Holmes, E. (1911). *What is and what might be.* London: Constable.

Hutchings, M. (2015). *Exam factories: The impact of accountability measures on children and young people.* London: NUT.

Institute of Directors (2016). *Lifelong learning: Reforming education for an age of technical and demographic change.* IoD Policy Report.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Koretz, D. (2015). Adapting educational measurement to the demands of test-based accountability, *Measurement, 13,* 1–25.

Leithwood, K. (2001). School leadership and educational accountability. *International Journal of Educational Leadership, 3*(4), 217–237.

Lobascher, S. (2011). What are the potential impacts of high-stakes testing on literacy education in Australia? *Literacy Learning: The Middle Years, 19*(2). 9–19.

Mac Iver, M. (2016). When minimum grading policies backfire: who decides whether to let students fail? In M.A. Gottfried & G.Q. Conchas (Eds.) *When school policies backfire. How well-intentioned measures can harm our most vulnerable students.* Cambridge Harvard Education Press.

McEachin, A. & Atteberry, A. (2016). Why summer learning loss affects student test scores. In M.A. Gottfried & G.Q. Conchas (Eds.) *When school policies backfire. How well-intentioned measures can harm our most vulnerable students.* Cambridge Harvard Education Press.

Morgan, N. (2015a). Why knowledge matters. Speech at the Carlton Club, 27 January 2015. https://www.gov.uk/government/speeches/nicky-morgan-why-knowledge-matters

Morgan, N. (2015b). The future of education in England. Speech at the *Sunday Times* Festival of Education, Wellington College, Berkshire,18 June 2015. https://www.gov.uk/government/speeches/nicky-morgan-discusses-the-future-of-education-in-england

Neelands, J., Belfiore, E., Firth, C., Hart, N., Perrin, L., Brock S. Holdaway, D. & Woddis, J. (2015). *Enriching Britain: culture, creativity and growth.* Coventry: The University of Warwick.

O'Neill, O. (2002). *A question of trust.* The BBC Reith lectures 2002. Cambridge: Cambridge University Press.

Rapple, B. (1992). A Victorian experiment in economic efficiency in education. *Economics of Education Review, 11*(4), 301–316.

Rasbash, J., Leckie, G., Pillinger, R. & Jenkins, J. (2010). Children's educational progress: Partitioning family, school and area effects. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 173*(3), 657–682.

Roberts-Holmes (2015). The 'datafication' of early ears pedagogy: 'If the teaching is good the data should be good and if there's bad teaching, there is bad data'. *Journal of Education Policy, 30*(3), 302–315.

Scott, S. (2015). Anxiety is on the rise. *Schools Week,* 21 December.

Sharp, A. (2013). Exam culture and suicidal behaviour among young people. *Education and Health, 31*(1), 7–11.

The Times (2015). True scale of child mental health crisis uncovered, *The Times,* 12 March 2015, p.1, p.8.

UKLA (2013). UKLA statement on teaching grammar. https://ukla.org/downloads/UKLA_statement_on_teaching_grammar_rev.pdf

Weisberg, D., Hirsh-Pasek, K. & Golinkoff, R. (2013). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education, 7,* 104–112.

Wespieser, K., Durbin, B. & Sims, D. (2015). *School choice: The parent view.* Slough: NFER.

Wheater, R., Ager, R., Burge, B. & Sizmur, J. (2014). Achievement of 15-year-olds in England: PISA 2012 National Report (OECD Programme for International Student Assessment) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/299658/programme-for-international-student-assessment-pisa-2012-national-report-for-england.pdf

White, E. (1886). *The elements of pedagogy.* New York: American Book Company. https://archive.org/stream/elementsofpedago01whit#page/n3/mode/2up

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist, 45*(2), 107–122.

Wolf, A. (2011). *Review of vocational education.* The Wolf Report. DfE.

World Health Organisation (2012). Social determinants of health and well-being among young people: Health behaviour in school-aged children study: International report from the 2009/2010 survey. http://www.euro.who.int/__data/assets/pdf_file/0003/163857/Social-determinants-of-health-and-well-being-among-young-people.pdf

Open dialogue peer review: A response to Professor Hutchings

Selma Babayiğit

DRAWING ON the key findings of her comprehensive nationwide survey involving over 7900 teachers, Hutchings highlights how the current system of assessment and accountability has come to drive the national education system in England and its negative effects on pedagogy, curriculum, and psychological wellbeing of teachers and children with far reaching societal consequences. The urgency of an informed debate on the issue of assessment and accountability cannot be emphasised further. The findings of the report, *Exam Factories? The impact of accountability measures on children and young people* (Hutchings, 2015) further underscores why the current systems of national assessment and accountability constitute the most challenging obstacle for fulfilling the key objectives of education; that is to prepare children for life by developing their critical and creative thinking skills and equipping them with the essential character strengths to be able to navigate through the challenging currents of uncertainty and rapid change, which have come to define our contemporary world.

Hutchings begins the open dialogue with an eloquent examination of the history of how the ideas around the notion of accountability have evolved over time. Indeed, it is not possible to contextualise Hutchings's analyses and the findings of her report without a more thorough historical analysis of how neoliberal market ideology[1] has been increasingly driving the education systems where the emphasis is on more market, less State, competition (league tables), a punitive system of regulation (e.g. Ofsted imposed sanctions), individual as opposed to collective social good, and data-driven economising measures. A similar data-driven accountability agenda is also associated with the current problems facing higher education in the UK (Ball, 2015; Radice, 2013). Hutchings further underlines that accountability in this context is 'largely management accountability'. This was also captured in the findings that the main concern of headteachers was the inspection, which tends to be primarily based on decontextualised data from test scores with significant consequences in the form of job loss, academisation along with an element of public humiliation. This form management does not require any educational expertise and devolves all responsibility to teachers, schools and even children and their parents.

As Hutchings argues externally imposed high stakes exams inevitably result in 'teaching to the test', hence serve to constrain the curriculum to what is being assessed. The first casualties of this system of assessment are creativity and reflexivity in a curriculum, which embeds character building in its core activities. These 'soft' learning outcomes are not always easily quantifiable. Broadfoot (2016) in the previous special issue in the *PER* eloquently expands on Claxton and Lucas's (2016) ideas, and argues that it is the national assessment, which is the main factor constraining the current curricula to prepare children for the contemporary challenges of life by helping them to develop skills in '7Cs', namely confidence, collaboration, communication, creativity, curiosity, commitment and craftsmanship. A balanced

[1] It is beyond the scope of this article to explain in detail the neo-liberal market view of education, for more information refer to Ball (2012) and Lingard & Lewis (2016).

curriculum with the Claxton and Lucas's '7 C's' is crucially important for addressing the social capital needs of children and specifically those from disadvantaged backgrounds. The *Exam Factories* report elucidates that the current assessment contributes further to the poverty-related educational achievement gaps by undermining the ability of the schools to equip children from disadvantaged backgrounds with the kinds of social capital, which will promote their social mobility and help them to fulfil their full potential during these vitally important foundational years.

There is so much schools can do to buffer disadvantaged children against the failings of successive governments to address the widening socio-economic gaps in society. Although 'value added' measures claim to address this problem, as Hutchings states progress measures are inevitably constrained by the same socioeconomic stresses. Evidence suggests that relative to their better-off peers, children from disadvantaged background tend to begin formal schooling with less well-developed linguistic and school readiness skills (Hart & Risley, 1995; Mistry et al., 2010). The consensus states that addressing the poverty-related gaps in education should begin during the prenatal period and requires investment in quality preschool and child care. A recent PISA (the Programme for International Student Assessment) report found that children from disadvantaged backgrounds are less likely to attend preschool and this factor alone directly contributes to socio-economic gaps in educational achievement at age 15 (PISA, 2012). In light of these findings, it is recommended that:

> Governments should ensure that quality pre-primary education is available locally, especially when disadvantaged families are concentrated in certain geographic areas. Governments should also develop fair and efficient mechanisms for subsidising pre-primary education to ease the financial burden on families (PISA, 2012, p.192).

Finally, the mental health problems among children and young people are increasing. This is clearly a complex and multifactorial issue. However, as Hutchings argues the pressures of assessment and accountability on both teachers and children cannot be ignored. This is clearly an area that requires further systematic study.

What is the way forward?

Hutchings cites two alternatives to assessment as a way forward, though acknowledges that none of them provides a sufficient or satisfactory solution. Inevitably the content of a test must be constrained by the pragmatics of developing a valid and reliable assessment measure. Hence even the most sophisticated assessment measures will not be able to overcome all the problems of assessment outlined in this article. Another suggestion is a monitoring system based on a subsample of schools with a more comprehensive content, which assesses skills beyond the core subjects. However, as Hutchings also notes this cannot provide information on failing schools, highlighted by the government as their major concern, which motivates the current assessment and accountability systems. It is notable that Hutchings also cites this approach in her recommendations in the *Exam Factories* report.

> If tests are used as accountability measures, they should be similar to the PISA international tests in that only a sample of schools should take them on any occasion. The results of these tests would not be communicated to parents, and should not be used for judging individual schools; rather, they would give a picture of the national pattern of attainment, and the variability of attainment across groups of pupils. This would therefore inform practice in all schools (Hutchings, 2015, p.7).

Although I agree with these suggestions, I would also like to echo the critical voices

in relation to PISA and similar international assessments (for a detailed critical analysis, Ball, 2012; Lingard & Lewis, 2016). The international comparisons tests like PISA are already dominating the national curriculum of an increasing number of nation states fired by the intense competition to achieve a higher ranking score. PISA yields data and governments can use this data as 'evidence' to justify any policy they desire. Whereas this can be a push for a positive change, it can also be used for imposition of a particular policy for political ends. When examining the data from PISA type of assessments, it is also all too easy to overlook the complexities of linguistic and sociocultural factors that may influence performance on these tests. We should be mindful that the international tests may also dominate authentic pedagogies shaped by careful considerations of local linguistic and sociocultural factors.

Hutchings also mentions briefly that the inspection used to take the form of an advisory role before the current punitive data-driven system. I believe there is scope to analyse this period in more detail. We need a system of accountability where there is a collective sense of responsibility; inspectors are experienced educationalists and there is a flexible system where local solutions can be developed to address local problems facing schools. In the main report, Hutchings also lists among the recommendations '… where there are serious concerns about a particular school, a team of advisors should be available to call in to support that school (along similar lines to the London Challenge advisors). They would be educational professionals with substantial experience of leading schools and of school improvement, who could provide on-going advice and support' (Hutchings, 2015, p.7). My question on this view is that why wait for *serious concerns to be raised* or schools to fail before acting. An expert advisory support could be the way forward to empower schools with an inspection model based on professional support, sharing of expertise, and collective responsibility and prevention of failure.

It is unfortunate that the *Exam Factories* report does not elaborate on the Finnish model. It is widely agreed that the Finnish education system owes its stellar success to the fact that it eschews the neoliberal market view of education, which has been at the driving seat of the UK national education policy (Sahlberg, 2007, 2011). This does not mean that we can import the Finnish model as it is. This never works and will not solve the problems of education in the UK. What seems to be missing in the debate are the core values that underpin the success of the Finnish system, which have been further supported in a recent report by the OECD (PISA, 2012). The PISA (2012) report highlighted the importance of teacher and school autonomy over the curriculum and assessment to improve the overall school system. The report also cites the counterproductive effects of school choice and competition, which lie at the heart of the market view of education, and does not promote academic success or overall success of school systems. Interestingly these issues are often overlooked by recent governments who tend to focus PISA scores.

Finally, what seems to be missing in Hutchings' report is the crucial gap in investment in excellence in professional training of teachers. Teachers should be equipped with an in-depth understanding of children's learning, cognitive and social-emotional development and common developmental disorders like Attention Deficit Hyperactivity Disorder. Teachers should have the research skills to conduct their own systematic evaluations. In this connection, it is noteworthy that teacher training programmes in Finland are research-based and trainees have to study for five to six years before they can teach on their own (Sahlberg, 2007, 2011).

To conclude, Hutchings's eloquent analysis of accountability should be at the forefront of any discussion on education for the 21st century. The school and teacher autonomy over assessment and curriculum is central in this debate. An advisory system, based on the notion of collective respon-

sibility and empowerment should replace the current punitive inspection system. In a global world, comparison of educational success via international assessment programmes like PISA might be a positive push for some countries and help spread the internationally validated practices of educational excellence. However, scores on an international assessment should not be a means to an end and caution should be exercised to ensure that international tests do not take the priority over local pedagogy. Finally, the way forward is an evidence-based teacher training and education system and a wider public debate across every possible platform to stop governments treating education as a political football. Given the perverse effects of the current market model of accountability on education, I share the sense of urgency expressed by Hutchings. We need to act now.

The author

Dr Selma Babayiğit, CPsychol,
Department of Health and Social Sciences,
Faculty of Health and Applied Sciences,
University of the West of England, Bristol,
Coldharbour Lane,
Bristol BS16 1QY.
Email: Selma.Babayigit@uwe.ac.uk

References

Ball, S.J. (2012). *Global education inc: New policy networks and the neo-liberal imaginary*. Abingdon: Routledge.

Ball, S.J. (2015). Education, governance and the tyranny of numbers. *Journal of Education Policy, 30*(3), 299–301.

Broadfoot, P. (2016). Open Dialogue peer review: A response to Claxton & Lucas, *Psychology of Education Review, 40*(1),13–16.

Claxton, G. & Lucas, B. (2016). The hole in the heart of education (and the role of psychology in addressing it). *Psychology of Education Review, 40*(1), 4–12.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes Publishing.

Hutchings, M. (2015). *Test Factories? The impact of accountability measures on children and young people*. London: National Union of Teachers (NUT). Retrieved 13 January 2017 from https://www.teachers.org.uk/files/exam-factories.pdf

Lingard, B. & Lewis, S. (2016) Globalisation of the Anglo-American approach to top-down, test-based educational accountability. In G.T.L. Brown & L.R. Harris (Eds), *Handbook of human and social conditions in assessment* (pp.387–403) (New York, Routledge).

Mistry, R.S., Benner, A.D., Biesanz, J.C., Clark, S.L. & Howes, C. (2010). Family and social risk, and parental investments during the early childhood years as predictors of low-income children's school readiness outcomes. *Early Childhood Research Quarterly, 25*(4), 432–449.

PISA (2012). *What makes schools successful? Resources, policies and practices – Volume IV*. OECD publishing. Retrieved 10 January 2017 from http://www.oecd.org/pisa/keyfindings/Vol4Ch6.pdf

Radice, H. (2013). How we got here: UK higher education under neoliberalism. *ACME: An International E-Journal for Critical Geographies, 12*(2), 407–418.

Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy, 22*(2), 147–171. doi: 10.1080/02680930601158919

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

It's the interpretation, not the tests

John Hattie & Janet Clinton

IT IS WORTH RECALLING that it has only been over the past 20 years that politicians have claimed to have some authority and legislative control over 'outcomes' of schooling. David Blunkett and Michael Barber were among the earliest UK politicians to switch the debate to outcomes in the UK (Tucker & Codding, 2005); and Bill Clinton in the US began the focus on accountability for effective student learning, standards based reforms, and creating national goals which specified outcomes (e.g. the high school graduation rate will increase to at least 90 per cent; all students will demonstrate competency over challenging subject matter, the US will be first in the world in math and science; McDonnell, 2005). There has been the rise and rise of PISA, and the belief among the voters and employers that schools should deliver outcomes. The outcomes debate is now entrenched, and the main mechanisms are accountability, tests, and inspection. Test scores have created a new audience (almost voyeuristic), a new narrative, and like most mechanisms they become the main game and not the reasons for which they were introduced. Hutchings' (2017) paper sums up typical reactions to the dominance of tests, and like many other such reports misses the main game – how to do we best enhance the outcomes of students in our schools. What are the alternative mechanisms to achieve this goal (or are there better goals)?

The answer to the current narrative about 'school outcomes' over the past two decades seems to be more tests, twiddle with the curriculum, create different forms of schools, blame the teachers, and invent non-productive metaphors – closing the gap, raising the tide, the year of impact, and now exam factories. How disparaging to the many excellent educators – who fortunately (or unfortunately) just get on and use their expertise to improve the learning lives of students. 'Unfortunate', as so often they too, when asked, these teachers ask for more resources to be left alone (when collaborative expertise is among the most powerful influences), and often deny their own expertise (Wiggins, 2016).

The current (over-) emphasis on testing is based on the belief that these 'tests' can best measure the desired outcomes – with the claims that 'yes, they can' (they are reliable, measure valuable attributes, we need to keep the schoolies honest) and 'no they can't' (they are narrow, stifle creativity, force teachers to be automatons, stress students). The binary answer seems alive in education; it seems we have never seen a dichotomy that educators cannot exploit. Hutchings' (2017) paper is another example of privileging the divide.

As she noted, during the last two decades there has been a continual backlash against the washback of testing; and nearly all this washback is claimed to be negative. Once again, we have yet another report with a sensationalist title aimed to make the negative case. Once again, the troops rally – leave us alone to do our important work; stop the stress on teachers and students; bring back arts and creativity, and so on – as if the presence of tests or removing or decreasing their impact are the prime catalysts. The throwback title to the 'factories' conjures up images of Dickens and Taylorism, and it seems intended to scare or mock; although if the author had visited a modern factory she would see many attributes that might make many schools envious – light, clean, collaborative teams, privileging expertise, seeking and using feedback to improve, pride, some well-paid, efficient, and effective. Our phone, Uber car, computer, and food are all produced using the factory model so often derided by the factory metaphor.

Yes, in many schools there is an over-reaction to the presence of high stakes tests. For example, we are conducting a series of studies in the UK whereby we gain transcripts of lessons. There is a dramatic shift in focus and teaching method about four weeks before the major external tests – in the expected way: More drill, more test prep, more silence, less anything but the test subjects. This is surely a distortion. We also note the increase in test focus in the UK over the past 20 years while also watch it decline in relative and absolute value on the PISA measures; and we note many countries that have similar if not more testing flourish in these same league tables (Hattie, 2016). We note the double whammy of the UK test focus and the seemingly annual flipflop of emphasis by an intrusive inspectorate (surely Ofsted should be held to same accountability they demand of schools – if their intrusions cannot be shown to improve the learning lives of students why have them). We note the major debates about infrastructure that has led to a narrative about school choice, different kinds of schools (free, academy, comprehensive, grammar) when all the evidence shows that the variance between schools is tiny relative to the variance within schools (and the mean scores of students entering some of these different types of schools are more influential on the test scores than what differently happens in these schools!). As noted in the *Politics of Distraction* (Hattie, 2015) we love to debate the things that do not matter. So, yes, the misinterpretation of test scores can distort the debates, the funding imperatives, and the work that happens in schools.

On the other hand, what would Hutchings do about schools where most students do not gain a year's growth for a year's input. How would she know which schools these were; and why is it so many schools who are not making these gains explain this failure by post-code, poor parenting, lack of resources while the schools down the road are making the difference with similar post-codes, parents, and resources? Who would know about the positive impact of the London Challenge if there were no measures to evaluate the progress of this programme (Kidson & Norris, 2014)? Is the alternative the testimonial porn that so pervades the debates about the outcomes of schooling? Why are there are many teachers who see value in the externally moderated tests who also have positive impacts on their students whereas many who have negative views are indeed less likely to have these impacts (Brown, 2004)? Why is it that schools are not transparent to parents about the achievement status of students – more than 99 per cent of students in New Zealand, based on their school reports are achieving well, putting in effort, and are a pleasure to teach (Hattie & Peddie, 2003). No wonder parents want external checks. As Ronald Reagan proclaimed, we need to Trust but Verify.

The revolution in measurement

In measurement, there was a revolution in the concept of the validity of tests 20 years ago. Messick (1989) outlined that the validity of tests was primarily a function of the preponderance of evidence to 'support the 'adequacy' and 'appropriateness' of inferences and actions based on test scores' (p.13). That is, validity shifted from the attributes of the test to the inferences from the test (and noted how the former was critical to the success of the latter; Hattie, 2014; Kane, 2016). This led to two major pressures. First, the system should defend the presence of testing in schools on the basis on the quality of inference and actions they lead to – and wow is this sorely missing. Second, the system should provide evidence that the teachers and school leaders are making appropriate inferences and actions because of the testing. Herein lies the problem.

Similarly, 40 years ago the stress researchers changed their focus when they noted that the same stressors led to different interpretations and actions. The stress research then moved from investigating stressors (like test anxiety) to the coping strategies to stressors.

Why do some students and teachers stress from the same testing regime that others cope with admirably? The problem is not one of narrowness of the tests. There are so many schools that do not narrow the curriculum, do enhance creativity, have excellent music and physical education programs and administer the national tests! Again, it is the interpretations made by teachers and students to a potential stressor, not the stressor itself that matters most.

The problem of so many of our tests in schools is the lack of interpretation value. Much of this problem comes from how tests have been developed. The typical process involves drawing up blueprints of test specifications, writing, piloting, and analysing items, then creating tests to meet reliability and validity claims (or more often, Rasch models are used to maximise reproducibility coefficients and information across the range of ability). This is a well-known process and from all the reports issued by the UK test authorities this is done to a very high level of psychometric sophistication. But this, however, is not attending to the core issue of validity.

The alternatives to the 'exam factory'

So, what is the alternative. Hutchings (2017) notes the inevitability of accountability measures – and well she should have, given this is what the voters, parents, and students (if they had the vote) seem to be demanding. Her answer is to keep seeking the intended and unintended consequences (hardly new as her paper demonstrates), use sample surveys that do not allow report at school or student level, use group tasks, privilege more teacher summative measures. That's it. Yes, sampling has many merits and is practiced in some Western countries (e.g. NAEP in US) but with little impact on much (Brookhart et al., 2016); yes, checking the unintended consequences; yes to group tasks; and yes teachers are already creating and administering many summative measures. But the answer is not tests or not tests; it is about how we use the information from the tests to make a difference to the learning lives of students. Here is where the article could be better titled 'we have the data, but not the information; we have the information but not the actions'. The imperative is how the test information is interpreted

Consider an alternative: a system where assessment system that provides rigorous, worthwhile interpretations to assist educators and students enhance their learning. A system that is longitudinal, constructively aligned with curricula, can allow for tests to be constructed on the fly specifically related to what teachers and students are doing right now, and judged by the value of its interpretations and inferences and not merely dominated by numbers and test scores alone. This was the basis of the model we developed for the New Zealand (NZ) assessment system (e-asTTle: Hattie & Brown, 2008). We started backwards. We devised reports that aimed to influence the decisions of teachers, school leaders, parents and students. We conducted hundreds of focus groups continually changing and refining the reports so that they met the two validity questions: (1) Did the educators made the correct interpretations (if they did not we changed the reports, we did not blame them for lack of assessment literacy); and (2) Was there a consequence in terms of how they thought about their impact or changed/reinforced their practice? It took us much investment to get our seven reports to pass these two tests (Hattie, 2010). Then we backfilled the reports with items and tests using the usual psychometric toolkit. The system is voluntary, and it is worth noting that most teachers and schools are still using this e-asTTle tool 16 years after it was introduced. Because they find the interpretations of the tests valuable. It's the interpretations not the tests that matter.

Given the way e-asTTle is designed there are many millions of test permutations, so a critic could claim NZ has the perfect test factory. But no, this factory provides services that those who truly make the difference in the learning lives of students want! Yes, the

government now has a sturdy, sampled, and non-intrusive thermometer of the health of the system based on not only the status of students at any point of time but also indicators of the growth over time (as it can follow students from age 8–18), but without the backlash outlined in Hutchings paper. But note, it is test based! What is necessarily wrong with tests. e-asTTle could be considered a 'factory farm.' It is accountability, and schools use the tool to evaluate their impact, their programmes, and their successes. It is both formative and summative, and there has been (in many schools) an excellent move to develop 'student assessment capabilities' using the tool (Absolum et al., 2009). That is, helping the students interpret the information from the test to enhance their next learning.

Let us stop the binary – tests are good, tests are bad. Let us move to asking about the quality of interpretations and actions; let us find schools that do this well and tell their story; let us realise that it is the moment by moment judgements teachers and school leaders make that matter most and let us provide resources (like e-asTTle) to help inform, calibrate, and evaluate these judgements; let us provide ways to help schools do the right work and not keep devising methods that put thermometers into schools with no consequences to improve the students it is measuring. We need both measures of status (where the student is) and growth (the gain over a year) and not one or the other. We need tests that not only evaluate what the student knows, but tests that aid diagnosis, prediction, and lead to optimal 'where to next' lessons. We need tests (of many forms) constructively aligned with and across the curricula. We need a 'basket of goods' measurement approach to evaluate schools not merely narrow excellence of literacy and numeracy (Zhao, 2015). We need to develop evaluative thinking in our schools based on the interpretations that teachers and school leaders are making from many sources including test scores (Clinton & Hattie, 2014). We need measures not only what they learn but how they learn (is it not ironic that such measures are so rare when this is surely a major purpose of schooling, to teach students how to learn; Hattie & Donoghue, 2016).

We commend Hutchings for so eloquently raising the issues and now invite debates about the alternatives and beg not another debate about tests or no tests. The question needs to be how to devise a measurement system that passes the two validity questions noted above – and it can be done, it has been done. It is the quality of the interpretations that teachers and school leaders make on a moment-by-moment basis that matters; it is how to inform these judgements; how to moderate them so that they are appropriate, challenging, and shared. This needs to be the purpose of investing in tests. We cannot have some teachers and schools with very low expectations of what can be realised when teachers in similar schools realise their very high expectations. This will not be resolved by getting rid of tests, developing a sampling model of tests, and the other alternatives in Hutchings paper. We already have so many excellent teachers and schools who are using test information to inform their decisions and we need good measurement to identify, privilege, and sustain this expertise.

The authors

Professor John Hattie &
Professor Janet Clinton
Melbourne Graduate School of Education
University of Melbourne
Parkville 3010
VICTORIA
Australia

Correspondance
Email: jhattie@unimelb.edu.au

References

Absolum, M., Flockton, L., Hattie, J., Hipkins, R. & Reid, I. (2009). *Directions for assessment in New Zealand (DANZ).* Wellington: Ministry of Education. http:// assessment.tki.org.nz/Assessment-in-the-classroom/Assessment-position-papers

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M. & Welsh, M.E. (2016). A century of grading research meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848.

Brown, G.T. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301–318.

Clinton, J.M. & Hattie, J.A.C. (2014). Teachers as evaluators – An empowerment evaluation approach. In D. Fetterman, S. Kaftarian & A. Wandersman, (Eds.). *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability.* Thousand Oaks, CA: Sage.

Hattie, J.A.C. (2010). The validity of reports. Online Educational Research Journal. http://www.oerj.org/View?action=viewPaper&paper=6

Hattie, J.A.C. (2014). Validity, types of. In D. Phillips (Ed.), *Encyclopaedia of educational theory and philosophy.* (pp.829–831). Thousand Oaks, CA: Sage Inc. doi: http://dx.doi.org/10.4135/9781483346229.n340

Hattie, J.A.C. (2015). *What doesn't work in education: The politics of distraction.* Open Ideas at Pearsons. https://www.pearson.com/hattie/distractions.html

Hattie, J.A.C. (2016). Shifting away from distractions to improve Australia's schools: Time for a Reboot. *ACEL Monograph Series, 54,* 1–24.

Hattie, J.A.C. & Brown, G.T.L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189–201. http://tinyurl.com/472um3k

Hattie, J.A.C. & Donoghue, G. (2016). Learning strategies: A synthesis and conceptual model. *Nature: Science of Learning. 1.* doi:10.1038/npjscilearn.2016.13 http://www.nature.com/articles/npjscilearn201613

Hattie, J.A.C. & Peddie, R. (2003). School reports: 'Praising with faint damns'. *Set: Research information for teachers, 3,* 4–9.

Hutchings M. (2017). Accountability measures: The factory farm version of education, *The Psychology of Education Review, 41*(1).

Kane, M.T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice, 23*(2), 309–311.

Kidson, M. & Norris, E. (2014). *Implementing the London challenge.* London: Institute for Government.

McDonnell, L.M. (2005). No child left behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education, 80*(2), 19–38.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement,* (3rd edn, pp.13–103). New York: American Council on Education, MacMillan.

Tucker, M. & Codding, J. (2005). *The case of England.* National Center on Education and the Economy, New Commission on the Skills of the American Workforce

Wiggins, (2016, 6th May). Teachers prefer smaller class sizes to pay rises. *Times Educational Supplement.* Retrieved 1 February 2017 from https://www.tes.com/news/tes-magazine/tes-magazine/teachers-prefer-smaller-class-sizes-pay-rises.

Zhao, Y. (2015). *Counting what counts.* USA: Solution Tree.

Commentary – 'Accountability measures: The factory farm version of education'

Pam Jarvis

HUTCHINGS raises many useful points in her article, commenting on her research findings with teachers, head-teachers and pupils. The comments that these participants made will surely come as no surprise to those in control of policy for schools in England, given that they have been receiving similar criticism throughout the past decade. The last truly independent review of primary education in England, the *Cambridge Review* (Alexander, 2010) raised similar points to many of those contained in *Exam Factories* and *Accountability Measures,* but the government of the day rejected it outright, particularly the suggestion that children should start school at six, proposing that this 'flies in the face of international evidence' (BBC, 2009).

However, seven years later, England has made no progress in the PISA comparisons between OECD nations, and the most commonly reported school starting age of nations at the top of the PISA charts is in fact six. Finland, a perennial high performer, has a school starting age of seven and no statutory assessment system at all (Jarvis, 2016). Moreover, as Hutchings comments, the fact that GCSE performance has apparently increased, while, over the same period, PISA performance has remained static is an issue that urgently merits further investigation. Such disparate findings raise many questions about who should be made accountable, and for what.

As Hutchings points out, the notion that schools and teachers should be 'accountable' for 'performance' against externally-imposed targets has been an obsession for the British government since the advent of mass schooling, resulting in the periodic application of over-simplistic metrics apparently intended to raise the quality of education, but which eventually result in depressing it. Consider, for example, the following:

> ...the standards themselves were defective because they were based not on an experimental enquiry into what children of a given age actually knew, but on an a priori notion of what they ought to know. They largely ignored the wide range of individual capacity, and the detailed formulations for the several ages were not always precise or appropriate. Gillard (2011)

It might seem that Gillard is discussing the current education system here, but in fact, he is referring to the system initiated by the 'Revised Code of Conduct' in 1862. This code and its infamous 'payment by results' regime held teachers and pupils in its iron grip for an astonishing 30 years until it was recognised as a failure and abandoned in 1892, following increasing concern about the narrowness of the curriculum brought about by the inevitable teaching to test. Extending the similarity to the current situation in state education, the inadvisability of such regimes was in fact well known for at least 40 years before the 'Revised Code of Conduct' was published. In the early 19th century, pioneer educator Robert Owen commented, based on what he had observed in the early factory school system, that he had concluded teaching narrowly to test resulted in children becoming mentally 'cramped and paralysed' which would 'render their moral character depraved and dangerous' to the extent that they could 'never become really useful subjects of the state' (Owen, 1991, p.163).

Owen also deplored mechanical, skills-based approaches to literacy instruction, pledging that his school would: '...avoid literacy becoming an end in its own right, to ensure that the ability to decode and print text was wedded to the capacity to comprehend and derive satisfaction from the act of reading' (Davis & O'Hagan, 2010, p.94). This has also recently re-arisen as a live issue in the light of current government insistence on a statutory phonics test for five year olds, and continues to be hotly debated by literacy experts and policy makers (Clark, 2016).

So why do we continue to endlessly circle around accountability issues, generation after generation? As Hutchings comments, the most logical way to move forward from today's fractured and factionalised narrative in state education 'is surely to tackle wider economic inequality in society rather than to blame schools and teachers'. However, in 2016, as in 1862, it is much easier for the government of the day to scapegoat teachers, schools and parents for children's 'failure to perform' than it is to engage with powerful echelons which have many vested interests in a national and international economy that depends upon maintaining a grossly unequal society. 'Performance measures' are therefore imposed upon schools as a smokescreen, spitting out a constant stream of highly questionable data (Roberts-Holmes & Bradbury, 2016) that obscures the devastating effects of economic inequality, blaming parents, teachers and schools for the differential achievements of the nation's children.

Alongside the creation of a range of highly questionable assessment metrics, including the eventually abandoned attempt to impose a so-called 'baseline test' upon the nation's four-and-a-half year olds, which according to the National Union of Teachers cost the nation 'millions [in the pursuit of] establishing a 'flawed' system' (Children and Young People Now 2016, online), the Conservative-led coalition government of 2010–2015 additionally introduced the rhetoric of 'troubled families' which underpinned an initiative into which they pumped £400,000,000. In 2016, an evaluation of this project found that its activities had principally been used as 'window dressing' to assure the public that 'something was being done', but that in fact, the project had no significant impact upon the lives of socio-economically deprived children or their families (DFE, 2016).

While the parents and teachers of the mid-19th century did not have the psychological vocabulary to fully discuss the unhappiness that existed in their schools with policy-makers (although writers like Charles Dickens and Charlotte Bronte described this vividly in novels such as *Hard Times*, *Nicholas Nickleby* and *Jane Eyre*), this aspect of childhood is of growing concern in the early 21st century. There has been a doubling of juvenile depression between the 1980s and 2000s (Young Minds, 2016a) and a huge rise in self harming, with increasing numbers of children and young people needing hospitalisation (*The Guardian*, 2014). Self-harming is a reaction to being placed under unbearable mental pressure, as physical injury releases endorphins that counteract the stress response (Students Against Depression, 2016). There are also a growing number of young people developing eating disorders and suicidal thoughts (Young Minds, 2015), a statistic which has risen rapidly since 2009, with a doubling of numbers presenting psychiatric problems to Accident and Emergency departments (Young Minds, 2016b). Two successive UNICEF reports on children's wellbeing in rich nations undertaken in 2007 and 2013 (UNICEF 2007, 2013) have additionally indicated that English children have a very low sense of wellbeing.

As Hutchings points out, quoting one of her participants 'you can't be counselling them for what you are putting them through at school', and it is becoming increasingly clear, as she also comments 'the target-driven school system is a factor in young people's mental health problems'. This comes down harder on young people whose lives are already blighted by poverty which continues to worsen within the

current austerity economy (Child Poverty Action Group, 2016), feeding into a spiral of failure, as Hutchings documents: 'teachers who reported most pressure on pupils and uniform (boring) lessons, came from schools that had low attainment and poor inspection grades... there is a strong correlation between these two factors and proportion of disadvantaged pupils in the school.'

Many of the issues Hutchings raises relate to policy makers' lack of understanding of 'human' and developmental issues, a point that I have raised continually in my own research (e.g. Jarvis et al., 2014; Jarvis, 2015, 2016). Hutchings' examples include complex and non-linear reasons for parental selection of schools for their children, the impossible pressure upon teachers and head teachers that turns schools into socio-emotional pressure cookers, damaging the mental health and wellbeing of both adults and children within them, the removal of professional decision making from teachers, resulting in their inability to address children's individual needs including developmental difficulties and delays, and a narrow definition of 'failure', particularly with respect to an over-reliance on 'pencil and paper' skills. As Hutchings comments, 'even if young people have other skills and attributes that could be useful in a job, they are likely to be rejected'.

Overall then, I am in complete agreement with Hutchings' thesis that there 'is substantial evidence that accountability measures have had a great many negative impacts on teachers, children and young people, and on the quality of education in England'. Throughout her text, Hutchings vividly demonstrates how the current regime of English state education wastes human talent, deskilling both teachers and pupils and creating vicious circles of disadvantage. The historical data indicates that she is also correct in her proposal that 'there is evidence that the same negative impacts have been experienced whenever high stakes testing has been imposed'.

There are however some points in the article that would have benefitted from slightly greater clarity. For example, the relationship between English and Maths and Literacy/Numeracy skills where the discussion of 'learning more things that were practically useful' arises, and the role of enjoyment, imagination and narrative in reading. As a psychologist who specialises in developmental issues, I would also dispute that 'guided play' is the 'most effective strategy' for supporting young children's learning, and would instead suggest that free, spontaneous interaction amongst adults and peers is the most appropriate learning environment for children between birth and seven, where the adult follows the child's lead rather than vice versa (Jarvis et al., 2014; Jarvis, 2015). The narrative of the article would have been enriched by reference to the highest quality early years practice, which is based in freely chosen activity on the child's part and sustained shared thinking on the adult's (Siraj-Blatchford, 2009), with particular reference to the fact that it is impossible to embed such a process in a regime where both teachers and children are judged on their abilities to address narrow, pre-defined targets.

Recent damning evidence has emerged to further illustrate Hutchings' thesis that performance targets in education spawn a dysfunctional system that discriminates against the most vulnerable. In November 2016, the outgoing Ofsted Chief Inspector, Michael Wilshaw, reported that young people with special needs that make them unlikely to achieve government targets are at risk of being excluded from school and channelled into so called 'alternative provision', some of which operates illegally, under the direction of untrained staff. He reported a dramatic rise in the number of excluded pupils being schooled in 'unsafe and unhygienic premises by staff who have not been properly checked' (*The Guardian*, 2016).

This then, is the final irony: the imposition of 'standards' apparently created to protect and enhance children's rights to education, which ultimately result in the complete exclusion from the system of

those in the greatest need, because they are not good quality commodities in terms of bolstering the school's standing in the national league tables. Instead of being safe at school, such children are instead constructed as 'damaged goods' to be siphoned off into 'potentially dangerous places' (*The Guardian,* 2016). In the light of such a dystopian nightmare, I emphatically support Hutchings' call for 'a comprehensive review of accountability strategies with a view to radically changing them'- indeed, our children, who have now become pawns in a cynical socio-political game, deserve nothing less. Like our ancestors in 1892, we must finally recognise that the accountability emperor is utterly naked, and move on from the deeply dysfunctional culture that currently dominates state education. As responsible adults, we must now act to protect both children currently in state education in England, and those yet to come. We must construct a new framework for state education which is rooted in 'policies... grounded on the best available evidence of what human beings are like' (Singer, 1999, p.61). A starting point for this conversation has already been waiting in the wings for the past seven years, encapsulated in the *Cambridge Primary Review.* Professionals and academics in education and child development await the government's call.

The author

Dr Pam Jarvis
Leeds Trinity University,
Brownberrie Lane,
Horsforth,
Leeds LS18 5HD.
Email: p.jarvis@leedstrinity.ac.uk

References

Alexander, R. (2010). *Children, their world, their education: Final report and recommendations of the Cambridge Primary Review*. Abingdon: Routledge.

BBC (2009). *Primary education too narrow*. Retrieved 11 December 2016 from http://news.bbc.co.uk/1/hi/education/7896751.stm

Bronte, C. (1847). *Jane Eyre*. Retrieved 11 December 2016. from http://www.gutenberg.org/ebooks/1260

Child Poverty Action Group (2016). *Child poverty facts and figures*. Retrieved 11 December 2016. from http://www.cpag.org.uk/child-poverty-facts-and-figures

Children and Young People Now (2016). *Government U-turns on testing of four-year-olds*. Retrieved 12 December 2016 from http://www.cypnow.co.uk/cyp/news/1156772/government-u-turns-on-testing-of-four-year-olds.

Clark, M. (2016). *Flawed arguments for phonics. The mismeasurement of learning* (pp.20–21). London: The National Union of Teachers. Retrieved 11 December 2016 from https://reclaimingschools.files.wordpress.com/2016/11/mismeasurement.pdf

Davis, R. & O'Hagan, F. (2010). *Robert Owen*. London: Bloomsbury.

Department for Education (2016). *National evaluation of the troubled families programme, final synthesis report*. London: DFE. Retrieved 11 December 2016 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/560499/Troubled_Families_Evaluation_Synthesis_Report.pdf

Dickens, C. (1854). *Hard Times*. Retrieved 11 December 2016 from http://www.gutenberg.org/ebooks/786

Dickens, C. (1835). *Nicholas Nickleby*. Retrieved 11 December 2016 from http://www.gutenberg.org/ebooks/967

Gillard, D. (2011). *Education in England: The history of our schools*. Retrieved 11 December 2016 from http://www.educationengland.org.uk/history/chapter03.html

Jarvis , P. (2015). *It's against human nature to send two-year-olds to school. The conversation*. Retrieved 11 December 2016 from https://theconversation.com/its-against-human-nature-to-send-two-year-olds-to-school-37180

Jarvis, P. (2016). *Too much, too young for PISA*. The Huffington Post. Retrieved 11 December 2016. from http://www.huffingtonpost.co.uk/pam-jarvis/too-much-too-young-for-pi_b_13479730.html

Jarvis, P., Newman, S. & Swiniarski, L. (2014). On 'becoming social': The importance of collaborative free play in childhood. *International Journal of Play, 3*(1), 53–68. doi:10.1080/21594937.2013.863440. Retrieved 11 December 2016 from http://research.leedstrinity.ac.uk/files/161043/Jarvis_Newman_Swiniarski_On_becoming_social_August_2013.pdf

Owen, R. (1991). *A new view of society and other writings*. London: Penguin Classics.

Roberts-Holmes, G. & Bradbury, A. (2016). *'Datafication' in the early years. The mismeasurement of learning* (pp.16–17). London: The National Union of Teachers. Available from https://reclaimingschools.files.wordpress.com/2016/11/mismeasurement.pdf

Singer, P. (1999). *A Darwinian left: Politics, evolution and cooperation*. London: Weidenfeld & Nicolson.

Siraj-Blatchford, I. (2009). Conceptualising progression in the pedagogy of play and sustained shared thinking in early childhood education: A Vygotskian perspective. *Education and Child Psychology, 26*(2), 77–89.

Students Against Depression (2016). *Understanding Self Harm*. Retrieved 12 December 2016 from http://studentsagainstdepression.org/get-support/check-suicide-and-self-harm/understanding-self-harm/

The Guardian (2014). Shock figures show extent of self-harm in English teenagers. Retrieved 12 December 2016 from https://www.theguardian.com/society/2014/may/21/shock-figures-self-harm-england-teenagers

The Guardian (2016). Ofsted chief warns schools over use of 'alternative provision' for challenging pupils. Retrieved 11 December 2016 from https://www.theguardian.com/education/2016/nov/08/ofsted-chief-warns-schools-over-use-of-alternative-provision-for-challenging-pupils

UNICEF (2007). *An overview of child well-being in rich countries*. Florence: UNICEF. Retrieved 12 December 2016 from https://www.unicef.org/media/files/ChildPovertyReport.pdf

UNICEF (2013). *Child well-being in rich countries: A comparative overview*. Florence: UNICEF. Retrieved 12 December 2016 from https://www.unicef-irc.org/publications/pdf/rc11_eng.pdf

Young Minds (2015) *Teenage eating disorders*. Retrieved 12 December 2016 from http://www.youngminds.org.uk/news/blog/2777_large_rise_in_uk_admissions_for_teenage_eating_disorders

Young Minds (2016a) *Mental health statistics*. Retrieved 12 December 2016 from http://www.youngminds.org.uk/training_services/policy/mental_health_statistics

Young Minds (2016b) *Mental Health Statistics*. Available at: http://www.youngminds.org.uk/about/whats_the_problem/mental_health_statistics Retrieved 12 December 2016

Response to 'Accountability measures: The factory farm version of education' by Merryn Hutchings

Christine Merrell

ACCOUNTABILITY has long been a feature of our education system and, I think, rightly so. However, past approaches to accountability have had damaging effects, as Hutchings' article has clearly illustrated. Yet we do not seem to respond to the increasing body of evidence and make changes for the better. Hutchings' article discusses changes that have been proposed by academics including Haertel (2013) and Harlen (2014), which suggest improving the validity of assessments and increasing the use of teachers' observations and judgements as evidence of the effectiveness of schools, but we could go further.

At the Centre for Evaluation and Monitoring (CEM), Durham University, we have run large-scale monitoring systems for primary and secondary schools to use for self-evaluation purposes for 30 years (Merrell, 2016; Tymms & Coe, 2003; Tymms & Albone, 2002; Fitz-Gibbon, 1996). These systems cover the 3 to 18 age range and schools, districts and, in some instances, jurisdictions, pay to use them. At the time of writing, approximately one million students are assessed each year and schools in over 70 countries make use of them. They provide high-quality information about pupils' attainment, progress, developed ability and attitudes to learning for use by teachers and head teachers. The feedback from these systems enables teachers and school managers to monitor the progress of each pupil. They can see whether their teaching strategies and interventions are having a positive impact and adjust as necessary. The assessments are computer-delivered, adaptive to the ability of each child, and quick to administer. They do not take vast amounts of time away from teaching, nor are teachers expected to spend time coaching their pupils for the assessments. Feedback about children's strengths and areas for development is rapid. Finding out what children know and can do, and monitoring their progress in response to teaching and interventions using a combination of large-scale monitoring systems alongside teacher assessment could form the basis of an accountability system for schools. Inspections could focus on the quality of the assessment system that each school has put in place, the way in which information is being used to inform practice tailored to the needs of each pupil, and how head teachers are using the data to monitor resourcing, including teachers' performance. Thus a school would be accountable for the effective implementation and use of a monitoring system, intelligent interpretation of data and use of research findings to inform their practice. They would be expected to know about every pupil's learning and be able to explain how they are catering for their needs. This is a quite different approach to that of judging the quality of schooling against a narrow set of pupil outcomes.

Teachers would need to have effective training in the use and interpretation of assessment data. They should also be familiar with research-based teaching and learning strategies, be able to match these to the needs of their own pupils and know how to rigorously evaluate their effectiveness through regular monitoring. They would need to be given sufficient time within the working week to be able to analyse information and use it to inform their teaching. Empowering teachers by providing them

with ongoing professional development to hone their skills in assessment and evaluation of their practice, the time to develop and perform assessments and to analyse the information on a continual basis to inform their practice could form a sophisticated approach to accountability and at the same time continually develop a high level of expertise within the profession.

At system level, the government needs to know if policies are effective and that could be achieved, as suggested by others, using annual assessments which are administered through a sampling approach. The sampling approach should be designed to collect sufficient data to allow for analysis by groups but does not need report school-level information. Thus there would be no pressure for teachers to coach their pupils for these system-level assessments.

Contrast this proposed approach with the current system that has prompted the description of 'the factory farm version of education'. Making significant changes to an established system is difficult and carries a risk of failure but is that risk more damaging than the effects of the current system?

The author

Professor Christine Merrell
Director of Research, Centre for Evaluation and Monitoring and Professor of Education
Durham University
Stockton Road
Durham DH1 3UZ
Email: christine.merrell@cem.dur.ac.uk

References

Fitz-Gibbon, C.T. (1996) *Monitoring education: Indicators, quality and effectiveness*. London: Cassell.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement, 11,* 1–18.

Harlen, W. (2014). *Assessment, standards, and quality of learning in primary education*. Cambridge Primary Review Trust.

Merrell, C. (2016) Understanding monitoring in the United Kingdom context. In V. Scherman, R. Bosker & S.J. Howie (Eds), *Monitoring the quality of education in schools*. Rotterdam/Boston/Taipei: Sense Publishers.

Tymms, P. & Coe, R. (2003). Celebration of the success of distributed research with schools: The CEM Centre, Durham. *British Educational Research Journal, 29*(5), 639–653.

Tymms, P. & Albone, S. (2002). Performance indicators in primary schools. In A.J. Visscher & R. Coe, (Eds.). *School improvement through performance feedback*. Lisse/Abingdon/Exton PA/Tokyo, Swetz & Zeitlinger, 191–218.

Response to 'Accountability measures: The factory farm version of education' by Merryn Hutchings

Dave Putwain

MERRYN HUTCHINGS sets out in this article (and the longer report on which it is based) to document the negative impact of accountability policy in English schools on education and well-being of children and adolescents.

Before commenting specifically on the article, I would like to clarify my position in order to contextualise my response. I do not see accountability in itself as a bad thing. Teachers and schools, in my opinion, should be held accountable to students, parents, governors, and the state. No-one could argue that we do not wish for 'good' teachers and 'good' schools. I see nothing wrong in a system that attempts to facilitate 'good' albeit what 'good' constitutes is politically motivated and highly contested. Furthermore, I do not see tests or exams, *per se,* as a bad thing. A test is just a means of evaluating student learning; a fairly blunt one, perhaps, that prioritises certain knowledge and cognitive styles.

The problem comes, in my view, when you base accountability on test scores and then incentivise teachers (and schools) using those test scores. As Merryn points out, narrowing the curriculum and spending time coaching students for tests is a rational response to this system. The incentives, or perhaps better positioned as sanctions, for teachers and schools are greater than ever before; league tables, Ofsted, and teacher-performance-related pay. There are no incentives for happy, healthy, inquisitive children who enjoy learning, and no incentives for teachers and schools to prioritise these areas other than from their own sense of humanity.

What makes this system even more perverse is that teachers can only partly account for the test outcomes of their students. Citing Stephen Gorard, Merryn makes this point. If one is to believe John Hattie's (2008) magnum opus, and I do know there are various criticisms of the methodology, once cognitive ability, family background, and socio-demographic circumstance is accounted for, the teacher can account for approximately 20 per cent of the variance in tested learning outcomes. This is not to minimise the importance of teachers, but I do not follow the logic of holding a person (the teacher) or the collective (the school) responsible for something that is only partly within their control.

Test results are not the only way to judge the quality or effectiveness of a teacher. There are myriad ways and possibilities of doing this. Indeed Ofsted use observations of teachers that include criteria such as use of questioning, assessment of learning during lessons, pace and depth of learning, and so on. One could also consider other tools such as AIMS, developed in the US by Roehrig and Christesen (2010), for judging atmosphere, instruction, management, and student engagement. The drawback with these and other such tools is that in comparison to using test scores they are time consuming to conduct and standardise. This for me is the hidden elephant in the room. It is cheaper and quicker to use test scores to judge teachers than to conduct quality, rigorous assessments of their teaching quality.

Having set out my own position, I would like to comment on three aspects of Merryn's

article. First, many of the negative consequences of accountability policies described are not new. The findings reported by Merryn mirror those from reports and reviews from the 2000s (e.g. Harlen & Deakin-Crick, 2002; Neill, 2002; Tymms & Merrell, 2007) as well as conceptual critiques from the more sociologically orientated educational literature (e.g. Ball, 2003; Torrance, 2004). On the one hand, it is somewhat saddening to hear the same issues being repeated a decade following these earlier reports. However, it is also important to keep accumulating evidence of the negative impact of this possibility and also be open to the possibility that there could also be benefits too. Although the findings contained within the report themselves are a considerable cause for concern, I am heartened that the evidence base continues to grow.

If there is to be any possibility of civil servants and policy makers making use of findings pertaining to accountability by testing policy we need a strong evidence base. This leads to my second point. Merryn articulates three areas of impact: curriculum, pedagogy, and wellbeing. From my reading of the literature, the evidence base concerning the impacts on curriculum and pedagogy is larger and more robust than the evidence for the negative impacts on wellbeing. There are three types of evidence that I would like to see to bolster the argument that this policy has a detrimental effect on children's wellbeing: longitudinal, comparative, and intra-individual.

Given the methodological constraints imposed by trying to answer a causal question from naturalistic data I would like to see more longitudinal studies. That is children tracked over their primary and secondary education to establish how their wellbeing changes in periods of testing (and test preparation) and how their wellbeing interacts with other salient educational factors (e.g. results, method of feedback, etc.). I would also like to see comparisons of English children with children in other education systems that either do not use tests for accountability purposes (e.g. Welsh primary schools) or systems with similar schooling (e.g. terminal secondary school examinations, such as the zhōng kǎo (中考) taken in China by students aged 15 at the end of junior secondary education, where accountability is judged differently).

Quantitative studies examining the impact of testing and accountability on student wellbeing, from both psychological and sociological perspectives, have followed the dominant approach of group-level analyses. For instance, one might set out to answer how the typical (i.e. group mean) self-esteem of children changes around periods of testing, or how the typical (i.e. group mean) self-esteem of children relates to the typical (i.e. group mean) test performance. The obsession with the mean can obscure how individual children can respond in very different ways and show different trajectories to others. Hence, I would also call for a greater study of the intrapersonal changes and trajectories that might occur.

Second, it is important to highlight that children and adolescents show marked individual differences in their responses to the pressure of high-stakes testing. Some thrive under pressure, some choke, and others are indifferent (e.g. Wang & Shah, 2013). Thus, it is possible that the performance of some children and adolescents is better than it would have been without the additional pressures brought to bear by the accountability policies. The factors that determine why some individuals respond to pressure as a challenge and others as a threat is a complex issue with many interacting personal and socio-demographic factors (e.g. Zeidner & Matthews, 2005). It is important, however, to highlight that it is not reducible to ability; it is not simply the fact that high ability children thrive and low ability children choke. Challenge and threat responses to pressure occur at all ranges of ability.

To date, the academic literature around accountability by high-stakes testing has been mainly, but not exclusively (e.g. Gandal & McGiffert, 2003) focused on the negative

impacts. While it is difficult to imagine how a narrowed curriculum could be anything other than bad, one could make a case that the improved test performance of those individuals who thrive under pressure is a possible benefit; one that is not enjoyed by the majority, and only applying to a thin slice of assessed learning (i.e. what is tested). I would like to see examples in the academic literature examining accountability policy of negative case analyses; those individuals who seem to buck the trend and show a beneficial response. I feel, perhaps naively, that civil servants and policy makers will be more inclined to consider evidence that considers benefits as well as drawbacks of accountability. Thus, one route to highlighting the negative could also be to highlight the positive.

Third, and finally, we should not forget the impacts on teachers; their professionalism and their wellbeing. Research from the US (von der Embse et al., 2016a, 2016b) has highlighted negative consequences of test-based accountability policies on teachers' stress and job satisfaction. When record numbers of teachers are leaving the profession in England it seems facile and irresponsible for the Department of Education to (mostly) place the blame on universities for selecting inappropriate students for teacher education courses. There must be some recognition that the pressures of accountability policies contribute in some way. This is not a criticism of Merryn's article and report. The impact on teachers was not the focus and the voice of the teacher is present. It is just a reminder that these issues must be recognised and that at present, at least in the UK, they are not being extensively researched.

So what is a possible solution or way forward with this issue? Merryn cites Harlen's approach to extend testing beyond core subjects and not report findings at pupil or school level. My preferred option would be to go one stage further and de-couple accountability from test scores completely. That is to continue with the accountability agenda, but one that is based on rigorous assessment of teaching quality, and which could still identify teachers and schools that are not providing quality education, rather than judging teachers by the performance of their students on tests and examinations. This not entirely different from the approach currently taken by Ofsted, although I would include those vastly underrated 'soft' aspects of teaching such building trusting and supportive relationships with students (which as every teacher knows is essential to successful learning but which does not appear of the National College of Teaching and Leadership core teaching standards or the Ofsted inspection criteria) and the wellbeing of students as a priority. Perhaps it could be called OfstedPLUS.

The author

Professor Dave Putwain, AfBPS C.Psychol,
School of Education,
Liverpool John Moores University,
IM Marsh Campus,
Mossley Hill Road,
Liverpool L17 6DB.
Email: d.w.putwain@ljmu.ac.uk

References

Ball, S.J. (2003). The teacher's soul and the terror of performativity. *Journal of Educational Policy, 18*(2), 215–228.

Gandal, M. & McGiffert, L. (2003). The power of testing. *Educational Leadership, 60*(5), 39–42.

Harlen, W. & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review version 1.1). In *Research Evidence in Education Library. Issue 1.* London: EPPI Centre, Social Science Research Unit, Institute of Education.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Oxon: Routledge.

Neill, S.R. (2002). *National curriculum tests: A survey analysed for the National Union of Teachers.* Warwick, Institute of Education: The University of Warwick.

Roehrig, A.D. & Christesen, E. (2010). Development and use of a tool for evaluating teacher effectiveness in grades K-12. In V.J. Shute & B.J. Becker (Eds.), *Innovative assessment for the 21st century* (pp.207–228). New York: Springer.

Torrance, H. (2004). Using action research to generate knowledge about educational practice. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education.* Maidenhead: Open University Press.

Tymms, P. & Merrell, C. (2007) *Standards and quality in English primary schools over time: The national evidence* (Primary Review Research Survey 4/1). Cambridge: The University of Cambridge Faculty of Education.

von der Embse, N.P., Sandilos, L.E., Pendergast, L. & Mankin, A. (2016a). Teacher stress, teaching-efficacy, and job satisfaction in response to test-based educational accountability policies. *Learning and Individual Differences, 50,* 308–317.

von der Embse, N.P., Pendergast, L.L., Segool, N., Saeki, E. & Ryan, S. (2016b). The influence of test-based accountability policies on school climate and teacher stress across four states. *Teaching and Teacher Education, 59,* 492–502. doi: 10.1016/j.tate.2016.07.013

Wang, Z. & Shah, P. (2013). The effect of pressure of high- and low-working-memory students: An elaboration of the choking under pressure hypothesis. *British Journal of Educational Psychology, 84*(2), 226–238.doi: 10.1111/bjep.12027

Zeidner, M. & G. Mathews. (2005). Evaluation anxiety. In A.J. Elliot & C.S. Dweck (Eds.), *Handbook of competence and motivation* (pp.141–163). London: Guilford Press.

Programming robots for factories: The impacts of baseline assessment measures upon young children

Guy Roberts-Holmes

MERRYN HUTCHINGS' *Exam Factory* research and article raises extremely important questions regarding the DfE's current obsession with accountability data and its negative impacts upon schools, families and children. What was particularly fascinating and disturbing at the same time was that Hutchings' conclusions regarding a narrowing of the curriculum, teacher and pupil anxiety, teaching to the test, various forms of cheating and a tacit exclusion of SEN children were all reflected in Bradbury and Roberts-Holmes' (2016) research with four- and five-year-old children in Primary school Reception classes. In an extraordinary linkage between the two studies, the teachers' description of accountability turning schools into 'exam factories' in Hutchings' study was remarkably similar to the Reception teachers stating that children had become reduced to 'robots' because of Baseline Assessment. It could be argued that there has been a reconfiguration of children as robots who are being prepared for exam factories. Early years accountability measures such as the Early Years Foundation Stage Profile (EYFSP) and the Baseline Assessment have increasingly been narrowed, leading to an intensification of 'school readiness' pressures and constrained by performativity demands to produce 'appropriate' data, particularly for narrowly defined literacy and maths.

This school readiness agenda (or turning young children into efficient exam producing robots for international sales in factories) in the early years was further exacerbated by the introduction of Baseline Assessment in 2015 (Bradbury & Roberts-Holmes, 2016). Baseline Assessment would have resulted in a *single numerical score* for each child; when they reach Year 6, each child in the cohort would have been measured against their Baseline Assessment score in order to judge the progress they had made while attending primary school. Baseline thus attempted to reduce children's learning to a single numerical score; in effect generating a 'data shadow' to govern the child. Baseline Assessment thus represented a major shift in approaches to accountability in primary education which involved the early years phase more than ever before. Reception Baseline Assessment was part of the Government's policy document *Reforming assessment and accountability for primary schools* (2014). The rationale for this policy was predicated upon an assumption that Primary schools, including Reception classes were underperforming in formal accountability measures and that 'current expectations for primary schools are set too low' and the policy described in detail the spurious linkage between primary school 'results' and GCSE 'results'. Given BA's evident inaccuracy to measure young children in any meaningful way, it was not surprising that the DfE withdraw BA in February 2016. Despite the entire early years community rejecting BA as inappropriate for young children there are indications at the time of writing (January 2017) that the DfE may try to re-introduce BA in some form.

Within BA learning and pedagogy are reduced to a numerical representation so that a single number can be compared and ranked with other children, classes and schools in an (international) competitive

race to achieve higher results. Working within this hyper scientific and positivist paradigm, assessment, accountability and therefore pedagogy, all too easily becomes reduced to governing children (and teachers) by numbers. Early years teachers were equally reduced to data collectors and 'grey technicians' whose professional judgements were 'hollowed out' whilst private companies did the data analysis, sending this information back to teachers to implement. Indeed, within BA teachers' professionalism is stripped away as they were reduced by policy discourse to 'scorers' (DfE, 2014). The notion of a teacher being a scorer again resonates deeply with the exam factory metaphor.

Baseline Assessment data was to form a key part of primary school accountability. That such an invalid number was subsequently expected to be used to predict children's scores across seven years, was problematic. DfE guidance for the BA stated that 'each assessment item must require a single, objective, binary decision to be made by the scorer' (DfE, 2014, p.1). As such BA policy was 'part of a broader drive to position policymaking as a technocratic exercise, to be undertaken by an elite band of experts who are immune to the influence of politics and ideology' (Morris, 2016, p.226). For the 'band of experts' who devised the BA, the notion was that everything about four-year-old children could be rendered to a single objective number located BA firmly within the hyper-positivist scientific paradigm in which 'reductionism is the name of the game' (Alexander, 2010, p.812). BA was particularly problematic because of the inaccuracy of BA's binary judgements; the negation of English as Additional Language (EAL) children's competencies in their first language; further curriculum narrowing upon literacy and maths and BA's negation of diverse individual children's lived experiences and chronological age differences. Moreover, the reductionist production of a single number for each child inevitably led some schools to use BA scores for prediction and ability grouping in the Reception classroom.

To attempt to reduce young children's complex learning, competencies and abilities to a single number largely based on maths and literacy is deeply disrespectful of young children's competent learning through sociable play. Socio-cultural research has demonstrated that children learn through sets of social relationships (Broadhead & Burt 2012; Fleer 2010). Authentic, holistic, and developmentally appropriate assessment, based upon teachers' observations over time in a range of contexts, makes visible what young children are capable of learning in supportive and collaborative relationships. A particularly useful time to engage in such observations and respectful listening to children is when they are participating in rich and meaningful play activities (Fleer & Richardson, 2009) and can be used to build up a 'learning journey' (Carr & Lee, 2012). Such formative and summative assessment practices aim to make children's learning ever more stimulating, rich and successful. A child's wellbeing and the characteristics of effective learning, such as resilience, perseverance and self-regulation learnt in the context of meaningful play are seen to be more reliable predictors of later academic achievement (Bodrova & Leong, 2007; Whitebread & Bingham, 2012) rather than 'short-term academic results' which may not last.

Bradbury and Roberts-Holmes (2016) reported that baseline assessment ignored the messy and inaccurate production of a single score with four-year-olds in their first six weeks of school. Such a reductionist score also negated the fact that many primary school children experience significant changes and challenges in their lives effecting any simplistic linear rising profile. Reception Baseline assessment was certainly not about children and their learning but rather was an attempt to further regulate early education. Baseline attempts to construct a linear relationship for progress

from age four to age 11, even though the content of the assessments are different; this was seen as a major flaw in the system.

> I don't think you should [use it to measure progress], I don't think you can, because they are children and they are not robots, not machines, they are children. You don't know what influences they have got from outside, what is going to happen in those seven years, so I think it is ridiculous.

The variation between children and their rates of progress meant that any reductionist and simple correlation between Reception and Key Stage 2 was impossible:

> Children's progress is going to be judged against how far they have gone in seven years. Now to my mind that is an almost impossible thing to do because you can't test children at 11 about the same things you were testing them at four. It just doesn't make sense.

The reductionist nature of baseline accountability lent itself to so-called 'ability' grouping at ever earlier ages. The Cambridge Primary Review (2013) were very clear that 'notions of fixed ability would be exacerbated by a baseline test in reception that claimed to reliably predict future attainment'.

> 'BA helps us to group the children in differentiated maths and phonics groups.'

> 'There is no time given to these poor little children to settle in before they are assessed and in our school they are put into ability groups based on these results!'

Such pedagogical differentiation can potentially constrain and limit a child's educational possibilities at ages four and five. Allocating differential resources according to a baseline, serves to make up and produce social inequality and later justify educational inequality of outcome. For example, if some children are constructed as a low 'one' whilst others are constructed as a high 'five', the 'ones' will not be expected to leave primary school performing at the level of 'five'. This is where algorithmic predictive profiling at age four becomes a potentially dangerous and malicious form of control. The potential for grouping and labelling children on Baseline Assessment accountability data is a worrying development especially given that many respondents queried the accuracy of the Baseline Assessment. Hutchings exam factory analysis referring to Victorian and early 19th century inspectors' reports is entirely justified as the Victorian residue of fixed notions of four- and five-year-olds 'ability' re-asserted itself through baseline assessment.

English children are already the most tested in the world and the associated stress of BA may further contribute to the low levels of wellbeing shown by children in the UK in international comparisons (UNICEF, 2011). Hutchings' findings on the damaging consequences of accountability upon children's wellbeing and mental health was again reflected in teachers' comments on baseline assessment.

> I feel that the Baseline Assessment has to be completed too early in the year and means that teachers are madly trying to collect evidence, rather than concentrating on the welfare of their new pupils and helping to create a calm and relaxing environment which is vital for a positive start to their school life.

This is ironic because the development of young children's wellbeing and learning dispositions are more important and reliable predictors of later academic achievement than early gains in the narrow skills involved in literacy and maths (Whitebread

& Bingham, 2013). Children whose experience in the early years has instead supported emotional well-being, cognitive development and self-regulation during play may score less well on early academic tests, but evidence indicates that these children show higher achievement benefits in the longer term (Goswami & Bryant, 2007). Children in Finland, for example, begin formal schooling up to three years later than in England, following active, play-based provision in their early years; they go on to out-perform British children in later attainment (Bodrova, et al., 2007) as this headteacher noted

> I think doing any sort of reputable assessment of very young children is dodgy because the children are so young. You know if those children were in Denmark they wouldn't have had to pick up a pencil yet.

Trying to assess children who had not yet sufficiently developed emotionally led to a deficit model of assessment showing what they can't do as opposed to what they can do. This means that the assessment itself provided a negative, inaccurate and detrimental measure: BA focused on what the children could not do as opposed to what they could do.

> It's ridiculous. It's not a fair representation of children. Many young children are not yet confident enough to show their new teacher what they can do when put on the spot.

Unfortunately, however, the strict DfE regulations meant that BA had to be carried out within six weeks of the children starting school regardless of whether or not the children had 'settled' in.

> I did have children that were crying and I just couldn't get anything out of them at all because they were too upset to do anything, even when I left it later on. Some children just refused or just weren't ready and I know they said you only assess them when they are ready, but some children, well, you got to the point where you had to assess them because it had to be done whether they were ready or not. And obviously then it is not accurate because they weren't at a stage when they wanted to say things.

This leads not only to inaccurate data being generated but was ethically inappropriate and potentially damaging for children's developing self-confidence, self-esteem and learner identity.

> Some children looked at me and said 'I can't read' when asked to read parts of the assessment. It was heartbreaking to see their reaction to it and I spent a lot of time reassuring children.

Here BA had the effect of demotivating and undermining young children's confidence in their reading abilities. Once again the approach of BA accountability was that 'everything can be reduced to a common outcome, standard and measure. What it cannot do is accommodate, let alone, welcome, diversity – of paradigm or theory, pedagogy or provision, childhood or culture' (Moss et al., 2016, p.348). So, BA negated young children's competencies, abilities and creativity as they were tested on a narrow band of academic skills.

The author

Guy Roberts-Holmes
University College London,
Institute of Education,
Department of Learning and Leadership,
20 Bedford Way,
London WC1H 0AL.
Email: g.roberts-holmes@ucl.ac.uk

References

Alexander, R.J. (2010). 'World Class Schools' – noble aspirations or globalised hokum? *Compare 40*(6), 801–817.

Bradbury, A. & Roberts-Holmes, G. (2016). *'They are children, not robots': The introduction of baseline asssessment'*. London: ATL/NUT.

Broadhead, P. & Burt, A. (2012). *Understanding young children's learning through play: Building playful pedagogies*. Abingdon: Routledge.

Bodrova, E. & Leong, D.J. (2007). *Tools of the mind: The Vygotskian approach to early childhood education* (2nd edn.). Upper Saddle River, NJ: Prentice-Hall.

Cambridge Primary Review (2013). *DfE, consultation response form: Primary assessment and accountability under the new national curriculum*. Available from http://cprtrust.org.uk/wp-content/uploads/2014/07/DfE CPRT primary assessment and accountability response.pdf

Carr, M. & Lee, W. (2012). *Learning stories: Constructing learner identities in early education*. London: Sage.

DfE (2014). *Reforming assessment and accountability for primary schools*. Retrieved 15 November 2016 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/297595/Primary_Accountability_and_Assessment_Consultation_Response.pdf.

Fleer, M. (2010). *Early learning and development: Cultural-historical concepts in play*. Cambridge: Cambridge University Press.

Fleer, M. & Richardson, C. (2009) Cultural-historical assessment: Mapping the transformation of understanding. In A. Anning J. Cullen & M. Fleer (Eds.), *Early childhood education* (2nd edn.) London: Sage.

Goswami, U. & Bryant, P. (2007). *Children's cognitive development and learning primary*. Cambridge: University of Cambridge Faculty of Education.

Morris, P. (2016). *Education policy, cross-national tests of pupil achievement and the pursuit of world-class schooling*. London: UCL Institute of Education Press.

Moss, P. et al. (2016). The organisation for economic co-operation and development's international early learning study: Opening for debate and contestation. *Contemporary Issues in Early Childhood, 17*(3), 343–351.

UNICEF (2011). *Child well-being in the UK, Spain and Sweden*. Paris: UNICEF.

Whitebread, D. & Bingham, S. (2013). *School readiness: A critical review of perspectives and evidence*. TACTYC: Association for the professional development of early years educators. Available from: http://tactyc.org.uk/occasional-paper/occasional-paper2.pdf

Response to 'Accountability measures: The factory farm version of education' by Merryn Hutchings

Wendy Symes

IN HER ARTICLE *Accountability measures: The factory farm version of education,* Hutchings outlines the accountability measures currently imposed on schools, and the impact of these measures on school curricula, teacher pedagogy and student emotional and mental health. Of particular interest to me was the notion that increased exposure to testing under the current educational regime potentially increases students' exposure to failure, and the belief that (by Government at least) such exposure is an important and necessary part of learning. In my response, I would like to discuss why emphasising performance, and highlighting failure in particular, is unlikely to result in improved academic outcomes. In fact, on the contrary, it is likely to result in a more negative academic self-concept (ASC), which in turn may lead to increased anxiety, reduced motivation and, ultimately, lower examination scores.

Academic self-concept (ASC) and achievement

ASC refers to the subjective perceptions students hold about their academic competence or ability (Marsh & Martin, 2011). Although subjective, students are likely to draw on their past academic performance when making judgements about their ability (Marsh & Martin, 2011). Once developed, research suggests that the relationship between ASC and performance is reciprocal: performance influences ASC, and ASC influences subsequent performance. Thus, ASC and performance mutually reinforce each other over time. This relationship has been found with both primary (e.g. Guay, Marsh & Boivin, 2003) and secondary (e.g. Marsh et al., 2005) school-age children. Thus, it is likely that early exposure to failure (through, for example, performing poorly on the phonics check at age six) may result in lower ASC, resulting in lower future academic performance in turn. Such findings suggest that, in the early years at least, students should be protected from failure to enable them to make more positive judgements about their ability.

Furthermore, the emphasis on performing well in two, relatively distinct subjects, English and Maths, may also present problems. Research suggests that ASC is not a global construct, but instead, students hold different perceptions of their ability in each subject (Marsh & Craven, 2006). When making these subject-specific competence judgements, students draw not only on their previous performance in a particular subject, but also compare this performance to their performance in other subjects. For example a child ascertaining their ability in Maths may compare their performance in Maths with their performance in English. If the comparison reveals that they are working at a higher level in Maths, then they may draw the conclusion that they are good at Maths, and, crucially, that are not good at English (Marsh & Hau, 2004). That is, even if a student is performing at a good level in English their ASC for their performance in that subject may be negatively impacted as a result of their superior performance in Maths. Consequently, their performance in English is likely to be reduced as a result. Thus, from a psychological perspective, emphasising achievement in both maths and English is likely to come at the detriment to performance in at least one of those subjects.

ASC and anxiety

There are a number of ways in which low ASC may negatively impact academic performance. One way may be through the emotions students experience in achievement settings, such as when completing assigned tasks that may be relevant to an exam, or preparing for an upcoming test. According to the Control Value theory, students' emotional experiences of such settings arise from their subjective judgements as to whether they believe they can complete a particular task (control), and whether that particular task is of importance (value) to them (Pekrun, 2006). Different emotions will result from different control and value appraisals, whereby high control and high value is likely to result in positive emotions, such as enjoyment and hope, and low control and high value is likely to result in negative emotions, such as anxiety (Pekrun, 2006). Consider then, a GCSE student with low ASC preparing for their upcoming Maths exam. They may view the outcome of the exam as important for their future life trajectory, but their low ASC may lead them to believe that they will not be able to pass the exam. As a result they will become anxious. Anxiety, and in particular anxiety about exams, has been linked to a number of negative educational outcomes including lower academic performance (Hembree, 1988).

ASC and fear appeals

Students with low ASC may not only interpret academic tasks as anxiety-provoking, but they may also appraise the actions and messages of those around them as threatening too. For example, prior to important exams, such as GCSEs, teachers may remind students of the importance of these exams, especially in relation to their future lives. A teacher, may, for example, emphasise the consequences of failure on future employment opportunities ('You will find it difficult to get a good job if you fail GCSE maths', Putwain & Roberts, 2009). Messages that emphasise failure in this way have been labelled 'fear appeals' (Sprinkle et al., 2006) and research has shown that different students can interpret these messages in different ways, depending on their ASC. Students with high ASC are more likely to appraise these messages as motivating, feeling inspired when their teachers use them (Putwain & Symes, 2014). They in turn will go on to be more engaged in their studies and perform better in exams (Putwain, Symes & Wilkinson, 2016). Students with low ASC, however, are likely to interpret such messages as threatening (Putwain & Symes, 2014). Being threatened by these messages has been linked to increased test anxiety (Putwain & Symes, 2011), decreased engagement and lower exam performance (Putwain, Symes & Wilkinson, 2016). Similar negative findings have also been found in younger students (Putwain & Best, 2011).

It is likely that if teachers are required to emphasise the levels at which students should be working, they may draw on fear appeals to highlight to students the consequences of failure. Indeed, teachers have indicated support for the use of fear appeals with students preparing for their GCSEs (Putwan & Roberts, 2012), and research suggests that teachers use them more as students get closer to taking their exams (Remedios & Putwain, 2013). Given the relationship between ASC and performance, whereby students with lower ASC have lower academic performance, it is also likely that teachers will tend to use these messages more with students with low ASC. This is particularly alarming given that increased frequency of fear appeal use has been linked to higher threat. That is, the more teachers use these messages, the more threatening students find them (Symes, Putwain & Remedios, 2015). Higher threat appraisals would exacerbate the negative outcomes highlighted in the preceding paragraph, and most likely reduce, rather than increase, academic attainment for those students with low ASC.

ASC and intrinsic motivation

Students who interpret fear appeals as threatening are also less likely to feel intrin-

sically motivated to complete academic tasks (Putwain & Remedios, 2014), and it is likely no coincidence that this is also the case for students with low ASC. This is a concern given that lower intrinsic motivation has been linked to poorer academic performance, particularly in exams (Vansteenkiste et al., 2004). According to Self-Determination Theory, intrinsic motivation can only occur if the three basic needs of competence, autonomy and relatedness are met (Ryan & Deci, 2000). As discussed above, the current educational climate in England is unlikely to be conducive to students developing feelings of competence. Furthermore, other school and teacher practices highlighted in Hutchings' report are likely to further restrict the extent to which the remaining two needs are met. For example, teachers spoke of their frustration at having to present content in a standardised way, and of not having the time to allow students to explore content in more detail. Both of these conditions are likely to reduce the extent to which students perceive themselves as having autonomy over their learning. Furthermore, teachers lamented that they did not have the time to get to know their students, or to think of them as individuals. Under such conditions it is unlikely that teachers can create the feelings of relatedness and connectedness that are also needed for intrinsic motivation to thrive.

In this response I have tried to show how emphasising performance, and failure specifically, may result in detrimental student outcomes. From a psychology of education perspective, the accountability measures currently in place in schools in England appear counter-productive. Emphasising failure, teaching to the test and increasing teacher workload are all likely to reduce the extent to which students can thrive. Whilst it cannot be denied that positive experiences and outcomes are possible for those students who perform at or above expected levels, this is most certainly not the case for those that do not. Furthermore, it is likely that students who already perform well, could achieve even more in a more supportive environment. As noted in Hutchings' report, school-related anxiety is recognised as playing a key role in the mental wellbeing of children and young people. Whilst it is unlikely that the current accountability culture will change, there are things schools could do to improve the emotional experiences of their students (and, no doubt, their teachers) and, in turn, improve outcomes. Placing emphasis on successes, rather than failures, and providing more opportunities for student autonomy and relatedness are all key. Whilst this may seem at odds with the accountability culture in which teachers find themselves, it is more likely a better route to improved exam scores than current practices.

The author

Wendy Symes
School of Education,
University of Birmingham,
Edgbaston,
Birmingham B15 2TT.
Email: w.symes@bham.ac.uk

References

Guay, F., Marsh, H.W. & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives. *Journal of Educational Psychology, 95,* 124–136.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58,* 47–77.

Marsh, H.W. & Craven, R.G. (2006). Reciprocal effects of a self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163.

Marsh, H.W. & Hau, K.T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal-external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96,* 56–67.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 39–416.

Marsh, H.W. & Martin, A.J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81,* 59–77.

Pekrun, R. (2006). The Control-Value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315–341.

Putwain, D.W. & Best, N. (2011). Fear appeals in the primary classroom: Effects on test anxiety and test grade. *Learning and Individual Differences, 21,* 580–584.

Putwain, D. & Remedios, R. (2014). The scare tactic: Do fear appeals predict motivation and exam scores? *School Psychology Quarterly, 29,* 503–516

Putwain, D.W. & Roberts, C.M. (2009). The development of an instrument to measure teachers' use of fear appeals in the GCSE classroom. *British Journal of Educational Psychology, 79,* 643–661.

Putwain, D.W. & Roberts, C.M. (2012). Fear and efficacy appeals in the classroom: The secondary teachers' perspective. *Educational Psychology, 32,* 355–372.

Putwain, D.W. & Symes, W. (2011). Teachers' use of fear appeals in the Mathematics classroom: Worrying or motivating students? *British Journal of Educational Psychology, 81,* 456–474.

Putwain, D.W. & Symes, W (2014). The perceived value of maths and academic self-efficacy in the appraisal of fear appeals used prior to a high-stakes test as threatening or challenging. *Social Psychology of Education, 17,* 229–248.

Putwain, D.W., Symes, W. & Wilkinson, H.M. (2016). Fear appeals, engagement, and examination performance: The role of challenge and threat appraisals. *British Journal of Educational Psychology.*

Remedios, R. & Putwain, D.W. (2013). *Under pressure: The relation between teacher fear appeals and student motivation.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Ryan, R.M. & Deci, E.L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25,* 54–67

Sprinkle, R., Hunt, S., Simonds, C. & Comadena, M. (2006). Fear in the classroom: An examination of teachers' use of fear appeals and students' learning outcomes. *Communication Education, 55,* 389–402

Symes, W., Putwain, D.W. & Remedios, R. (2015). The enabling and protective role of academic buoyancy in the appraisal of fear appeals used prior to high stakes examinations. *School Psychology International, 36,* 605–619.

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K.M. & Deci, E.L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology, 87,* 246–260.

The psychological and instructional consequences of American educational accountability policies: A response to Hutchings

Nathaniel von der Embse

IN THE ARTICLE, *Accountability measures: The factory farm version of education,* Hutchings describes the results of a survey investigation drawn from nearly 8000 teachers in England, with additional interviews from teachers, pupils, and administrators. Throughout the article, Hutchings argued that accountability measures have had overwhelmingly negative impacts on teachers and students, leading to poor educational outcomes. The author frames this argument with the following logic model: government accountability policy (via tests) leads to greater teacher stress and poorer instructional practices (e.g. narrowing of the curriculum) leading to increased student mental health problems (e.g. test anxiety) leading to negative educational outcomes. Much of this argument is supported with research from the US. Thus, I reflect upon the model put forth by Hutchings through an American lens and maintain while many of the proposed relationships are consistent with accountability research from the US, there exists much variability at each stage calling into question the overall conclusion that accountability measures will inevitably lead to poor educational outcomes. I focus my accountability commentary on instructional consequences and psychological consequences, highlighting differences from England and the US, and offering suggestions for an alternative accountability system to meet the emotional, behavioral, and academic needs of all students.

American educational accountability

As noted in the Hutchings article, testing, and subsequent accountability policies attached to test results, have long been a staple of American educational policy. Since the early 1900s, large-scale tests were often used as high school entrance exams and measures of student performance (Resnick, 1980). In the 1960s, laws were passed mandating that schools report educational progress to receive government funds, resulting in a shift in the purpose of testing to minimum competency assessment. The minimum competency testing movement sought to ensure a basic level of competency for all students; this also marked the beginning of measurement-driven instruction, which was shaped by the belief that outcomes of instruction should meet set standards measurable by a test (Resnick, 1980). Tests had become the primary indicator of school success (Koretz & Hamilton, 2006). However, these policies were implemented on a state by state basis and were unequal in implementation and influence on student performance (Hanushek & Raymond, 2005). The 1983 report, *A Nation at Risk,* galvanised public opinion on the perceived failings of American schools and led to increased stakes of large-scale tests, including financial rewards or sanctions to schools based on test scores (Koretz & Hamilton, 2006). Test-based accountability systems typically involve four elements including: standards, measures of performance, targets for performance, and consequences attached to schools' success or failure at performance targets (Hamilton,

2003). Modern accountability policies, including No Child Left Behind (NCLB) and its successor the Every Student Succeeds Act (ESSA), have emphasised both excellence (i.e. reaching ambitious performance standards) and equity (i.e. all students achieving) yet have taken different approaches to achieve these goals. NCLB had focused on meeting annual performance targets resulting in 100 per cent student proficiency with failure resulting in often punitive consequences for schools (e.g. prescriptive use of government funds, closing schools). Much has been written about the negative consequences, yet there were also positive consequences including disaggregated data on student test performance allowing for comparisons amongst student subgroups (Ysseldyke et al., 2004). ESSA was recently signed into law and improves upon NCLB by removing some of the punitive consequences, offers greater flexibility in meeting performance standards, and includes a more comprehensive accountability system.

Importantly, there are fundamental differences in American and English accountability systems that may somewhat limit the comparison of research on related outcomes. For example, the 'stakes' of a test may vary quite significantly amongst schools, teachers, and students depending on the *consequence* attached to test performance. A test may be relatively low stakes for a student if performance is not tied to grade promotion whereas the same test may be high stakes for a teacher if annual evaluation is tied to student test performance. Given the 50 different state departments of education, there is substantial variability in the interpretation and application of federal accountability policy. In addition, there is no US national curriculum or accountability test despite recent efforts (e.g. Common Core State Standards; see Saeki et al., 2016). Thus, caution is urged when generalising the results of American accountability policies (Ysseldyke et al., 2004), and the following discussion of the impacts of said policy is preliminary.

Impact of accountability on instruction

Accountability policies and testing, across states, have largely been perceived by teachers to be highly stressful (Jones & Egley, 2006). In a recent study, nearly 30 per cent of educators reported a clinically significant level of stress in response to accountability testing, with most pressure stemming from school administrators (von der Embse et al., 2015). As such, teachers experiencing high levels of stress are more likely to engage in negative instructional practices (e.g. teaching to the test, reduction in instructional breadth, use of negative motivational tactics; Menken, 2006; von der Embse et al., in press). Instructional consequences of accountability policies also include less time to teach critical thinking and analytical skills (Fantuzzo et al., 2012) and targeting instruction to those students most likely to pass the test (i.e. 'bubble students') at the expense of very high and very low performing students (Diamond & Spillane, 2004). Hamilton and colleagues (2007) have reported that upwards of 76 per cent of teachers had altered their instruction to focus more on topic areas emphasized on accountability tests. However, all the aforementioned research employs correlational (limiting causal inference) and cross-sectional (limiting generalization across time) methodology. It is generally not possible to isolate the direct influence of accountability policy on teacher wellness and instructional practice, and research has noted there are many determinants of both (von der Embse et al., in press). Research will be necessary across various accountability systems to better understand these relationships.

Impact of accountability on student wellbeing

Hutchings postulates that accountability policy has had significant (negative) influence on student emotional health and wellbeing, indicating that exam pressures are one of the biggest causes of stress and anxiety among children and young people.

Teachers in the *EF* study said that test anxiety had affected a large range of students, and this anxiety 'had been the immediate trigger for mental health problems such as self-harming or anorexia.' Indeed, research in US schools have indicated that students report higher levels of anxiety on accountability tests than conventional classroom tests (Segool et al., 2013), and that test anxiety negatively relates to test performance (von der Embse & Witmer, 2014). However, there is little empirical research to suggest that accountability pressures have directly resulted in a greater number of mental health problems in students. Rather, schools are increasingly recognizing the prevalence of mental health issues as nearly 20 per cent of students have a diagnosable mental health disorder (Merikangas et al., 2010), yet only 20 per cent will eventually receive intervention services (Hoagwood & Johnson, 2003). Students with internalizing disorders (e.g. depression, anxiety) are much less likely to receive service (Kalberg et al. 2011). These base rates of mental health problems are not specific to or varied amongst schools under different levels of accountability pressures. Thus, the argument put forth by Devon and cited by Hutchings, that pressure to achieve is a key cause of mental health problems is not substantiated within the empirical research literature, and it is not likely that removing accountability policies would significantly reduce the presence of student mental health problems. Schools instead may consider promoting student wellness and mental health via evidenced-based assessment and intervention, thus also reducing stress associated with testing.

Moving accountability forward

As noted earlier, test-based accountability policies have had a long history in American and English educational systems. For almost as long, researchers and commentators have offered critiques of said policies by noting the often negative and unintended consequences for schools, teachers, and students. The White (1886) quote cited by Hutchings is particularly revealing by stating that 'tests have narrowed instruction, and caused much of the overpressure charged upon the schools…' Given this long history, it begs the question, why have there not been credible replacements for test-based accountability policies? Part of the answer may be found in why tests are used for accountability purposes. Linn (2000) has attributed the emphasis on test-based accountability to its ability to be externally (i.e. government) mandated, implemented rapidly, relative low cost (as compared to decreasing class size or implementing curricular changes), and provision of results visible to educators, parents, and policymakers. While the success of each attribute remains questionable at best, alternatives have failed to demonstrate superiority. For example, portfolio assessments or group tasks have never achieved either adequate psychometric evidence or scalability (see *Standards for Educational and Psychological Testing;* AERA, APA, NCME, 2014). Other alternatives such as sample surveys (Harlen, 2014) do not meet the critical accountability goal of ensuring *equity* as results are not reported at the school level and do not permit a comparison of relative effectiveness or student outcomes. Thus, rather than ridding schools of accountability measures and tests, it is important to consider improvements to better meet school and student needs.

Policymakers, educators, and researchers alike have long sought the ideal formula to promote excellence and equity in education. Hutchings asks a relevant and important question, have accountability policies resulted in schools becoming exam factories? The evidence for the negative impacts of accountability is compelling, and there is a clear need for additional research across time and context to better understand these relationships. Yet, I remain forever optimistic that schools can achieve excellence *and* equity, and I offer four considerations to move accountability forward. First, accountability policies should recognize the current onerous emphasis on test

performance and preparation by reducing total testing time. Shorter testing should be employed at regular intervals (i.e. formative testing) to better inform instructional practice. Second, accountability testing should integrate value-added models that take into consideration student demographic characteristics as well as growth over time; these models are not without limitation (see Rothstein, 2010) but offer improvement over current testing regimes. Third, accountability *consequences* should inform additional service provision, rather than punish and remove funding from underperforming schools. This process depends on testing that is reflective of and informative to school need. Fourth, accountability policy and by proxy testing, should reflect a more comprehensive view of educational performance. Academic success does not always equate to educational success—socio-emotional learning and meeting the mental health needs of students are necessary conditions for educational success and adequate preparation to enter the workforce or university (Adelman & Taylor, 2010). For example, the ESSA law allows state accountability systems to utilize nonacademic indicator to account up to 49 per cent of a school's annual evaluation such as school climate, positive behavior intervention supports, student wellness, and student-teacher relationships. Continued research, such as that reported by Hutchings, will be necessary to engage and inform policymakers to reconsider traditional forms of accountability to better meet the educational needs of all students.

The author

Nathaniel P. von der Embse, Ph.D., NCSP
College of Education,
Temple University,
Philadelphia, PA 19122,
USA.
Email: nate@temple.edu

References

Adelman, H.S. & Taylor, L. (2010). *Mental health in schools: Engaging learners, preventing problems, and improving schools.* Thousand Oaks, CA: Corwin Press.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (2014). *The standards for educational and psychological testing.* Washington, DC: AERA.

Diamond, J.B. & Spillane, J.P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106,* 1145–1176.

Fantuzzo, J., Perlman, S., Sproul, F., Minney, A., Perry, M.A. & Li, F. (2012). Making visible teacher reports of their teaching experiences: The early childhood teacher experiences scale. *Psychology in the Schools, 49*(2), 194–205.

Hamilton, L.S. (2003). Assessment as a policy tool. *Review of Research in Education, 27,* 25–68.

Hamilton, L., Stecher, B., Marsh, J., McCombs, J.S., Robyn, A., Russell, J., Saftel, S. & Barney, H. (2007). *Standards-based accountability under no child left behind: Experiences of teachers and administrators in three states.* Santa Monica, CA: Rand Corporation.

Hanushek, E.A. & Raymond, M.F. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis & Management, 24*(2), 297–327.

Harlen, W. (2014). *Assessment, standards, and quality of learning in primary education.* Cambridge Primary Review Trust.

Hoagwood, K. & Johnson, J. (2003). School psychology: A public health framework: From evidence-based practices to evidence-based policies. *Journal of School Psychology, 41,* 3–21.

Jones, B.D. & Egley, R.J. (2006). Looking through different lenses: Teachers' and administrators' views of accountability. *Phi Delta Kappan,* 87, 767–771.

Kahlberg, J.R., Lane, K.L., Driscoll, S. & Wehby, J. (2011). Systematic screening for emotional and behavioral disorders at the high school level: A formidable and necessary task. *Remedial and Special Education, 32,* 506–520.

Koretz, D.M. & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennen (Ed.), *Educational Measurement* (4th edn., pp.531–542). Westport, CT: American Council on Education/Praeger.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–17.

Menken, K. (2006). Teaching to the test: How no child left behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal, 30,* 521–546.

Merikangas, K.R., He, J.P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L. & Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: Results from the National comorbidity survey replication–adolescent supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(10), 980–989.

Resnick, D. (1980). Minimum competency testing historically considered. *Review of Research, 77,* 579–593.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125,* 175–214.

Saeki, E., Pendergast, L., Segool, N. & von der Embse, N.P. (2015). Psychosocial and instructional consequences of the Common Core State Standards: Implications for research and practice. *Contemporary School Psychology, 19* (2), 89–97.

Segool, N., Carlson, J., Goforth, A., von der Embse, N. & Barterian, J. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*(5), 489–499.

von der Embse, N.P., Kilgus, S.P., Bowler, M., Solomon, H. & Curtiss, C. (2015). Initial development and factor structure of the Educator Test Stress Inventory. *Journal of Psychoeducational Assessment, 33*(3), 223–237.

von der Embse, N.P., Schoemann, A., Wicoff, M., Kilgus, S.P. & Bowler, M. (in press). The influence of test-based accountability policies on teacher stress and teaching practices: A moderated mediation model. *Educational Psychology.*

von der Embse, N.P. & Witmer, S. (2014). High-stakes accountability: Student anxiety and large-scale testing. *Journal of Applied School Psychology, 2,* 1–24.

White, E. (1886). *The elements of pedagogy.* New York: American Book Company

Ysseldyke, J., Nelson, J.R., Christenson, S., Johnson, D.R., Dennison, A., Triezenberg, H. et al. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71,* 75–95.

Commentary – 'Accountability measures: The factory farm version of education'

Christina Wikström

THIS ESSAY presents a brief commentary to the paper *Accountability measures: The factory farm version of education* by Hutchings (2017). The paper is argumentative and quite critical towards educational accountability and high stakes testing. The paper summarises the findings from a previous investigation by the author (Hutchings, 2015) into how the accountability system is experienced by teachers and pupils in England, but also describes the context and the history behind the current system, and illuminates important issues in this system. Findings from previous research on accountability systems and high stakes testing, predominately from an English and US context, are included in the discussion to support the claims being made. In short, Hutchings concludes that accountability systems in education, and high stakes testing in particular, have a negative impact on curriculum, the pedagogy teachers are using, and also on pupils learning, well-being and self-esteem.

The paper gives a good description of the educational accountability system in England and Wales, reviews some of the important research on educational accountability, and illuminates a number of key problems that are linked to such systems. Hutchings focuses the main part of her argument on the problems with typical performance measures that are used in accountability systems for monitoring pupils', teachers' and school performances, in the English case these are tests in English and mathematics, and lists a number of consequences that follow the use of such tests. This leads to a discussion on problems with high stakes testing in general, and the negative effects on pupils, but also on pedagogy and learning; Since the test will signal what is important, this will direct the attention to what the test measures. Schools are then prone to use different strategies maximising their scores, such as putting a lot of emphasis on the knowledge and skills measured by the test, and the format of the test, known as teaching to the test, which, in turn, leads to consequences for pupils' learning. The increased focus on testing then causes stress and anxiety for both teachers and students.

Although the criticism is strong, the conclusions are fairly uncontroversial, as similar findings have been reported also in other systems over the years, particularly on the effects of the accountability system that was implemented in the US after the No Child Left Behind Reform (see, for instance, Stecher, Hamilton & Klein, 2002). It is also well known that that all pupils do not respond well to being tested when there are high stakes attached, and if testing becomes a dominating feature in education, this problem will increase.

The arguments put forward are convincing, and the paper makes a number of important points, that are particularly relevant for a policy discussion. Some of the findings are likely to be generalisable also to other similar systems, but with some caution, since school systems are complex and culturally different. What works or not, and the consequences from implementing different models are likely to vary. Below, I will address and discuss issues I find particularly interesting in the paper.

What does educational accountability mean?

An interesting aspect addressed in Hutch-

ings' paper is where accountability systems seem to fail, and how this can be addressed. However, first of all the matter of definition should be mentioned. It is not entirely clear how educational accountability is defined, and when a school system can be categorised as such. It is easy to assume that there is a common understanding and a common international discourse, but this is not really the case. Educational accountability is seldom defined in a thorough way, and it is clear that when it is, it can be conceptualised differently (Heim, 1996), but there seems to be a general agreement that it has to do with responsibility, authority, evaluation and control (see, for instance, Adams & Kirst, 1999; Stecher & Kirby, 2004). A general definition is made by Adams and Kirst (1999), that models of accountability are determined by the nature of the relationship between those who establish an expectation and those who are accountable, the nature of the accountability expectation, the type of mechanism employed to ensure accountability, and the nature of the incentives used to compel the actions of those who are accountable. This is described in the economic literature as the Principal and Agent Theory. It describes the relationship between a principal and an agent who performs some service on the principal's behalf, where the agent has the authority to make his own decisions, and stands the risk while the principle gives the incentives (Jensen & Meckling, 1976).

Hutchings mentions two types of accountability systems when describing the system adopted in England and Wales: a market-driven approach, characterised by performance measures and competition, where the market rewards successful schools, and management accountability, where the latter seems to be the type that is well known from a US context; it involves testing, targets, incentives and sanctions. This is in line with the definition by Stecher, Hamilton and Klein (2002) that test-based accountability system is built on the strategy that all students take a standardised achievement test, while attaching rewards for improved scores and sanctions in case they do not improve. However, it is not necessarily so that these systems are very different in practice, as will be discussed below.

Different model, similar consequences

The theoretical definitions addressed above can be useful when describing or evaluating the implementation of educational accountability in a system, but sometimes this becomes more complex. The Swedish system can for instance be defined as a market-driven approach to accountability, in line with the former system in England and Wales, as described in the paper, but has also elements of what is described as management accountability. During the 1990s, the Swedish school system changed through reforms where the assumption was that the school system could be made more efficient by putting the schools on the market, and introducing a combination of competition and performance targets (Lundahl, 2002). However, this model also involves performance measures and inspections. School inspectorates collect information such as school documentation (policies, administrative reports, etc.), observe the schools' daily work though on-site visits, take part in lectures and collect information how the schools are assisting students with special needs. Information is also collected through surveys and sometimes interviews where teachers, students and parents can express their opinions on the school, the study climate, etc. The inspectorates also evaluate information on school performances in terms of national test scores and grades, but not necessarily with the focus on high performances. When evaluating school performance the student body taken into consideration, as it is known to explain a large part of the variation in educational performance. Still, it is expected that all students should be given the support to meet the basic standard for each subject. The inspectorates' task is primarily to determine if the assessment in the schools is carried out according to regulations and

consequently in a reliable and valid way. There are no rewards for high scores or grades, but underperformance in an unexpected or conspicuous way will be investigated further. If the inspectorate finds that a school is not meeting the requirements in some way, the school will be asked to adjust their methods or administration within a certain time-frame, and then be re-assessed. If the issues are regarded as serious, and a school does not make sufficient efforts to make adequate improvements, it can be sanctioned, closed down or restructured (Eklöf, Andersson & Wikström, 2009).

The assumption is that when collecting extensive and various kinds of information, the risk that negative consequences that are listed in Hutchings paper will occur should decrease. This strategy is also supported in the literature (see, for instance, Stecher, Hamilton & Klein, 2002). Ideally, the mechanisms will function in balance – a school that is too focused on tests, or prioritising subjects measured by the test, will underperform on the other measures. To this point, the Swedish accountability system should work rather well. Schools are accountable for the work they do, but the information collected should give information on the schools' objectives in a wide perspective. However, in this system, the additional market-driven approach changes the balance. Since schools are competing, they are dependent on being attractive to potential pupils and their parents. Like in the historical background to the system in England and Wales, as described by Hutchings, future pupils and their parents do not always value performance as expected, and their choices may be based on other variables than traditional performance measures. But it is also important for the schools to show good results and high grades, as it signals quality and high grades are important for students who are on academic track. The pressure for high performance has also increased from policy level with the negative results on international studies such as PISA, as schools are pressured to improve their results, and show serious effort to do so. The increased pressure from stakeholders for high performance gives incentives for manipulation, in line with the strategies used in the English system when it comes to finding ways to maximise test scores. In the Swedish model, teachers have the sole responsibility for grading their pupils, a model that has been seen as a more valid way to assess performance, since they can collect various kinds of evidence, and observe the pupils in the classroom. This would, in theory, be a good performance indicator, in line with the suggestions addressed in paper. However, when the stakes are high, the measures tend to be compromised, irrespective of format. Teachers report on being pressured for lenient or high grading, not only by the pupils and their parents, but also by school leaders, which has caused a heated debate in the media and the public discussion recently (Sveriges Television, 2016). The problems with grade inflation have, in turn, led to an increased focus on tests, as these are seen as a way to calibrate grades and make these more comparable. There is a growing belief among policy makers and in the public debate that performances should be monitored more thoroughly with standardised tests. As a result, the stakes of the national tests have increased and additional uses of the tests have been added over time (Eklöf, Andersson & Wikström, 2009). They are now, among other things, to be used more systematically as evidence in the grading process and to evaluate school performances. The tests are still valued as informative tools for guiding the teachers when grading the students, and so far there are no direct sanctions attached to low performance on the tests. But it can also be noted that the stakes have increased, and teachers and students are often complaining about increased pressure, that the testing and test administration has increased, and there is a stronger focus on performance indicators that often is in competition with their teaching task. In all, there are many resemblances between the English and the

Swedish system, although the focus on tests is still stronger in England, and also their consequences for curriculum, pedagogy and the pupils.

Concluding comments

Hutchings' paper shows that there is a complexity embedded in accountability systems, that often leads to a number of negative consequences, some even counter-productive to the original intentions when implementing the system. A question I have raised it if is educational accountability or how the accountability model is designed that is most problematic, and it seems it is the latter. It is of key importance to have valid performance measures, but also the balance in the system is crucial (Jensen & Meckling, 1976). When there is unbalance, there will be openings for strategies that can lead to negative consequences. When it comes to valid measures, performance measures such as standardised tests can serve as pieces of information, but they will not be a valid measure of school quality, especially not alone, which can be described as construct underrepresentation. A complexity in this context is also the multiple purposes of tests and similar performance measures since this often lead to problems and validity is known to vary with how the outcome is used. Systems and performance measures must be validated, which also is pointed out in the paper. Unfortunately this is often ignored from policy level, maybe due to ignorance. There is plenty of theory on validity and validation (see, for instance, Eggen & Stobart, 2016; Newton & Shaw, 2014), and also models for validation, such as the well known, but somewhat complicated validity theory by Messick (1989), that can be especially useful when guiding systematic investigations of the effects of assessment instruments or systems that are complex and have an impact on many levels.

To conclude, I find Hutchings' paper to be an important contribution to the discussion on accountability in education and its consequences, as it presents a good description of how the implementation of a system that is intended to increase efficiency in fact can be negative both for educational quality and for the wellbeing of teachers and pupils. The research presented in the paper indicates that if relieving the system from strong focus on narrow performance measures, some of the consequences that have been observed will probably be less prominent and it is possible that such a system can work more as intended. The paper also illustrates the importance of basing decisions on established theory and research, but also how we fail to learn from history. It is important to continue to raise awareness of the necessity to continuously validate systems and instruments used for high stakes decisions, to identify and also foresee unwanted consequences for schools, teachers and students, and consequently society at large.

The author

Christina Wikström
Department of Applied Educational Science,
Umeå University,
Johan Bures väg 16,
90187 Umeå,
Sweden.
Email: christina.wikstrom@umu.se

References

Adams, J. & Kirst, M. (1999). New demands and concepts for educational accountability: Striving for results in an era of excellence. In J. Murphy & K. Louis (Eds.), *Handbook of research in education administration* (pp.463–489). Washington, DC: American Association of Educational Researchers.

Eggen, T. & Stobart, G. (Eds.) (2016). *High-stakes testing in education: Value, fairness and consequences.* Abingdon: Routledge.

Eklöf, H., Andersson, E. & Wikström, C. (2009). The concept of accountability in education: Does the Swedish school system apply? *Cadmo, 2,* 55–66.

Heim, M. (1996). *Accountability in education: A primer for school leaders.* Honolulu, HI: Pacific Resources for Education and Learning, Hawaii State Department of Education.

Hutchings, M. (2015). *Test factories? The impact of accountability measures on children and young people.* London: National Union of Teachers (NUT). Retrieved 16 January 2017 from https://www.teachers.org.uk/files/exam-factories.pdf

Hutchings, M. (2017). Accountability measures: The factory farm version of education, *The Psychology of Education Review, 41*(1), 3–15.

Jensen, M.C. & Meckling, W.H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305–360.

Lundahl, L. (2002). Sweden: Decentralization, deregulation, quasi-markets and then what? *Journal of Education Policy, 17*(6), 687–697.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd edn.) (pp.13–103). New York: Macmillan.

Newton, P. & Shaw, S. (2014). *Validity in educational and psychological assessment.* London: Sage.

Stecher, B., Hamilton, H. & Klein, S.P. (2002). *Making sense of test-based accountability in education.* Santa Monica, CA: RAND Organization.

Stecher, B. & Kirby, S.N. (2004). *Organizational improvement and accountability: Lessons for education from other sectors.* Santa Monica, CA: RAND Organization.

Sveriges Television (2016) *Rektorer pressar lärare att sätta oskäligt höga betyg.* [School leaders are pressuring teachers for high grading] Retrieved 16 January 2017 from http://www.svt.se/ug/rektorer-pressar-larare-att-satta-oskaligt-hoga-betyg

Author's response

Merryn Hutchings

FIRSTLY I would like to thank everyone who took the trouble to read and respond so thoughtfully to my paper. I am grateful for the opportunity to respond to a few of the specific points that have been raised.

The majority of respondents indicated overall support for the arguments in the paper, agreeing that the use of tests for accountability has negative consequences for the curriculum, for teaching and for the well-being of children and young people. Several noted that these are not new arguments, which is certainly true. There is a wealth of previous research and commentary about the negative effects of both current and nineteenth century accountability structures in England and the US. But while these arguments, and the considerable body of evidence, are familiar to many in the education research community, they continue to be rejected by policy-makers. In view of the damage being done to children, teachers and educational outcomes, most respondents indicated that it is vital that researchers continue to amass evidence and present it in different forms.

Hattie and Clinton were exceptions here; they characterise the paper as taking sides in a debate about tests or no tests, 'another example of privileging the divide' summing up 'typical reactions to the dominance of tests'. I wonder if they have misunderstood the main argument, which is about the impacts of *testing used as a form of accountability;* it is not a debate about whether tests should be used. I stated in the paper that teachers need to assess what has been learned, and plan future learning. Such assessment can be done in a range of ways including using tests (though as Putwain comments, tests are a 'fairly blunt' means of evaluating student learning which prioritise 'certain knowledge and cognitive styles'. The responses from both Merrell and Hattie and Clinton describe test systems developed to support teachers with such assessment; the results can be reported to parents and children as part of schools' accountability to them. Such tests are completely different from the English national tests used by the government for accountability purposes; the latter are not used by teachers to plan future teaching (partly because the children move on to other schools, and partly because so much teaching to the test has been done that teachers prefer to use other tests to assess what the child has learned and plan future teaching); their main purpose is to give a picture of school performance which can be used to identify so-called 'failing' schools.

The *Exam Factories (EF)* report was written as an attempt to pull together the range of evidence from previous research along with new evidence to show the problems are current and ongoing, both to inform policy debates, and also to inform teachers and parents who have generally not read all the research evidence. Following its publication, I have spoken at meetings of teachers and parents around the country who are concerned about the direction schooling has taken, and who have welcomed research which lays out the overall picture.

As I explained in the paper, the 'exam factories' metaphor is one that many teachers used in their written comments and interviews. Hattie and Clinton argue that it is inappropriate because modern factories have attributes that teachers would envy – 'light, clean, collaborative teams, privileging expertise, seeking and using feedback to improve' and so on. But in this they miss the essential point of the metaphor; factories (and factory farms), whether modern or Victorian, aim to produce a uniform and standard product.

The point that teachers made repeatedly in the *EF* research is that children are not all the same; and their concerns about the production of 'identical little robots' and 'identikit children' illustrated this. Children have different interests, enthusiasms, ambitions, and different aptitudes. They do not respond to the same incentives; Symes and Putwain both show that students vary in their academic self-concept and response to pressure. They do not all learn at the same rate, as progress measures imply they should. Therefore I believe it is not appropriate to expect all children to follow the same academic curriculum and progress at the rate the government has decided.

Hattie and Clinton talk about the need for students to 'gain a year's growth for a year's input', and while I would want to ensure that every child is learning (and is motivated to learn and is enjoying learning), I am not convinced that we can set out how much ground should be covered in a year. In England, the government sets levels of what should be expected of different age groups, and appears to believe that if ever-higher levels are expected, children will learn more. The change in expectations of 10 to 11 year-olds resulted in a fall in numbers reaching the expected level from 80 per cent in 2015 to 53 per cent in 2016. Thus the amount a child is expected to learn in a year has been increased, and this is for political rather than educational reasons. In the light of research about academic self-concept described by Symes, it is clearly worrying that half the children in this age group have been told that they have failed. But one of the inevitable consequences of using tests for accountability purposes is that the pass rate will be adjusted to ensure that some children do fail; an accountability measure would not be successful if it indicated that every child reached the expected level.

One of the challenges in writing the paper was that I was trying to present an overview of the impacts of current accountability structures on children and young people in England. But to present all the relevant issues in proper depth would have required very much more space than I had either in the paper – or even in my original report. Moreover, I am certainly not an expert in all the relevant areas – testing, curriculum, mental health, etc. Thus I am very grateful to those who have added to my arguments by providing details of other research, including their own – for example, on early years (Roberts-Homes & Jarvis), and the relationship between test anxiety, academic self-concept and test outcomes (Symes), special needs (Jarvis), and so on. Some respondents also added interesting international perspectives (Babayiğit, Hattie & Clinton, von der Embse, Wikstrom).

I agree with those who critique PISA tests (Babayiğit), and the way they are encouraging governments to put pressure on schools for higher results (Wikstrom). In this light, I am particularly concerned by the UK government's support for the new PISA cross-national assessment of early learning outcomes. The point I was making in the paper was simply that if two sets of tests (PISA and GCSEs) give such contrasting pictures of how standards are changing over time, one has to doubt the validity of using any tests for this purpose.

Respondents also identified areas where they believe that further research is needed. In particular, Putwain and Babayiğit commented that further research is needed into the relationship between accountability testing and mental health, and von der Embse reports that in the US there is little empirical evidence that 'accountability pressures have directly resulted in a greater number of mental health problems in students.' I agree that more research would be useful. However, this is fraught with difficulty: Where is the borderline between normal anxiety or stress and a mental health problem? How can the extent to which different factors have contributed to the problem be teased out? There is evidence that many children say they worry about their school work, and in England, the numbers appear to be increasing; for example, in a

recent survey of 10 to 11 year-olds, around two in five reported that they worried about school work and tests 'all the time' or 'a lot' (Place2Be, 2017) and the ChildLine Annual Review 2015–16 showed increasing numbers of children calling helplines about school work, tests and exams. The ChildLine (2016) review states:

> Exam pressures manifested themselves in many ways, but the most frequently talked about were feeling depressed, panic attacks, excessive crying, low self-esteem, self-harming and suicidal thought. …Exam stress also sometimes triggered relapses in young people who had not been cutting for some time. …For some, the pressures of exams and school work had become so great they had started to have suicidal thoughts and feelings. (pp.28–29)

While we cannot pinpoint exactly how many young people are affected, or how much tests and exams are contributing to mental health problems, I think it would be irresponsible to ignore the evidence we have.

Putwain points out that another area for further research is the impact of accountability measures on teachers. This is clearly important in view of the current teacher shortage and numbers leaving the profession. While teachers were not the focus of the *EF* research, many of the comments written on the survey attested to the pressure that some were experiencing, and to their unhappiness – which inevitably affects children's experience of school.

While I believe that the current structures for accountability in England are damaging, this does not mean that I am opposed to accountability. I believe that schools should be accountable – to children, parents, and government. In the *EF* research, I was asked to investigate the impacts on children and young people of current accountability structures in England. The focus of the research (and consequently, of the report and the paper) was not on alternative ways of ensuring accountability. Clearly this is also important, and I welcome the suggestions of Merrell who argues that schools should be accountable for 'the effective implementation and use of a monitoring system, intelligent preparation of data and use of research findings to inform their practice, and of Putwain, whose preferred option would be to base accountability on 'rigorous assessment of teaching quality' including aspects of teaching such as 'building trusting and supportive relationships' and 'the well-being of students'. As Merrell argues, such approaches would also make increased demands on teacher education and professional development, and teachers would need more time during their working week to assess children and analyse the information. However, as Putwain points out, 'it is quicker and cheaper to use test scores to judge teachers than to conduct quality, rigorous assessments of their teaching quality.' More expensive approaches will not be appealing to politicians.

In the previous Open Dialogue, Claxton and Lucas expressed their impatience with 'the slow and faltering nature of education reform', and attributed much of this to 'the pusillanimous attitude of politicians and the simplistic headline-grabbing reportage by the media'. They argued that change would not happen 'until a substantial fraction of the electorate is clamouring for it'. I agree, and I believe therefore that education researchers need to make their findings available in many different forms, and need to involve themselves in campaigning alongside parents and teachers. Campaigns for change have to emphasise the damage being done to children, and the consequences that this has for the future of our society and economy. In my paper I quoted the arguments about this put forward by the CBI and Institute of Directors. Their voices, together with those of other employers, and of university leaders are clearly important in any effort to bring about change. Parents' voices are also critical, as are those of a

whole range of education experts, subject associations, and psychologists. Therefore I strongly welcome the formation of the campaign More than a Score (UK), a coalition of parents' groups, teachers' organisations, education researchers and other experts formed to campaign against the use of testing as a form of accountability in English primary schools. Both the Association of Child Psychotherapists and Association of Educational Psychologists form part of this coalition, as does the British Educational Research Association. I also welcome further evidence of the damage that tests are doing – for example, in *The Mismeasurement of Learning: How tests are damaging children and primary education,* a collection of papers from Reclaiming Schools, a network of academics and researchers. I believe that those who oppose the current use of tests as a form of accountability need to continue to present the arguments, because at the moment, the government shows no sign of accepting them – on the contrary, there are indications that they may re-impose baseline testing, and they have agreed to take part in the OECD's cross-national assessment of early learning outcomes, which will involve testing of four- and five-year-old children. These moves have to be resisted.

In the 19th century, it took 30 years for the problems associated with payment by results to be recognised, and a further period of time for schools to recover. We cannot afford to allow the damage created by current accountability policies to continue for that long.

The author

Merryn Hutchings
Emeritus Professor of Education,
Institute for Policy Studies in Education,
London Metropolitan University,
166–220 Holloway Road,
London N7 8DB.
Email: m.hutchings@londonmet.ac.uk

References

ChildLine (2016). It turned out someone did care: *ChildLine annual review 2015–2016.* Retreived 13 February 2017 from https://www.nspcc.org.uk/services-and-resources/research-and-resources/2016/childline-annual-review-2015-16-turned-out-someone-did-care/

Place2Be (2017). *Children's Mental Health Week Survey Results.* Retrieved 13 February 2017 fromhttps://www.place2be.org.uk/media/587987/childrens-survey-factsheet.pdf

Notes for contributors

The Psychology of Education Review is published twice yearly (Spring/Autumn). The aim is to publish material in the area of Psychology and Education. Submissions in the following form are welcomed.

The Open Dialogue: This is a mechanism whereby there is simultaneous exchange of views on an issue of substantial interest. It includes: An Initial Paper, outlining a distinctive position; Peer Review, in which peers comment on the position; Author's Reply, offering a final response. Anyone wishing to contribute should contact and discuss preliminary ideas with the Editors.

Individual Papers: We welcome individual papers in any aspect of Psychology and Education. Papers should be 2000 to 3000 words and may be of a theoretical or empirical nature. Individual papers are peer refereed.

Work in Progress: Graduate students, researchers and others are invited to describe, discuss and identify areas of their current research in Psychology and Education. This section is divided into two sub-sections:

(1) peer-refereed reports on on-going research by established researchers (750 to 2000 words);
(2) reports on on-going research from research students (up to 1000 words).

Book Reviews: *The Psychology of Education Review* aims to provide reviews of relevant books as soon as possible after their publication. Authors should alert the Editor (see inside front cover) if they wish to see a particular book reviewed – either of their own writing or if they feel it is relevant to our readers. Authors may be invited to respond to reviewed books.

In Brief: This includes: Information; Letters; News; Short Reports.

Submission of material
An electronic copy should be submitted to the Editor. Submissions must follow the British Psychological Society's guidelines for journal submission and should state clearly within which section of *The Psychology of Education Review* they are to be considered.

Annual subscription
Free to members of the Psychology of Education Section. Individual electronic versions can be purchased from www.bps.org.uk – £3 to non-members of the Psychology of Education Section or £4.20 to non-members of the British Psychological Society.

Editorial address
Individual Papers and Work in Progress submissions should be sent to the Editor.

Editor: Katy Smart
Email: katy.smart@bristol.ac.uk

Contents

St Andrews House, 48 Princess Road East, Leicester LE1 7DR, UK
Tel 0116 254 9568 Fax 0116 227 1314 Email mail@bps.org.uk www.bps.org.uk

ISSN 1463-9807